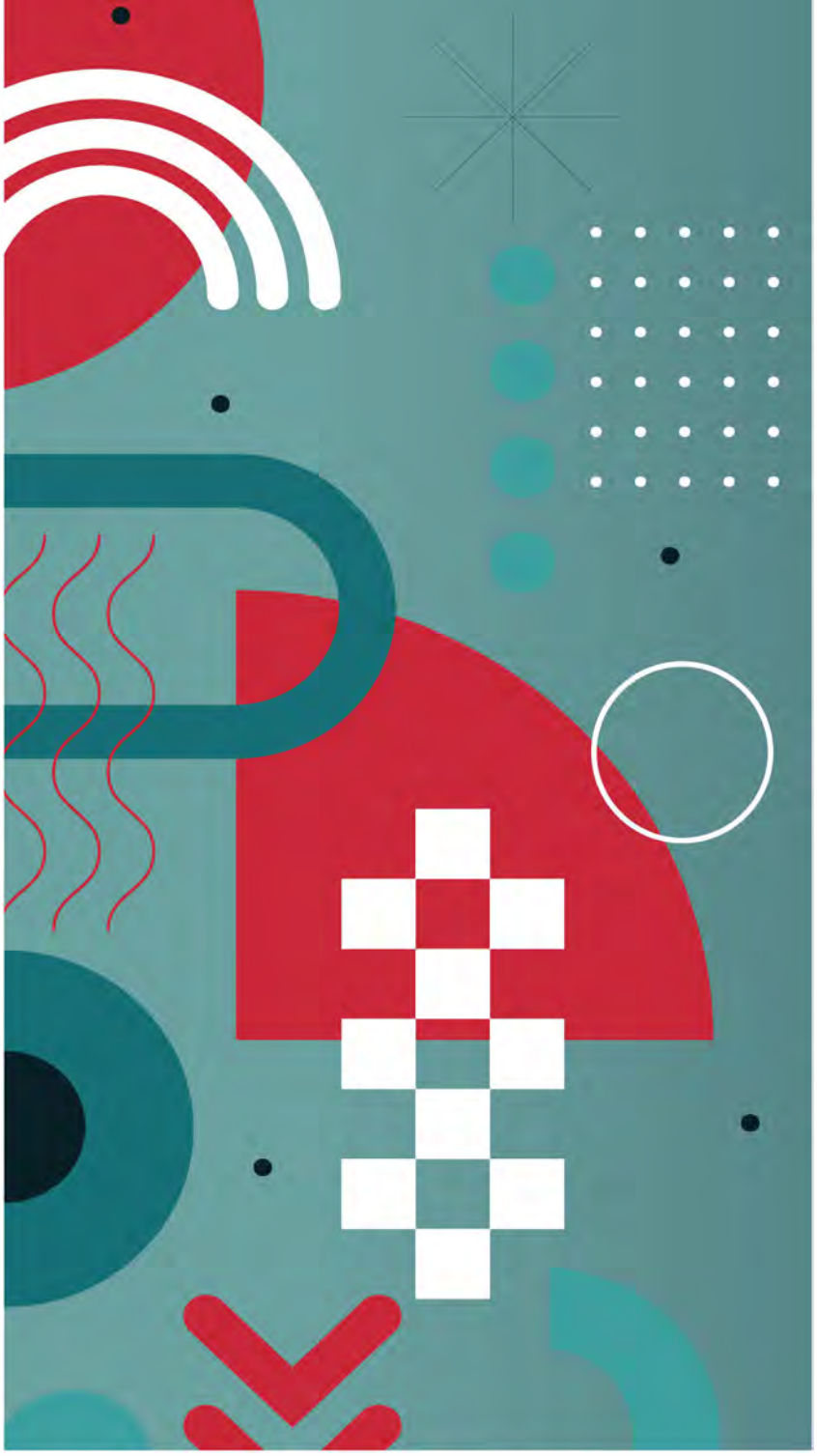


cacic2022

XXVIII CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN

3 al 6 de octubre - La Rioja Capital

LIBRO DE ACTAS

Red UNCI

UNLaR DACEFyN | Universidad Nacional de La Rioja

UNLaR Universidad Nacional de La Rioja

ISBN: 978-987-1364-31-2

Título: Libro de Actas : XXVIII Congreso Argentino de Ciencias de la Computación- CACIC 2022

Razón Social: Universidad Nacional de La Rioja. Departamento de Ciencias Exactas, Físicas y Naturales.

Sello Editor: Universidad Nacional de La Rioja - EUDELAR

Menciones:

Diseñado por Reus Quinteros, Debora Daiana ; Diseño e Ilustración de cubierta por Oro Gregoriadis Rita Mariana.

Compilación: Rodríguez, Sandra Isabel ; Giménez, Mónica Noemi, Molina, Miguel Ángel.

Fecha Publicación: 01/2023

Páginas: 1008

Soporte: Libro Digital (EN INTERNET)

Formato: PDF

Tipo Soporte: Digital: descarga y online

Tipo de Contenido: Texto (legibles a simple vista)

ISBN 978-987-1364-31-2

9 789871 364312

XXVIII CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN

03 al 06 de octubre de 2022

LA RIOJA – ARGENTINA

UNIVERSIDAD NACIONAL DE LA RIOJA

DEPARTAMENTO ACADÉMICO
DE CIENCIAS EXACTAS, FÍSICAS Y NATURALES

https://cacic2022.unlar.edu.ar/

Comité Científico

Abásolo, María José (Argentina)
Aciti, Claudio (Argentina)
Alfonso, Hugo (Argentina)
Ardenghi, Jorge (Argentina)
Arroyo, Marcelo (Argentina)
Astudillo Hernán (Chile)
Baldasarri, Sandra (España)
Balladini, Javier (Argentina)
Barbosa, Luis (Portugal)
Bertone, Rodolfo (Argentina)
Bría, Oscar (Argentina)
Brisaboa, Nieves (España)
Buckle, Carlos (Argentina)
Cañas, Alberto (EE.UU)
Chavez, Edgar (México)
Coello Coello, Carlos (México)
Dix, Juerguen (Alemania)
Doallo, Ramón (España)
Docampo, Domingo (España)
Casali, Ana (Argentina)
Castro, Silvia (Argentina)
Cechich, Alejandra (Argentina)
Cuckierman, Uriel (Argentina)
De Giusti, Armando (Argentina)
De Giusti, Laura (Argentina)
De Vincenzi, Marcelo (Argentina)
Deco, Claudia (Argentina)
Depetris, Beatriz (Argentina)
Diaz, Javier (Argentina)
Dujmovic Jozo (USA)
Estayno, Marcelo (Argentina)
Estevez, Elsa (Argentina)
Eterovic, Jorge (Argentina)
Falappa, Marcelo (Argentina)
Fillottrani, Pablo (Argentina)
Finochietto, Jorge (Argentina)
Frati Emmanuel (Argentina)
Fridlender Daniel (Argentina)
García Garino, Carlos (Argentina)
García Villalba, Javier (España)
Género, Marcela (España)
Gomez, Sergio (Argentina)
Gröller, Eduard (Austria)
Guerrero, Roberto (Argentina)
Ierache, Jorge (Argentina)
Janowski, Tomasz (Naciones Unidas)
Kuna Horacio (Argentina)
Lanzarini, Laura (Argentina)
Leguizamón, Guillermo (Argentina)
Lopez Gil, Fernando (España)
Loui, Ronald Prescott (EEUU)
Luque, Emilio (España)
Madoz, Cristina (Argentina)
Malberti, Alejandra (Argentina)
Manresa Yee, Cristina (España)
Marín, Mauricio (Chile)
Mas Sansó, Ramón (España)
Micolini, Orlando (Argentina)
Mon Alicia (Argentina)
Motz, Regina (Uruguay)
Naiouf, Marcelo (Argentina)
Navarro Martín, Antonio (España)
Pardo, Álvaro (Uruguay)
Pasini, Ariel (Argentina)
Pesado, Patricia (Argentina)
Piattini, Mario (España)
Piccoli, María Fabiana (Argentina)
Printista, Marcela (Argentina)
Puppo, Enrico (Italia)
Olivas Varela, José Angel (España)
Ramón, Hugo (Argentina)
Rexachs, Dolores (España)
Reyes, Nora (Argentina)
Roig Vila, Rosabel (España)
Rossi, Gustavo (Argentina)
Rosso, Paolo (España)
Rueda, Sonia (Argentina)
Russo, Claudia (Argentina)
Salto, Carolina (Argentina)
Sanz, Cecilia (Argentina)
Simari, Guillermo (Argentina)
Steinmetz, Ralf (Alemania)
Suppi, Remo (España)
Tarouco, Liane (Brasil)
Tirado, Francisco (España)
Thomas, Pablo (Argentina)
Vénere, Marcelo (Argentina)
Velho, Luiz (Brasil)
Vendrell, Eduardo (España)
Villagarcía, Horacio (Argentina)
Zanarini, Dante (Argentina)

Comité Académico

UBA – Cs. Exactas – Ceria, Santiago
UBA – Ingeniería – Echeverria, Adriana
UN La Plata – Pesado, Patricia
UN Sur – Rueda, Sonia
UN San Luis – Printista, Marcela
UN CPBA – Aciti, Claudio
UN Comahue – Grosso, Guillermo
UN La Matanza – Eterovic, Jorge
UN La Pampa – Alfonso, Hugo
UN Tierra del Fuego – Koremblit, Gabriel
UN Salta – Gil, Gustavo
UN Patagonia Austral – Lasso, Marta
UN SanJuan – Rodriguez, Nelson
UADER – Mengarelli, José Luis
UN Patagonia SJB – Buckle, Carlos
UN Entre Ríos – Tugnarelli, Mónica
UN Nordeste – Dapozo, Gladys
UN Rosario – Casali, Ana
UN Misiones – Caballero, Sergio
UN NOBA – Russo, Claudia
UN Chilecito – Carmona, Fernanda
UN Lanús – Azcurra, Diego
UN Santiago del Estero – Figueroa, Liliana
Esc. Sup. Ejército – Arroyo Arzubi, Alejandro
UN Litoral – Loyarte, Horacio
UN Rio IV – Arroyo, Marcelo
UN Córdoba – Fridlender, Daniel
UN Jujuy – Herrera Cognetta, Analía
UN Rio Negro – Vivas, Luis

UN Villa María – Prato, Laura
UN Lujan – Fernandez, Juan Manuel
UN Catamarca – Poliche Maria Valeria
UN La Rioja – Molina, Miguel
UN Tres de Febrero – Oliveros, Alejandro
UN Tucumán – Luccioni, Griselda María
UNAJ – Morales, Martín
UN Chaco Austral – Zachman Patricia
UN del Oeste – Foti, Antonio
UN de Cuyo – García Garino, Carlos
UN de Mardel Plata- Ríos, Carlos
UN de Quilmes
UN Hurlingham – Puricelli, Fernando
UN SAdA – Ramón, Hugo
UN SAM – Estayno, Marcelo
U Morón – Chapperon, Gabriela
UAI – De Vincenzi, Marcelo
U Belgrano – Guerci, Alberto
Ukennedy Panizzi, Marisa
U Adventista del Plata – Bournissen Juan
UCAECE – Malbernat, Lucía
UP alermo – Alvarez Adriana
UCA Rosario – Grieco, Sebastián
U Salvador – Zanitti, Marcelo
U Aconcagua – Giménez, Rosa
U Gastón Dachary – Ruidías, Hector Javier
UADE – Feijó, Daniel
UCEMA Guglianone, Ariadna
U Austral – Cosentino, Juan Pablo
U Atlántida Argentina – Rathmann, Liliana
U CA La Plata – Bertone, Rodolfo
ITBA – Bolo, Mario
U Champagnat – Brachetta Mariana

Comité Organizador

Responsable

Miguel Ángel Molina

Equipo de Trabajo

Mónica Giménez

María Rosa Díaz

Sandra Isabel Rodríguez

Natalia Alvarez Gómez

Isaías Díaz

Dalila Varas

Andrea Agüero

Andrea Cabañas

Claudia Inés Romero

Eduardo Escobar

Cristina Gramajo

Valeria Paez

Adriana Moreno

Gabriela Oviedo

Andrés Tejerina

Juan Pablo Millicay

Alberto Leguiza

Debora Reus

Sofía Cortez

Axel Valor

Cecilia Nuñez

Maricel Almada

Alejandro Agüero

Adrian Saenz

Autoridades REDUNCI

Coordinador Titular

Pesado Patricia (UNLP) 2020-2022

Coordinador Alterno

Estayno Marcelo (UNLZ) 2020-2022

Junta Directiva

Aciti Claudio (UNCPBA) 2021-2023

Lasso Marta (UNPA) 2021-2023

Panizzi Marisa (UK) 2021-2023

Printista Marcela (UNSL) 2020-2022

Tugnarelli Mónica (UNER) 2020-2022

Caballero Sergio (UNaM) 2020-2022

Arroyo Marcelo (UNRC) 2021-2022

Malbernat Lucia (CAECE) 2021-2022

Miembro Honorario

De Giusti Armando (UNLP)

Secretarías

Secretaría Administrativa: Jorge Eterovic

Secretaría Académica: Russo Claudia

Secretaría de Ciencia y Técnica: Rodríguez Nelson

Secretaría de Asuntos Reglamentarios: De Vincenzi Marcelo

Secretaría de Vinculación Tecnológica y Profesional: Gil Gustavo

Secretaría de Congresos, Publicaciones y Difusión: Thomas Pablo

ÍNDICE

XXIII Workshop Agentes y Sistemas Inteligentes **(WASI)**

Coordinadores

Guillermo Leguizamón (UNSL)

Carolina Salto (UNLPam)

Daniel Fridlender (UNC)

Predicción de incendios forestales mediante modelos de Machine Learning

Ana Martínez Saucedo[1] y Pablo Ezequiel Inchausti[1]

Universidad Argentina de la Empresa (UADE). Instituto de Tecnología (INTEC).
Buenos Aires, Argentina
{anmartinez,pinchausti}@uade.edu.ar

Resumen La severidad de los incendios forestales ha llegado a niveles preocupantes tanto a nivel internacional como nacional. No obstante, gracias al avance de la tecnología es posible predecir su ocurrencia y magnitud a través de modelos de Machine Learning especialmente desarrollados para tal fin. En línea con diversas investigaciones realizadas en materia de predicción espaciotemporal de incendios forestales, en el presente trabajo el objetivo fue desarrollar un modelo de Machine Learning que contribuya a la prevención de incendios forestales en el Partido de Pinamar. Para ello se entrenaron diversos modelos utilizando registros de incendios históricos de la zona, alcanzando una sensibilidad del 88.4 % para predecir la ocurrencia de incendios forestales a través de un árbol de decisión. Gracias al desarrollo de un pipeline de datos y el entrenamiento automatizado de modelos se sentaron las bases necesarias para posibilitar la predicción de incendios forestales en localidades vecinas.

Palabras claves: machine learning, aprendizaje supervisado, incendios forestales, medioambiente.

1. Introducción

Las consecuencias ambientales, económicas y sociales que los incendios forestales provocan en el mundo llevaron a gobiernos e investigadores a estudiar diversas maneras de prevenirlos, sobre todo en los últimos tiempos donde cada vez se torna más difícil controlarlos. Tan solo en el 2020 en Argentina se quemaron más de 1,1 millones de hectáreas a causa de incendios forestales. Además, se registraron en ese año más de 74.113 focos activos, una cifra récord que representa un incremento del 251,9 % con respecto al año anterior [1].

Esta tendencia nacional también se vio reflejada en distintos puntos del país. La ciudad balnearia de Pinamar, conocida por sus bosques de pino y dunas de arena, ha perdido según datos de incendios provistos por los bomberos locales más de 3,5 kilómetros cuadrados de bosques en los últimos seis años a causa de incendios forestales. Esta cifra representa aproximadamente el 10 % de la superficie total cubierta por vegetación del partido, confirmando así la tendencia creciente de incendios forestales en la zona detectada por bomberos y autoridades locales. Si bien en Argentina el 95 % de los incendios forestales son originados

por el hombre [2], su magnitud y desarrollo dependen en gran medida de las condiciones climáticas y ambientales del momento y lugar donde se desarrollan.

En la literatura se han desarrollado modelos de Machine Learning (ML) que utilizan todas o combinaciones de algunas de las variables que conforman el denominado Triángulo de Comportamiento del Fuego (TCF) para predecir incendios forestales, siendo estas la topografía, el combustible y la meteorología. En efecto, las mismas definen bajo qué condiciones es más propenso que se produzca un incendio forestal y cómo se desarrollará el mismo. Además, el Índice Meteorológico de Peligro de Incendio (FWI) [3] permite estimar la probabilidad de que se produzca un incendio a partir de variables meteorológicas y de combustible [4,5]. Asimismo, el Índice de Vegetación de Diferencia Normalizada (NDVI) ha sido utilizado en varios estudios ya que es un indicador de la salud de la vegetación [6,7]. Los investigadores han recurrido a fuentes como satélites, sensores remotos, mapas y estaciones meteorológicas para obtener estos datos y entrenar modelos predictivos. Entre los algoritmos aplicados para modelar la ocurrencia de incendios forestales se encuentran árboles de decisión, máquinas de vectores de soporte, redes neuronales artificiales y redes bayesianas [8].

No obstante, la mayor dificultad a la hora de estudiar la predicción de ocurrencia y magnitud de incendios forestales radica en la obtención de datos históricos. En Argentina son pocos los cuarteles de bomberos que cuentan con un registro informatizado de incendios, y las investigaciones realizadas en el país han recurrido al uso de datasets de otros países, como el del parque Montesinho de Portugal que abarca los incendios ocurridos en el mismo entre los años 2000 y 2003 [9]. Por ello, y en línea con diversos trabajos realizados con el fin de determinar la relación que existe entre la superficie final quemada a causa de incendios forestales y las condiciones ambientales circundantes, el presente trabajo tuvo como objetivo desarrollar y evaluar modelos de ML para predecir la ocurrencia y magnitud de incendios forestales en la ciudad de Pinamar, provincia de Buenos Aires, utilizando datos de incendios locales recopilados por bomberos de la zona y no sólo variables meteorológicas, sino también topográficas y de combustible.

El presente artículo se organiza de la siguiente manera: en la sección 2 se delimita el área de estudio del trabajo y las variables explicativas utilizadas. En la siguiente sección 3 se describen los distintos algoritmos de ML utilizados y las métricas empleadas para su evaluación. Luego, en la sección 4 se explica la metodología adoptada en el trabajo. A continuación, en la sección 5 se expone y analiza el rendimiento de los modelos entrenados. Por último, en la sección 6 se resumen los resultados obtenidos y se presentan futuras líneas de investigación.

2. Área de estudio y datos utilizados

El partido de Pinamar está ubicado en la zona sudeste de la provincia de Buenos Aires y se caracteriza por una gran predominancia de coníferas implantadas a lo largo de 40 kilómetros cuadrados, representando el 63 % de la superficie total del partido. Según los datos provistos los bomberos locales, los incendios forestales de mayor magnitud (>1 hectárea) ocurren en primavera y verano durante la franja horaria de 8 de la mañana a 3 de la tarde. Coincidentemente, las

zonas donde mayor cantidad de incendios forestales han ocurrido corresponden a aquellas donde la densidad poblacional o concentración de actividades turísticas en temporada alta y fines de semana largos es mayor.

Esta información permitió determinar qué variables pueden explicar la ocurrencia de incendios forestales en Pinamar. Para construir el dataset se desarrolló un pipeline de datos automatizado que extrae, transforma y preprocesa las variables meteorológicas, topográficas y de combustible como el FWI, el código de sequía (DC), el código de humedad del mantillo (DMC), la velocidad de propagación del incendio (ISI) y el combustible disponible (BUI) (Tabla 1) para un rango de fechas y par de coordenadas establecidas. En el presente estudio el dataset creado corresponde a los incendios forestales ocurridos en el área descrita por las coordenadas 56°57′ O, 37°12′ S, 56°48′ O, 37°3′ S y desde el 01/01/2015 hasta el 01/01/2020.

Tabla 1. Descripción de variables utilizadas en el dataset creado.

Descripción	**Variables**	**Origen**	**Resolución espacial**	**Resolución temporal**	**Cobertura geográfica**
Incendios forestales	Día Mes Día no laboral Hora X, Y	Asociación Bomberos Voluntarios de Pinamar	N/A	Horaria	Partido de Pinamar
Meteorología	Temperatura Humedad Viento Precipitaciones	Servicio Meteorológico Nacional	N/A	Horaria	Partidos de Pinamar y Villa Gesell
Temperatura de superficie de suelo (LST)	LST	NASA / MYD11C1 v006 [10]	0.05°× 0.05°	Diaria	Global
Índice de Vegetación de Diferencia Normalizada	NDVI	NASA / MYD13Q1 v006 [11]	250 metros × 250 metros	Quincenal	Partido de Pinamar
Índice meteorológico de peligro de incendio e índices derivados	DC DMC FFMC ISI BUI FWI	NASA / GFWED GEOS-5 - GPM Late v5 [12]	0.1° × 0.1°	Diaria	60° S - 60° N

3. Métodos

Tal como se describió anteriormente, en la literatura se ha abordado la predicción de ocurrencia y magnitud de incendios forestales utilizando diversos algoritmos. En particular, en este estudio se han implementado algunos de ellos para la predicción de tanto la ocurrencia de incendios forestales (clasificación),

como de la superficie final quemada (regresión). A su vez se han variado los hiperparámetros utilizados en cada modelo hasta encontrar una configuración que provea los mejores resultados. A continuación se describe brevemente cada uno.

3.1. Regresión logística o *Logistic Regression* (LR)

Los modelos de LR explican la relación que existe entre una variable dependiente y varias independientes o explicativas. Este es uno de los algoritmos de clasificación supervisados más sencillos y utilizados en ML. Los modelos de LR son lineales, por lo que se obtienen buenos resultados cuando los datos son linealmente separables a través de una frontera de decisión. No obstante, para modelos más complejos en donde hay múltiples fronteras de decisión los modelos de LR no pueden capturar la complejidad de las relaciones en los datos [13].

En este trabajo se aplicó la regularización L2 para evitar que el modelo de LR se sobreajuste a los datos de entrenamiento y se varió el hiperparámetro C en 100 valores espaciados en el intervalo comprendido entre 0.00002 y 1.

3.2. Máquinas de Vectores Soporte o *Support Vector Machines* (SVM)

SVM busca determinar los puntos (o vectores de soporte) que separan al máximo dos clases, aunque se puede generalizar a múltiples clases. Para ello cada ejemplo se representa como un vector en un espacio n-dimensional, donde n es la cantidad de atributos o características. Los vectores de soporte definen de esta forma un hiperplano que separa los datos linealmente [14].

Entre los hiperparámetros que se calibraron para tanto el clasificador como el regresor se encuentran la función kernel (lineal, polinómico y *Radial Basis Function* o RBF), el valor gamma del kernel (4 valores espaciados entre 0.0001 y 1) y C (10 valores espaciados entre 0.1 y 10000).

3.3. Árboles de decisión o *Decision Tree* (DT)

Un DT se representa mediante un árbol binario donde cada nodo representa un atributo de entrada. En caso de que la variable de entrada sea numérica, el nodo también es un punto de división sobre esa variable. Por otro lado, las hojas del árbol representan las salidas, que pueden ser tanto etiquetas para problemas de clasificación como un valor continuo para problemas de regresión. El objetivo del mismo es capturar las relaciones entre las variables de entrada y salida mediante el árbol más pequeño posible, de forma tal de evitar caer en un estado de sobreajuste de los datos de entrada [15].

En el presente estudio se variaron hiperparámetros como la profundidad máxima del árbol (10, 6, 3, sin límite), la cantidad máxima de características a considerar (33, 26, 20, 10 o 5), el número mínimo de ejemplos en el cual un nodo se considera hoja (1, 3, 5, 9 o 10), y el criterio de ganancia de información (Gini y Entropía en el caso de los clasificadores, y MAE y Friedman para los regresores).

3.4. Redes Neuronales Artificiales o *Artificial Neural Networks* (ANN)

Una ANN es un modelo que simula el comportamiento de una red neuronal biológica. La misma está compuesta por capas de nodos, particularmente una capa de entrada, una o más capas ocultas y una capa de salida. Cada nodo está conectado con otro, y posee un peso y umbral, especificando si el mismo está activado y envía datos a la siguiente capa de la red. Por el contrario, si está desactivado no se envían datos a la siguiente capa [16].

La arquitectura de red adoptada en los experimentos llevados a cabo es de dos capas ocultas, variando la cantidad de nodos de cada una entre 4 y 25 según la experiencia manifestada en la literatura [6,17]. Otros hiperparámetros ajustados fueron la tasa de aprendizaje α (5 valores espaciados entre 0.00001 y 0.001), la dilución (6 valores espaciados entre 0.15 y 0.5) y la función de activación (ReLu, Tangente Hiperbólica y Sigmoide).

3.5. Métodos de ensamble

Los métodos de ensamble de algoritmos de ML combinan las predicciones de múltiples modelos con el fin de obtener predicciones más precisas. Entre las técnicas de ensamble más comunes se encuentran:

Bagging: se construyen independientemente varios estimadores (ya sean clasificadores o regresores) utilizando el mismo algoritmo y distintos subconjuntos del conjunto de datos de entrenamiento. Una vez entrenados, las predicciones de cada modelo se promedian (en problemas de regresión) o votan (en problemas de clasificación) [18]. En este trabajo se utilizó el modelo de bosques aleatorios o *Random Forests* (RF) que conforman un ensamble de distintos árboles de decisión. Para ello se ajustaron hiperparámetros como la profundidad máxima de cada estimador (80, 90, 100, 110, 120 o 150), la cantidad máxima de características a considerar (4, 6, 12, 24 o 33), el número mínimo de ejemplos para dividir un nodo interno (8, 10 o 12), el número mínimo de ejemplos para determinar si un nodo es hoja (3, 4 o 5), y la cantidad de árboles en el bosque (100, 200, 500 o 1000).

Boosting: se desarrollan secuencialmente varios *weak learners* o modelos sencillos, buscando corregir en cada paso los errores del modelo predecesor. Un ejemplo de esta técnica es el algoritmo *Gradient Boosted Trees* (GB), en el que varios árboles de decisión individuales simples y con pocas ramificaciones se ajustan secuencialmente [19]. Además de los hiperparámetros ajustados en el modelo de DT, en GB se han ajustado la tasa de aprendizaje (0.01, 0.025, 0.05, 0.075, 0.1, 0.15 o 0.2), la función de calidad de división (Friedman, RMSE), y la cantidad de etapas de boosting (10, 50, 100, 200, 500 o 1000).

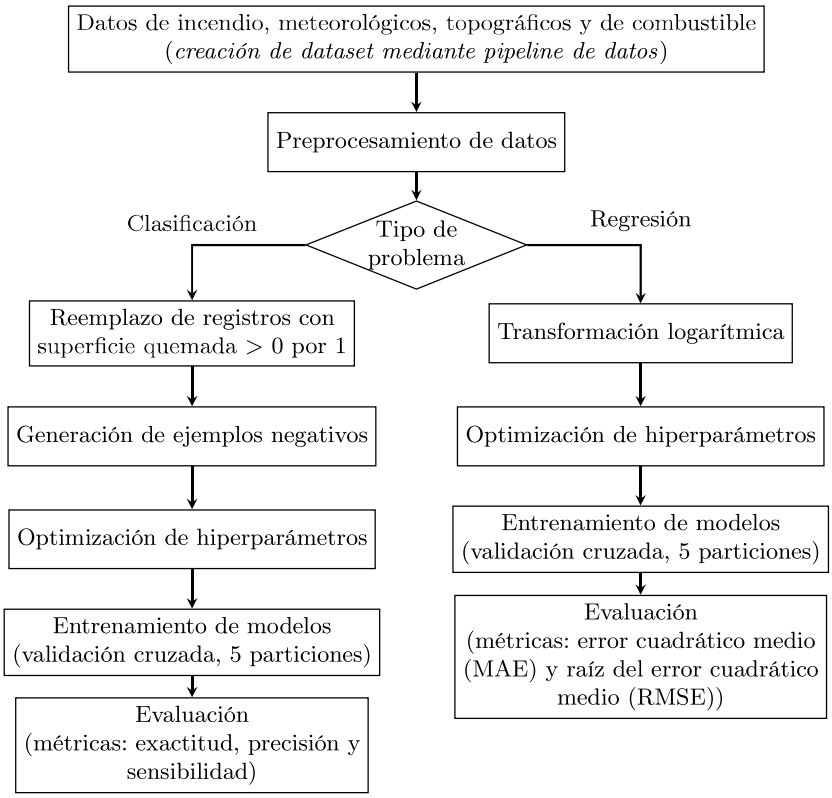
4. Metodología de modelado

La metodología adoptada en el presente trabajo se detalla en la Figura 1. En primer lugar se generó el dataset de incendios forestales de los años 2015 a 2020 para el área de interés. El mismo posee un total de 750 registros, aunque luego de la etapa de preprocesamiento sólo 597 se utilizaron para entrenar modelos

mientras que los restantes 163 registros se descartaron por no contar con alguno de los atributos de la Tabla 1 requeridos. Entre las tareas de preprocesamiento que se llevaron a cabo se destacan:

- La conversión de coordenadas geográficas (latitud y longitud) a coordenadas (X, Y) de 250 metros x 250 metros.
- La aplicación de técnicas de escalado de datos como *MinMax*.
- La limpieza de registros con datos meteorológicos, topográficos o de combustible faltantes.
- La conversión de variables categóricas a numéricas.

Figura 1. Diagrama de flujo de la metodología propuesta.



La experiencia manifestada en la literatura y en el presente trabajo demuestran que la distribución de la magnitud de incendios forestales es altamente despareja: la mayoría de los incendios forestales ocurridos en Pinamar han sido pequeños, aunque representan el 37 % de la superficie quemada históricamente. Por el contrario, el 61 % de la superficie quemada fue producida por el 6 % del total de incendios forestales. Esta disparidad afecta directamente el rendimiento de modelos que, en este tipo de distribuciones, se verán influenciados por lo que se

refleja en la mayoría de los casos, ya sea la cantidad de incendios o el promedio de superficie quemada.

Con el fin de abordar este problema, y en línea con trabajos como el de [20], el enfoque adoptado es el de dividir en dos pasos la predicción de incendios forestales: en primer lugar, se predice por cada coordenada de la grilla si se producirá o no un incendio forestal (clasificación binaria); y en segundo lugar, para aquellas coordenadas en las que se haya predicho la ocurrencia de un incendio forestal (con una probabilidad mayor a 0.5), se predicen las hectáreas que podrían quemarse (regresión). A partir de esta división, dependiendo del tipo de problema los pasos en la metodología difieren. Por un lado, para el problema de clasificación se reemplazan los registros cuyo valor de superficie quemada es mayor a 0 por 1. Luego, como los datos que se cuentan corresponden enteramente a la clase 1 (Incendio), se genera aleatoriamente una cantidad proporcionada de registros de la clase 0 (No Incendio) (Algoritmo 1) en concordancia con algoritmos propuestos en la literatura [21]. El objetivo de este tipo de algoritmos es que las ubicaciones de los puntos en donde se produjeron incendios y donde no estén espacial y temporalmente relacionados con los atributos. Por otro lado, para abordar el problema de regresión se aplica una transformación logarítmica sobre la variable objetivo (superficie quemada) para reducir la asimetría de la distribución. A continuación, en ambos problemas se optimizan los hiperparámetros de los distintos algoritmos seleccionados según lo descrito en la Sección 3 utilizando la técnica de búsqueda exhaustiva [22], y se entrenan modelos aplicando la técnica de validación cruzada utilizando 5 particiones sobre el conjunto de datos de entrenamiento. Por último, se evalúan los modelos entrenados a través de las métricas correspondientes al tipo de problema.

Algoritmo 1 Generación de ejemplos negativos ("no incendio")

$dias \leftarrow 3$
$kilometros \leftarrow 1{,}5$
$P \leftarrow \varnothing$ ▷ Puntos de "no incendio"
for $puntoDeIncendio$ **in** $incendios$ **do**
    $fechaIncendio \leftarrow incendio.fecha$
    $incendiosRecientes \leftarrow seleccionarIncendiosEnLapso(dias, fechaIncendio)$
    $R \leftarrow \varnothing$ ▷ Región donde se produjeron incendios
    for $incendioReciente$ **in** $incendiosRecientes$ **do**
        $coordenadas \leftarrow incendioReciente.coordenadas$
        $R \leftarrow R \cup crearRegionEn(kilometros, coordenadas)$
    end for
    $P \leftarrow P \cup seleccionarCoodenadasEnArea(R')$
end for

5. Resultados y discusión

Los resultados obtenidos en los modelos de clasificación y regresión se presentan en las Tablas 2 y 3 respectivamente. En lo que respecta a la predicción de incendios forestales (problema de clasificación), el modelo con mayor exactitud es el de RF con un valor del 82.4 %. Coincidentemente este modelo fue también el que obtuvo el valor de precisión más alto (82 %). No obstante, el modelo con

mayor sensibilidad es el de DT con un 88.4 %, convirtiéndolo en aquel que mejor distingue la ocurrencia de incendios. Tanto los modelos correspondientes a LR como SVM han tenido un rendimiento inferior en comparación al resto de los modelos.

La disparidad observada en el rendimiento de los distintos modelos puede tener relación con la complejidad y alta dimensionalidad del dataset, sumado a que la distribución de los datos no es normal. En efecto, DT es un algoritmo no paramétrico que no asume la distribución de los datos. No obstante, los algoritmos no paramétricos requieren datasets con mayor cantidad de registros, por lo que los resultados podrían mejorar en caso de añadir al dataset los incendios forestales registrados en los últimos dos años.

Tabla 2. Resultados de la evaluación de modelos de predicción de incendios forestales.

Modelo	Exactitud	Precisión	Sensibilidad	F1 score
ANN	74.2 %	71.6 %	79.3 %	75.3 %
SVM	61.5 %	62.4 %	56.2 %	59.1 %
GB	77.9 %	77.7 %	77.7 %	77.7 %
LR	62.3 %	63.6 %	56.2 %	59.7 %
DT	80.7 %	76.4 %	**88.4 %**	82 %
RF	**82.4 %**	**82 %**	82.6 %	**82.3 %**

En lo que respecta a modelos de predicción de superficie quemada, los mejores modelos en términos de valores de MAE y RMSE son ANN. Particularmente, el MAE más bajo corresponde a un valor de 0.255 mientras que el mejor RMSE es de 0.178. No obstante, los hiperparámetros de ambas redes difieren entre sí: mientras la red n°1 tiene 18 nodos en la capa oculta n°1 y 4 en la n°2, la red n°2 posee 9 y 18 respectivamente. Por otro lado, la red n°1 utiliza como función de activación la tangente hiperbólica, mientras que la n°2 utiliza la ReLu. También se puede observar que el RMSE obtenido a partir de los modelos de ANN es considerablemente mejor a los demás modelos.

En este sentido, una de las ventajas de las ANN radica en que su rendimiento no se ve alterado por altas correlaciones entre las variables de entrada. Efectivamente, a partir del análisis de datos previo al entrenamiento de modelos se observó una tendencia de que los incendios de mayor magnitud se producen en las primaveras y los veranos (correlación temporal). Asimismo, las localidades donde han ocurrido más incendios forestales corresponden a Ostende y Valeria del Mar, donde la densidad poblacional es mayor (correlación geográfica).

Luego del entrenamiento de estos modelos se pudo analizar cuáles son las variables que demostraron ser importantes a la hora de explicar la ocurrencia de incendios forestales en Pinamar. Correspondientemente a la experiencia manifestada por bomberos locales, la ubicación es un factor importante ya que los incendios son más comunes en zonas de alta densidad poblacional. En efecto, la variable longitud tuvo una importancia del 83 %. En segundo lugar, la variable elevación resultó en una importancia del 8 %. Por último, el NDVI obtuvo una importancia del 6 %. Estos resultados demuestran que el componente humano es el más importante a la hora de predecir incendios en Pinamar.

Tabla 3. Resultados de la evaluación de modelos de predicción de superficie quemada.

Modelo	MAE	RMSE
ANN (n°1)	0.295	**0.178**
ANN (n°2)	**0.255**	0.215
SVM	0.274	0.438
DT	0.294	0.46
RF	0.299	0.448

6. Conclusiones

En este estudio se desarrollaron y entrenaron varios modelos de ML para predecir tanto la ocurrencia como magnitud de incendios forestales en Pinamar. Estas predicciones pueden ser de gran importancia para los bomberos, ya que posibilitan estimar tempranamente los recursos humanos y materiales que deben ser empleados para combatir incendios forestales lo más rápido posible y así evitar pérdidas materiales, ecológicas y humanas. Los resultados obtenidos son variados, no obstante los mejores modelos alcanzaron una sensibilidad del 88.4 % (DT, clasificación) y un RMSE del 0.178 (RNA, regresión).

Como futuras líneas de investigación se contempla variar la arquitectura de las RNA para disminuir el RMSE asociado a la magnitud de incendios forestales. Asimismo, incrementar el tamaño del dataset para que incluya una cantidad balanceada de incendios de baja y alta magnitud podría mejorar el rendimiento de los modelos, ya que los datos utilizados en este estudio tienen una distribución despareja. En este sentido, gracias a la automatización de la metodología propuesta realizar estas tareas insumiría poco tiempo, dando lugar a la posibilidad de realizar experimentos más extensos. No obstante, resulta fundamental contar con un sistema informático que permita registrar fácilmente los incendios forestales ocurridos para poder refinar los modelos obtenidos.

Agradecimientos Los autores agradecen a la Universidad Argentina de la Empresa (UADE) y al Instituto de Tecnología (INTEC) por el apoyo brindado en el presente trabajo realizado en el marco del Proyecto Final de Ingeniería en Informática "AQUA: Desarrollo de un modelo de Machine Learning para prevenir incendios forestales en Pinamar", articulado en el ACyT "Aplicaciones de Machine Learning para mejorar el uso de Recursos Naturales" (A21T03).

Referencias

1. Instituto Nacional de Pesquisas Espaciais: Monitoramento dos Focos Ativos por País - Programa Queimadas - INPE, `https://queimadas.dgi.inpe.br/queimadas/portal-static/estatisticas\_paises`, last accessed 8 May 2021
2. Argentina.gob.ar: Causas de los incendios forestales, `https://www.argentina.gob.ar/sinagir/incendio-forestal/causas`, last accessed 28 Mar 2021
3. Taylor, S.W., Alexander, M.E., Taylor, S.W., Alexander, M.E.: Science, technology, and human factors in fire danger rating: the Canadian experience. Int. J. Wildland Fire. 15, 121–135 (2006). https://doi.org/10.1071/WF05021
4. Xie, Y., Peng, M.: Forest fire forecasting using ensemble learning approaches. Neural Computing and Applications. 31 (2019). https://doi.org/10.1007/s00521-018-3515-0
5. Castelli, M., Vanneschi, L., Popovič, A.: Predicting Burned Areas of Forest Fires: an Artificial Intelligence Approach. fire ecol. 11, 106–118 (2015). https://doi.org/10.4996/fireecology.1101106

6. Jafari Goldarag, Y., Mohammadzadeh, A., Ardakani, A.S.: Fire Risk Assessment Using Neural Network and Logistic Regression. J Indian Soc Remote Sens. 44, 885–894 (2016). https://doi.org/10.1007/s12524-016-0557-6
7. Sayad, Y.O., Mousannif, H., Al Moatassime, H.: Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Safety Journal. 104, 130–146 (2019). https://doi.org/10.1016/j.firesaf.2019.01.006.
8. Jain, P., Coogan, S.C.P., Subramanian, S.G., Crowley, M., Taylor, S., Flannigan, M.D.: A review of machine learning applications in wildfire science and management. Environ. Rev. 28, 478–505 (2020). https://doi.org/10.1139/er-2020-0019
9. Cardenas, M., Castillo, J., Medel, R., Casco, O., Navarro, M., Gutierrez, S., Curti, A.: Sistema de predicción de incendios forestales para la provincia de Córdoba. Presented at the Congreso Nacional de Ingeniería en Informática / Sistemas de información , Salta (2016).
10. Wan, Zhengming, Hook, Simon, Hulley, Glynn: MYD11C1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006, https://lpdaac.usgs.gov/products/myd11c1v006/, (2015). https://doi.org/10.5067/MODIS/MYD11C1.006
11. Didan, Kamel: MYD13Q1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V006, https://lpdaac.usgs.gov/products/myd13q1v006/, (2015). https://doi.org/10.5067/MODIS/MYD13Q1.006
12. Field, R.D., Spessa, A.C., Aziz, N.A., Camia, A., Cantin, A., Carr, R., de Groot, W.J., Dowdy, A.J., Flannigan, M.D., Manomaiphiboon, K., Pappenberger, F., Tanpipat, V., Wang, X.: Development of a Global Fire Weather Database. Natural Hazards and Earth System Sciences. 15, 1407–1423 (2015). https://doi.org/10.5194/nhess-15-1407-2015.
13. Hosmer, D.W., Lemeshow, S.: Introduction to the Logistic Regression Model. In: Applied Logistic Regression. p. 1. John Wiley & Sons (2004).
14. Luger, G.: Machine Learning: Connectionist. In: Artificial Intelligence: Structures and Strategies for Complex Problem Solving. pp. 482–484. , Boston (2008).
15. Brownlee, J.: Classification and Regression Trees. In: Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. pp. 72–74. Machine Learning Mastery (2016).
16. Haykin, S.: Introduction. In: Neural Networks and Learning Machines. pp. 21–24. New York (2008).
17. Vega-Garcia, C.: Applying neural network technology to human-caused wildfire occurrence prediction. AI Applications. 10, 9–18 (1996).
18. Casal, R.F., Bouzas, J.C., Fuente, M.O. de la: 3.1 Bagging. Aprendizaje Estadístico. (2021)
19. Casal, R.F., Bouzas, J.C., Fuente, M.O. de la: 3.4 Boosting. Aprendizaje Estadístico. (2021)
20. Wang, Q., Zhang, J., Guo, B., Hao, Z., Zhou, Y., Sun, J., Yu, Z., Zheng, Y.: CityGuard: Citywide Fire Risk Forecasting Using A Machine Learning Approach. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 3, 1–21 (2019). https://doi.org/10.1145/3369814.
21. Stojanova, D., Kobler, A., Ogrinc, P., Ženko, B., Džeroski, S.: Estimating the risk of fire outbreaks in the natural environment. Data Min Knowl Disc. 24, 411–442 (2012). https://doi.org/10.1007/s10618-011-0213-2.
22. Krauß, J.: Fundamentals in the Selection of Hyperparameter Optimization Techniques. In: Optimizing Hyperparameters for Machine Learning Algorithms in Production. pp. 41–42. Apprimus Wissenschaftsverlag (2022).

Software inteligente para la digitalización de placas espectroscópicas

Franco Ronchetti[1,4] Facundo Quiroga[1,5] Nehuén Pereyra[1] Joaquín Miranda[1] Santiago Ponte[1] Yael Aidelman[2,3] Roberto Gamen[2,3] and Laura Lanzarini[1]

[1] Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata (UNLP), La Plata, Argentina
[2] Facultad de Ciencias Astronómicas y Geofísicas, UNLP, La Plata, Argentina
[3] Instituto de Astrofísica de La Plata, CONICET–UNLP, La Plata, Argentina
[4] Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC-PBA), La Plata, Argentina
[5] Becario Postdoctoral UNLP

{fronchetti, fquiroga}@lidi.info.unlp.edu.ar

Resumen La Facultad de Ciencias Astronómicas y Geofísicas (FCAG) de la Universidad Nacional de La Plata (UNLP) tiene un rico historial de observaciones astronómicas, geofísicas y meteorológicas. En particular, su colección de observaciones incluye 15.000 registros espectroscópicos en placas de vidrio obtenidas entre 1920 y 1980. La recuperación de los registros espectroscópicos implica un proceso complejo que incluye varias etapas: la recopilación de las placas fotográficas y sus respectivos metadatos; el escaneo de las placas y su conversión a archivos con formatos útiles a la astronomía y la extracción de los espectros y su calibración en longitud de onda.
Dada la cantidad de registros y las dificultades de su procesamiento, usando métodos manuales la recuperación del patrimonio completo llevaría una cantidad de tiempo prohibitiva y podría introducir diferencias en la sistematización y el análisis de los registros.
En este artículo presentamos un software inteligente para ayudar en la recuperación de la información contenida en las placas. El software contiene un módulo que detecta de forma automática los espectros, agiliza la carga de los metadatos y verifica su estandarización. El uso del software permitirá reducir significativamente las horas-persona necesarias y los errores de procesamiento.

Keywords: Espectros · YOLO · Placa espectroscópica · ReTrOH · Universidad Nacional de La Plata · Astronomía

1. Introducción

En el año 2000 la Unión Astronómica Internacional (IAU) emitió una resolución (Nro. B3 *Safeguarding the information in photographic plates*) [11], donde

solicita que se tomen medidas para conservar los datos históricos de todas las fuentes astronómicas, ya que, de no hacerlo, se perderán para las futuras generaciones de astrónomos. Agrega, además, que se debe procurar la transferencia de los datos históricos a medios modernos, los cuales deberán proveer su acceso a toda la comunidad internacional para el bien de toda la investigación astronómica. Desde ese entonces, en distintas partes del mundo se ha comenzado con la labor de digitalización de placas fotográficas, por ejemplo: la colección de placas astronómicas de Harvard[6] o la colección del Observatorio Maria Mitchell[7], por mencionar algunas. Particularmente en Argentina, por esos años, se comenzaron a digitalizar placas fotográficas creando el Primer archivo digital de placas fotográficas del Observatorio Astronómico de Córdoba [1].

El interés por la digitalización no es solo histórico; a partir de estos procesos de recuperación se han obtenido resultados importantes. Por ejemplo, se detectaron variaciones de largo período en blazares [13], se obtuvieron posiciones y magnitudes estelares [7], se detectaron y clasificaron nuevos objetos [5], se construyeron curvas de luz de estrellas T Tauri [4], se descubrieron eventos eruptivos en la estrella Variable Luminosa Azul R71 ocurridos a comienzos del siglo pasado y nunca reportados [12], se detectó un cúmulo abierto entre las estrellas de campo de la región de Collinder 132 y se calcularon el movimiento propio medio y las probabilidades de pertenencia de las estrellas de la región [8], por mencionar algunos.

La FCAG cuenta con una gran colección de placas espectrográficas y fotográficas en formato de vidrio (Figura 1). Estas observaciones fueron realizadas entre las décadas del '20 y del '80 por renombrados astrónomos argentinos como R. Barbá, V. Niemela, E. Brandi, L. García, O. Ferrer, N. Morrell, H. Levato, J. Sahade y A. Ringuelet entre otros. Se estima que en la FCAG hay más de 15.000 espectros registrados en placas, tomados con instrumentos instalados en el Observatorio de La Plata (OALP), el Observatorio Astronómico de Córdoba (OAC) y el Observatorio Interamericano de Cerro Tololo (CTIO) en Chile [3].

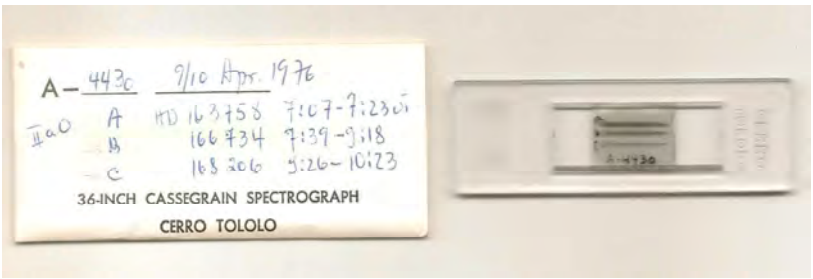


Figura 1. Placa espectroscópica, conteniendo tres espectrogramas, junto a sus anotaciones en formato papel.

[6] http://dasch.rc.fas.harvard.edu/project.php
[7] https://www.mariamitchell.org/astronomical-plates-collection

Sin embargo, la forma en que estos datos están disponibles no es útil, ya que no permite acceder a ellos con herramientas y software modernos. Además, el instrumental destinado a estos procesos ya no se encuentra operativo. Por esta razón, el 1 de mayo de 2019 se institucionalizó en la FCAG la creación de un repositorio científico para la preservación de este acervo histórico-científico-cultural mediante el llamado proyecto de Recuperación del Trabajo Observacional Histórico (ReTrOH)

Junto al Nuevo Observatorio Virtual Argentino (NOVA[8]), con sede en la FCAG, se ha comenzado el proceso de recuperación de datos históricos utilizando un escáner Nikon 9000ED. Hasta hoy se han digitalizado alrededor de 250 placas que se encuentran disponibles en el Repositorio Institucional del Servicio de Difusión de la Creación Intelectual (SeDiCI) de la UNLP. Se han analizado de forma manual dos espectros, que fueron procesados y analizados para validar la calidad de la recuperación de los mismos con fines científicos. Estos resultados dieron lugar a una tesis de licenciatura [6]. Si bien el proceso fue exitoso en cuanto a la validez del producto científico final, i.e. espectro digital, se concluyó que tal procedimiento no podría extrapolarse a las miles de placas existentes debido a la gran cantidad de tiempo necesario. Además, la naturaleza manual del procesamiento puede introducir errores sistemáticos o criterios subjetivos de procesamiento.

Por otro lado, los métodos de Inteligencia Artificial permiten automatizar varias de las tareas repetitivas comúnmente realizadas por humanos, particularmente aquellos basados en datos como el Aprendizaje Automático. Las Redes Neuronales son los modelos de aprendizaje automático con mejor desempeño en la actualidad en una gran variedad de problemas, particularmente en áreas de Visión por Computadora. En los últimos años, se ha conseguido entrenar Redes Neuronales con múltiples capas mediante un conjunto de técnicas que suelen denominarse Aprendizaje Profundo (*Deep Learning*) [2].

En este artículo presentamos un sistema que permite agilizar y automatizar el proceso completo de digitalización de estas placas. El objetivo del sistema es detectar los diferentes espectros que se encuentran en cada placa, recortarlos y exportarlos al formato FITS (por sus siglas en inglés *Flexible Image Transport System*), estándar en astronomía. La detección se realizó con un modelo de *Deep Learning* entrenado específicamente para el problema. El sistema además permite ingresar y verificar los metadatos asociados a cada espectro de la placa.

En el resto de esta sección, introducimos conceptos básicos sobre las placas espectrográficas. En la sección 1.1, detallamos los aspectos principales del desarrollo del software, dejando para la sección 3 los detalles de la detección automática de espectros. La sección 4 presenta las conclusiones y el trabajo futuro esperado.

[8] https://nova.conicet.gov.ar

1.1. Placas Espectroscópicas

En tiempos previos a la era digital, se utilizaban diversos sistemas para registrar las observaciones. Los más comunes eran las placas fotográficas. Estas consisten de una base de vidrio con una emulsión fotosensible adherida a una de sus caras. Mediante efecto fotoquímico, la emulsión es capaz de reaccionar a los fotones, portadores de información astrofísica. La intensidad de la señal queda visibilizada en el oscurecimiento de la emulsión. La figura 1 muestra un ejemplo de estas placas junto con anotaciones manuscritas realizadas por el observador.

Las placas fotográficas se utilizaron desde fines del siglo XIX hasta bien entrada la década del '80 del siglo pasado y si bien actualmente se encuentran obsoletas, es importante remarcar su simplicidad y el gran tamaño que pueden llegar a alcanzar (característica ideal para realizar relevamientos de grandes áreas del cielo).

Generalmente, las placas se encuentran almacenadas dentro de sobres de papel. En su mayoría, los mismos cuentan con la información de la observación como el nombre del objeto observado, nomenclatura de placa, fecha y hora de observación, tipo de lámpara de comparación, tipo de emulsión, observador, observatorio e instrumental utilizado, etc.

Espectros El objetivo general de la digitalización de las placas consiste en tener los espectros de ciencia disponibles para la comunidad científica. La figura 2 muestra un ejemplo de un espectro digitalizado y recortado. Un espectro es una imagen en la cual se observa la cantidad de energía recibida por longitud de onda. La manera de obtener una imagen de este estilo consiste en hacer incidir la luz sobre un objeto dispersor, como por ejemplo un prisma o una red de difracción, logrando así descomponer la luz "blanca" en todos sus "colores" (diferentes longitudes de onda).

Este "arco iris" así formado, contiene información sobre el emisor de la luz, como su temperatura y los elementos químicos que lo componen, y se denomina espectro. Los espectros no son una mera secuencia de colores sino que en ellos se pueden identificar una serie de líneas oscuras superpuestas. Estas líneas son las que contienen la información sobre los elementos químicos presentes, ya que cada elemento de la tabla periódica tiene asociada una serie de líneas particulares y exclusivas (como las huellas digitales en los seres humanos). El espectro de una estrella trae consigo esas huellas.

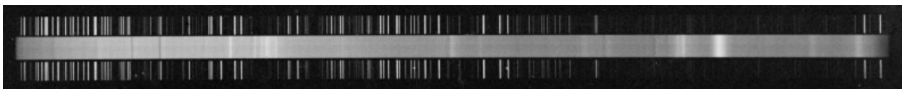


Figura 2. Espectro digitalizado. En la región central se observa el espectro de ciencia acompañado de los espectros de la lámpara de comparación (arriba y abajo).

Los espectros de los objetos de estudio adquiridos, llamados espectros de ciencia, siempre van acompañados de "espectros de comparación". Estos últimos corresponden a un espectro de una lámpara con gases cuya secuencia de líneas es bien conocida (generalmente se utilizan lámparas de hierro, de neón y argón, o de helio-neón-argón, entre otras) lo que permite relacionar sus posiciones en la placa con la longitud de onda que les corresponde. Este proceso se denomina calibración en longitud de onda. En la figura 2 se puede observar un espectro estelar en la región central[9], y arriba y abajo, los espectros de la lámpara de comparación (espectro de líneas brillantes).

2. Software desarrollado

El software desarrollado permite al equipo científico realizar las digitalizaciones de las placas de un modo más ágil y ordenado. Está disponible públicamente para su uso, y su desarrollo es bajo el modelo de software libre [10]

El objetivo del sistema es que una vez generado un archivo imagen con la placa digitalizada, el sistema permita el recorte de los diversos espectros que haya y la carga de los metadatos de cada uno. Cada espectro debe ser exportado en formato FITS. El software posee los siguientes componentes:

- Frontend HTML/Javascript desarrollado con el framework Svelte que puede ejecutarse en cualquier navegador web. Este permite la definición de un directorio de trabajo donde es posible seleccionar la imagen a procesar. Se visualizan los diferentes espectros detectados en la imagen y permite la carga de los metadatos correspondientes a cada uno (ver figura 3).
- Backend Python desarrollado con el framework Flask que realiza el vínculo entre la detección de los espectros y la aplicación web.
- Detector de espectros (ver sección 3) que permite detectar automáticamente los espectros en la imagen digitalizada, con el consiguiente ahorro del tiempo de procesamiento de las imágenes.

En el frontend, el sistema permite la definición de un directorio de trabajo, donde se podrá elegir cada imagen a tratar. El sistema persiste el estado en el que se encuentra cada archivo del directorio: sin procesar, en curso, o terminado. Al estar en curso, es posible recargar la sesión de trabajo del archivo, junto con los espectros detectados y los metadatos que se hayan cargado de ellos. El sistema permite además calibrar el brillo, contraste y filtro de color de la imagen cargada, para su mejor manipulación. Estos cambios no se ven reflejados en el recorte final de cada objeto.

Un aspecto fundamental en el sistema es el manejo adecuado de los metadatos. Por un lado, los archivos FITS que se generan deben tener ciertos metadatos convenidos mundialmente [9], incluso si no se tiene información sobre estos para

[9] En este caso particular se observa un espectro combinado, con líneas brillantes (más blancas) y líneas oscuras (más negras).

[10] https://github.com/midusi/spectrogram

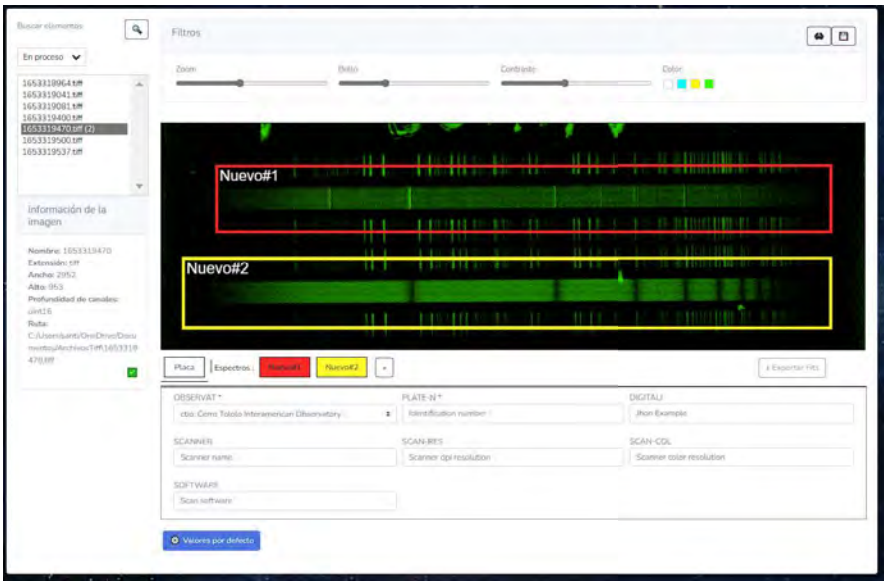


Figura 3. Frontend del sistema desarrollado. A la izquierda puede verse el directorio de trabajo e información del archivo. A la derecha puede verse la imagen escaneada junto con los espectros detectados y la carga de los metadatos.

un espectro particular. Por otro lado, el sistema permite diferenciar entre dos tipos de metadatos, los referentes a la placa donde estaba el espectro y los referentes al objeto observado. La información de la placa debe ser replicada en cada archivo FITS, ya que es común a todos los espectros (por ejemplo, número de placa, nombre del observador, lugar de observación, entre otros). Por ende, el sistema tiene facilidades para realizar esta carga de metadatos de forma completa de manera manual.

No obstante, para alivianar la carga, también permite buscar en forma automática en la base de datos SIMBAD[11] en el caso en que el objeto a cargar sea conocido previamente. Para esto, se debe indicar tres campos obligatorios: nombre del objeto, día de observación y la hora en formato universal. Con esta información el sistema puede completar, en la mayoría de los casos, una gran parte de los metadatos como el tipo de objeto y sus coordenadas.

3. Detección automática de Espectros

Con el fin de agilizar el trabajo del profesional que debe digitalizar las placas, desarrollamos un detector automático de espectros dentro de la placa. Para esto, confeccionamos una base de datos y entrenamos un modelo Convolucional, como se menciona en las siguientes secciones.

[11] http://simbad.cds.unistra.fr/simbad

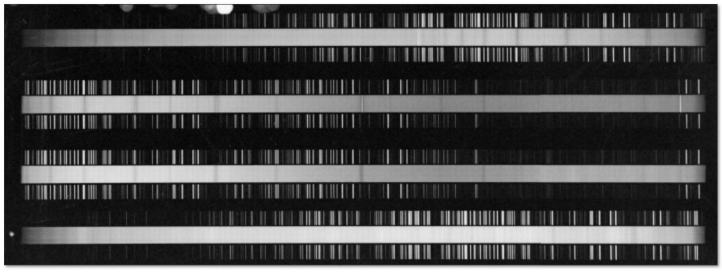


Figura 4. Ejemplo de placa digitalizada con cuatro espectros distintos.

3.1. Desafíos

Si bien la identificación manual resulta ser relativamente simple, existen diferentes desafíos a tener en cuenta para realizar la detección de forma automática, como se comentan a continuación:

- **Varios espectros por imagen**. Como puede verse en la figura 4, lo más usual es tener diversos espectros en una placa. Esto no sería un gran problema, pero sí es necesario abordar la estrategia como un problema de detección dentro de una imagen.
- **Placas deterioradas**. Se encontraron diversas placas con suciedad o manchas inherente al tiempo de almacenamiento o a diferentes problemas al realizar la observación. En la figura 5 pueden verse algunos ejemplos.
- **Falta de lámparas de comparación**. En algunas ocasiones el espectro carece de sus lámparas de comparación, o incluso existen lámparas sin el espectro correspondiente. Estas situaciones deben ser descartadas ya que no permite hacer un análisis de los datos de forma apropiada. La figura 5 muestra un ejemplo de esto. Estos ejemplos fueron caracterizados como *negativos* al momento de realizar la base de datos (ver sección siguiente).

3.2. Base de datos

En primer lugar, construimos una base de datos para el problema particular de la detección de los espectros, ya que no hemos encontrado una similar que pueda ser utilizada. Para esto, utilizamos 256 imágenes de placas digitalizadas donde cada una contiene un promedio de 3 espectros cada una, dando un total de aproximadamente 800 espectros etiquetados de forma manual y validadas por expertos. Además de estas placas, utilizamos un conjunto de imágenes digitalizadas en el Instituto de Ciencias Astronómicas, de la Tierra y del Espacio (ICATE; CONICET y Universidad Nacional de San Juan)[12]. El etiquetado se realizó utilizando el software Label Studio.

[12] https://icate.conicet.gov.ar/

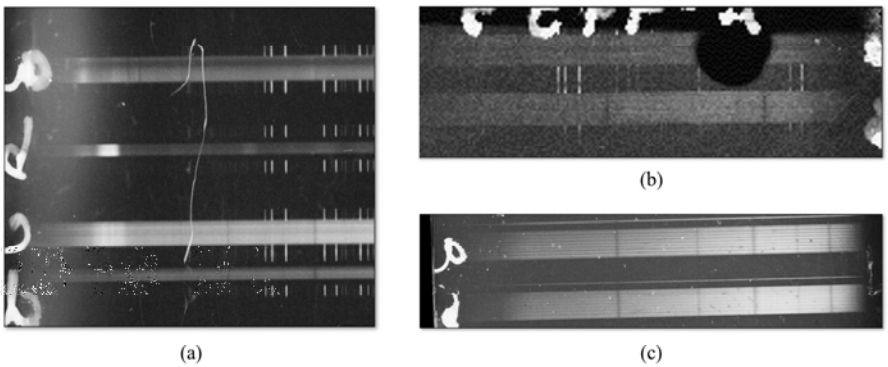


Figura 5. Ejemplos encontrados para la identificación automática de espectros. (a) y (b) suciedad o placas deterioradas. (c) espectros sin lámparas de comparación.

3.3. Modelo de detección

Como técnica de detección de los espectros entrenamos un modelo YOLO [10]. Este modelo, basado en Redes Neuronales Convolucionales, permite el entrenamiento en un dominio particular para detectar regiones de interés dentro de una imagen. Para realizar el entrenamiento redujimos las imágenes de sus tamaños originales a una resolución de $512x512$ pixeles con una profundidad de color de 8 bits. Dividimos las imágenes en un 80 % para realizar el entrenamiento y 20 % para la evaluación. El modelo fue entrenado con tamaño de lotes de 64 y un total de 250 épocas. Utilizamos la técnica de *data augmentation* con los siguientes parámetros:

- Cambio de brillo ±30 %
- Ruido gaussiano del 10 % al 50 %
- Volteo completo de la imagen de forma vertical
- Rotación ±3°
- Escalado ±20 %
- Mosaico

Los rangos operacionales de los parámetros fueron validados por un profesional de la astronomía para verificar que mantengan realismo. Luego del entrenamiento del modelo se logró obtener un *f-measure* cercano al 98 %, dando excelentes resultados de detección. La figura 6 muestra un ejemplo de detección en algunas placas del conjunto de imágenes para evaluación. Como puede apreciarse, en este caso la detección funcionó de manera aceptable. Los dos espectros no detectados (ver segunda placa de la figura) son espectros que no son válidos para analizar por estar corruptos, con lo cual el modelo se comportó como era deseado.

Figura 6. Validación del modelo entrenado en algunas imágenes del conjunto de Evaluación.

4. Conclusiones y trabajos futuros

En este trabajo presentamos el desarrollo de un software novedoso para asistir en la digitalización de placas espectroscópicas de vidrio. El software permite la digitalización de grandes cantidades de placas en menor tiempo, evitando errores y automatizando gran parte del proceso. Así, se posibilita efectivamente la conservación del patrimonio observacional, como así también el almacenamiento en formato adecuados de los espectros de ciencia para un futuro procesamiento. El software fue desarrollado utilizando Javascript para el frontend y Python para el backend, dando posibilidad de ejecutar el sistema desde cualquier navegador web.

Para agilizar el trabajo de las personas encargadas de la digitalización desarrollamos un detector automático de los espectros de dentro de cada placa digitalizada. Para esto, entrenamos un modelo basado en YOLO, creando una base de datos específica para este dominio. El modelo mostró excelentes resultados tanto con datos de validación como en nuevas placas digitalizadas.

Como trabajos futuros, proponemos:

- Validar el uso del software y del modelo de detección al digitalizar el total de placas existentes. Este trabajo ya se encuentra en proceso.
- Calibrar las lámparas de comparación es una tarea que el profesional debe hacer como primer paso en el procesamiento de los espectros de ciencia. Esta es una tarea manual y lenta que aplicarla a los miles de espectros existentes en la FCAyG sería inviable a corto plazo. Se está trabajando en la creación de modelos de aprendizaje automático para calibrar las lámparas de forma automática.
- Igual que para el inciso anterior, el procesamiento posterior del espectro es una tarea manual, donde se identifican las frecuencias de interés de un objeto particular. Como paso siguiente a la calibración de lámparas, se espera poder procesar la información existentes en las observaciones, obteniendo las diferentes curvas espectroscópicas para cada objeto.

Referencias

1. Calderón, J.H., Calderón, J.H., Bustos Fierro, I.H., Bustos Fierro, I.H., Melia, R., Willimoës, C., Willimoës, C., Giuppone, C., Giuppo-

ne, C.: The Digital Archive of the Photographic Images of the Córdoba Observatory Plates Collections. Ap&SS**290**(3), 345–351 (2004). https://doi.org/10.1023/B:ASTR.0000032547.59015.7b
2. Chollet, F., et al.: Deep learning with Python, vol. 361. Manning New York (2018)
3. Cydale, L., Gamen, R., Aidelman, Y., Ronchetti, F., Quiroga, F., Rodriguez, J., López, M., Peralta, R., Meilán, N., Alessandroni, M.d.R., Colazzo, S., Pereyra, N.: Recuperación del trabajo observacional histórico (retroh). la facultad de ciencias astronómicas y geofísicas. unlp. https://retroh.fcaglp.unlp.edu.ar/, accedido el: 2/6/2022
4. Heines, A.: Longterm Variability - First Results from Digitised Photographic Plates. In: Guenther, E., Stecklum, B., Klose, S. (eds.) Optical and Infrared Spectroscopy of Circumstellar Matter. Astronomical Society of the Pacific Conference Series, vol. 188, p. 171 (1999)
5. Hudec, R., Kopel, F., Macsics, R., Hadwige, M., Heber, U., Cayé, W.: Classification of Variable Objects for Search for GRB Candidates on Bamberg Photographic Plates. Acta Polytechnica **53**(3), 27 (2013)
6. Meilán, N., Collazo, S., Alessandroni, M.R., López Durso, M., Peralta, R.A., Aidelman, Y., Cidale, L.S., Gamen, R.: Proyecto de digitalización de placas espectrográficas del Observatorio de La Plata. Boletin de la Asociacion Argentina de Astronomia La Plata Argentina **61B**, 251–253 (2020)
7. Muminov, M., Yuldoshev, Q., Ehgamberdiev, S., Kahharov, B., Relke, H., Protsyuk, Y., Pakuliak, L., Andruk, V.: Astrometry of the h and χ Persei clusters based on the processing of digitized photographic plates. Bulgarian Astronomical Journal **26**, 3 (2017)
8. Orellana, R.B., de Biasi, M.S., Bustos Fierro, I.H., Calderón, J.H.: A revisit to the region of Collinder 132 using Carte du Ciel and Astrographic Catalogue plates. A&A**521**, A39 (2010). https://doi.org/10.1051/0004-6361/200913741
9. Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A., Stobie, E.: Definition of the flexible image transport system (fits), version 3.0. Astronomy and Astrophysics **524**, A42 (2010). https://doi.org/10.1051/0004-6361/201015362
10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
11. Rickman, H.: Unión astronómica internacional. b3 - safeguarding the information in photographic plates. https://www.iau.org/static/publications/ib88.pdf, accedido el: 2/6/2022
12. Walborn, N.R., Gamen, R.C., Morrell, N.I., Barbá, R.H., Fernández Lajús, E., Angeloni, R.: Active Luminous Blue Variables in the Large Magellanic Cloud. AJ**154**(1), 15 (2017). https://doi.org/10.3847/1538-3881/aa6195
13. Wertz, M., Horns, D., Groote, D., Tuvikene, T., Czesla, S., Schmitt, J.H.M.M.: Hamburger Sternwarte plate archives: Historic long-term variability study of active galaxies based on digitized photographic plates. Astronomische Nachrichten **338**(1), 103–110 (2017). https://doi.org/10.1002/asna.201613201

Detección de Intrusiones en Redes Industriales

Evaluación Experimental de Algoritmos de Aprendizaje de Máquina

Aldo Insfrán[1,4], Fabio López-Pires[2], Benjamín Barán[3], Eustaquio Martínez[1]

[1]Facultad Politécnica, Universidad Nacional del Este
[2]Facultad de Posgrados, Universidad Internacional Tres Fronteras
[3]Facultad de Informática, Universidad Comunera
[4]División de Ingeniería Electrónica y Sistemas de Control, Dirección Técnica, ITAIPU Binacional

100113 Ciudad del Este, Paraguay
{aldo.insfran, amartinez}@fpune.edu.py
fabio.lopez@uninter.edu.py
bbaran@ucom.edu.py
ajid@itaipu.gov.py

Resumen. Ataques cibernéticos a sistemas industriales de infraestructura crítica son una realidad en la actualidad y sus consecuencias constituyen un riesgo a la continuidad de los negocios, la economía y el bienestar de la población. En este sentido, este trabajo presenta un análisis de implementaciones de sistemas de detección de intrusiones para sistemas industriales y una evaluación experimental de un conjunto de algoritmos, utilizados en dicho tipo de sistemas, aplicando un conjunto de datos obtenido de un sistema industrial de infraestructura crítica. Dicho análisis da énfasis a cuestiones como, algoritmos y conjuntos de datos de evaluación utilizados, parámetros de entrenamiento, ataques ensayados y métricas de evaluación. La evaluación experimental se lleva a cabo sobre un conjunto nueve algoritmos de aprendizaje de máquina utilizando un conjunto de datos con siete tipos de ataques cibernéticos a la red de un sistema industrial del tipo gasoducto en el que se utiliza el protocolo de comunicaciones *modbus* para la supervisión y el control. Los resultados experimentales mostraron que los algoritmos basados en árboles de decisión arrojan los mejores resultados de clasificación para la métrica de *F1-Score*.

Palabras Claves: *Sistemas de Control Industrial*, *Sistemas de Detección de Intrusiones*, *Modbus*, *Aprendizaje de Máquina*, *Ciberseguridad*.

1 Introducción

Las plantas de generación de energía, las subestaciones eléctricas y ciertas manufacturas, consideradas infraestructuras críticas, tienen como componentes principales a los sistemas de control industriales (*Industrial Control Systems - ICS*).

Originalmente, la ciberseguridad en este tipo de sistemas era confiada al hecho de permanecer desconectados (*airgapped*) de los demás sistemas. En la actualidad, la integración cada vez mayor entre ambientes empresariales (*Information Tecnologies - IT*) e industriales (*Operational Tecnologies - OT*) y la necesidad de conexiones remotas representan desafíos al mantenimiento de la ciberseguridad.

En este escenario, considerando los daños y las consecuencias que podría acarrear el comprometimiento de una infraestructura crítica, numerosas iniciativas han surgido para proteger los mencionados *ICS*. Una de las estrategias propuestas es la defensa en profundidad [1] cuya implementación se establece en el estándar ISA/IEC 62443 a través de la creación de zonas y conductos. En las zonas correspondientes al proceso industrial se sugiere el uso de sistemas no invasivos de detección de intrusiones, que a diferencia de los sistemas de control de código malicioso convencionales, como los antivirus, no ejercen una acción posterior a la detección de una anomalía. Acorde con esta filosofía, varios trabajos han propuesto sistemas de detección de anomalías (*Anomaly Detection Systems - ADS*) para redes industriales basados en el análisis del tráfico de red, los llamados sistemas de detección de intrusiones (*Intrusion Detection Systems - IDS*), los cuales han mostrado su potencial en la detección de ciberataques a los *ICS*. Estos *IDS* utilizan normalmente algoritmos de aprendizaje de máquina (*Machine Learning - ML*) entre los que se diferencian los de aprendizaje clásico (e. g., árboles de decisión), los basados en redes neuronales (*Neural Networks - NN*) y otros de tipo combinado.

En el presente trabajo se lleva a cabo primeramente una revisión y análisis de trabajos relacionados a *IDS* utilizados en sistemas de supervisión y control industriales del tipo *Supervisory Control and Data Adquisicion* (*SCADA*). La composición genérica de este tipo de sistemas comprende: servidores principales de aplicación (*Main Terminal Units – MTU*), unidades terminales de agregación y control local (*Remote Terminal Units – RTU*), interfaces humano-máquina (*Human Machine Interfaces – HMI*), estaciones de ingeniería, y los dispositivos de planta como controladores de máquina, los sensores y actuadores. Seguidamente se realiza la evaluación experimental de un grupo de algoritmos de *ML* utilizando un conjunto de datos (*dataset*) extraído de una red en un sistema de control de un gasoducto (*pipeline*) [2]. El *dataset* es obtenido en condiciones de operación normal del sistema así como durante la ejecución de una serie de ataques de distintos tipos.

La organización del documento es como sigue: la Sección 1 introduce los conceptos básicos que explican de manera general el contexto y el propósito del trabajo; la Sección 2 presenta el análisis de los trabajos relacionados, los cuales se toman como base en la selección del *dataset* utilizado, los algoritmos a evaluar y la métrica de evaluación; la Sección 3 describe los componentes del análisis experimental; el procedimiento de generación de *datasets*, las pruebas y los resultados obtenidos; así también se señalan los principales hallazgos relacionados a dichos resultados planteando una discusión acerca de las características del *dataset* y las condiciones de parametrización de los algoritmos en cada caso. Finalmente, en la Sección 4 se presentan las conclusiones generales y se sugieren posibles trabajos futuros.

2 Trabajos Relacionados

En una revisión de la literatura sobre trabajos relacionados a *IDS* para sistemas industriales fueron encontradas varias propuestas de implementación de variantes de algoritmos de *ML*, las cuales son evaluadas utilizando *datasets* en ocasiones capturados específicamente para tal propósito y en otros casos obtenidos de otros trabajos, cuyo objetivo era el de crearlos con las debidas consideraciones [3] que permitan su uso para el entrenamiento/evaluacion de los *IDS*. Otros trabajos presentan estudios comparativos de estas propuestas, con los que, se orientan futuros esfuerzos de investigación.

En los siguientes apartados se analiza un conjunto de trabajos tanto de implementación de *IDS*, como de desarrollo de *datasets*. Dicho análisis delimita la selección de algoritmos, *datasets*, parámetros de entrenamiento (*features*) y métricas utilizadas en este trabajo.

2.1 *Datasets* para Evaluación de *IDS*

En cuanto a los protocolos de comunicaciones, directamente relacionado con la aplicación industrial del sistema del cual se genera el *dataset*, se ha encontrado que en la mayor parte de los trabajos, como en [2], [4] y [5], se utilizan variantes del protocolo *modbus* (*RTU* y *Transport Control Protocol – TCP*) ampliamente aplicado en varios sectores industriales como manufactura, *utilities* y energía. En la selección de los *features* que finalmente conforman estos *datasets,* se observa que todos presentan los campos de las unidades de datos de protocolo (*Protocol Data Unit – PDU*) además de etiquetas que identifican los mensajes maliciosos. Otros trabajos como [6] y [7] tratan sobre sistemas eléctricos en los que se hace uso de protocolos IEC 60870-5-104, donde los *datasets* están compuestos también por algunos campos de las PDU de las capas 2, 3, 4, y 7 de dichos protocolos, además de las etiquetas.

Entre los ataques que son ensayados, el más frecuente es el de tipo reconocimiento a nivel de aplicación, también conocido como ataque de *fingerprinting*, como se muestra en [2], [4], [6] y [7]. Como parte de este tipo de ataque se determinan las direcciones, códigos de función utilizados y el contenido de memoria (*COILs* y *Registers*) de los nodos. Otras variantes encontradas, relacionadas al reconocimiento, son el escaneo de direcciones de capa de red y de puertos de capa de transporte como en [6] y [7]. También, en [2], [4], [6] y [7], se han encontrado ataques de inyección de comandos y respuestas, todos llevados a cabo a través de una estrategia *man-in-the-middle* (*MITM*) con excepción de [4], en el que se ataca primeramente un nodo pivote (controlador) al que se carga (*load*) código malicioso. Los ataques de reconocimiento en [6] y [7] se llevan a cabo haciendo *spoofing* mediante la inyección de mensajes de lectura. En [2], [5], y [6] se ejecutan ataques de denegación de servicio (*Denial of Service – DOS*). En [2] se fabrican mensajes con valores incorrectos del código de redundancia (*Cyclic Redundancy Check – CRC*) de *modbus*, en [5] y [6] tales ataques se llevan a cabo usando *icmp*[1] y *tcp-syn flooding*, en [5] se utiliza además *modbus-query flooding*. Todos los trabajos describen básicamente la forma en la que son llevados los

[1] *Internet Control Message Protocol – ICMP*

ataques: [2] a través del uso de un *datalogger*; [4], [5], [6] y [7] muestran comandos de consola y en ocasiones otras herramientas utilizadas como *metasploit*[2].

Vale destacar la clasificación realizada en [8], en la que se diferencian cuatro clases de ataques: reconocimiento (*reconnaissance* - *RECON*), inyección de respuestas (*response injection* - *RI*), inyección de comandos (*command injection* - *CI*) y denegación de servicios (*denial-of-service* - *DOS*). También se muestran subclases para los *RI*; *Naive Malicious RI* (*NMRI*) y *Complex Malicious RI* (*CMRI*); y, subclases para los *CI*; *Malicious State CI* (*MSCI*), *Malicious Parameter CI* (*MPCI*), *Malicious Function-Code CI* (*MFCI*). Esta clasificación se utilizará como referencia para las discusiones sobre los tipos de ataque en este trabajo.

De todos estos trabajos que presentan *datasets*, en [2] se considera la mayor diversidad de ataques practicados a un *IDS*. Los 35 escenarios de ataque ensayados corresponden a los 7 tipos que forman parte de la clasificación llevada a cabo en [8]: *NMRI, CMRI, MSCI, MPCI, MFCI, RECON* y *DOS*. Todos estos corresponden a ataques llevados a cabo directamente sobre el protocolo industrial *modbusRTU*. Los datos capturados son: valores de campos del *PDU modbus*, estampas de tiempo y etiquetas de tipo de ataque. Tales características motivaron a la selección de este *dataset* como base para la evaluación experimental en este trabajo.

2.2 Propuestas de *IDS* Estudiados

De todos los trabajos estudiados que presentan implementaciones de *IDS* para *ICS*, el protocolo *modbus* es el más utilizado. En menor medida fueron encontrados casos en los que se analizan otros protocolos como los del IEC 60870-5-104 o del IEC 61850, correspondientes a aplicaciones en el sector eléctrico. Así mismo, en la mayor parte de los trabajos, al tiempo de proponer algoritmos o modelos de *IDS*, se generan *testbeds* y *datasets* propios. De los trabajos analizados, sólo [9], [10] y [11] utilizan datos generados en otros trabajos como los presentados en la Sección 2.1.

En cuanto a los *features* analizados en el proceso de entrenamiento de los algoritmos o modelos utilizados para la detección, se observa que en la mayoría de los trabajos se generan flujos de comunicaciones agrupando los mensajes obtenidos en los *datasets*. Tales flujos contienen *features* que son valores calculados a partir de los campos de los mensajes intercambiados en la red. En [10] y [11] se utilizan datos correspondientes a la capa de red como la cantidad y el tamaño de los paquetes, en [12] son utilizados valores obtenidos de los campos del protocolo de capa de aplicación del IEC 60870-5-104 (*Application Protocol Data Unit* - *APDU*). Adicionalmente, en [11] se considera la secuencia de transición de los mensajes. Otros trabajos calculan estos *features* directamente del conjunto de mensajes, sin agruparlos en flujos. En [13] se calculan longitudes e intervalos entre paquetes, en [14] además se hace un recuento de paquetes del tipo *ARP* (*Address Resolution Protocol*), con el mismo origen o destino. En [15] se analizan tiempos de respuesta y de retransmisión, mientras que en [16] se utilizan, entre otros, los tiempos de ida y vuelta. Algunos trabajos se centran exclusivamente en *features* calculados a partir de, valores obtenidos del protocolo industrial de capa de

[2] *https://www.metasploit.com/*

aplicación, parámetros del sistema de automatización (e. g., registros de un *Programmable Logic Controller – PLC*) y de mediciones de la planta. En [9] que hace uso del *dataset* de [2], estos *features* se seleccionan de entre los parámetros *modbus.* En [16] se utilizan valores de medición de voltaje mientras que en [17] se utiliza el código de función. Todos estos trabajos utilizan datos de red. En [18] sin embargo se utilizan datos obtenidos a partir del *Windows Performance Monitor* (*WPM*) en la estación de ingeniería.

Parte de los trabajos presenta escenarios de ataque a protocolos no industriales, o de capas inferiores a la de aplicación, en ellos se llevan a cabo ataques de tipo *MITM* con los que se hace *spoofing* de *ARP* para hacer *fingerprint* de los activos de red industriales como en [18]**,** o escaneos de red (*RECON*) con el uso de mensajes *TCP-FIN* [14]. También se observan ataques de tipo *DOS* a través de inundación con *TCP-SYN*. Otro conjunto de trabajos presenta además, ataques directamente ejecutados sobre los protocolos industriales. Algunos de ellos también ejecutados con una estrategia de *MITM* y *spoofing*; *fingerprinting* de los controladores a través de comandos de lectura de *modbus*, fabricación e inyección de comandos (*MFCI*) mediante comando de escritura *modbus* [17]; y, alteraciones de los valores leídos de campo a través de inyección de respuestas maliciosas (*NMRI* y *CMRI*) [16]. También se consideran ataques de exfiltración de datos [12] manipulando el campo *COT* del *ASDU* del IEC 60870-5-104. A diferencia del enfoque *MITM*, en [4] y [13] se ensayan cargas de código malicioso a un activo de red, desde el cual posteriormente se practican otros ataques.

2.3 Algoritmos de *IDS* Estudiados

La mayor parte de los trabajos revisados utiliza variantes de *Support Vector Machine* (*SVM*) como algoritmo base para sus implementaciones de *IDS* como por ejemplo [13]. También se utiliza en conjunto con procesos de optimización en la obtención de *features* como en [17], y en *clusters* con sus resultados agregados mediante otros algoritmos como el *k-means* en [14]. Otro tipo de algoritmos muy utilizados y que muestran buen desempeño son los árboles de decisión, encontrado en variantes con el uso de poda como *Reduced Error Pruning* (*REPTree*) en [15], o con métodos de agregación como *bagging* (*Random Forest – RF*) en [18] y *boosting* en [10] y [15]. También se encuentran algoritmos bayesianos, *Multinomial Naive Bayes* en [15] y de redes bayesianas en [9]. Por otro lado, [10] y [16] hacen uso de redes neuronales de tipo *Multi-Layer Perceptron* (*MLP*) y [17] además prueba funciones del tipo *Radial Basis Function* (*RBF*). A diferencia de estos trabajos, en [12] se propone la construcción de un autómata probabilístico y en [11] la definición de un modelo ciber-físico.

La Tabla 1 resume los algoritmos encontrados en las implementaciones de *IDS* examinadas. La Tabla 2 relaciona tales algoritmos con los trabajos donde fueron encontrados.

La métrica de evaluación más utilizada es la precisión de detección (*Accuracy*), seguida por los valores de *Precision*, *Recall* y la media armónica de estos valores *F1-Score*. En [18] se adopta directamente los valores de la matriz de confusión *TP* (*True Positives*), *TN* (*True Negatives*), *FP* (*False Positives*), *FN* (*False Negatives*), mientras

que [17] y [14] dan mayor importancia a los falsos positivos utilizando *false positive rate*. Adicionalmente en [17] se evalúa el tiempo de clasificación y en [15] el tiempo de creación del modelo.

FA[3]	**Cod.**	**Nombre**	**Cod.**	**Nombre**
***Decision Trees* (*DT*) [FA1]**	A1	*C4.5*	A2	*REPTree*
	A3	*Decision Stump Tree*	A4	*Ripple Down Rule*
	A5	*Degged Decision Stump*		
***Instance-Based* [FA2]**	A6	*KNN*	A7	*Linear SVM*
	A8	*One Class SVM*	A9	*Logically-Deep SVM*
***Probabilistic* [FA3]**	A10	*Logistic Regression*	A11	*Multinomial Naive Bayes*
	A12	*Bayes Net*	A13	*Bayes Point Machine*
***Dimensionality Reduction* [FA4]**			A14	*Linear Discriminant Analysis*
***Ensemble* [FA5]**	A15	*Decision Forest*	A16	*Bagged REPTree*
	A17	*Decision Jungle*	A18	*Boosted DT*
	A19	*k-means + SVM*		
***Neural Netw.* [FA6]**	A20	*MLP*	A21	*RBF-MLP*
	A22	*Averaged Perceptron*		
***Otros* [FA7]**	A23	*Finite State Machine*	A24	*Deterministic Probabilistic Automata*

Tabla 1. Algoritmos encontrados en las implementaciones de *IDS* estudiadas

Con respecto a los mejores resultados de detección obtenidos, [15] y [10] indican a los algoritmos de tipo *ensemble* de árboles como los de mejor precisión de detección. A su vez [9] muestra un resultado de 100% en todas sus métricas con el uso de redes bayesianas analizando, al igual que los anteriores un *dataset modbus*. En [18] se muestra mejores resultados para el algoritmo *K-Nearest Neighbors* (*KNN*) pero con un *dataset* conformado por parámetros del *WPM*. Con lo anterior es difícil hacer una comparativa de resultados dado que los *features* utilizados en la construcción de los modelos que son evaluados son distintos.

Entre otras cuestiones, se tiene que solo [9] realiza experimentos de clasificación multiclase, supervisada y no supervisada. Todos los demás trabajos llevan a cabo básicamente clasificaciones binarias. Solo [10] y [16] muestran experimentalmente el efecto de la variación del desbalance entre las clases objetivo de detección (anomalías) y la clase mayoritaria (normal).

En la segunda parte del presente trabajo se lleva a cabo un análisis experimental de un conjunto de algoritmos de clasificación y aprendizaje supervisado cuya selección ha sido inspirada en la Tabla 1. Han sido seleccionados representantes de cinco de las familias de algoritmos presentadas: *C4.5* y *REPTree*, de la FA1; *KNN* y *SVM*, de la FA2; *Multinomial Naive Bayes* (*MNB*) y *Logistic Regression* (*LR*), de la FA3; *MLP*, de la FA6; y, *RF* junto con *Gradient Boosted Tree* (*GBT*), como algoritmos de tipo *ensemble* (FA5). Se han omitido los algoritmos de reducción dimensional, cuyo uso

[3] FA: Familia de Algoritmos

podría aprovecharse mejor en clasificaciones no supervisadas, y a los modelos de máquinas de estado y autómatas determinísticos.

	[17]	[14]	[15]	[13]	[9]	[10]	[18]	[6]	[16]	[12]	[11]
A1			FA1		FA1						
A2			FA1								
A3			FA1								
A4			FA1								
A5			FA1								
A6							FA2	FA2			
A7				FA2		FA2		FA2			
A8	FA2	FA2									
A9						FA2					
A10			FA3			FA3					
A11			FA3								
A12					FA3						
A13						FA3					
A14								FA4			
A15						FA5	FA5				
A16			FA5								
A17						FA5					
A18						FA5					
A19		FA5									
A20						FA6			FA6		
A21	FA6										
A22						FA6					
A23									FA7		FA7
A24										FA7	

Tabla 2. Relación de Familias de Algoritmos e implementaciones de *IDS*

3 Evaluación Experimental Propuesta

3.1 Preparación de *datasets*

El *dataset* propuesto en [2], consta de 35 tipos diferentes de ataque pertenecientes a las 7 categorías descritas en [8]. La preparación de los datos para la evaluación experimental consiste en la creación de subconjuntos de datos (*subset*), a partir del *dataset* original, para una clasificación binaria. De esta manera se crean 7 *subsets*, cada uno conteniendo una sola categoría de ataque, sin discriminación entre escenarios individuales de ataques dentro de la misma categoría. Por tanto, en los *subsets*, cada fila está marcada con una de dos posibles etiquetas; "0" para tráfico normal, y "1" para tráfico anómalo. Adicionalmente, para evaluar los efectos del desbalance entre las clases (definido aquí como el cociente entre la cantidad de filas etiquetadas como

ataque y el número total de filas) se divide nuevamente cada *subset* de detección binaria en tres *subsets* con desbalances diferentes. El primero de ellos conserva el desbalance que se obtuvo al separar el *dataset* original en cada uno de los 7 *subsets*. Para la creación del segundo *subset*, del primero se van eliminando de forma aleatoria filas etiquetadas como tráfico normal hasta conseguir el mismo desbalance que se tiene en el *dataset* global multiclase provisto en [2], que corresponde a aproximadamente un 22%. El tercer *subset* se obtiene del primero aumentando el número de filas eliminadas hasta conseguir un desbalance del 50%. Finalmente, cada *subset* se divide en dos, con una relación de 70/30, para el entrenamiento y las pruebas respectivamente.

Entre las columnas (*features*) de los *subsets* se tienen, campos del *PDU* de *modbus*, un indicador de que el mensaje se trata de un comando o respuesta y una estampa de tiempo. La Tabla 3 muestra los mismos junto con una breve descripción. A diferencia de otros trabajos como [10], en el presente, los mensajes intercambiados en la red no son previamente agrupados en flujos. Como paso siguiente a la formación de los *subsets* se trata cada una de las columnas estableciendo para ellas tipos específicos de datos, en general, diferenciándolas entre valores numéricos y etiquetas.

Ord.	***Feature***	**Descripción**	**Ord.**	***Feature***	**Descripción**
1	*Address*	Dirección del esclavo *modbus*	10	*System Mode*	Oper. manual/ automática
2	*Function*	Código de función *modbus*	11	*Control Scheme*	Control por Bomba/ Válvula
3	*Lenght*	Longitud de mensaje	12	*Pump*	Activ. manual de bomba
4	*Setpoint*	Punto de operación de bomba	13	*Solenoid*	Activ. manual de válvula
5	*Gain*	Parámetros de ajuste del controlador PID (Proporcional Integral Derivativo)	14	*Pressure measurement*	Medición de Presión
6	*Reset rate*		15	*CRC rate*	Cod. chequeo de error *modbus*
7	*Deadband*		16	*Command/ Response*	Bandera de identificación
8	*Cylce time*		17	*TimeStamp*	Estampa de tiempo
9	*Rate*		18	*Categ. Result*	Result. de Clasif.

Tabla 3. *Features* para el entrenamiento de algoritmos evaluados

Otra cuestión importante que surge en el tratamiento de los datos para la formación de los *subsets* es el manejo de los valores desconocidos o "*missing values"*. Las tablas que forman el *dataset* tienen como columnas a cada uno de los *features* seleccionados (campos del *PDU modbus*). De esta manera, cada fila constituye un mensaje *modbusRTU*. Existen varios tipos de mensajes *modbus*, el tipo de un mensaje particular es determinado por el campo de código de función. No todos los tipos de mensajes transportan los mismos campos *modbus*, por esta razón, no todas las filas de la tabla

contendrán valores para todas las columnas (*features*). Este es el motivo por el cual se originan los "*missing values*".

Pudo observarse experimentalmente que la presencia de tales valores desconocidos empobrecía los resultados de predicción de los algoritmos. Por esta razón, se optó por una estrategia de relleno. Cabe destacar que de todos los trabajos explorados en la Sección 2, sólo [6] aborda el tema proponiendo también el relleno.

Todos los valores faltantes están relacionados con registros de memoria de esclavos *modbus*. Estos valores solo pueden cambiar en respuesta a comandos enviados desde el maestro o por decisión de una lógica de control interna del *PLC*. Los comandos donde se modifican registros del esclavo están representados por mensajes en los que se identifican campos *modbus* con los nuevos valores para tales registros. Las modificaciones que eventualmente sean realizadas por la propia lógica de control del *PLC* son finalmente reportadas en las respuestas a los mensajes de *polling* del maestro. Por lo anterior, para este trabajo, la estrategia de relleno fue la de mantener el último valor del registro del esclavo hasta leer un mensaje con un nuevo valor para el mismo.

La manipulación del *datset*, la creación de los *subsets*, la parametrización y evaluación de los algoritmos se llevan a cabo con las funcionalidades ofrecidas por la aplicación *knime*[4].

3.2 Entrenamiento y Evaluación de los Algoritmos

Para la evaluación de los algoritmos del conjunto, se crean flujos de trabajo en *knime*, en los que primeramente se normalizan los valores que luego pasan a entrenar a cada uno de los algoritmos. Como se comenta más adelante, según sea el caso, se utilizan dos tipos de normalizaciones lineales; calculadas a partir del mínimo y máximo del conjunto analizado y a partir de una distribución gaussiana con media 0 y desviación estándar 1. Al igual que con el desbalance, se llevaron a cabo pruebas cambiando la cantidad de *features*, utilizando u obviando las estampas de tiempo. La razón de esto es apreciar la dependencia de los resultados con la secuencia de ocurrencia o aparición de los mensajes en el proceso (industrial). Tanto los flujos de trabajo creados en *knime* como el *dataset* utilizado se han hecho de acceso público[5]. A continuación se presentan las parametrizaciones para cada uno de los algoritmos evaluados.

Para los algoritmos de tipo árbol de decisión, que incluyen los de tipo *ensemble*; *C4.5*, *REP* de *C4.5*, *RF*, y *GBT*; los valores de los *subsets* pasaron previamente por un proceso de normalizados lineal de tipo mínimo-máximo. Para el caso de *C4.5*, *REP*, y *RF* fueron utilizadas dos variantes de medidas de "calidad" de la información, el *Gini Index* y el *Gain Ratio* [19].

Para las variantes *MNB* y *LR* de los clasificadores estadísticos y *KNN* de aprendizaje basado en instancias, también fue utilizada normalización lineal de mínimo-máximo. Adicionalmente, para *LR* fueron ensayados dos métodos propuestos en *knime*; pequeños cuadrados con modificación iterativa de pesos (*Iteratively Reweighted Least Squares – IRLS*), y gradiente de media estocástica (*Stochastic Average Gradient –*

[4] *https://www.knime.com/*
[5] *https://github.com/AldoJavInsD/eval-IDS-ICS.git*

SAG). Para la selección de *K* en *KNN*, como en [20] se consideran valores impares (evitando empates) entre 1 y $\sqrt{N}$ (raíz cuadrada el número de filas en el *subset* de entrenamiento). Dado el desbalance en dichos *subsets* (llegando al 1% para el *subset DOS*), fueron tomados valores pequeños de *K* para evitar la predominancia de la clase mayoritaria [21]. De este modo fueron realizados cálculos para distancias a 3 y 5 vecinos. Cabe comentar que en el desarrollo de los experimentos, probando valores mayores de *K*, cercanos a $\sqrt{N}$, no se observaron grandes diferencias en los resultados de las predicciones.

En el caso del *MLP*, fue utilizada la implementación *RProp* de *knime* [22]. La normalización fue llevada a cabo con las dos variantes lineales previamente mencionadas; con valores en distribución gaussiana (*z-score*) y con valores mínimo-máximo. La incorporación del segundo tipo se debe a que en las pruebas se observaron mejoras en las métricas usando esta normalización para algunos casos. De igual forma se llega a una configuración de 5 capas, 17 neuronas por capa y 750 iteraciones, posterior a ensayos en los cuales se observaron los efectos del cambio de cada uno de los valores, seleccionando finalmente los que ofrecían mejor compromiso entre el tiempo de generación del modelo (usando el *subset* de entrenamiento) y los resultados obtenidos al evaluar el *subset* de pruebas.

Finalmente para el algoritmo *SVM*, fueron también utilizadas las dos formas de normalización, y además se exploraron las tres opciones de *kernel* que ofrece *knime*: *Polynomial*, *Radial Basis Function* (*RBF*) e *HyperTangent*.

	C4.5	REP	RF	GBT	KNN	SVM	MNB	LR	MLP
S1	**0,998**	0,996	0,997	0,997	0,985	0,975	0,994	0,997	0,994
S2	0,903	0,902	0,937	**0,977**	0,898	0,906	0,905	0,900	0,905
S3	0,916	0,916	0,938	**0,976**	0,897	0,906	0,902	0,901	0,904
S4	0,819	0,818	0,987	**0,988**	0,975	0,972	0,814	0,783	0,928
S5	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
S6	0,902	0,904	0,988	**0,993**	0,951	0,844	0,734	0,789	0,934
S7	0,835	0,839	**0,978**	0,963	0,949	0,823	0,828	0,807	0,940

Tabla 4. Mejores valores de *F1-Score* para los algoritmos evaluados con los 7 *subsets*

Los resultados obtenidos en los ensayos para todo el conjunto de algoritmos y *subsets* pueden verse en la Tabla 4, en la que por comodidad, se han nominado a los *subsets* como S1 (*RECON*), S2 (*NMRI*), S3 (*CMRI*), S4 (*MSCI*), S5 (*MFCI*), S6 (*MPCI*) y S7 *(DOS*), respectivamente.

3.3 Principales Hallazgos (PH)

- *PH1*: En el conjunto de algoritmos relacionados con árboles de decisión, como se esperaba, los mejores rendimientos fueron para los de tipo *ensemble* (*RF* y *GBT*), prácticamente para todos los tipos de ataques. En general para todos estos algoritmos los resultados se mantuvieron por encima del 93,7%, dándose los resultados más bajos para los ataques de tipo *RI* (inyección de respuestas

maliciosas), y los mejores para el *MFCI*, con el cual la mayoría de los algoritmos presentó buenos resultados. Lo anterior condice con la forma de ejecución del ataque, los códigos de función utilizados en los mensajes malicioso no son encontrados en comandos normales. Fuera del caso del *MFCI*, el mejor valor obtenido fue 99,74% para *GBT* (con un límite de 4 niveles para los árboles componentes) con *RECON* (con un desbalance del 50%, sin el uso de estampas de tiempo) y el peor 81,77% para *REP* (utilizando *Gini Index*) en *MSCI* (50% de desbalance, sin el uso de estampas de tiempo).

- *PH2*: Entre los algoritmos de clasificación probabilísticos, *MNB* y *LR*, el primero presentó los mejores resultados. Para este tipo de algoritmos el ataque más difícil de reconocer fue el de inyección de comando *MPCI* y al igual que los anteriores, el más fácil de detectar fue el de *MFCI*. Obviando este último, el mejor valor obtenido fue 99,70% para *LR* usando *SAG* con *RECON* (2% de desbalance, sin estampas de tiempo) y el peor 73,44% para *MNB* en *MPCI* (50% de desbalance, sin estampas de tiempo).
- *PH3*: Para el caso de los algoritmos basados en redes neuronales artificiales, no se aprecia una diferencia considerable entre el uso de normalización mínimo-máximo y con valores de distribución de probabilidad. El mejor valor obtenido, fuera del *MFCI*, fue 99,43% para normalización mínimo-máximo con *RECON* (desbalance del 2%, con estampas de tiempo) y el peor 89,06% para el mismo tipo de normalización en *MSCI* (50% de desbalance, con estampas de tiempo).
- *PH4*: Para los clasificadores de aprendizaje basado en instancias o memoria, *KNN* y *SVM*, el valor más alto de la métrica fue obtenido por *KNN* (3 vecinos) para el conjunto *RECON* (desbalance del 50%, sin el uso de estampas de tiempo) con 98,46%. A su vez, el menor valor fue para *SVM* (utilizando *RBF*) también con el conjunto *RECON* (50% de desbalance con estampas de tiempo) con 61,47%. Entre las variantes de *kernel* para *SVM*, el de tipo polinomial fue el que en general mejor desempeño mostró. La normalización con valores gaussianos no presentó mejoras respecto de la normalización de valores mínimo-máximo.
- *PH5*: La mayor parte de los mejores valores obtenidos en la métrica de clasificación pertenecen a situaciones en las que fue utilizado el conjunto desbalanceado a un 50% con un 82% de los casos. Cabe destacar que 65% de los mejores valores obtenidos en la métrica fueron para los casos en los que la estampa de tiempo fue utilizada en el conjunto de entrenamiento, siendo los probabilísticos, y en particular *LR* el menos afectado por este *feature*.
- *PH6*: Con respecto al tiempo de entrenamiento de los algoritmos, el *SVM* seguido del *MLP*, son los de mayor tiempo de entrenamiento.
- *PH7*: El algoritmo menos afectado por el desbalance fue el *KNN* para el cual en 3 de las evaluaciones de los 7 *subsets* de ataque, los mejores resultados no fueron para el conjunto balanceado al 50%.
- *PH8*: *GBT* obtuvo los mejores resultados de clasificación para 5 de los 7 *subsets*, los cuales corresponden a ataques de inyección de respuestas maliciosas, simple *NMRI* (S2) y compleja *CMRI* (S3), así como para todos los subtipos de inyección

de comandos, alteración de estado *MSCI* (S4), código de función malicioso *MFCI* (S5), y alteración de parámetros *MPCI* (S6). Si bien en los demás casos dicho algoritmo no obtuvo los valores más altos, los obtenidos fueron bastante cercanos a los mejores.

- *PH9*: Para el ataque de reconocimiento (*RECON*) el algoritmo de árbol simple C4.5 obtiene el mayor valor de *F1-Score* con un 99,8% seguido por *GBT* y *RF* con un 99,7%.
- *PH10*: La mejor clasificación para el *subset* con denegación de servicio (*DOS*) fue para el *RF* con un 97,8% también seguido como segundo mejor valor por el *GBT* con un 96,3%.

4 Conclusiones

En cuanto al análisis de trabajos relacionados, fueron revisados 4 trabajos que proponen *datasets* de evaluación para evaluación de *IDS* de entre los cuales uno fue seleccionado para la evaluación experimental. Así también fueron revisados más de 11 trabajos de implementación de *IDS* para *ICS*, los cuales fueron analizados y comparados aspectos como algoritmos, *datasets* de evaluación (tipos ataques y desbalance de clases), y métricas. De entre los algoritmos utilizados en estos trabajos, un conjunto de 9 algoritmos de 5 familias distintas fue seleccionado para la evaluación experimental.

La evaluación experimental arrojó como mejores resultados, para la métrica *F1-Score*, a los presentados en la Tabla 4. Analizando tales resultados fueron presentados los principales hallazgos en la parte final de la Sección 3. Como resumen de dicho análisis puede concluirse que se encontró que *MFCI*, seguido de *RECON*, son los dos tipos de ataques con mayor facilidad de detección. Así también, los algoritmos basados en árboles de decisión son los que en un mayor número de ocasiones superaron el 99% de *F1-Score* constituyéndose así en los de mejor desempeño de clasificación. En particular, dicho valor fue superado en un 32% de todas las evaluaciones. Los algoritmos probabilísticos y de redes neuronales superaron este umbral en un 29% de sus evaluaciones, y por último los basados en instancias solo superaron este porcentaje en la evaluación del conjunto *MFCI*.

Estas evaluaciones fueron llevadas a cabo con variantes no complejas de los algoritmos de base tomando ventaja de las funcionalidades ya implementadas en la herramienta *knime*. Sin embargo, se considera que los resultados obtenidos permiten apreciar suficientemente el rendimiento de los algoritmos, motivando así a la experimentación de mejoras que optimicen su desempeño con base en alguna dimensión en particular, como la cantidad de falsos positivos y el tiempo de creación de los modelos.

Queda como propuesta para trabajos futuros, la experimentación con protocolos industriales que representen sistemas de otra naturaleza como los especificados en los estándares IEC 62870-5-104 e IEC 61850, ampliamente utilizados en sistemas de generación y subestaciones eléctricas, así como otras variantes de los diferentes algoritmos de aprendizaje de máquina (*ML*).

5 Referencias

[1] Industrial Control Systems Cyber Emergency Response Team, "Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies," 2016.

[2] I. Turnipseed, Z. Thornton and T. Morris, "Industrial Control System Simulation and Data Logging for Intrusion Detection System Research," in *7th Annual Southeastern Cyber Security Summit*, Huntsvile, AL, 2015.

[3] P. Nevavuori and T. Kokkonen, "Requirements for Training and Evaluation Dataset of Network and Host Intrusion Detection System," in *New Knowledge in Information Systems and Technologies*, Springer, 2019, pp. 534-546.

[4] A. Lemay and J. Fernandez, "Providing SCADA Network Data Sets for Intrusion Detection Research," in *9th USENIX Workshop on Cyber Security Experimentation and Test*, Austin, TX, 2016.

[5] I. Frazão, P. Abreu, T. Cruz, H. Araújo and P. Simões, "Denial of Service Attacks: Detecting the Frailties of Machine Learning Algorithms in the Classification Process," in *13th International Conference, CRITIS*, Kaunas, Lithuania, 2018.

[6] M. Egger, E. Gunther and E. Dominik, "Comparison of approaches for intrusion detection in substations using the IEC 60870-5-104 protocol," in *9th DACH+ Conference on Energy Informatics ONLINE*, Sierre, Switzerland, 2020.

[7] P. Maynard, K. McLaughlin and S. Sezer, "An Open Framework for Deploying Experimental SCADA Testbed," in *In 5th International Symposium for ICS & SCADA Cyber Security Research*, University of Hamburg, Germany, 2018.

[8] T. Morris and W. Gao , "Industrial Control System Traffic Data Sets for Intrusion Detection Research," in *Critical Infrastructure Protection VIII*, vol. 441, J. Butts and S. Shenoi, Eds., Berlin, Heidelberg: Springer, 2014, pp. 65-78.

[9] I. Ullah and Q. H. Mahmoud, "A Hybrid Model for Anomaly-based Intrusion Detection in SCADA Networks," in *IEEE International Conference on Big Data (BIGDATA)*, Boston, MSA, USA, 2017.

[10] G. Vasquez, R. S. Miani and B. B. Zarpelao, "Flow-Based Intrusion Detection for SCADA networks using Supervised Learning," in *XVII Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais - SBSeg*, Brasília, DF, Brasil, 2017.

[11] C. Sheng, Y. Yao, Q. Fu and W. Yang, "A cyber-physical model for SCADA system and its intrusion detection," *Computer Networks,* vol. 185, 2021.

[12] P. Matoušek, O. Ryšavý, M. Grégr and V. Havlena, "Flow based monitoring of ICS communication in the smart grid," *Journal of Information Security and Applications,* vol. 54, 2020.

[13] A. Terai, S. Abe, S. Kojima and Y. Takano, "Cyber-Attack Detection for Industrial Control System Monitoring with Support Vector Machine based on Communication Profile," in *IEEE European Symposium on Security and Privacy Workshops*, Paris, France, 2017.

[14] L. Maglaras, J. Jiang and T. J. Cruz, "Combining ensemble methods and social network metrics for improving accuracy of OCSVM on intrusion detection in SCADA systems," *Journal of Information Security and Applications,* no. 30, p. 15–26, 01 10 2016.

[15] S. Ponomarev and T. Atkison, "Industrial Control System Network Intrusion Detection by Telemetry Analysis," *IEEE Transactions on Dependable and Secure Computing,* vol. 13, no. 2, pp. 252-260, 2016.

[16] P. Kreimel, O. Eigner, F. Mercaldo, A. Santone and P. Tavolato, "Anomaly detection in substation networks," *Journal of Information Security and Applications,* vol. 54, 2020.

[17] P. Zeng, W. Shang, M. Wan, L. Li and P. An, "Intrusion detection algorithm based on OCSVM in industrial control system," *SECURITY AND COMMUNICATION NETWORKS,* vol. 9, no. 10, p. 1040–1049, 2015.

[18] F. Zhang, H. A. Dias Edirisinghe Kodituwakku, J. W. Hines and J. Coble, "Multilayer Data-Driven Cyber-Attack Detection System for Industrial Control Systems Based on Network, System, and Process Data," *IEEE Transactions on Industrial Informatics,* pp. 4362 - 4369, 07 01 2019.

[19] S. Tangirala, "Evaluating the Impact of GINI Index and Information," *International Journal of Advanced Computer Science and Applications,* vol. 11, no. 2, pp. 612-619, 2020.

[20] A. Hassanat, M. Ali Abbadi, G. A. Altarawneh and A. A. Alhasanat, "Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach," *International Journal of Computer Science and Information Security,* vol. 12, no. 8, pp. 33-39, 2014.

[21] I. Chelliah, "An Introduction to K-Nearest Neighbors Algorithm," Medium, 23 November 2020. [Online]. Available: https://towardsdatascience.com/an-introduction-to-k-nearest-neighbours-algorithm-3ddc99883acd. [Accessed 28 07 2022].

[22] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: the RPROP algorithm," in *IEEE International Conference on Neural Networks (ICNN)*, San Francisco, CA, USA, 1993.

Entorno de Simulación basado en DEVS para Agentes de Aprendizaje por Refuerzo aplicado a la Generación y Administración de Energías Renovables

Ezequiel Beccaria, Veronica Bogado, and Jorge A. Palombarini

Facultad Regional Villa María - UTN, Villa María, X5900 HLR Argentina; CIT Villa María - CONICET-UNVM, Villa María, Córdoba, X5900 HLR Argentina
{ebeccaria,vbogado,jpalombarini}@frvm.utn.edu.ar
http://www.frvm.utn.edu.ar

Resumen La dinámica y complejidad de los entornos industriales actualmente han llevado a la necesidad de soluciones que permitan capturar la interacción en tiempo real para tomar decisiones sobre el control de los procesos involucrados. El Aprendizaje por Refuerzo es un enfoque promisorio, que se aplica en problemas de decisión secuencial, donde la complejidad radica en la interacción agente-entorno y la incertidumbre subyacente del entorno, pero requiere de una simulación que refleje el proceso bajo control (entorno) y su dinámica para entrenar el agente. En particular, el uso de energías alternativas representa un problema con enormes desafíos, donde existen procesos críticos que necesitan monitoreo y control en tiempo real de un ambiente altamente incierto. En este trabajo, se presenta una solución para entrenar este tipo de agentes con entornos modelados y simulados usando DEVS. El mismo se aplica al problema de generación y administración de una energía alterna, biogás producido por un digestor y usado por diferentes perfiles de consumidores industriales.

Keywords: Aprendizaje por Refuerzo, DEVS, Problemas de Decisión, Energías Renovables

1. Introducción

El *Reinforcement Learning (RL)* [24] se ha convertido en uno de los campos de más rápido crecimiento como metodología para brindar capacidades de aprendizaje a los agentes de *Inteligencia Artificial (IA)* que deben encontrar políticas de acción para diferentes *Problemas de Decisión Secuencial (PDS)* complejos. Un aspecto limitante en el uso de RL es la necesidad de contar con un entorno de entrenamiento para llevar a cabo el aprendizaje de los agentes [24]. Este entorno permite al agente RL experimentar estados diferentes del PDS y la consecuencia de las acciones disponibles en ellos, y así poder inferir una política de acción que maximice la recompensa del agente.

Para aplicar soluciones de RL a nivel industrial, no basta pensar en mecanismos que aseguren el correcto desempeño del agente, sino también, en el nivel de correlación entre el entorno de entrenamiento, donde se realiza el aprendizaje, y el proceso real, es decir, es necesario reducir la brecha entre el proceso real y el entorno simulado. En la actualidad, no se ha prestado mucha atención al desarrollo de entornos de entrenamiento formales para el entrenamiento de agentes RL, con el fin de evitar la incertidumbre, el riesgo y las posibles pérdidas económicas de una mala correlación del mismo.

Este trabajo, se propone un entorno de entrenamiento para agentes RL basado en Discrete EVent System Specification (DEVS) [9]. Dicho entorno general se aplica al problema de generación y administración de energía generada a través del proceso de digestión anaeróbica, y así reducir el consumo eléctrico tradicional para distintos perfiles de consumidores industriales.

El resto de este trabajo se organiza de la siguiente forma. En la sección 2, se realiza una descripción detallada de los trabajos relacionados. Luego, en la sección 3 se presenta la definición del problema. En la sección 4, se desarrolla un caso de estudio describiendo la implementación computacional y los parámetros utilizados para las simulaciones y, por ultimo, en la sección 6, se presentan algunas conclusiones sobre el actual trabajo y las expectativas a cubrir en trabajos futuros.

2. Trabajos Relacionados

Respecto de la temática abordada en la presente propuesta, no existen antecedentes directos en los cuales se aborde de manera integral el modelado y simulación generativa del proceso de generación, almacenamiento y administración del consumo de energía eléctrica a partir de biogás. Sin embargo, en [6] se define un marco para la definición de entornos de entrenamiento no formales para agentes RL, una interface común entre estos (agente y entorno) y una gran cantidad de entornos de prueba pre-implementados. En en [4] se define un marco metodológico para el desarrollo de entornos de entrenamiento formales para el entrenamiento de agentes RL, el cual es utilizado para el desarrollo del caso de estudio propuesto en este trabajo.

Existen algunos antecedentes relacionados con el modelado y simulación del proceso de generación de biogás a partir de desechos orgánicos, de manera acotada a la evaluación de la producción de metano en relación a la variabilidad del sustrato empleado. En ese sentido, en [18] se emplea un modelo basado en Redes Neuronales (RN) para determinar la performance de un biorreactor a escala de laboratorio. En [20], [26] y [5], un modelo similar se emplea para desarrollar una metodología de análisis del proceso de producción de biogás, donde se utiliza un algoritmo de optimización basado en en Algoritmos Genéticos [12] para identificar las variables relevantes en la predicción del flujo de biogás. Con el mismo objetivo, en otros trabajos se han desarrollado modelos analíticos enfocados en mejorar la estabilidad y el desempeño del proceso de producción de biogás para aumentar la eficiencia de operación de planta. Así, en [21], se ha desarrollado

un modelo simple con el objetivo de representar adecuadamente las dinámicas de la Digestión Anaeróbica (DA) a partir del ajuste de tres sustratos principales (proteínas, carbohidratos y lípidos). En [15], se emplea un enfoque basado en simulación para estudiar los efectos del uso de residuos sólidos municipales incinerados en la producción de biogás empleando la ecuación de Gompertz para simular el rendimiento obtenido. Por otra parte, en [10], se predice el comportamiento observado de la DA a partir de un modelo bio-cinético basado en los balances de masa del sustrato, la biomasa y la producción de metano. En [27], se utilizan procesos de optimización aplicados al sustrato con el cuál se alimenta el biodigestor para obtener mejores rendimientos de metano. En [23] y [2], se aplica un modelo matemático con parámetros variables en el tiempo para describir el comportamiento dinámico de la DA de residuos animales. Por otra parte, en [17], se realiza la simulación de un proceso de DA de biomasa agrícola para analizar costos de producción. En [1], se lleva a cabo un estudio teórico de la DA para predecir la cantidad de biogás generado a partir de desechos orgánicos agrícolas, y en [19], se simula el proceso de DA a partir de un modelo de primer orden, para predecir el rendimiento de metano empleando como sustrato residuos orgánicos municipales y barro biológico. En [11], se desarrolla un complemento en MATLAB para realizar la optimización, el análisis de estabilidad y el control de una planta industrial de biogás.

Por otro lado, con respecto al modelado y simulación de gestión energética, en [13] se realiza el modelado y simulación de los distintos componentes de una red eléctrica inteligente de manera tal de tomar decisiones de diseño adecuadas y en [7] se emplea DEVS para simular la operación de una granja de turbinas de viento y asistir en la toma de decisiones relativa a su mantenimiento. En [8] se define un modelo de generación de energía solar y en [3] se define un modelo de simulación para la generación de energía a partir de biogás producido a partir de desechos biológicos. Ambos modelos basados en DEVS, serán utilizados como parte del entorno de simulación propuesto en este trabajo.

3. Definición del Problema

En la actualidad, el paulatino agotamiento de los combustibles fósiles y la necesidad global de una reducción en las emisiones de gases de efecto invernadero han atraído especial interés a los procesos de producción de energía no contaminante de manera sustentable y sostenible en el tiempo, como es el caso de la digestión anaeróbica de residuos biológicos [25]. El uso de energías renovables tiene como condicionante la incertidumbre a la hora de contar con la disponibilidad de la energía generada de manera constante (un generador eólico no puede generar energía en días de poco viento). El biogás producido a partir de desechos animales no esta exento de este problema.

A diferencia de la energía eólica, el costo de almacenar el biogás para ser utilizado en el momento adecuado no es prohibitivo, pero esto presenta el inconveniente de tener que administrar el uso del mismo, para contar con esa energía en el momento indicado para optimizar costos de producción. La dinámica de

los distintos procesos industriales hace que esta tarea no sea trivial, presentando la necesidad de contar con metodologías de control en tiempo real para la administración de estos procesos.

Para poder utilizar *RL* como marco de referencia para la solución de este tipo de problemas, es necesario contar con un entorno de entrenamiento con el cual el agente interactúe y aprenda (ver Fig. 1). El *Entorno* implica todo lo que no es directamente controlable por el Agente. Ambas entidades interactúan continuamente, el Agente ejecutando acciones A_t y el Entorno respondiendo a dichos estímulos con nuevos estados S_t y premios o castigos R_t por sus acciones, que el agente debe tratar de maximizar con el paso del tiempo [24].

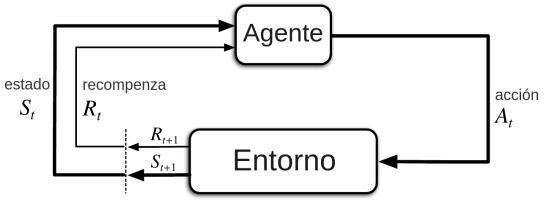


Figura 1. Interacción entre Agente y Entorno en RL

Debido a la gran cantidad de interacciones que necesita un agente RL para poder aprender una política cercana a la óptima y el riesgo de experimentar acciones que pueden implicar perjuicios económicos y sociales, desarrollar el entrenamiento del agente en un entorno real se vuelve una tarea sumamente inviable. Esta situación, presenta la necesidad de contar con entornos de entrenamiento simulados que tengan un alto grado de correlación con el entorno real y, así, reducir la incertidumbre y el riesgo de sufrir posibles perdidas económicas.

El formalismo DEVS se presenta como un gran candidato para lograr esto, ya que permite definir modelos específicos teniendo en cuenta su adaptación a cualquier ambiente en donde existen componentes que interactúan entre si siguiendo un esquema de caja negra. Asimismo, DEVS es una herramienta para el modelado y simulación basado en la teoría de sistemas y provee conceptos bien definidos de acoplamiento entre componentes, construcción modular y jerárquica de componentes siguiendo el paradigma orientado a objetos, el cual facilita el reuso de componentes de manera natural. Para realizar este trabajo, se sigue una metodología *bottom-up* donde primero se definieron los DEVS atómicos, las relaciones entre ellos y, finalmente, los DEVS acoplados [9].

En este contexto, utilizar DEVS para modelar y simular un entorno de entrenamiento para agentes RL implica el mapeo de cada uno de los componentes definidos en la Fig. 1 a uno, o mas, modelos DEVS. Los detalles de este mapeo se pueden encontrar en [3, 4], estos son omitidos por cuestiones de espacio.

4. Caso de Estudio

A la hora elegir la metodología para llevar a delante el caso de estudio, lo cual implica el desarrollo del entorno de simulación, la verificación del mismo, la selección y entrenamiento del agente RL, y la posterior evaluación del mismo, se siguieron los paso definidos en [4].

Como caso de estudio, se presenta un esquema de oferta-demanda de energía eléctrica, donde distintos perfiles de consumidores industriales utilizan energía generada mediante medios renovables (biogás y solar), para reducir el consumo eléctrico de red. En este esquema, un agente RL es el responsable de administrar el biogás generado, para reducir el costo del consumo eléctrico total de los distintos perfiles de consumidores industriales (figura 2).

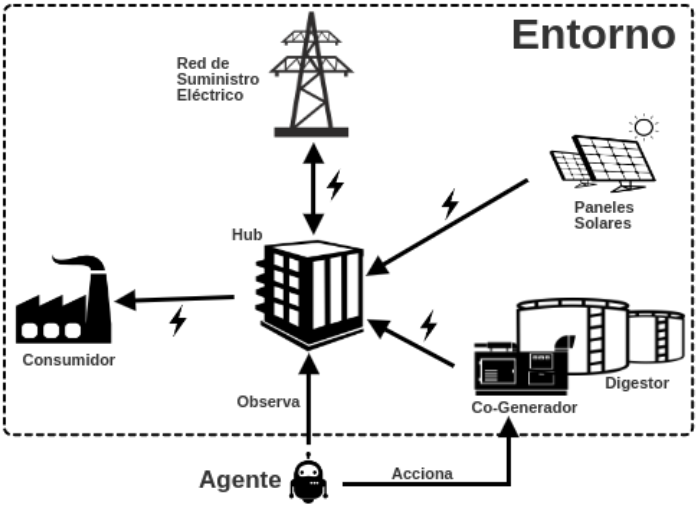


Figura 2. Entorno de control para la gestión de energías renovables.

El *consumidor* es una de las piezas variables de este esquema, dado que cada uno de estos, presenta variaciones en el nivel de la energía que requieren para operar, como así también en los horarios donde desarrollan su actividad. El componente *Hub/Controlador*, es donde llega la energía generada por los medios renovables y en caso de existir un excedente, vende el mismo a la red de suministro eléctrico. En cuanto a los medios de generación de energía renovables, la granja de paneles solares [8] no tiene ningún mecanismo de control, y el nivel de energía generado es dependiente del clima, la fecha y hora simulada. En cuando a la energía generada a partir de biogás, se simula un arribo diario de materia biológica al digestor y la correspondiente producción del biogás utilizando el modelo definido en [3]. Este biogás almacenado es utilizado para encender un co-generador que produce energía eléctrica para ser utilizada por el consumidor.

El objetivo del agente es determinar el mejor momento para encender el co-generador, en base al estado global del entorno, y así utilizar el gas almacenado para minimizar el consumo eléctrico de red total del consumidor. El estado del entorno de entrenamiento esta compuesto por:

- La energía consumida por el consumidor CE_{Wh}.
- La energía pico consumida por el consumidor CP_W.
- La energía producida por la granja de paneles solares S_{Wh}.
- La energía producida por el co-generador a partir del biogás almacenado cuando este esta encendido CO_{Wh}. En caso contrario el valor es igual a 0 (cero).
- El volumen de biogás almacenado (m^3).
- Fecha y hora de la simulación.

La recompensa percibida por el agente en cada etapa de decisión (cada 1 hora de simulación) es:

$$r_t = b_1 \times (S_{Wh} + CO_{Wh}) - (c_1 \times CE_{Wh} + c_2 \times max(CP_{Wh})) \quad (1)$$

donde b_1 es cuanto paga el distribuidor de electricidad por la energía inyectada a la red, c_1 es el costo de venta de electricidad por parte del distribuidor y c_2 un costo adicional por el máximo de energía pico utilizada por el consumidor. Se poseen 2 años de datos de consumo de los distintos perfiles de consumidores industriales, por lo que se utiliza el primer año de los datos para llevar a cabo el entrenamiento del agente, y el segundo año es utilizado como conjunto de prueba para evaluar la política de acción aprendida por el mismo.

4.1. Parámetros Experimentales

Los distintos perfiles de consumo industrial utilizados son: una fábrica de alimento balanceado para mascotas, un molino harinero y una fábrica metalúrgica. La cantidad de energía consumida por cada uno de los perfiles de consumidores utilizados en la simulación, fueron provisto por el ente de distribución eléctrica de la zona "Empresa Provincial de Energía de Córdoba" (EPEC)[1].

La cantidad de material bio-degradable que llega al digestor para la producción de biogás (una media diaria de 2,5 T) es simulada a partir de los datos de desperdicios diarios que genera una planta de procesamiento de carne porcina. Adicionalmente, se simula la producción de energía renovable a partir de una planta de generación de energía fotovoltaica de 100 m^2, que complementa la energía generada con el biogás producido por el digestor.

Los parámetros utilizados a la hora de determinar la capacidad de generación y almacenamiento de biogás, como así también, la capacidad de generación de energía eléctrica a partir del biogás producido son los definidos en [3]. En cuanto a la producción de energía fotovoltaica, el modelo utilizado esta definido en [8]. Los datos climáticos para determinar la cantidad de energía solar producida se obtuvieron de una estación meteorológica local.

[1] https://www.epec.com.ar/

Como algoritmo de aprendizaje para el entrenamiento del agente RL, se utiliza Proximal Policy Optimization (*PPO*) [22]. PPO ha demostrado a lo largo de su uso en numerosos problemas, un excelente compromiso entre capacidad de aprendizaje y eficiencia en la cantidad de tiempo necesario para encontrar una política de acción cercana a la optima. Los parámetros utilizados por el algoritmo de aprendizaje se encuentran definidos en la tabla 1.

Tabla 1. Definición de Hiper-parámetros para PPO

Hiper-parámetros	*PPO*
Capas ocultas (Tamaño)	2 (32)
Activación	ReLU - [16]
Tamaño de salida	2+1
Activación salida	Softmax(2) + Identity(1)
Método de optimización estocástica	*ADAM* - [14]
Hilos	16
Ratio de aprendizaje	3e−4
Ratio de descuento	0.99
Normalización de entrada	Si
Normalización de recompensa	Si

Como linea base de desempeño, también se evaluaron un agente aleatorio y un agente con una política base. En esta última, cada vez que el consumo energético del consumidor se dispara, y existan reservas de biogás, encenderá el cogenerador para disminuir el consumo energético de red.

5. Resultados

Al evaluar la política de acción aprendida luego de llevar a cabo el entrenamiento del agente PPO, el mismo demostró tener un desempeño superior al resto de los agentes en todos los escenarios. En la Figura 3 y en la Tabla 2, se pueden visualizar los resultados obtenidos. Al aprender una política de acción para el perfil *Fábrica Metalúrgica*, la mejora de desempeño relativa del agente PPO es de un 22 % en el costo del consumo energético con respecto al agente de menor desempeño (Linea Base) y de un 17 % con respecto al agente aleatorio. En los restantes perfiles de consumo, la mejorar relativa de desempeño es de un 0.5 % y 0.19 % respectivamente. Esto es debido a que los niveles de consumo bajos y los horarios de operatoria diurnos de los perfiles hacen que se reduzca el margen de acción posible por parte de un agente debido a la existencia de la planta de generación de energía fotovoltaica complementaria.

6. Conclusiones y Trabajo Futuro

En este trabajo, se propuso un entorno de simulación basado en DEVS para entrenar agentes RL aplicado a la generación y administración de biogás con

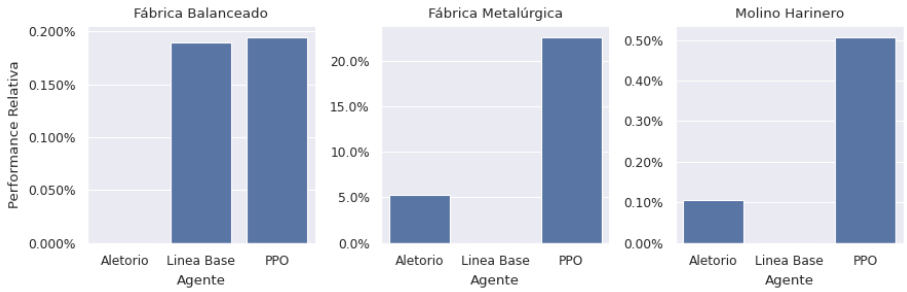


Figura 3. Desempeño relativo por agente para cada perfil de consumidor.

Tabla 2. Mejora de desempeño relativa entre agentes.

Perfil Consumidor	**Random**	**Fixed**	**PPO**
Fábrica Balanceado	0.0 %	0.0018 %	**0.0019** %
Fábrica Metalúrgica	0.05 %	0.0 %	**0.22** %
Molino Harinero	0.001 %	0.0 %	**0.005** %

el fin de facilitar el proceso de toma de decisiones en tiempo real. Esta solución permite acelerar los tiempos de entrenamiento del agente RL, mejorando su performance a menor costo, combinando las ventajas de usar RL en un proceso donde la toma de decisiones secuencial se da en tiempo real con DEVS como herramienta para especificar y simular dicho proceso y su dinámica. En particular, en el contexto de energías renovables, esta propuesta permite analizar diferentes escenarios de uso de energías alternativas, en este caso biogás, modificando las demandas de consumo, la capacidad de producción de metano, como así también pudiendo escalar el sistema al incorporar un mayor número de componentes, por ejemplo, componentes que representen la producción de biogás a partir de materia orgánica diversa. Asimismo, permite mejorar la gestión de dichos procesos mediante la modificación de los parámetros.

Si bien existen propuestas basadas en modelos más precisos para predecir la cantidad de biogás producido por digestión anaeróbica, el objetivo principal de este trabajo es desarrollar un modelo de simulación que permita capturar lo mejor posible el proceso que se está simulando, es decir, su complejidad, considerando no solo una variable sino varias simultáneamente como, por ejemplo, la producción de biogás, el consumo, factores externos (eventos) que pueden afectar, entre otros. Es importante destacar ésto ya que este entorno de simulación es el entorno con el que interactúa el agente RL, del cual aprende. Al ser un enfoque basado en DEVS provee ventajas como la definición de elementos de simulación específicos del dominio (problemática energética), el desacople entre el modelo, el simulador (interno) y el marco experimental, facilitando la reutilización de componentes de simulación y su evolución en el tiempo. Esto permite una construcción modular y jerárquica del proceso que se está controlando, pudiendo capturar tanta complejidad como se necesite para que el agente aprenda

una política. Todas estas características definen una herramienta flexible para tomar decisiones relacionadas a la generación, almacenamiento y uso de energías alternativas.

En trabajos futuros, se pretende desarrollar una herramienta de software que incluya el marco general de entrenamiento de agentes RL usando modelos de simulación basados en DEVS para facilitar su usabilidad en diferentes problemas. En particular, respecto a la problemática de energías renovables, se pretende trabajar con otros algoritmos de RL, para así lograr una mejor política de administración de la energía renovable generada.

Referencias

[1] Spyridon Achinas y Gerrit Jan Willem Euverink. "Theoretical analysis of biogas potential prediction from agricultural waste". En: *Resource-Efficient Technologies* 2.3 (2016), págs. 143-147.

[2] BK Bala. "System dynamics modelling and simulation of biogas production systems". En: *Renewable energy* 1.5-6 (1991), págs. 723-728.

[3] Ezequiel Beccaria, Veronica Bogado y Jorge A Palombarini. "A devs-based simulation model for biogas generation for electrical energy production". En: *2018 IEEE Biennial Congress of Argentina (ARGENCON)*. IEEE. 2018, págs. 1-8.

[4] Ezequiel Beccaria, Veronica Bogado y Jorge Andres Palombarini. "A DEVS Based Methodological Framework for Reinforcement Learning Agent Training". En: *IEEE Latin America Transactions* 19.4 (2021), págs. 679-687.

[5] Tetyana Beltramo y col. "Artificial neural network prediction of the biogas flow rate optimised with an ant colony algorithm". En: *Biosystems Engineering* 143 (2016), págs. 68-78.

[6] Greg Brockman y col. *OpenAI Gym*. 2016. eprint: `arXiv:1606.01540`.

[7] Eunshin Byon y col. "Simulation of wind farm operations and maintenance using discrete event system specification". En: *Simulation* 87.12 (2011), págs. 1093-1117.

[8] Carlos M. Chezzi y col. "Modelo DEVS para Evaluación de Asignación de Energía Solar para Vivienda Estándar". En: *V Congreso Nacional de Ingeniería Informática - Sistemas de Información*. Facultad Regional Santa Fe - Univercidad Tecnologíca Nacional. 2017, págs. 890-898.

[9] A. I. Concepcion y B. P. Zeigler. "DEVS Formalism: A Framework for Hierarchical Model Development". En: *IEEE Transactions on Software Engineering* 14.2 (feb. de 1988), págs. 228-241. ISSN: 0098-5589. DOI: `10.1109/32.4640`.

[10] M Fedailaine y col. "Modeling of the anaerobic digestion of organic waste for biogas production". En: *Procedia Computer Science* 52 (2015), págs. 730-737.

[11] Daniel Gaida y col. "MATLAB toolbox for biogas plant modelling and optimization". En: *Progress in Biogas II-Biogas production from agricultural biomass and organic residues* (2011), págs. 67-70.

[12] David E Goldberg. "Genetic algorithms in search, optimization, and machine learning, 1989". En: *Reading: Addison-Wesley* (1989).
[13] Moath Jarrah. "Modeling and simulation of renewable energy sources in smart grid using DEVS formalism". En: *Procedia Computer Science* 83 (2016), págs. 642-647.
[14] Diederik P Kingma y Jimmy Ba. "Adam: A method for stochastic optimization". En: *arXiv preprint arXiv:1412.6980* (2014).
[15] HM Lo y col. "Modeling biogas production from organic fraction of MSW co-digested with MSWI ashes in anaerobic bioreactors". En: *Bioresource Technology* 101.16 (2010), págs. 6329-6335.
[16] Andrew L Maas, Awni Y Hannun y Andrew Y Ng. "Rectifier nonlinearities improve neural network acoustic models". En: *Proc. icml*. Vol. 30. 1. 2013, pág. 3.
[17] Maizirwan Mel y col. "Simulation study for economic analysis of biogas production from agricultural biomass". En: *Energy Procedia* 65 (2015), págs. 204-214.
[18] Vijay V Nair y col. "Artificial neural network based modeling to evaluate methane yield from biogas in a laboratory-scale anaerobic bioreactor". En: *Bioresource technology* 217 (2016), págs. 90-99.
[19] A Nielfa, R Cano y M Fdz-Polanco. "Theoretical methane production generated by the co-digestion of organic fraction municipal solid waste and biological sludge". En: *Biotechnology Reports* 5 (2015), págs. 14-21.
[20] H Abu Qdais, K Bani Hani y N Shatnawi. "Modeling and optimization of biogas production from a waste digester using artificial neural network and genetic algorithm". En: *Resources, Conservation and Recycling* 54.6 (2010), págs. 359-363.
[21] Anna Schneider. "Dynamic modeling and simulation of biogas production based on anaerobic digestion of gelatine, sucrose and rapeseed oil". Tesis doct. Jacobs University Bremen, 2016.
[22] John Schulman y col. "Proximal Policy Optimization Algorithms". en. En: (2017), pág. 12.
[23] Iv Simeonov, V Momchev y D Grancharov. "Dynamic modeling of mesophilic anaerobic digestion of animal waste". En: *Water Research* 30.5 (1996), págs. 1087-1094.
[24] *Sutton & Barto Book: Reinforcement Learning: An Introduction*. 2018.
[25] Peter Weiland. "Biogas production: current state and perspectives". En: *Applied microbiology and biotechnology* 85.4 (2010), págs. 849-860.
[26] Kaan Yetilmezsoy y col. "Development of ann-based models to predict biogas and methane productions in anaerobic treatment of molasses wastewater". En: *International journal of green energy* 10.9 (2013), págs. 885-907.
[27] Martin Zaefferer, Daniel Gaida y Thomas Bartz-Beielstein. "Multi-fidelity modeling and optimization of biogas plants". En: *Applied Soft Computing* 48 (2016), págs. 13-28.

Evaluación de Variantes de la Metaheurística VNS para el Problema de Planificación de Máquinas Paralelas

Claudia Ruth Gatica, Silvia Marta Molina y Guillermo Leguizamón
LIDIC, Universidad Nacional de San Luis
Ejército de Los Andes 950 - Local 106, San Luis, Argentina,
{crgatica, smolina, legui}@unsl.edu.ar

Resumen. VNS (*Variable Neighborhood Search*) es una metaheurística de trayectoria y usa diferentes estructuras de vecindarios siguiendo algún criterio pre-establecido para realizar la búsqueda. En este trabajo se proponen variantes de VNS estándar (o simplemente VNS) para mejorar su desempeño introduciendo cambios en las secuencias de vecindarios utilizadas y/o mecanismos de exploración considerando el problema de Planificación de Máquinas Paralelas. Las variantes propuestas son: VNS+R (VNS *Random*) con selección de vecindario aleatoria; VNS+LHS (VNS *Latin Hypercube Sample*) con preselección de vecindarios a través de Cuadrados Latinos; VNS+E (VNS *Exploratory*) que intensifica la exploración del espacio de búsqueda y por último, VNS+ER (VNS *Exploratory&Random*) que combina aspectos funcionales de VNS+R y VNS+E. Los resultados muestran que las variantes que intensifican la exploración en el espacio de búsqueda con selección aleatoria de estructuras de vecindario, mejoran al desempeño de VNS, variante representada por el algoritmo VNS+ER.

Palabras claves: Planificación de Máquinas Paralelas, Tardanza Máxima, Búsqueda en Vecindarios Variables, Metaheurísticas.

1 Introducción

Las metaheurísticas de trayectoria (S-metaheurísticas, de aquí en adelante) son algoritmos de búsqueda usados para resolver problemas NP-duros tal como son los problemas de planificación. En particular VNS (Mladenović y Hansen [6]) realiza una búsqueda sistemática de estructuras de vecindarios, en donde la secuencia de exploración y tamaño de cada vecindario es crucial en el diseño del algoritmo [5].

La planificación (*scheduling*) de actividades es un proceso de toma de decisión que tiene un papel importante en los sistemas de producción y multiprocesadores, en los entornos de fabricación y distribución de información, y de transporte. En particular, este artículo considera el problema de planificación o *scheduling* de máquinas paralelas idénticas sin restricciones con el objetivo de minimizar la Tardanza Máxima

(*Maximum Tardiness*). El entorno de máquinas paralelas se ha estudiado durante varios años debido a su importancia tanto a nivel académico como a nivel industrial.

En una revisión bibliográfica se observa que la metaheurística VNS ha sido aplicada a diversos problemas académicos y del mundo real, aplicada sola o en forma híbrida en combinación con otras metaheurísticas. En [2] se presenta un VNS para proveer las soluciones iniciales a otras metaheurísticas implementadas para resolver el problema de planificación de procesos dinámicos y asignación de fechas de vencimiento (DIPPSDDA), en donde la función objetivo fue minimizar la puntualidad y la tardanza (E/T). En [4] se aplica al problema de planificar un conjunto de trabajos independientes, con tiempos de configuración dependientes de la secuencia y restricciones de compatibilidad de tareas, en un conjunto de máquinas paralelas, con el objetivo de minimizar el tiempo máximo de finalización (*makespan*). En [9] se aplica un algoritmo híbrido VNS-GSA (VNS y el *Algoritmo de búsqueda gravitacional*) para la planificación de máquina única y de máquinas paralelas, los objetivos de minimizar el *maximum earliness* y el número de tareas (*jobs*) tardías con efecto de aprendizaje basado en la posición (donde el tiempo real de procesamiento del trabajo es una función de su posición) y los tiempos de procesamientos dependientes de la configuración. En [17] se presenta VNS para el problema de planificación de máquinas paralelas idénticas, con el objetivo de minimizar el tiempo total de finalización. Para la configuración de parámetros, en [3] se presenta un VNS para el diseño de los parámetros de los controladores de amortiguamiento en sistemas de potencia multi-máquina: tal como estabilizadores del sistema de potencia (PSS) y el controlador de flujo de potencia interlínea-amortiguación de oscilación de potencia (IPFC-POD). Para el problema del mundo real llamado despacho de energía en redes distribuidas inteligentes se propone en [14] el algoritmo híbrido VNS-DEEPSO (*Differential Evolutionary Particle Swarm*), para obtener soluciones para minimizar los costos operacionales y maximizar los ingresos de la red inteligente. Este algoritmo fue diseñado con algunas mejoras que permiten la evaluación de ecuaciones formadas a partir de ecuaciones algebraicas no lineales y ecuaciones diferenciales en la que es imposible obtener su derivada con un modelo matemático exacto [3].

En el presente trabajo se presentan y analizan diferentes variantes del algoritmo VNS. Dichas variantes (VNS+R, VNS+LHS, VNS+E y VNS+ER) surgen a partir de cambios en el esquema básico de búsqueda y selección de vecindarios, con el fin de mejorar su desempeño sobre el problema minimización de la tardanza máxima.

La organización del presente trabajo es la siguiente. En la sección 2 se describe y formula el problema de planificación de máquinas paralelas. En la sección 3 se describen las variantes propuestas de VNS. En la sección 4 se detalla el diseño de los experimentos. En la sección 5 se muestran y analizan los resultados obtenidos. Finalmente, en la sección 6 se presentan las conclusiones.

2 EL Problema de Planificación de Máquinas Paralelas

El problema de planificación (*scheduling*) de máquina paralelas sin restricciones es un problema común en los sistemas reales de manufacturación y producción, y es también un problema de interés desde el punto de vista teórico y práctico.

En la literatura, el problema estudiado se denota como: $P_m||T_{max}$. El primer campo describe el ambiente de las máquinas P_m, el segundo contiene las restricciones, aquí se puede notar que el problema no tiene restricciones, por lo tanto, el campo está vacío, y el tercero provee la función objetivo a ser minimizada T_{max} [12], [16].
Este problema de planificación se puede expresar de la siguiente manera: existen *n* tareas para ser procesadas sin interrupción en algunas de las *m* máquinas idénticas que pertenecen al sistema P_m; cada máquina no puede procesar más de una tarea a la vez. La tarea t_j, (j=1, 2, 3, …, *n*) está disponible en el tiempo cero. La misma requiere un tiempo de procesamiento p_j positivo e ininterrumpido sobre una máquina y también tiene una fecha de vencimiento d_j en la cual su procesamiento debería estar finalizado. Para una orden (*schedule*) de procesamiento de tareas dada, el tiempo de completitud más temprano C_j y el tiempo máximo de tardanza T_j= {C_j – d_j, 0}, son fácilmente calculados. El problema es encontrar una orden (*schedule*) óptima que minimice el valor de la función objetivo de la ecuación:

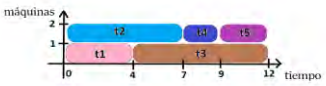
$$\textit{Maximum Tardiness}: T_{max} = max_j (T_j) \qquad (1)$$

Este problema es considerado NP-duro cuando $2 \leq m \leq n$ [12], [16]. Una instancia del problema con *n=5* tareas y *m=2* máquinas es mostrada en el Ejemplo 1.

Ejemplo 1: Instancia con *n*=5 tareas t_j, *m*=2, con tiempos de procesamiento p_j y fechas de vencimientos d_j; para j=1, …, 5.

t_1	t_2	t_3	t_4	t_5	t_j
(4, 15)	(7, 20)	(8,19)	(2, 6)	(3, 8)	(p_j, d_j)

Figura 1: Diagrama de Gantt de una posible solución para el Ejemplo 1.



El diagrama de Gantt de la Figura 1 describe una planificación **S** (*por schedule*) de tareas en dos máquinas m_1 y m_2, en donde S= [{m_1, (t_1, p_1, d_1), (t_3, p_3, d_3)}, {m_2, (t_2, p_2, d_2), (t_4, p_4, d_4), (t_5, p_5, d_5)}] y la distribución de las tareas= {t_1, t_2, t_3, t_4, t_5} es: t_1 y t_3 se ejecutan en m_1 y t_2, t_4 y t_5 en m_2. Por consiguiente, los tiempos de completitud respectivos son C_j=((t_1,4), (t_2,7), (t_3,12), (t_4, 9), (t_5, 12)) y los tiempos de tardanza computados en cada tarea T_j= (0, 0, 0, 3, 4), donde la tardanza máxima es T_{max}= 4.

3 La S-metaheurística VNS y las variantes propuestas

La S-metaheurísticas han mostrado su eficacia para abordar varios problemas de optimización en diferentes dominios [2], [3], [9], [14], entre otros. Éstas realizan una exploración a través de estructuras de vecindarios [5] mediante un procedimiento iterativo a partir de una solución inicial, generalmente aleatoria, y aplicando operadores que definen la trayectoria de búsqueda.

VNS considera de manera secuencial un conjunto de distintas estructuras de vecindarios N_i donde i=1, 2, 3, …, k_{max}, son definidas en el espacio de soluciones del problema. La secuencia previa es realizada en forma sistemática o aleatoria a fin de obtener soluciones de alta calidad a los efectos de escapar de los óptimos locales. Una estructura de vecindario N en el espacio de soluciones X de una solución dada x, es denotada como $N(x) \subset X$, en donde $N: X \rightarrow \wp(X)$, siendo $\wp(X)$ el conjunto potencia de x [15].

En nuestro trabajo, se entiende como secuencia de estructuras de vecindarios al orden de inspección de cada una de las estructuras de vecindarios N_i definidas como la entrada del algoritmo VNS. Un aspecto interesante de VNS es que presenta un esquema básico que sirve como marco flexible para implementar variantes heurísticas [15] y además demuestran un buen desempeño en resolver problemas complejos.

3.1 Algoritmo estándar VNS

VNS (`Algoritmo 1`) recibe como entrada un conjunto de estructuras de vecindarios definidas de antemano N_i, i=1, …, k_{max}. Cada iteración del algoritmo está compuesta de tres pasos: (1) sacudida (*shaking*), (2) búsqueda local y (3) moverse (*move*). La solución inicial S_0 es sacudida en el vecindario actual N_k, una búsqueda local (`Algoritmo 2`) es aplicada sobre la solución S_1 (se obtiene S_1 mediante el operador de perturbación N_k) para obtener una solución candidata S_2.

Algoritmo 1 VNS()	**Algoritmo 2 LS() {Búsqueda Local}**
1: **Entrada**: un conjunto de estructuras de vecindarios N_i i=1,…,k_{max} 2: $S=S_0$ 3: k=1 4: **Repetir** 5: **Mientras** $k \leq k_{max}$ **Hacer** 6: $S_1=N_k(S)$ {Shaking} 7: S_2=**LS**(S_1) {Búsqueda Local en N_k} 8: **Si** $f(S_2) \leq f(S)$ **Entonces** 9: $S = S_2$ 10: k = 1 11: **Sino** 12: k = k + 1 {Move} 13: **Fin Si** 14: **Fin Mientras** 15: **Hasta Un Criterio de Parada** 16: **Salida**: La mejor solución encontrada S	1: **Entrada**: $S=S_0$ {Solución Inicial y una estructura de vecindarios N_k} 2: **Repetir** 3: $S_1= N_k(S)$ {Genera un vecino de S_1 a partir de S aplicando N_k} 4: **Si** $f(S_1) \leq f(S)$ **Entonces** 5: $S=S_1$ 6: **Sino** 7: Detener 8: **Fin Si** 9: **Hasta Detener** 10: **Salida**: Solución final S

La solución actual S es reemplazada por el nuevo óptimo local S_2, si y sólo si, S_2 es mejor que S y el mismo procedimiento es recomenzado en N_1, si S_2 no es mejor, la búsqueda se mueve al siguiente vecindario N_{k+1}, y se genera una nueva solución en N_{k+1} que se intenta mejorar hasta agotar las estructuras de vecindario disponibles.

El estudio de las variantes del algoritmo VNS se basó en la hipótesis que un cambio en la elección de la secuencia de vecindarios podría ayudar a escapar eficientemente de mínimos locales y, por ende, mejorar su desempeño. En las

siguientes subsecciones se describen las variantes implementadas y estudiadas en el presente trabajo.

3.2 Variante VNS+R

El algoritmo VNS+R (Algoritmo 3) utiliza una secuencia de estructuras de vecindarios de entrada aleatoria, la selección de la siguiente estructura de vecindario N_i se obtiene de algunas de las N_r (Algoritmo 3, líneas 12-13).

Algoritmo 3 VNS+R()	**Algoritmo 4 VNS+E()**
1: **Entrada**: un conjunto de estructuras de vecindarios N_i i=1,…,k_{max} 2: S=S_0 3: k=1 4: **Repetir** 5: **Mientras** k ≤ k_{max} **Hacer** 6: S_1=N_k(S) 7: S_2=**LS**(S_1){Búsqueda Local en N_k} 8: **Si** f(S_2) ≤ f(S) **Entonces** 9: S = S_2 10: k = 1 11: **Sino** 12: r = random(1, k_{max}) (r =! k) 13: k = r; 14: **Fin Si** 15: **Fin Mientras** 16: **Hasta Un Criterio de Parada** 17: **Salida**: La mejor solución encontrada S	1: **Entrada**: Un conjunto de estructuras de vecindarios N_i i=1,…,k_{max} 2: S=S_0 3: k=1 4: **Repetir** 5: **Mientras** k ≤ k_{max} **Hacer** 6: S_1 = N_k(S) 7: S_2 = **LSE**(S_1) {Búsqueda Local Explorativa} 8: **Si** f(S_2) ≤ f(S) **Entonces** 9: S = S_2 10: k = 1 11: **Sino** 12: k = k + 1 13: **Fin Si** 14: **Fin Mientras** 15: **Hasta Un Criterio de Parada** 16: **Salida**: La mejor solución encontrada S

Algoritmo 5 LSE() {Búsqueda Local Explorativa}	**Algoritmo 6 VNS+ER()**
1: **Entrada**: S=S_0 {Solución Inicial y estructura de vecindario N_k} 2: **Repetir** 3: genera$_n$(S) {Genera n vecinos a partir de S aplicando N_k} 4: S_1 = rank-best-neighbor() {Ordena y selecciona el mejor vecino de S} 5: **Si** f(S_1) ≤ f(S) **Entonces** 6: S=S_1 7: **sino** 8: Detener 9: **Fin Si** 10: **Hasta** Detener 11: **Salida**: Solución final S	1: **Entrada**: un conjunto de estructuras de vecindarios N_i i=1,…,k_{max} 2: S=S_0 3: k=1 4: **Repetir** 5: **Mientras** k ≤ k_{max} **Hacer** 6: S_1 = N_k(S) 7: S_2 = **LSE**(S_1) {Búsqueda Local Explorativa en N_k} 8: **Si** f(S_2) ≤ f(S) **Entonces** 9: S = S_2 10: k = 1 11: **Sino** 12: r = random(1, k_{max}) 13: k = r; 14: **Fin Si** 15: **Fin Mientras** 16: **Hasta Un Criterio de Parada** 17: **Salida**: La mejor solución encontrada S

3.3 Variante VNS + LHS

VNS+LHS tiene una etapa previa donde se procesa un conjunto de estructuras de vecindarios $\{N_1, ..., N_{kmax}\}$ para generar la secuencia a aplicar durante la búsqueda. Ésta se selecciona a partir de un conjunto de posibles secuencias de estructuras de vecindarios construidas con cuadrados latinos. Por tanto, a fin de construir dicho conjunto de posibles secuencias de estructuras de vecindarios y luego elegir la más conveniente, se propuso: 1) generar una combinación de secuencias uniformes en un espacio de diseño con el enfoque de cuadrados latinos, 2) evaluar las secuencias con la función objetivo del problema sobre un conjunto acotado de instancias y aplicando pruebas estadísticas apropiadas, y finalmente 3) elegir aquella que haya resultado más conveniente. A partir de allí, VNS+LHS se comporta como VNS.

3.4 Variante VNS+E

La variante VNS+E (`Algoritmo 4`) se implementó extendiendo en la búsqueda local (`Algoritmo 4`, línea 7) a un número mayor que 1, respecto a los posibles vecinos de la solución actual. Se realiza luego un ranking mediante la evaluación de la función objetivo y se selecciona a la solución de mejor calidad como la mejor solución con respecto a la solución actual (LSE, `Algoritmo 5`).

3.5 VNS+ER

El algoritmo VNS+ER (`Algoritmo 6`) es la última variante que se propone, la cual es obtenida a través de la combinación del VNS+E y el VNS+R, éste utiliza también la búsqueda LSE (`Algoritmo 5`).

4 Diseño de experimentos

Todos los algoritmos fueron implementados en lenguaje C++, en la biblioteca MALLBA [1]. Se establecen 20 instancias del problema, para cada una de estas instancias los algoritmos se ejecutan 30 veces y se configura como criterio de parada al número máximo de evaluaciones en 300.000. Los experimentos fueron ejecutados en un sub-cluster formado por once nodos cuyas características son las siguientes: CPUs de 64 bits cada uno con Intel Q9550 Quad Core 2.83GHz, 4GB DDR3 1333Mz de memoria, 160 Gb SATA disco duro y Asus P5Q3 placa madre.

4.1 Instancias del problema

Con el fin de encontrar instancias de prueba del problema, se realizó una revisión de la literatura en donde se aborda el problema de planificación de máquinas paralelas [3], [9] y [17] entre otros. Si bien se observó que existen instancias para el problema de planificación de máquinas paralelas, las mismas en general involucran más información dependiendo de la función objetivo estudiada y de las restricciones del

problema como, por ejemplo, tiempos dependientes de la configuración (*setup times*), o tiempos de lanzamientos al sistema (*release times*), tiempos de deterioro de tareas (*jobs*). Sin embargo, para nuestro problema sin restricciones (en las asignaciones) no fueron halladas ninguna, por tal motivo el conjunto de instancias fue construido a partir de datos obtenidos en la OR-Library [7], del problema *weighted tardiness problem*. Estos datos consistieron en pares (p_j, d_j), donde p_j es el tiempo de procesamiento y d_j es el (*due date*) la fecha de vencimiento o expiración de la tarea t_j para instancias de tamaño de 100 tareas, tales instancias son numeradas con #i (i=1, …, 125). Para los experimentos se seleccionaron 20 instancias no consecutivas en forma aleatoria por diferentes grupos. Además, debemos tener en cuenta que éstas fueron generadas con un factor de *tardiness* con dificultad incremental, de manera tal que, a mayor índice de identificación #i, es mayor el factor de *tardiness* con el que fueron generadas, es un motivo por el cual las instancias de mayor índice tienden a ser más difíciles de optimizar mediante la función objetivo T_{max}. Tales están disponibles previa solicitud (email: crgatica@email.unsl.edu.ar) .

La función de evaluación de la solución toma como entrada una instancia y calcula el valor de *maximum tardiness* de acuerdo a la ecuación (1) de la sección 2.

4.2 Ajustes de parámetros

Una de las ventajas de los algoritmos VNS consiste en su escaso número de parámetros para configurar, más allá de establecer el criterio de parada y de definir el conjunto de estructuras de vecindarios de entrada, estos son: cantidad de estructuras de vecindarios k_{max}=6, número máximo de iteraciones en la búsqueda local es igual 1.000, número de vecinos adicionales en la búsqueda local exploratoria es igual a 10, número máximo de evaluaciones de la función objetivo es igual a 300.000.

Las estructuras de vecindarios N_k están dadas por los operadores de perturbación N-swap (N_1), 2-opt (N_2), 3-opt (N_3), 4-opt (N_4), Shift (N_5) y Scramble (N_6). Una descripción de tales operadores se encuentra en la bibliografía [5]. La representación de cada solución es una permutación de enteros.

4.3 Métricas de evaluación del desempeño de los algoritmos

Las métricas de evaluación definidas para el estudio experimental fueron: *Bench*, el valor de referencia u óptimo conocido hasta el momento [18]; *Best*, el mejor valor de la función objetivo obtenido; *Ebest*=(*Best*-*Bench*/*Bench*)*100, el error porcentual; *Mbest*, el valor medio de *Best*; *Mebest*, el valor medio de *Ebest*; *Mediana*, es el valor central del conjunto de valores *Best* ordenados de menor a mayor; *Mevals*, el valor medio de ejecuciones donde se encontró el valor *Best*; *M. Times Runs*, el tiempo promedio de ejecución en nanosegundos.

5 Análisis de resultados

En esta sección se presenta el análisis de resultados de los experimentos computacionales realizados considerando las 20 instancias de tamaño *n*=100 tareas y

m=5 máquinas del problema de *planificación de máquinas paralelas idénticas sin restricciones* para minimizar la función objetivo de *maximum tardiness* (T_{max}).

La Tabla 1 sintetiza los resultados obtenidos por VNS y las variantes VNS+R, VNS+LHS, VNS+E y VNS+ER. Se muestran los valores *Bench, Best* y el error porcentual *Ebest.* Las entradas en la tabla marcadas en negrita corresponden a valores iguales o menores a los valores de referencia *Bench*, en estos casos los valores del error porcentual *Ebest* son cero o negativo y están marcados en itálica y subrayados.

Si miramos en las columnas correspondientes a VNS+R los valores (*Best* y *Ebest*) y los comparamos con los valores de VNS, vemos que no se obtuvieron mejores valores, salvo en las instancias del problema 6 y 41. La columna de la variante VNS+LHS, muestra mejores valores en varias instancias del problema con respecto a VNS, de la misma manera con las variantes VNS+E y VNS+ER notamos mejores valores, por lo tanto, se cumple la hipótesis planteada.

Utilizamos el test estadístico no paramétrico *Ranking de Friedman* [10] y [11] para realizar una comparación de las medias experimentales obtenidas tomando como métrica de evaluación al valor promedio de la *Mediana*, de esta manera obtuvimos un *ranking* de los algoritmos.

Tabla 1: Comparación de los valores *Bench* vs. *Best* y *Ebest* de los algoritmos propuestos.

Variantes		**VNS**		**VNS+R**		**VNS+LHS**		**VNS+E**		**VNS+ER**	
Inst.	*Bench*	*Best*	*Ebest*	*Best*	*Ebest*	*Best*	*Ebest*	*Best*	*Ebest*	*Best*	*Ebest*
1	**548**	**547**	*-0,182*	553	0,912	**542**	*-1,095*	**542**	*-1,095*	**539**	*-1,642*
6	**1594**	**1576**	*-1,129*	**1584**	*-0,627*	**1571**	*-1,443*	**1574**	*-1,255*	**1574**	*-1,255*
11	**2551**	**2548**	*-0,118*	2560	0,353	**2544**	*-0,274*	**2547**	*-0,157*	**2549**	*-0,078*
19	**3703**	3728	0,675	3741	1,026	3712	0,243	3717	0,378	**3703**	*0,000*
21	**5187**	**5187**	*0,000*	5207	0,386	**5184**	*-0,058*	**5181**	*-0,116*	**5177**	*-0,193*
26	**84**	99	17,857	113	34,524	99	17,857	**84**	*0,000*	**75**	*-10,714*
31	**1134**	1171	3,263	1170	3,175	1135	0,088	**1135**	*0,088*	**1132**	*-0,176*
36	**2069**	2111	2,030	2081	0,580	**2061**	*-0,387*	**2071**	*0,097*	**2061**	*-0,387*
41	**3651**	**3626**	*-0,685*	**3649**	*-0,055*	**3608**	*-1,178*	**3647**	*-0,110*	**3608**	*-1,178*
46	**4439**	4445	0,135	4456	0,383	4443	0,090	4440	0,023	**4443**	*0,090*
56	**617**	667	8,104	708	14,749	**609**	*-1,297*	**617**	*0,000*	**609**	*-1,297*
61	**1582**	1620	2,402	1743	10,177	1589	0,442	1615	2,086	**1580**	*-0,126*
66	**2360**	2403	1,822	2474	4,831	2364	0,169	**2359**	*-0,042*	**2359**	*-0,042*
71	**3786**	3829	1,136	3897	2,932	3797	0,291	3802	0,423	**3787**	0,026
86	**1194**	1259	5,444	1289	7,956	1214	1,675	1225	2,596	1209	1,256
91	**2204**	2284	3,630	2397	8,757	2244	1,815	2252	2,178	**2221**	*0,771*
96	**3185**	3196	0,345	3214	0,911	3196	0,345	3196	0,345	**3186**	*0,031*
111	**1365**	1621	18,755	1689	23,736	1486	8,864	1462	7,106	1437	5,275
116	**2222**	2392	7,651	2513	13,096	2279	2,565	2318	4,320	2288	2,970
121	**2999**	3230	7,703	3265	8,870	3150	5,035	3065	2,201	3046	1,567

Los resultados indican que existen diferencias significativas entre los algoritmos comparados donde el algoritmo VNS+ER muestra mejor desempeño que los restantes. Para comparar cada algoritmo entre sí y verificar si existen diferencias entre ellos, utilizamos un procedimiento *post-hoc*, que al igual que los test antes mencionados éste también es proveído por la herramienta *Controltest*, [8].

Al aplicar el procedimiento post-hoc de comparaciones de a pares entre el algoritmo de control (VNS+ER) respecto a VNS y VNS+R los valores *p-values* son menores a 0,05 por lo tanto, las parejas son significativamente diferentes, en cambio los *p-values* del VNS+LHS y VNS+E son mayores a 0,05 lo que implica que tales

algoritmos no son diferentes estadísticamente al algoritmo de control. Para finalizar el análisis de los resultados experimentales concluimos que la variante VNS+R no mejora el desempeño de VNS, en cambio las otras tres variantes VNS+LHS, VNS+E y VNS+ER si lo hacen, siendo la de mejor desempeño la variante VNS+ER en primer lugar, VNS+E en segundo y por último VNS+LHS según el ranking de *Friedman*. Finalmente, se observan (aunque no reportados aquí) tiempos de ejecución promedio de manera incremental según el siguiente orden: VNS+LHS, VNS+R, VNS, VNS+E y VNS+ER.

6 Conclusiones

En este trabajo presentamos la S-metaheurística VNS para el problema de *planificación de máquinas idénticas paralelas sin restricciones* que minimiza la función objetivo de *Tardanza Máxima* T_{max}. El objetivo del estudio consistió en proponer variantes heurísticas de la versión VNS estándar que mejoren el desempeño de la misma. Los resultados experimentales obtenidos muestran que el algoritmo VNS+ER fue el que obtuvo mejores resultados, seguido VNS+E y evidenciando diferencias estadísticamente significativas con VNS. A pesar de los buenos resultados obtenidos se observó una gran variabilidad de VNS+ER en el número promedio de evaluaciones de la función objetivo usados para alcanzar el óptimo. Es decir, estos dependen de la instancia del problema y de esta manera puede usar muy pocas o todas las evaluaciones permitidas dentro del máximo establecido por el criterio de parada. Esto nos permite concluir que en trabajos futuros será necesario realizar nuevos experimentos para analizar tanto un conjunto más grande de instancias del problema y de mayores dimensiones, como así también un análisis profundo del comportamiento de los algoritmos propuestos, en especial de VNS+ER.

Referencias

1. Alba, E., Almeida F., Blesa M., Cotta C., Diaz M., Dorta I., Gabarró J., González J., León C., Moreno L., Petit J., Roda J., Rojas A., Xhafa F., MALLBA: A library of skeletons for combinatorial optimization. Proceedings of the Euro-Par, pp. 927-932, 2002.
2. C. Erden, H. I. Demir and A. H. Kökçam: Solving Integrated Process Planning, Dynamic Scheduling, and Due Date Assignment Using Metaheuristic Algorithm. Hindawi Mathematical Problems in Engineering Volume 2019, 2019.
3. E. L. Franca Senne and A. A. Chaves: A VNS Heuristic for an Industrial Parallel Machine Scheduling Problem, ICIEOM – CIO, Valladolid, Spain, 2013.
4. E. Fortesa, L.H. Macedob Percival, B. de Araujo, R. Romero: A VNS algorithm for the design of supplementary damping controllers for small signal stability analysis. International Journal of Electrical Power & Energy Systems, vol. 94, no. 1, pp. 41–56, 2018.
5. E. G. Talbi: Metaheuristics from design to implementation, by John Wiley & Sons, C., 2009.
6. P. Hansen and N. Mladenovic, VNS: principles and applications. Eur. J. Oper. Res., 130, 449–467,1999.
7. J. E. Beasley, OR-Library: distributing test problems by electronic mail, Journal of the Operational Research Society, pp 1069-1072, 1990.

8. J. Derrac, S. García, D. Molina, and F. Herrera: A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. Swarm and Evolutionary Computation, 2011.
9. J. Pei, B. Cheng, X. Liu, P. M. Pardalos and M. Kong: Single-machine and parallel-machine serial-batching scheduling problems with position-based learning effect and linear setup time. A. Oper Res. 272:217–241, 2019.
10. M. Friedman: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of American Statistical Association, 3:674–701, 1937.
11. M. Friedman: A comparison of alternative test of significance for the problem of the rankings. Annals of Mathematical Statistics, 11:86–92, 1940.
12. M. Pinedo, Scheduling: Theory, Algorithms and System, Prentice Hall, 1995.
13. Muestreo de hipercubo latino, https://es.wikipedia.org/wiki/Muestreo_de_hipercubo_latino.
14. P. J. García-Guarín; J. Cantor-López; C. Cortés-Guerrero; M. A. Guzmán-Pardo; S. Rivera-Rodríguez: Implementación del algoritmo VNS-DEEPSO para el despacho de energía en redes distribuidas inteligentes, INGE CUC, vol. 15, no. 1, pp. 142-154, 2019.
15. P. Hansen, N. Mladenović, R. Todosijević, S. Hanafi: Variable neighborhood search: basics and variants, EURO J. on Comp. Opt, 2016.
16. T. Morton and D. Pentico: Heuristic Scheduling Systems. John Wiley and Sons, NY, 1993.
17. W. Cheng, P. Guo, Z. Zhang, M. Zeng, and J.Liang: VNS for Parallel Machines Scheduling Problem with Step Deteriorating Jobs. Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 7, 2012.
18. E. Ferretti and S. Esquivel: A Comparison of Simple and Multirecombinated Evolutionary Algorithms with and without Problem Specific Knowledge Insertion, for Parallel Machines Scheduling, International Transaction on Computer Science and Engineering, volume 3, number 1, 207-221, 2005.

Marco Metodológico para el Desarrollo de un Sistema de Reconocimiento Biométrico Mediante Técnicas de Machine Learning

Silvia Estela Ruiz[1] y Carlos Eduardo Alvez[2]
Universidad Nacional de Entre Rios
[1] {silvia.ruiz, [2]carlos.alvez}@uner.edu.ar

Abstract: El uso del Aprendizaje Automático en los sistemas de reconocimiento biométrico supone un gran paso en la evolución tecnológica. En el presente trabajo se exploran diferentes papers en los cuales se aplican distintos mecanismos del aprendizaje automático en los sistemas de reconocimiento biométrico mediante iris, los cuales han logrado resultados exitosos en los últimos años. Debido a que en la actualidad no existe una metodología estándar formal de proyectos que utilicen Aprendizaje Automático, se analizaron varias metodologías y modelos de procesos adoptadas en la actualidad. Por consiguiente, el presente trabajo de tesis, tiene como principal objetivo proponer un marco metodológico para el desarrollo de un sistema biométrico de reconocimiento de iris mediante Aprendizaje Automático.

Keywords: reconocimiento de iris, aprendizaje automático, metodología.

Fecha de inicio: Abril 2022.

1 Introducción

Se observa en la actualidad una tendencia muy marcada en la incorporación de técnicas biométricas, tales como rostro, huellas digitales, geometría de la mano, iris, patrones de retina, firma y voz, entre las más destacadas en los sistemas de identificación/autenticación de personas. Todas estas técnicas presentan diferentes grados de singularidad, permanencia, mensurabilidad, desempeño, aceptación del usuario y robustez.

Los sistemas de reconocimiento biométrico utilizan características fisiológicas o de comportamiento propias de cada individuo para identificarlo, es decir, se reconoce al usuario por lo que es en lugar de por lo que posee o sabe [1]. Los rasgos fisiológicos presentan una reducida variabilidad a lo largo del tiempo, pese a que su adquisición es más invasiva y requiere de la cooperación de los sujetos. Por el contrario, los rasgos de comportamiento resultan menos invasivos aunque la exactitud de la identificación es menor debido a la variabilidad de los patrones de comportamiento.

De esta forma, el objetivo de todos ellos será obtener, a partir de la captura de un rasgo biométrico, una representación de cada individuo que resulte lo suficientemente discriminante respecto a las de otros usuarios del sistema. En lo que respecta

a los distintos rasgos biométricos, el reconocimiento del iris se considera uno de los sistemas biométricos más confiables y precisos, a pesar de ser un campo de estudio muy reciente en relación a otros rasgos [2]. En diferentes trabajos se ha demostrado [1, 3] que el rasgo del iris tiene una serie de ventajas con respecto a otros rasgos biométricos (por ejemplo, cara, huella digital), que lo hacen comúnmente aceptado para su aplicación en sistemas biométricos precisos y de alta confiabilidad. Entre las distintas ventajas que presenta el iris se destacan: un patrón único en cada persona, la estructura permanece invariable durante toda la vida del ser humano y la adquisición de la imagen se ha simplificado con el avance tecnológico de distintos dispositivos de captura. Por los motivos precedentes el desarrollo de este trabajo de tesis se enfoca en sistemas de reconocimiento basados en iris. En la Figura 1, se representa un esquema con el proceso de un sistema biométrico de reconocimiento de iris. El primer paso captura la imagen y el segundo la reconoce, realiza el preprocesamiento y la extracción de características. La etapa de normalización tiene como objetivo obtener una imagen del iris que sea independiente del tamaño de la pupila y permita la comparación entre diferentes iris. Los valores de las características son almacenados en la base de datos como plantillas (templates) biométricas. Luego, la toma de decisión se realiza clasificando la entrada dada con las plantillas biométricas en la base de datos.

Fig. 1. Esquema de un sistema de reconocimiento biométrico de iris.

A pesar de las ventajas que presentan los sistemas de reconocimiento biométrico enfrentan diversos desafíos en cuanto a la precisión de los datos. En este sentido, a lo largo de los últimos años, se han comenzado a utilizar diferentes técnicas de Aprendizaje Automático o Machine Learning (en adelante ML) en las etapas que conforman el proceso de reconocimiento de iris [4]. En [5,6] se destacan principalmente las técnicas de ML tales como SVM (Máquina de Soporte Vectorial), WPNN (Redes Neuronales Probabilísticas Ponderadas), BPNN (Redes Neuronales de Retro Propagación) y RBFNN (Redes Neuronales con Función de Base Radial), las cuales son utilizadas generalmente en la etapa de clasificación de los patrones del iris, así como también para la normalización, segmentación y extracción de características.

En las últimas dos décadas, el término Aprendizaje Automático se ha convertido en una herramienta común en casi cualquier tarea que requiera la extracción de información de grandes conjuntos de datos. Estamos rodeados de una tecnología basada en el Aprendizaje Automático: motores de búsqueda, software antispam, software para detección de fraudes, etc. Una característica común de todas estas aplicaciones es que, en contraste con los usos más tradicionales de las computadoras, en estos ca-

sos debido a la complejidad de los patrones que necesitan ser detectados, un programador humano no puede proporcionar información explícita y detallada de cómo se deben ejecutar tales tareas. Muchas de las habilidades del ser humano se adquieren o refinan a través del aprendizaje de la experiencia (en lugar de seguir instrucciones explícitas). Las herramientas de Aprendizaje Automático se interesan en dotar a los programas de la capacidad de "aprender" y adaptarse.

Por otro lado, en toda implementación de un sistema es conveniente seguir un modelo de proceso o metodología [7] a fin de obtener mayor probabilidad de éxito en los resultados obtenidos. Según Pressman [7], un modelo de proceso define un conjunto de actividades organizadas para llevar a cabo una tarea determinada, mientras que una metodología además de definir las actividades y tareas de un proceso, define cómo llevar a cabo las mismas.

Al respecto, si bien en la actualidad se están desarrollando distintas líneas de investigación para definir una metodología estándar específica para utilizar en sistemas que utilicen ML [8] aún no existe una definición estándar formal. En este sentido, existen varias metodologías y modelos de procesos de minería de datos y ML, tales como KDD (Knowledge Discovery in Databases) y SEMMA (Sample, Explore, Modify, Model and Assess), no obstante, se suele adoptar como base el estándar de facto CRISP-DM (Cross Industry Standard Process for Data Mining) [9]. Este modelo de referencia es utilizado habitualmente como marco metodológico en distintos tipos de proyectos [10], principalmente de minería de datos.

En este sentido, el presente trabajo de tesis tiene como principal objetivo proponer un marco metodológico para el desarrollo de un sistema biométrico de reconocimiento de iris mediante ML. Para esto, se realizará un estudio de los fundamentos teóricos de la biometría, los sistemas biométricos y de las distintas técnicas de ML. Además, será necesario realizar una revisión sistemática de los distintos modelos de procesos de trabajo y/o metodologías adoptadas en el desarrollo de sistemas que utilicen ML, analizando las ventajas y desventajas de las mismas e identificando los desafíos que se encuentran en su aplicación. En base a la información recolectada, se realizará un análisis comparativo y se elaborará una propuesta de marco metodológico que pueda ser aplicado en un proyecto real de un sistema de reconocimiento de iris mediante ML.

2 Aportes realizados

A continuación se presentan los aportes parciales obtenidos en pos del objetivo general del trabajo de tesis. Se propone un marco metodológico específico para proyectos de sistemas de reconocimiento de iris mediante el uso de técnicas de ML. En la Fig. 2 se presenta el esquema metodológico propuesto y se detallan brevemente las fases generales.

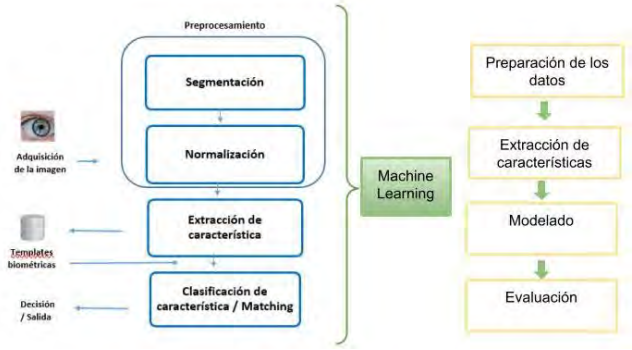


Fig. 2 Marco metodológico propuesto.

- **Preparación de los datos**

Implica las tareas de análisis, descripción, verificación y preprocesamiento de los datos. Se debe preprocesar la imagen para localizar los límites de la pupila y el iris y así poder normalizar el iris. Para ello se utilizan diferentes algoritmos como por ej. filtro gaussiano de paso bajo el cual permite suavizar y agudizar la imagen del ojo. Todo este proceso permite la adaptación de los datos para el uso posterior de las técnicas de ML seleccionadas.

- **Extracción de características**

En esta fase se procede a la extracción de características y su correspondiente codificación. Es importante destacar que, el éxito de cualquier sistema biométrico definido como un sistema de clasificación y reconocimiento depende principalmente de la eficiencia y robustez de las etapas de extracción y clasificación de características.

- **Modelado**

Esta fase abarca el diseño, construcción y evaluación del modelo. Aquí es necesario seleccionar las técnicas de modelado más apropiadas para la construcción del modelo. Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de la precisión que se desea lograr con el modelo. En general, se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados.

- **Evaluación**

Una vez que se tienen los datos preparados, se puede empezar a entrenar el modelo. Para ello, se realizan diferentes entrenamientos con distintos algoritmos y parámetros de los mismos. En lo que respecta a los sistemas de reconocimiento de iris, en trabajos recientes [12,13] se ha demostrado que el uso Deep Learning y Redes Neuronales logran tasas de precisión significativamente alta mejorando aspectos de eficiencia y confiabilidad. Los resultados del modelo podemos verlos mediante la utilización de distintas herramientas que permitan la visualización del desempeño del algoritmo de aprendizaje utilizado.

3 Posibles líneas de investigación a futuro

Las actividades recomendadas en cada fase de este marco metodológico surgen de prácticas usuales de distintos proyectos de ML y minería de datos. Sin lugar a dudas, la adopción de un marco metodológico en este tipo de proyectos favorece la calidad del proceso para la obtención de datos confiables en forma eficiente. Como trabajo futuro se propone profundizar las fases de preparación de los datos y modelado, consideradas esenciales para el éxito de un sistema de reconocimiento biométrico, avanzando en una descripción más detallada de las tareas que las conforman definiendo sus objetivos específicos y los documentos entregables de cada una. Por otro lado, se pretende profundizar en la aplicación de diferentes formas de medir su éxito, es decir, utilizando métricas específicas y criterios de aceptación.

4 Bibliografía básica

1. A. K. Jain, A. Ross and S. Prabhakar. “An introduction to biometric recognition”. IEEE Transactions on Circuits and Systems for Video Technology., 14(1), 2004.
2. J. Daugman. “High confidence visual recognition of persons by a test of statistical independence”. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993.
3. A. Basit. “Iris Recognition: An Identification Biometric System”. Lap Lambert Academic Publishing. 2010.
4. M. De Marsico, A. Petrosino, y S. Ricciardi. “Iris recognition through machine learning techniques: A survey”. Pattern Recognit. Lett., vol. 82, pp. 106-115, oct. 2016, doi: 10.1016/j.patrec.2016.02.001.
5. A. A. Khan, S. Kumar, y M. Khan, “Iris Pattern Recognition using Support Vector Machines and Artificial Neural Networks”. IJIREEICE, pp. 2208-2211, dic. 2014, doi: 10.17148/IJIREEICE.2014.21203.
6. K. Saminathan and T. Chakravarthy and M. Chithra Devi. “Iris Recognition based on kernels of support vector machine”. ICTACT J. Soft Comput., vol. 05, n.o 02, pp. 889-895, ene. 2015, doi: 10.21917/ijsc.2015.0125.]
7. R. Pressman y B. Maxim, “Software engineering: a practitioner’s approach”. Novena Edición. McGraw-Hill Education. New York. 2020.
8. S. Studer, T. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, K. Mueller, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology”. ArXiv200305155 Cs Stat, mar. 2020, Accedido: ene. 29, 2022. [En línea]. Disponible en: http://arxiv.org/abs/2003.05155.
9. C. Shearer, “The CRISP-DM model: the new blueprint for data mining”. J Data Wareh., vol. 5, pp. 13-22, ene. 2000.
10. V. G. Cortina, “Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario”. p. 120.
11. J. Daugman, “How iris recognition works”. Proceedings of 2002 International Conference on Image Processing, Vol. 1, 2002.
12. A. Al-Waisy, R. Qahwaji, S. Ipson, A. Shumoos, T. Nagem “A multi-biometric iris recognition system based on a deep learning approach”. Pattern Analysis and Applications. 21(3): 783-802. 2017.
13. S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, D. Zhang , “Biometric Recognition Using Deep Learning: A Survey”. 2020.

Data Cleansing en entornos Big Data: Mapeo Sistemático de la Literatura

Caffetti Yanina A.[1], Eckert Karina B. [1,2], Ruidias Héctor J.[1,2] and Vera Laceiras María Silvia[1]

[1] Universidad Nacional de Misiones, Misiones, Argentina
[2] Universidad Gastón Dachary, Misiones, Argentina

yanina.caffetti@fcf.unam.edu.ar, karinaeck@gmail.com, chandra149@gmail.com, vlhsilvia@fceqyn.unam.edu.ar

Resumen. La tecnología Big Data tiene por objetivo la gestión de grandes volúmenes de datos e información de manera inteligente que ayude a una correcta toma de decisión. Las etapas del trabajo en Big Data incluyen muchas decisiones que deben ser tomadas por el usuario, para garantizar el éxito del objetivo propuesto, entre ellas la limpieza y pre-procesamiento de datos. El presente artículo es un Mapeo Sistemático de la Literatura, que busca identificar las metodologías, técnicas o herramientas utilizadas para el tratamiento de datos basura (dirty data) o limpieza de datos (data cleansing), en entornos Big Data. Existe, en la literatura actual, cierta escasez de publicaciones específicas, aun siendo un tema de suma relevancia para el éxito de este tipo de proyectos, donde se requiere procesamientos que cumplan con las características propias del entorno Big Data.

Keywords: Data Cleansing, Dirty Data, Big Data, Method, Treatment.

1 Introducción

El crecimiento de los datos es considerado exponencial, algunos autores indican que el volumen de los datos digitales se duplicará cada dos años [1]. Precisamente, a diario se generan grandes volúmenes provenientes de diferentes fuentes y formatos, especialmente en entornos de Big Data (BD), lo que aumenta el desafío de verificar la calidad de estos datos, que pueden ser imprecisos, tener anomalías o no ser adecuados para el análisis o procesamiento afectando así la precisión de los resultados obtenidos [2].

Entre los inconvenientes asociados a la calidad de los datos crudos, se pueden mencionar los siguientes problemas típicos: ausencia de datos, valores ficticios o predeterminados, ruido, datos erróneos, datos inconsistentes, datos crípticos, claves primarias duplicadas, identificadores no únicos, campos multipropósito y violación de reglas comerciales [3]. La limpieza de datos, incluye diferentes técnicas de detección y representa un proceso complejo que requiere mucho tiempo para poder garantizar que los datos limpios tengan una mejor calidad. Este pre-procesamiento es útil y fundamental para garantizar la fase siguiente o de análisis [2], [4]. El proceso de limpieza

de datos se define en cinco fases; (1) análisis de datos, (2) definición de flujo de trabajo de transformación y regla de mapeo, (3) verificación, (4) transformación y (5) reflujo de datos limpios. Dicho procedimiento consiste básicamente en identificar errores y corregirlos, para así obtener resultados que colaboren en el proceso de toma de decisiones. Por estas razones, el proceso de detección y limpieza de anomalías dentro de los datos recopilados se convierte en un factor crítico. Según el Data Warehousing Institute en su web https://tdwi.org, los datos sucios o erróneos conocidos como Dirty Data cuestan más de $600 mil millones por año en negocios de EEUU; los mismos son causados por diversos factores (datos obsoletos, incompletos, duplicados y/o sin formato, etc.) [1], [2], [4], [5].

El objetivo del presente artículo es indagar sobre los procesos de limpieza de datos, metodologías, técnicas y/o herramientas para el tratamiento de datos sucios en el ámbito de BD, mediante un Mapeo Sistemático de la Literatura (MSL) del tema. El artículo está estructurado de la siguiente manera, en la sección 2 se describe la metodología utilizada (MSL), luego en la sección 3 se exponen el análisis de los resultados obtenidos, para finalmente en la sección 4 presentar las conclusiones del trabajo realizado.

2 Mapeo Sistemático de la Literatura

Como primera fase del MSL [6] se definió el problema a través del planteo de la siguiente pregunta de investigación: ¿Cuáles son las metodologías, métodos, técnicas y herramientas utilizadas para la limpieza de datos en entornos de Big Data?

Las subpreguntas de investigación derivadas orientan las fases subsecuentes, desde la búsqueda hasta el análisis de la información; posibilitando la apropiación y producción del tema, se plantearon dos:

SPI1: ¿Qué tipos de estudios hay respecto a la temática?

SPI2: ¿Qué tipo de tratamientos son los más utilizados en la actualidad?

Tabla 1. Fuentes de datos

Fuentes de datos	Sitio Web	Artículos
ACM Digital Library (ACM)	https://dl.acm.org	1712
IEEE Xplore (IEEEX)	https://ieeexplore.ieee.org	34
Springer Link(SL)	https://link.springer.com	129
ScienceDirect (SD)	https://www.sciencedirect.com	599
DSpace MIT (DMIT)	https://dspace.mit.edu/	51
Scopus (SC)	https://www.scopus.com/home.uri	66

Como segunda fase, se estableció la estrategia de búsqueda y el proceso de selección de los trabajos de investigación. Para ello, se determinaron las cadenas de búsqueda con la finalidad de identificar la literatura relevante. Luego de un proceso de refinamiento en la combinación términos y operadores lógicos utilizados para la confección de las cadenas de búsqueda, se seleccionó la siguiente: *"(Data Cleansing OR*

Dirty Data) AND (Methodology OR Method OR Treatment OR Tool) AND Big Data" que se aplicó a las fuentes de datos indicadas en la Tabla 1, y en donde también se refleja la cantidad de artículos obtenidos por cada fuente de datos. Siendo un total de 2591 artículos seleccionados. En cada buscador se aplicaron filtros específicos, tales como los de restringir la búsqueda específicamente a artículos de investigación ("research articles" por ejemplo en el caso de SD) o por campo de conocimiento que permitieran la mayor especificidad posible ("computer science" fue el filtro que se pudo emplear en plataformas tales como SL, SD) con la finalidad de reducir la muestra.

Posteriormente se eliminaron los estudios duplicados, se definieron y aplicaron los criterios de inclusión y exclusión necesarios para realizar un refinamiento de la búsqueda e identificar aquellos que proporcionan relación directa con la pregunta de investigación.

Criterios de Inclusión (CI):

CI1: Documentos publicados en los últimos 5 años.

CI2: Publicaciones realizadas en revistas, libros, conferencias o workshop.

CI3: Trabajos que refieran específicamente a metodologías, métodos, técnicas o herramientas para el tratamiento o limpieza de datos en los títulos y/o resúmenes.

Criterios de Exclusión (CE):

CE1: Publicaciones previas al 2017.

CE2: Trabajo en sitios que no permiten la disponibilidad del texto completo.

3 Análisis y Discusión

Al aplicar los criterios de inclusión (CI) y exclusión (CE) planteados, se redujo considerablemente la cantidad, obteniendo un total de 11 documentos. Se entiende entonces que el marco de referencia acerca de trabajos realizados buscando una normalización en los datos y el tipo de limpieza de los mismos es un campo aún en desarrollo especialmente en entornos de BD; precisamente en la Fig. 1 se puede notar la tendencia existente entre cantidad de artículos divulgados en los últimos 5 años y los tipos de publicaciones detectados (revistas y conferencias, en este estudio).

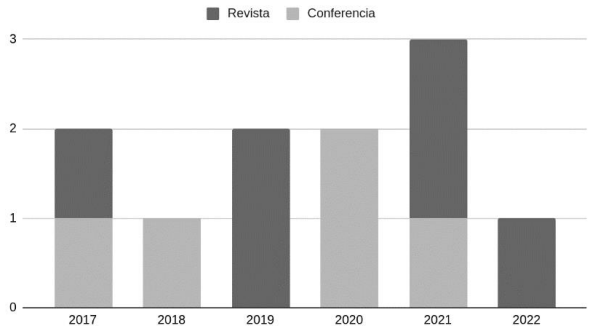


Fig. 1. Relación artículos por tipo y año.

Los artículos resultantes coinciden en que si bien el proceso de limpieza de datos puede ser abordado con herramientas convencionales (tales como hojas de cálculo), en entornos de BD al considerar el volumen de datos y los tiempos de procesamiento requerido, resultan inviables recurrir a tales técnicas tradicionales y se requiere un enfoque sistemático. Además se detecta la necesidad de automatizar y adecuar a los objetivos de los clientes la detección de datos relevantes y ofrecer la confiabilidad necesaria para suponer una mejora en procesos ante un escenario de pérdida de precisión debido al dirty data. Algunas de las técnicas se asientan más en métodos que herramientas, apoyándose fuertemente en la confianza que implica contar con conocimiento experto. Se destacan los métodos basados en técnicas de Inteligencia Artificial, como ser técnicas de aprendizaje automático (machine learning), agrupamiento, especificación y selección de reglas, base de conocimiento y crowdsourcing [2], [3], [4], [7], [8], [9], [10], [11], [12], [13], [14].

4 Conclusiones

En la era de Big Data, la cantidad de datos sigue aumentando, a su vez que se vuelven más complejos debido a su diversidad y combinación (datos estructurados, semiestructurados y no estructurados), lo que hace que la calidad de los mismos disminuya en muchos casos dado que los datos recopilados están sucios (fenómeno conocido como dirty data); donde los procesos tradicionales de limpieza no son adecuados en términos de cantidad, complejidad y velocidad requeridos en proyectos de BD.

En los artículos recuperados y definidos como relevantes, se encontraron ciertas especificaciones en cuanto a metodologías y herramientas en data cleansing. A su vez se identifica la necesidad de investigaciones sobre limpieza de datos centradas en los criterios de BD, donde se cubran las características conocidas como las “5 V” de BD: Volumen, Variedad, Velocidad, Valor y Veracidad. Además de contar con expertos en el área que validen y verifiquen los datos a procesar en las siguientes etapas.

A partir del MSL presentado en esta investigación, se recuperaron artículos relevantes en cuanto a propuestas de tratamiento y limpieza de datos en Big Data, sin embargo no se encuentran del todo estandarizados. En cambio, se han visto como una variación de técnicas más tradicionales como ETL (Extract, Transform and Load) son utilizados. Queda como futura línea de investigación, el planteo de formular un estándar para la aplicación de la data cleansing en entornos BD.

Referencias

[1] M. I. Hossen, M. Goh, A. Hossen, and Md. A. Rahman, “A Study on the Aspects of Quality of Big Data on Online Business and Recent Tools and Trends Towards Cleaning Dirty Data,” in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, Aug. 2020, pp. 209–213. doi: 10.1109/ICSGRC49013.2020.9232648.

[2] F. Ridzuan and W. M. N. Wan Zainon, “A Review on Data Cleansing Methods for Big Data,” *Procedia Computer Science*, vol. 161, pp. 731–738, Jan. 2019, doi: 10.1016/j.procs.2019.11.177.

[3] K. Stöger, D. Schneeberger, P. Kieseberg, and A. Holzinger, “Legal aspects of data cleansing in medical AI,” *Computer Law & Security Review*, vol. 42, p. 105587, Sep. 2021, doi: 10.1016/j.clsr.2021.105587.

[4] T. K. Dang, D. K. Nguyen, and L. M. Tuan, “OpenK: An Elastic Data Cleansing System with A Clustering-based Data Anomaly Detection Approach,” in *2021 15th International Conference on Advanced Computing and Applications (ACOMP)*, Nov. 2021, pp. 120–127. doi: 10.1109/ACOMP53746.2021.00023.

[5] Z. Opršal and J. Harmáček, “Clean aid or dirty aid? The environmentalization of Czech foreign aid.,” *Journal of Cleaner Production*, vol. 224, pp. 167–174, Jul. 2019, doi: 10.1016/j.jclepro.2019.03.198.

[6] B. A. Kitchenham, D. Budgen, and O. Pearl Brereton, “Using mapping studies as the basis for further research – A participant-observer case study,” *Information and Software Technology*, vol. 53, no. 6, pp. 638–651, Jun. 2011, doi: 10.1016/j.infsof.2010.12.011.

[7] F. Ridzuan and W. M. N. Wan Zainon, “Diagnostic analysis for outlier detection in big data analytics,” *Procedia Computer Science*, vol. 197, pp. 685–692, Jan. 2022, doi: 10.1016/j.procs.2021.12.189.

[8] H. A. Sulistyo, T. F. Kusumasari, and E. N. Alam, “Implementation of Data Cleansing Null Method for Data Quality Management Dashboard using Pentaho Data Integration,” in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, Nov. 2020, pp. 12–16. doi: 10.1109/ICOIACT50329.2020.9332030.

[9] P. Petrova, V. Jotsov, and V. Sgurev, “Puzzle Methods for Automatic Selection of Data Cleansing Techniques,” in *2018 International Conference on Intelligent Systems (IS)*, Funchal - Madeira, Portugal, Sep. 2018, pp. 820–826. doi: 10.1109/IS.2018.8710580.

[10] Y. Lei, X. Zhou, X. Xu, and F. Jia, “A dirty data recognition method for machinery condition monitoring in big data era,” in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Beijing, China, Oct. 2017, pp. 7061–7066. doi: 10.1109/IECON.2017.8217235.

[11] J. G. Lawson and D. A. Street, “Detecting dirty data using SQL: Rigorous house insurance case,” *Journal of Accounting Education*, vol. 55, p. 100714, Jun. 2021, doi: 10.1016/j.jaccedu.2021.100714.

[12] W. Fan and C. Hu, “Big Graph Analyses: From Queries to Dependencies and Association Rules,” *Data Sci. Eng.*, vol. 2, no. 1, pp. 36–55, Mar. 2017, doi: 10.1007/s41019-016-0025-x.

[13] D. C. Setyawan, T. F. Kusumasari, and E. N. Alam, “Data Cleansing Processing using Pentaho Data Integration: Case Study Data Deduplication,” in *2020 6th International Conference on Science and Technology (ICST)*, Sep. 2020, vol. 1, pp. 01–05. doi: 10.1109/ICST50505.2020.9732824.

[14] S. Salloum, J. Z. Huang, and Y. He, “Exploring and cleaning big data with random sample data blocks,” *J Big Data*, vol. 6, no. 1, p. 45, Jun. 2019, doi: 10.1186/s40537-019-0205-4.

XXIII Workshop Procesamiento Distribuido y Paralelo **(WPDP)**

Coordinadores

Fabiana Piccoli (UNSL)

Laura De Giusti (UNLP)

Carlos García Garino (UNCu)

Comparación de Rendimiento y Esfuerzo de Programación entre Numba y Cython para una Aplicación Multi-hilada de Alto Rendimiento

Andrés Milla[1] and Enzo Rucci[2][0000−0001−6736−7358] ✉

[1] Facultad de Informática, UNLP.
La Plata (1900), Bs As, Argentina
andressmilla@gmail.com
[2] III-LIDI, Facultad de Informática, UNLP – CIC.
La Plata (1900), Bs As, Argentina
erucci@lidi.info.unlp.edu.ar

Resumen En la actualidad, Python es uno de los lenguajes más utilizados en diversas áreas de aplicación. Sin embargo, éste presenta limitaciones a la hora de poder optimizar y paralelizar aplicaciones debido a limitaciones de su intérprete oficial (CPython), especialmente para aplicaciones *CPU-bound*. Para solucionar esta problemática han surgido traductores alternativos, aunque cada uno con un enfoque diferente y con su propia relación de costo-rendimiento. Este trabajo es una continuación de otros previos, donde se presenta una comparación de rendimiento más justa y actualizada de los traductores Numba y Cython para el caso de estudio *N-Body* (un problema popular con alta demanda computacional). Además, se realiza un análisis comparativo del esfuerzo de programación requerido por ambas soluciones, lo que brinda un segundo criterio a la hora de optar entre ellos.

Keywords: N-body · CPU-bound · Programación paralela · Costo de programación · Python

1. Introducción

En la actualidad, a pesar de que Python se ha convertido en uno de los lenguajes más populares [20], sigue siendo considerado "lento" en comparación a otros lenguajes compilados como C, C++ y Fortran; especialmente para que aquellas aplicaciones intensivas en CPU (*CPU-bound*) [3]. Entre las causas de su pobre rendimiento, se encuentran su naturaleza de lenguaje interpretado y sus limitaciones al momento de implementar soluciones multi-hiladas [15]. En particular, el principal problema es la utilización de un componente llamado *Global Interpreter Lock* (GIL) en el intérprete oficial CPython. Este último sólo admite que un único un hilo se ejecute a la vez, lo que lleva a que su ejecución sea

[3] Programas que realizan una gran cantidad de cálculos utilizando la CPU de manera exhaustiva.

de forma secuencial. Para solucionar esta limitación, se suele utilizar procesos en vez de hilos, pero hay que tener en cuenta que el consumo de recursos es mayor y que aumenta el costo de programación por tener un espacio de direcciones distribuido [11].

Aunque existen intérpretes alternativos a CPython, algunos de estos también presentan el mismo problema, como es el caso de PyPy [10]. En sentido opuesto, existen intérpretes que optan por no utilizar el GIL en sus implementaciones, como por ejemplo Jython [5]. Lamentablemente, Jython emplea una versión discontinuada de Python, lo que limita el soporte a futuro para sus programas y la posibilidad de aprovechar las características que proveen las versiones posteriores del lenguaje. Por otro lado, otros traductores optan por ofrecerle al programador desactivar este componente, tal como es el caso de Numba, un compilador JIT que traduce Python en código de máquina optimizado [8]. Numba utiliza una característica de Python conocida como decoradores [1], para intervenir lo menos posible en el código del programador. Por último, se puede mencionar a Cython, un compilador estático que permite transpilar [4] Python a C, y luego compilarlo a código objeto [3]. También permite desactivar el GIL y utilizar librerías de C como OpenMP [4], lo cual resulta de suma utilidad para desarrollar programas multi-hilados.

Al momento de implementar una aplicación en Python, se debe seleccionar qué traductor utilizar. Esta elección es fundamental ya que no sólo impactará en el rendimiento del programa sino también en el tiempo requerido para desarrollo como también en el costo de mantenerlo a futuro. Para no tomar una decisión "a ciegas", resulta fundamental revisar la evidencia al respecto. Lamentablemente, la literatura disponible en la temática no es exhaustiva. Si bien existen estudios que contemplan comparaciones de traductores, lo llevan a cabo usando versiones secuenciales [22,18], lo que no permite evaluar sus capacidades de procesamiento paralelo. En sentido contrario, en el caso de sí usar paralelismo, lo hacen entre lenguajes y no entre traductores de Python [14,23,13,21]. Por otra parte, en la mayoría de ellos no se hace un estudio sobre la productividad y el esfuerzo de programación que contrae desarrollar cada solución, una cuestión que día a día se torna más importante [7,2].

En base a lo anterior, resulta fundamental conocer las ventajas y desventajas de diferentes traductores del lenguaje Python tanto en un paradigma secuencial como multi-hilado. Por lo tanto, este artículo se enfoca en su comparación considerando no sólo el rendimiento, sino también su productividad y esfuerzo de programación asociados. Este artículo es una continuación de trabajos previos [17,16], presentado las siguientes nuevas contribuciones:

- Una re-evaluación de los rendimientos de las implementaciones Numba y Cython considerando que Intel ha adoptado recientemente a LLVM [5] como *backend* de su compilador de C/C++ [6]. De esta manera, se ofrece una

[4] Proceso que realiza una clase especial de compilador en la cual se genera código fuente en un lenguaje a partir del correspondiente a otro lenguaje.

[5] The LLVM Compiler Infrastructure. `https://llvm.org`

comparación más justa y actualizada (ambos operan con el mismo *backend* ahora).
- Un análisis comparativo del esfuerzo de programación requerido para las soluciones Numba y Cython desarrolladas, el cual permite considerar un segundo criterio a la hora de optar entre ellos.

Mediante este estudio comparativo se espera contribuir a programadores de Python para que conozcan fortalezas y debilidades al momento de implementar aplicaciones multi-hiladas de alto rendimiento. El resto del artículo se organiza de la siguiente forma. La Sección 2 introduce el marco teórico para esta investigación. Luego, la Sección 3 describe las implementaciones realizadas. A continuación, la Sección 4 analiza los resultados experimentales mientras que la Sección 5 resume las conclusiones junto al trabajo futuro.

2. Marco teórico

2.1. Numba

Numba es un compilador JIT que permite traducir código Python a código de máquina optimizado a través de LLVM [6]. De acuerdo con su documentación, es capaz de alcanzar aceleraciones similares a las de lenguajes compilados como C, C++ y Fortran [8], sin necesidad de re-escribir su código gracias a un enfoque de anotaciones llamados decoradores [1].

Compilación JIT La librería ofrece dos modos de compilación: (1) modo objeto, el cual permite compilar código que haga uso de objetos; (2) modo *nopython*, que le permite a Numba generar código evitando a la API de CPython. Para indicar dichos modos, se utilizan los decoradores `@jit` y `@njit` (ver Fig. 1), respectivamente [8].

```
1  from numba import njit
2
3  # Equivalent to indicating
4  # @jit(nopython=True)
5  @njit
6  def f(x, y):
7      return x + y
```

Figura 1: Compilación en modo *nopython*.

Por defecto, cada función será compilada al momento de ser invocada y se mantendrá en la caché para futuras llamadas. Sin embargo, la inclusión del parámetro `signature` provocará que la función sea compilada al momento de la declaración. Además, también posibilitará indicar los tipos de datos que usará la función y controlar la organización de los datos [8] en memoria (ver Fig. 2).

Multi-hilado Numba permite activar un sistema de paralelización automática estableciendo el parámetro `parallel=True`, como también indicar una paralelización explícita mediante la función `prange` (ver Fig. 3), la cual distribuye las iteraciones entre los hilos de manera similar a la directiva `parallel for`

[6] The LLVM Compiler Infrastructure, `https://llvm.org/`

```python
from numba import njit, double

@njit(double(double[::1, :],
             double[:, ::1]))
def f(x, y):
    """
    x:  Vector 2D of the "Double" type
        organized in columns.
    y:  Vector 2D of the "Double" type
        organized in rows.
    Returns the sum of the product
    of vectors x and y.
    """
    return (x * y).sum()
```

Figura 2: Compilación en modo *nopython* con el parámetro `signature`

```python
from numba import njit, double, prange

@njit(double(double[::1]), parallel=True)
def f(x):
    """
    x: 1D Vector.
    Returns the sum of vector x
    through a reduction.
    """
    N = x.shape[0]
    z = 0

    for i in prange(N):
        z += x[i]

    return z
```

Figura 3: Compilación en modo *nopython* con el parámetro `parallel`

de OpenMP. Además, también soporta reducciones y se encarga de declarar las variables como privadas a cada hilo si son declaradas dentro del alcance de la zona paralela. Lamentablemente, Numba aún no soporta primitivas que permitan controlar la sincronización de los hilos, como pueden ser semáforos o *locks* [8].

Vectorización Numba delega en LLVM la autovectorización del código y la generación de instrucciones SIMD, pero le permite al programador controlar ciertos parámetros que podrían influir en esta tarea, como la precisión numérica mediante el argumento `fastmath=True`. También ofrece la posibilidad de utilizar *Intel SVML* en caso de estar disponible en el sistema [8].

Integración con NumPy Cabe destacar que Numba soporta un gran número de funciones de NumPy, lo cual le permite al programador controlar la organización de memoria de los arreglos y realizar operaciones entre ellos [9,8].

2.2. Cython

Cython es un compilador estático para Python creado con el objetivo de escribir código en C aprovechando la sintaxis simple y clara de Python [3]. En otras palabras, Cython es un *superset* de Python que permite interactuar con funciones, tipos y librerías de C.

Compilación Tal como se puede apreciar en la Fig. 4 el flujo de programación de Cython es muy diferente al que el programador de Python está habituado. La principal diferencia es que el archivo que contendrá el código fuente tendrá extensión `.pyx` a diferencia de Python, cuya extensión es `.py`. Luego, este archivo se podrá compilar a través de un archivo `setup.py`, en donde se indican los flags de compilación para dar como salida (1) un archivo con extensión `.c`, el cual corresponde al código transpilado de Cython a C y (2) un archivo binario con extensión `.so`, el cual corresponde a la compilación del archivo de C descrito previamente. Este último nos permitirá importar el módulo compilado en cualquier script de Python.

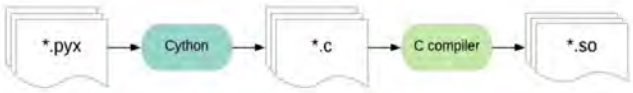


Figura 4: Flujo de programación en Cython.

Tipos de datos Cython permite declarar variables utilizando los tipos de datos de C a partir de la sentencia `cdef` (ver Fig. 5). Si bien esto es opcional, la documentación lo recomienda para optimizar la ejecución del programa, ya que se evita la inferencia de tipos de CPython en tiempo de ejecución. Además, Cython permite indicar la organización de memoria de los arreglos al igual que Numba [3].

Multi-hilado Cython provee soporte para utilizar OpenMP a través del módulo `cython.parallel`. El mismo contiene la función `prange`, la cual posibilita paralelizar bucles mediante el constructor parallel for de OpenMP. A su vez, esta función permite desactivar el GIL e indicar el *scheduling* de OpenMP a través de los argumentos `nogil` y `schedule` respectivamente.

Cabe destacar que todas las asignaciones declaradas dentro del bloques `prange` son transpiladas como `lastprivate`, mientras que las reducciones solo son identificadas si se utiliza un operador *in-situ*. Por ejemplo, la operación estándar de la suma ($x = x+y$) no será identificada como reducción, pero la operación *in-situ* de la suma ($x+ = y$) sí (ver Fig. 6).

Vectorización Cython delega la vectorización en el compilador de C que se esté utilizando. Si bien existen soluciones alternativas para forzar la vectorización, nativamente no es soportado por Cython.

Integración con NumPy Lamentablemente, las operaciones entre vectores de NumPy no son soportadas por Cython. Sin embargo, como se mencionó anteriormente se puede utilizar NumPy para controlar la organización de memoria de los arreglos.

```
1  cdef int x, y, z
2  cdef float a, b[100], *c
3
4  cdef struct Point:
5      double x
6      double y
```

Figura 5: Variables declaradas con tipos de datos de C en Cython.

```
1  from cython.parallel import prange
2
3  cdef int i
4  cdef int N = 30
5  cdef int total = 0
6
7  for i in prange(N, nogil=True):
8      total += i
```

Figura 6: Reducción utilizando el bloque `prange` de Cython.

2.3. Caso de estudio: N-Body

En esta sección se presenta el caso de estudio elegido: la simulación de N cuerpos computacionales (*N-Body*), un problema *CPU-bound* de complejidad computacional $O(n^2)$ y que resulta popular en la comunidad HPC.

El problema consiste en simular la evolución de un sistema compuesto por N cuerpos durante una cantidad de tiempo determinada. Dados la masa y el estado inicial (velocidad y posición) de cada cuerpo, se simula el movimiento del sistema a través de instantes discretos de tiempo. En cada uno de ellos, todo cuerpo experimenta una aceleración que surge de la atracción gravitacional del resto, lo que afecta a su estado.

```
1    for t in range(PASOS):
2        for i in range(N):
3            for j in range(N):
4                Calcular la fuerza ejercida por C_j sobre C_i
5                Totalizar las fuerzas ejercidas sobre C_i
6            Calcular el desplazamiento de C_i
7            Mover C_i
```

Figura 7: Pseudo-código del algoritmo N-Body

La simulación se realiza en 3 dimensiones y la atracción gravitacionales entre dos cuerpos C_i y C_j se computa de acuerdo con la mecánica Newtoniana. Más información se puede encontrar en [19].

El pseudo-código de la versión directa se muestra en la Fig. 7. Este problema presenta dos dependencias de datos que se pueden observar en la figura anterior. En primer lugar, ningún cuerpo puede moverse hasta tanto el resto haya computado sus interacciones. En segundo lugar, ningún cuerpo puede avanzar al siguiente paso hasta tanto el resto haya alcanzado el paso actual.

3. Implementaciones N-Body

En esta sección se describen las diferentes implementaciones propuestas.

3.1. Implementación con Numba

A continuación se describen las diferentes optimizaciones que fueron consideradas en la implementación con Numba (ver Fig. 8).

Arreglos de NumPy Como estructura de datos se optó por arreglos de NumPy, debido a que permiten controlar la organización de memoria de los datos.

Opciones de compilación Se indicó que el código fuera compilado con los siguientes parámetros:

- `signature` (línea 1): A través de este parámetro se indicó que los arreglos sean contiguos en memoria para ayudar a Numba a detectar un mayor número de instrucciones vectoriales.
- `parallel=True` (línea 9): Activa la paralelización.

- `fastmath=True` (línea 9): Activa la relajación de precisión.
- `error_model="numpy"` (línea 9): Permite utilizar el modelo de división de NumPy, el cual evita la verificación de división por cero, y por ende, menos comparaciones en tiempo de ejecución.

Multi-hilado Se introdujo paralelismo a nivel de hilos a través de la sentencia `prange`. Particularmente, se crearon dos zonas paralelas; la primera se encarga de calcular la ley de atracción gravitacional de Newton y la integración de Verlet (ver Fig. 8a), mientras que la segunda simplemente actualiza la posición de los cuerpos (ver Fig. 8b).

Operaciones con tipos de datos simples Aunque las operaciones vectoriales de NumPy (*broadcasting*) simplifican la codificación, impactan negativamente en términos de rendimiento [16].

Vectorización Si bien se le indicó a Numba la utilización de instrucciones AVX-512 como parámetro de compilación, se constató que ya hacía uso correctamente de este conjunto de extensiones.

Threading layer Se varió la API de hilos a través de las *threading layers* que utiliza Numba para traducir las regiones paralelas. En particular, se seleccionó *omp* (OpenMP) por presentar mejores rendimientos.

3.2. Implementación con Cython

A continuación se describen las diferentes optimizaciones que fueron consideradas en la implementación con Cython (ver Fig. 9).

Opciones de compilación Inicialmente, se indicaron los siguientes parámetros de compilación:

- `boundcheck` (línea 1): Evita verificaciones de errores de índices sobre los arreglos.
- `wraparound` (línea 2): Evita que los arreglos se puedan indexar en relación con el final. Por ejemplo, en Python si `A` es un arreglo, con la sentencia `A[-1]` se obtiene el último elemento.
- `nonecheck` (línea 3): Evita verificaciones por variables que puedan llegar a tomar el valor *None*.
- `cdivision` (línea 4): Realiza la división a través de C evitando la API de CPython.

Tipado explícito Se especificaron los tipos de datos de Cython, y particularmente, se indicó que los arreglos sean contiguos en memoria. Esta optimización permite que las variables sean transpiladas utilizando tipos de datos de C, evitando la API de CPython y el *overhead* generado por el último.

```python
@njit(
    void(
        int64,
        double[::1], double[::1], double[::1],
        double[::1],
        double[::1], double[::1], double[::1],
        double[::1], double[::1], double[::1],
    ),
    fastmath=True, parallel=True, error_model="numpy",
)
def calculate_positions(
    N,
    positions_x, positions_y, positions_z,
    masses,
    velocities_x, velocities_y, velocities_z,
    dp_x, dp_y, dp_z,
):
    # For every body that experiences a force
    for i in prange(N):
        # Initialize the force of the body i
        forces_x = forces_y = forces_z = 0.0
        # Calculate the forces exerted on body i
        # by every other body
        for j in range(N):
            # Newton's Law of Universal Gravitation
            dpos_x = positions_x[j] - positions_x[i]
            dpos_y = positions_y[j] - positions_y[i]
            dpos_z = positions_z[j] - positions_z[i]
            dsquared = (
                (dpos_x ** 2.0) + (dpos_y ** 2.0) +
                (dpos_z ** 2.0) + SOFT
            )
            gm = GRAVITY * masses[j] * masses[i]
            d32 = dsquared ** -1.5
            # Sum the forces
            forces_x += gm * d32 * dpos_x
            forces_y += gm * d32 * dpos_y
            forces_z += gm * d32 * dpos_z

        # Calculate acceleration of body i
        aceleration_x = forces_x / masses[i]
        aceleration_y = forces_y / masses[i]
        aceleration_z = forces_z / masses[i]
        # Calculate new velocity of body i
        velocities_x[i] += aceleration_x * DT / 2.0
        velocities_y[i] += aceleration_y * DT / 2.0
        velocities_z[i] += aceleration_z * DT / 2.0
        # Calculate new position of body i
        dp_x[i] = velocities_x[i] * DT
        dp_y[i] = velocities_y[i] * DT
        dp_z[i] = velocities_z[i] * DT
```

(a) Función que calcula las posiciones de los cuerpos.

```python
@njit(
    void(
        int64,
        double[::1], double[::1], double[::1],
        double[::1], double[::1], double[::1],
    ),
    fastmath=True,
    parallel=True,
    error_model="numpy",
)
def update_positions(
    N,
    positions_x, positions_y, positions_z,
    dp_x, dp_y, dp_z,
):
    # For every body that experienced a force
    for i in prange(N):
        # Update position of body i
        positions_x[i] += dp_x[i]
        positions_y[i] += dp_y[i]
        positions_z[i] += dp_z[i]
```

(b) Función que actualiza las posiciones de los cuerpos.

Figura 8: Implementación con Numba

Multi-hilado Se introduce paralelismo a nivel de hilos a través de la sentencia `prange` que provee Cython. Particularmente, se les indicó utilizar la política `static` como *schedule* para distribuir equitativamente la carga de trabajo entre los hilos considerando la regularidad del cómputo. Por último, se desactivó el GIL a través del argumento `nogil` para permitir que los mismos se ejecuten de forma paralela.

4. Resultados Experimentales

4.1. Diseño experimental

Todas las pruebas fueron realizadas en un sistema equipado con 2×Intel Xeon Platinum 8276 de 28 núcleos (2 hilos hw por núcleo) y 256 GB de memoria RAM.

```python
@boundscheck(False)
@wraparound(False)
@nonecheck(False)
@cdivision(True)
cpdef void nbody(
    int N, int D, int T,
    double[::1] positions_x, double[::1] positions_y, double[::1] positions_z,
    double[::1] masses,
    double[::1] velocities_x, double[::1] velocities_y, double[::1] velocities_z,
    double[::1] dp_x, double[::1] dp_y, double[::1] dp_z,
):
    cdef double forces_x, forces_y, forces_z
    cdef double aceleration_x, aceleration_y, aceleration_z
    cdef double dpos_x, dpos_y, dpos_z
    cdef double dsquared, gm, d32
    cdef int i, j

    # For each discrete instant of time
    for _ in range(D):
        # For every body that experiences a force
        for i in prange(N, nogil=True, schedule="static", num_threads=T):
            # Initialize the force of the body i
            forces_x = forces_y = forces_z = 0.0

            # Calculate the forces exerted on body i by every other body
            for j in range(N):
                # Newton's Law of Universal Gravitation
                dpos_x = positions_x[j] - positions_x[i]
                dpos_y = positions_y[j] - positions_y[i]
                dpos_z = positions_z[j] - positions_z[i]
                dsquared = (dpos_x ** 2.0) + (dpos_y ** 2.0)
                    + (dpos_z ** 2.0) + SOFT
                gm = GRAVITY * masses[j] * masses[i]
                d32 = dsquared ** -1.5
                # Sum the forces
                forces_x = forces_x + gm * d32 * dpos_x
                forces_y = forces_y + gm * d32 * dpos_y
                forces_z = forces_z + gm * d32 * dpos_z

            # Calculate acceleration of body i
            aceleration_x = forces_x / masses[i]
            aceleration_y = forces_y / masses[i]
            aceleration_z = forces_z / masses[i]
            # Calculate new velocity of body i
            velocities_x[i] += aceleration_x * DT / 2.0
            velocities_y[i] += aceleration_y * DT / 2.0
            velocities_z[i] += aceleration_z * DT / 2.0
            # Calculate new position of body i
            dp_x[i] = velocities_x[i] * DT
            dp_y[i] = velocities_y[i] * DT
            dp_z[i] = velocities_z[i] * DT

        # For every body that experienced a force
        for i in prange(N, nogil=True, schedule="static", num_threads=T):
            # Update position of body i
            positions_x[i] += dp_x[i]
            positions_y[i] += dp_y[i]
            positions_z[i] += dp_z[i]
```

Figura 9: Implementación con Cython.

El sistema operativo fue Ubuntu 20.04.2 LTS y el intérprete utilizado fue Python v3.8.10 junto con Numba v0.52.0 y NumPy v1.20.1. Por el otro lado, se utilizó Cython v0.29.22 con el compilador Intel v2022.0.0.20211123.

Para la evaluación de las implementaciones, se varió la carga de trabajo al usar diferentes números de cuerpos: $N = \{65536, 131072, 262144, 524288\}$ mientras que tanto el número de hilos (T=112) como el número de pasos de simulación (I=100) se mantuvieron fijos.

Por último, cabe destacar que la versión de Cython fue compilada utilizando el *flag* `-O3` junto con los siguientes *flags* adicionales: `march=native` (indica el uso de los flags más adecuados para el procesador subyacente) y `fp-model fast=2` (indica relajar la precisión de punto flotante).

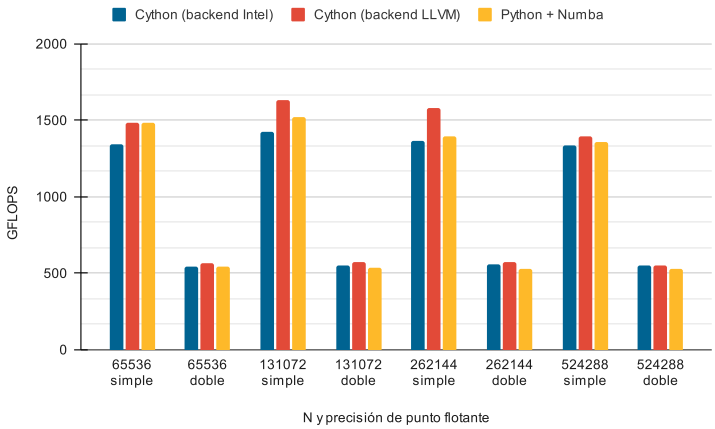


Figura 10: Comparación de rendimiento de las versiones finales entre Numba y Cython variando el tipo de dato y *N*.

4.2. Rendimiento

Al igual que en trabajos previos, se emplea la métrica GFLOPS para evaluar el rendimiento [16]. En la Fig. 10 se presenta una comparación entre las optimizaciones finales de Numba y Cython al variar la carga de trabajo y el tipo de dato. En el mismo, se puede observar que la incorporación de LLVM como *backend* mejoró las prestaciones de la versión anterior (*backend* propio de Intel); en particular, el rendimiento se incrementó (en promedio) 10 % y 3 % en simple y doble precisión, respectivamente. Por otro lado, resulta importante notar que la versión previa de Cython sólo lograba superar a Numba en precisión doble. Gracias a la mejora anterior, la nueva versión Cython resulta superior con ambos tipos de datos, logrando incrementos (en promedio) de 5 % y 6 % en simple y doble precisión, respectivamente.

4.3. Esfuerzo de Programación

En este trabajo el esfuerzo de programación se estimó en función del indicador SLOC (*Source Lines Of Code*), el cual permite contabilizar las líneas de código [12]. Sin embargo, la subjetividad de este indicador dificulta medir de manera exacta el costo de programación empleado. Por lo tanto, adicionalmente se realiza una comparación cualitativa con el fin de complementar al primer análisis y permitirle al lector un mejor entendimiento del esfuerzo de programación requerido por cada solución.

En la Fig. 11 se puede observar la cantidad de líneas de código de cada versión discriminando instrucciones, comentarios y líneas en blanco (medidas con la herramienta *cloc*[7]). La solución con Numba requirió 92 líneas de código, lo que

[7] `https://github.com/AlDanial/cloc`

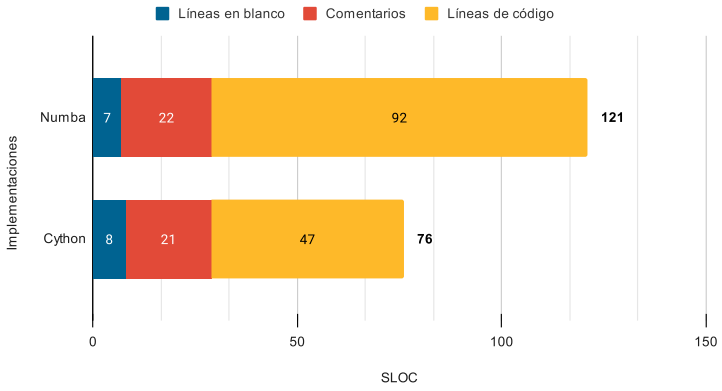


Figura 11: Cantidad de líneas de código de las implementaciones Numba y Cython para N-Body.

representa casi el doble que implementación de Cython. Sin embargo, hay que tener en cuenta que la primera contiene las opciones de compilación declaradas en la propia implementación (representa 32.6 % del código total) y, además, contiene llamadas a 2 funciones adicionales que actúan como zonas paralelas, mientras que en la implementación de Cython no se encuentra dicha modularización.

A continuación, se describen los aspectos cualitativos que fueron identificados al momento de desarrollar cada solución. Se puede mencionar que Cython requirió menos líneas de código, pero a su vez, se necesitó de un conocimiento adicional de C y OpenMP a la hora de paralelizar el problema. Mientras que en Numba, simplemente se necesitó indicar la sentencia `prange`, y luego la librería se encargó de traducir las zonas paralelas a la *threading layer* correspondiente. Sin embargo, a favor de Cython, esto permite un mayor grado de control sobre la sincronización de los hilos, ya que para llevarla a cabo podemos interactuar con la API de OpenMP. En sentido opuesto, Numba no permite nativamente la sincronización explícita de los hilos y se deben utilizar librerías externas para lograrlo (por ej. `threading`).

Por otra parte, en la solución de Cython se tuvo que modificar el código de la simulación para declarar los tipos, mientras que en Numba se los declaró a través del parámetro `signature` en el decorador `njit` sin interferir en el código de la función ya desarrollada.

Para finalizar, se debe destacar que la implementación de Cython necesitó 3 archivos para llevar a cabo la solución: (1) El archivo de compilación *setup.py*. (2) El archivo fuente de la simulación *cynbody.pyx*. (3) El ejecutable *run.py*. Por lo tanto, al momento de ejecutar cada simulación, fue necesario compilar el código fuente, luego importarlo a través de un *script* de Python y finalmente ejecutarlo mediante la línea de comandos; mientras que por otra parte, en Numba simplemente se precisó un archivo para almacenar el código fuente y un comando para ejecutar la simulación.

5. Conclusiones y Trabajo Futuro

En este trabajo se realizó una comparación más justa, actualizada y amplia de las prestaciones (rendimiento y esfuerzo de programación) de los traductores Numba y Cython para el caso de estudio *N-Body*.

Por un lado, los resultados muestran que la incorporación de LLVM como *backend* del compilador Intel para C repercute positivamente en el rendimiento de la versión Cython en comparación al *backend* clásico. A diferencia del estudio anterior, esta mejora lleva a que Cython resulte levemente superior a Numba tanto en doble como en simple precisión.

Por otro lado, en cuanto a esfuerzo de programación, Cython resultó con menos líneas de código, pero requirió un conocimiento adicional de C+OpenMP junto con un proceso más laborioso de ejecución. En sentido opuesto, Numba tomó un mayor número de líneas de código debido a la especificación de opciones de compilación en el mismo código y la modularización empleada; sin embargo, Numba resultó más simple a la hora de desarrollar gracias a su enfoque de decoradores y a la posibilidad de mantener el mismo código que la implementación original.

En base a lo anterior, se puede afirmar que en contextos similares a los de este estudio tanto Numba como Cython pueden ser potentes herramientas para acelerar aplicaciones *CPU-bound* desarrolladas en Python. La elección entre uno y otro estará mayormente determinada por el enfoque que el equipo de desarrollo encuentre más conveniente, considerando las características propias de cada uno.

Como trabajos futuros, resulta de interés extender este estudio mediante los siguientes aspectos:

- Replicar el estudio realizado considerando: (1) otros casos de estudio que sean computacionalmente intensivos pero cuyas características sean diferentes a las de *N-Body*; (2) otras arquitecturas multicore distintas a la usada en este trabajo. Ambas extensiones contribuirían a robustecer los resultados encontrados.
- Dado que existen otras tecnologías que permitan implementar paralelismo a nivel de procesos en Python, realizar una comparación entre ellas considerando no sólo el rendimiento sino también el costo de programación.

Referencias

1. 7. Decorators — Python Tips 0.1 documentation, `https://book.pythontips.com/en/latest/decorators.html`
2. Coiling Python Around Hybrid Quantum Systems, `https://www.nextplatform.com/2021/05/19/coiling-python-around-hybrid-quantum-systems/`
3. Cython: C-Extensions for Python, `https://cython.org/`
4. Home - OpenMP, `https://www.openmp.org/`
5. Home | Jython, `https://www.jython.org/`
6. Intel c/c++ compilers complete adoption of llvm, `https://www.intel.com/content/www/us/en/developer/articles/technical/adoption-of-llvm-complete-icx.html`

7. Microsoft's new research lab studies developer productivity and well-being | VentureBeat, https://venturebeat.com/2021/05/25/microsofts-new-research-lab-studies-developer-productivity-and-well-being/
8. Numba documentation — Numba 0.53.1-py3.7-linux-x86_64.egg documentation, https://numba.readthedocs.io/en/stable/index.html
9. NumPy, https://numpy.org/
10. PyPy, https://www.pypy.org/
11. What Is the Python Global Interpreter Lock (GIL)? – Real Python, https://realpython.com/python-gil/
12. Bhatt, K., Tarey, V., Patel, P., Mits, K.B., Ujjain, D.: Analysis of source lines of code (sloc) metric. International Journal of Emerging Technology and Advanced Engineering **2**(5), 150–154 (2012)
13. Cai, X., Langtangen, H.P., Moe, H.: On the Performance of the Python Programming Language for Serial and Parallel Scientific Computations. Scientific Programming **13**(1), 31–56 (2005). https://doi.org/10.1155/2005/619804
14. Gmys, J., Carneiro, T., Melab, N., Talbi, E.G., Tuyttens, D.: A comparative study of high-productivity high-performance programming languages for parallel metaheuristics. Swarm and Evolutionary Computation **57**, 100720 (Sep 2020). https://doi.org/10.1016/j.swevo.2020.100720
15. Marowka, A.: Python accelerators for high-performance computing. The Journal of Supercomputing **74**(4), 1449–1460 (Apr 2018). https://doi.org/10.1007/s11227-017-2213-5
16. Milla, A., Rucci, E.: Performance comparison of python translators for a multithreaded cpu-bound application. In: Pesado, P., Gil, G. (eds.) Computer Science – CACIC 2021. pp. 21–38. Springer International Publishing, Cham (2022)
17. Milla, A., Rucci, E.: Acelerando Código Científico en Python usando Numba. XXVII Congreso Argentino de Ciencias de la Computación (CACIC 2021) p. 12 (Oct 2021), http://sedici.unlp.edu.ar/handle/10915/126012
18. Roghult, A.: Benchmarking Python Interpreters: Measuring Performance of CPython, Cython, Jython and PyPy. Master's thesis, School of Computer Science and Communication, Royal Institute of Technology, Sweden (2016)
19. Rucci, E., Moreno, E., Pousa, A., Chichizola, F.: Optimization of the n-body simulation on intel's architectures based on avx-512 instruction set. In: Computer Science – CACIC 2019. pp. 37–52. Springer International Publishing (2020)
20. TIOBE Software BV: TIOBE Index for November 2021 (11 2021), https://www.tiobe.com/tiobe-index/
21. Varsha, M., Yashashree, S., Ramdas, D.K., Alex, S.A.: A Review of Existing Approaches to Increase the Computational Speed of the Python Language. International Journal of Research in Engineering, Science and Management (2019)
22. Wilbers, I., Langtangen, H.P., Odegard, A.: Using cython to speed up numerical python programs. In: Proceedings of MekIT. pp. 495–512 (2009)
23. Wilkens, F.: Evaluation of performance and productivity metrics of potential programming languages in the HPC environment. Bachelor's thesis, Faculty of Mathematics, Informatics und Natural Sciences, University of Hamburg, Germany (2015)

Optimización del código y las dependencias de las funciones Serverless para mejorar el rendimiento de las aplicaciones

Nelson Rodríguez, Martín Gómez, María Murazzo, Ana Laura Molina, Lorena Parra

Departamento de Informática, Facultad de Ciencias Exactas Físicas y Naturales, Universidad Nacional de San Juan, San Juan, Argentina
nelson@iinfo.unsj.edu.ar, martinsj0811@gmail.com, maritemurazzo@g,mail.com, almm95@gmail.com, lorenaparra152@yahoo.com.ar,

Resumen. Serverless Computing es una reciente arquitectura para Cloud Computing que presenta ventajas considerables para los usuarios. Sin embargo, debido a su reciente aparición, muchas de sus limitaciones o desventajas no están totalmente resueltas. Si bien el desarrollo serverless presenta condiciones comunes al desarrollo para otro tipo de plataformas o entornos, de las cuales se pueden obtener buenas prácticas de tipo "genéricas", también serverless presenta aspectos propios debido a que son funciones distribuidas ejecutándose en una plataforma Cloud, cuya ejecución es conducida por eventos, sin estado y sin responsabilidades operativas por parte del usuario, entre otras. Se debe considerar además el hecho de que determinadas prácticas pueden reducir costos como aquellas que conducen a minimizar el arranque en frío, otras apuntan a aspectos de la seguridad o a la gestión del BackEnd. En el presente trabajo se realizó un análisis de las buenas prácticas para serverless y se trabajó en especial en aquellas que impactan en la performance, en particular en el uso de recursos provistos por la plataforma y en optimizar las dependencias de las funciones, realizando una serie de pruebas y análisis en diversos lenguajes de programación, que permiten emitir conclusiones sobre el impacto que causan estas buenas prácticas en la mejora de los tiempos de ejecución.

Keywords: Serverless Computing, FaaS, Best Practices, Cloud Computing

1 Introducción

Cloud Computing es un modelo de provisión de servicios y una arquitectura con varios años de utilización por parte de la industria y la academia [1].

A partir de la publicación de un artículo de Google en 2003, la computación en la nube ha evolucionado del sistema de TI interno al servicio público y ya está arraigada en el diseño de plataformas de motores de búsqueda [2].

Siguiendo la evolución observada en la historia de la contenerización, los servicios en la nube se han adaptado para ofrecer contenedores de mejor ajuste que requieren menos tiempo para cargar (arranque) y proporcionar mayor automatización en el manejo (orquestación) de contenedores en nombre del cliente [3]. La Computación Serverless promete lograr la automatización completa en la gestión de contenedores.

El modelo de computación serverless es impulsado por eventos en el que los recursos informáticos se proporcionan como servicios escalables. En el modelo tradicional se cobra un costo fijo y recurrente por los recursos informáticos del servidor, independientemente de la cantidad de trabajo realizado por el servidor. Sin embargo, la implementación Serverless ha superado esta deficiencia, ya que se paga solo por el uso del servicio y no se cobra por el tiempo de inactividad.

En este paradigma emergente, las aplicaciones de software se descomponen en múltiples funciones independientes sin estado [4] [5]. Las funciones solo se ejecutan en respuesta a acciones desencadenantes (como interacciones de usuario, eventos de mensajería o cambios en la base de datos), y se pueden escalar de forma independiente y pueden ser efímeras (pueden durar una invocación) y están completamente administrados por el proveedor de Cloud.

Los principales proveedores de nube han propuesto diferentes plataformas informáticas sin servidor como AWS Lambda, Microsoft Azure Functions, Google Functions, IBM Cloud Functions, Cloudflare Worker, Alibaba Function Compute. Dichas plataformas facilitan y permiten que los desarrolladores se centren más en la lógica de negocios, sin la sobrecarga de escalar y aprovisionar la infraestructura [8].

Castro et al [6], ofrecen una definición basada en las características: “La informática serverless se puede definir por su nombre, que es pensar (o preocuparse) menos por los servidores. Los desarrolladores no necesitan preocuparse por los detalles de bajo nivel de administración y escalado de servidores, y solo pagan cuando procesan solicitudes o eventos”. Luego la define como: “La informática serverless es una plataforma que oculta el uso del servidor a los desarrolladores y ejecuta código que escala bajo demanda automáticamente y facturado solo por el tiempo que se ejecuta el código”.

En la mayoría de los casos, se pueden escribir funciones en el lenguaje que el programador considere más adecuado (Node.js, Python, Go, Java y más) y utilizar herramientas de contenedor y serverless, como AWS SAM o la CLI de Docker, para compilar, probar e implementar las funciones.

Por otro lado, especialistas de Expert Market Research pronostican que el mercado global de computación serverless crecerá en el período de pronóstico de 2022-2027 a una tasa compuesta anual del 22.2% [10].

Un modelo basado en funciones es adecuado para ráfagas, uso de CPU intensivo, cargas de trabajo granulares. Actualmente, los casos de uso de FaaS varían ampliamente, incluido el procesamiento de datos, el procesamiento de flujo, la computación de borde (IoT) y la computación científica [3].

Serverless cubre una amplia gama de tecnologías, que se pueden agrupar en dos categorías: Backend-as-a-Service (BaaS) y Functions-as-a-Service (FaaS).

Backend-as-a-Service permite reemplazar los componentes del lado del servidor con servicios listos para usar.Algunos ejemplos son los sistemas de autenticación remota, la administración de bases de datos, el almacenamiento en el cloud.

Existen diversos desafíos, oportunidades y problemas a resolver, entre ellos la experiencia del desarrollador [7], Interoperabilidad, testing, composición de funciones, seguridad, administración del ciclo de vida, administración de requerimientos no funcionales, performance, optimización del overhead, ingeniería para costo-performance, entre otros [6].

2 Trabajos relacionados

Existen solo dos trabajos que analicen las buenas prácticas aplicadas a Serverless Computing. Si bien existen variadas publicaciones de la industria, algunas de ellas no tienen la rigurosidad necesaria y otras presentan conclusiones parciales.

La publicación de Greg Wilson et at [9], trata solo sobre las mejores prácticas en computación científica. En el resumen indica que: Los científicos pasan cada vez más tiempo construyendo y usando software. Sin embargo, a la mayoría de los científicos nunca se les enseña cómo hacer esto de manera eficiente. Como resultado, muchos desconocen las herramientas y prácticas que les permitirían escribir código más confiable y mantenible con menos esfuerzo.

El trabajo de Rajdeep Mukherjee et al [10], presenta un framework de fuerza industrial para el análisis estático preciso de aplicaciones Python que utilizan los servicios en la nube de AWS. No trata sobre Serverless y se enfoca más en las características de Python.

3 Mejores Prácticas

Las mejores prácticas son un conjunto de directrices, ética o ideas que representan el curso de acción más eficiente o prudente en una situación empresarial determinada.

Las mejores prácticas pueden ser establecidas por autoridades, como reguladores, organizaciones autorreguladoras (SRO) u otros órganos de gobierno, o pueden ser decretadas internamente por el equipo directivo de una empresa. [11]

Según Diccionario Cambridge, Mejores prácticas es un método de trabajo, o conjunto de métodos de trabajo, que se acepta oficialmente como el mejor para usar en un negocio o industria en particular, la aplicación del término puede ser:

La compañía identificó las mejores prácticas que han llevado a un desarrollo de productos más exitoso.

Muchos empleadores quieren adoptar las mejores prácticas en la gestión de su gente, pero no saben qué es esto.

La empresa lleva adelante una política/programa de mejores prácticas [12]

Las buenas prácticas se pueden clasificar en: a) para el desarrollo de software en general, b) para el desarrollo serverless en general c) específicas para algún frameworks d) específicas para alguna plataforma Serverless como AWS o Cloud Function.

Debido a que el desarrollo serverless comparte algunas características comunes a todo desarrollo, se debe aplicar (aunque en efecto la mayoría de las veces aplican) los principios comunes al desarrollo como: KISS (mantener su código lo más simple posible), DRY (que significa "no repetir", y su idea subyacente es que tienes que evitar y reducir las repeticiones y la redundancia reemplazándolas con abstracciones o utilizando la normalización de datos), SOLID (representado por 5 prácticas que son: principio de responsabilidad única, Principio abierto-cerrado (OCP), Principio de sustitución de Liskov , Principio de segregación de interfaz, Principio de inversión de dependencia) [13].

En cuanto a las de desarrollo Serverless se puede mencionar a: diseñar las aplicaciones pensando en escalado automático, aplicar estrategias para minimizar el arranque en frío como reducir en lo posible el número de dependencias de las funciones o utilizar WebPack que optimiza el arranque en frío.

Específicas para AWS se puede mencionar: aprovechar la reutilización del entorno de ejecución para mejorar el rendimiento de su función utilizar una directiva keep-alive para mantener conexiones persistentes y minimizar el tamaño del paquete de implementación según sus necesidades de tiempo de ejecución [14].

Específicas para Cloud Function: escribir código de función en un estilo sin estado para asegurarse de que el código no se someterá a ningún mantenimiento de estado, crear instancias de cualquier objeto que pueda reutilizarse y se sugiere utilizar servicios de administración de código externo (como Git) para fines de administración de versiones y auditorías del código principal [15].

Específicas para Serverless Framework: implementar almacenamiento en caché y reducir los tiempos de inicialización [16].

Se ha probado una de estas mejores prácticas que se indican a continuación.

4 Desarrollo de las pruebas

Una de las mejores prácticas que se decidió probar es la que establece que se debe optimizar el código y las dependencias de las funciones sin servidor para mejorar el rendimiento de las aplicaciones. Por ejemplo, las funciones escritas en un lenguaje interpretado como Node.js y Python tienen tiempos de invocación inicial significativamente más rápidos que las funciones escritas en un lenguaje compilado como Go.

Para realizar esto se crearon dos funciones sin servidor utilizando el servicio Lambda de la plataforma Amazon Web Services. Las mismas resuelven un mismo problema de integración numérica de la misma manera, aunque codificadas en distintos lenguajes, una de ellas codificada en el lenguaje interpretado Python y la otra en el lenguaje compilado Go. Además se realizó una tercera prueba utilizando una función codificada en lenguaje Java.

A continuación, en la Figura 1 se muestra el código de la función en Python.

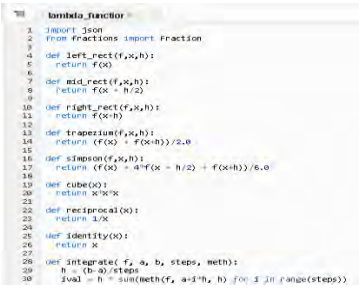


Figura 1

Los códigos utilizados fueron extraídos del sitio web rosettacode.org, el cual contiene algoritmos que son desarrollados en distintos lenguajes de programación a fin de realizar distintas comparaciones entre estos.

En las siguientes imágenes se muestran los resultados de las ejecuciones

Función en Go en su primera prueba

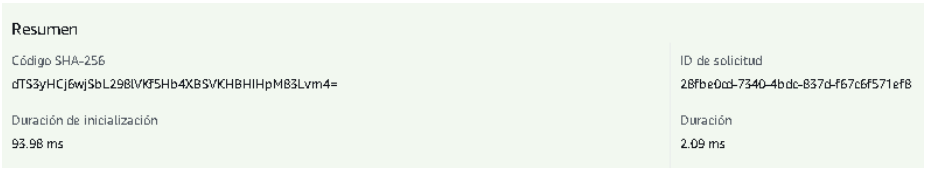


Figura 2

Función en Python en su primera prueba

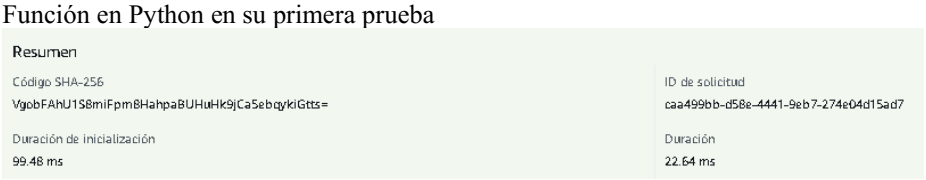


Figura 3

Función en Java en su primera prueba

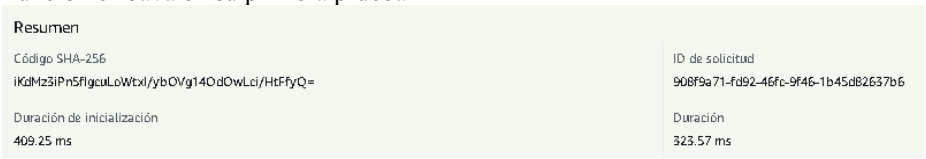


Figura 4

5 Resultados obtenidos

Al ejecutar las funciones se obtuvieron tiempos de inicialización similares entre la función desarrollada en Python y la desarrollada en Go, pero la gran diferencia se obtuvo en la duración de la ejecución, donde la función desarrollada en Go supera en 10 veces a la función desarrollada en Python. En el caso particular de Java, se observa que su inicialización y duración de primera ejecución supera enormemente a las otras dos implementaciones, esto se debe a que Java aún no se encuentra optimizado completamente para obtener una inicialización rápida. Es importante destacar que luego de la primera ejecución, las tres implementaciones disminuyen notablemente su duración de ejecución, debido a que luego del arranque en frío, la función Lambda queda inicializada en espera de otra ejecución y también utiliza una caché propia para ejecutarse rápidamente.

Los resultados de las observaciones anteriores se resumen en la Tabla 1, donde se puede observar el tiempo de inicialización de la función y la duración de su ejecución para cada uno de los lenguajes analizados.

		Python	Go	Java
Primera prueba	Tiempo de inicialización	99.48ms	93.98ms	409.25ms
	Duración primera ejecución	22.64ms	2.09ms	323.57ms
	Duración segunda ejecución	13.43ms	2.31ms	2.36ms
Segunda prueba	Tiempo de inicialización	101.95ms	72.02ms	419.35ms
	Duración primer ejecución	20.13ms	1.73ms	350.08ms

Tabla 1

6 Conclusiones y Futuros trabajos

El objetivo del presente trabajo es analizar las buenas prácticas de desarrollo Serverless que impacten directa o indirectamente en la performance, en particular se probó la optimización del código y las dependencias de las funciones para mejorar la performance. En una publicación de 2021[18] se pudo probar las ventajas de usar algunos recursos provistos por la plataforma Serverless como es Step Function en AWS, pero por razones de espacio no se muestran estos gráficos. También es ventajoso usar los Frameworks disponibles como Serverless Frameworks, Restaría probar en ambos casos, cuanto es la mejora y bajo qué condiciones. Por otro lado no se ha llegado a evaluar todas las prácticas que consideramos que van a impactar en la performance y creemos como la mayoría están indicadas por la industria, que es necesario hacerlo cuidadosamente porque muchas de ellas pueden presentar mejoras pero solo bajo determinadas situaciones. Existen otras buenas prácticas, que por el momento no han sido analizadas dado que tratan de resolver problemas no vinculados a la performance.

Referencias

1.M. Armbrust, et al., Above the clouds: A Berkeley view of cloud computing, Tech. Rep. No. UCB/EECS-2009-28, 2009.

2.S. Ghemawat, H. Gobioff and S. T. Leung, The Google file system, *SOSP*, 2003.
3.E. van Eyk, L. Toader, S. Talluri, L. Versluis, A. Uță and A. Iosup, "Serverless is More: From PaaS to Present Cloud Computing," in IEEE Internet Computing, vol. 22, no. 5, pp. 8-17, Sep./Oct. 2018, doi: 10.1109/MIC.2018.053681358.
4.Adzic, G., Chatley, R.: Serverless computing: economic and architectural impact. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (pp. 884-889). ACM.(2917)
5.AWS Lambda: aws.amazon.com/es/lambda/
6.Castro p., Ishakian v., Muthusamy v., Slominski a.: The rise of serverless computing. In: Communications of the ACM | Dec. 2019 | VOL. 62 | NO. 12 (2019)..
7.Rodríguez N., Atencio H. et al: Interoperabilidad de funciones en el Modelo de Programación de Serverless Computing. IV CICCSI. Universidad Champagnat (2020).
8.Bermbach D., Karakaya A., Buchholz S.: Using Application Knowledge to Reduce Cold Starts in FaaS Services. In: SAC '20, March 30-April 3, 2020, Brno, Czech Republic (2020).
9.Greg Wilson et al. Best Practices for Scientific Computing. https://doi.org/10.1371/journal.pbio.1001745 (2014)
10.10. EMR: Global Serverless Computing Market Outlook https://www.expertmarketresearch.com/reports/serverless-computing-market (2021).
11. Investopedia: Best Practices. Best Practices Definition (investopedia.com)
12.Cambridge Bussiness Dictionary. BEST PRACTICE | meaning in the Cambridge English Dictionary
13.What are Software Engineering Best Practices?. Software Engineering Best Practices. Principles of Programming & Development | LITSLINK Blog
14.Best practices for working with AWS Lambda functions. https://docs.aws.amazon.com/lambda/latest/dg/best-practices.html
15.Serverless Cloud Function Best Practice. Product Documentation. Tencent Cloud
16.Yan CuiTop 10 Serverless best practices. https://www.datree.io/resources/serverless-best-practices
17.Mukherjee R. et al. Static Analysis for AWS Best Practices in Python Code. https://arxiv.org/abs/2205.04432 (2022)
18.Rodríguez N.,Atencio H. et al. Análisis de ejecución múltiple de Funciones Serverless en AWS. XXVII CACIC.(2021).

Análisis de desempeño de Serverless para problemas HPC

Joaquín Lebeti[4] María Murazzo[1], Nelson Rodríguez[1], Ana Laura Molina[1]
[1] Departamento de Informática - F.C.E.F. y N. - U.N.S.J.

[4] Alumno Avanzado Licenciatura en Cs. de la Computación - F.C.E.F. y N. - U.N.S.J.
Complejo Islas Malvinas. Cereceto y Meglioli. 5400. Rivadavia. San Juan, 0264 4234129
lebejoaquin@gmail.com, marite@unsj-cuim.edu.ar, nelson@iinfo.unsj.edu.ar,
almm95@gmail.com

Resumen. Se ha demostrado que la ejecución de aplicaciones de HPC en el cloud es una opción viable a las arquitecturas paralelas o distribuidas convencionales, las cuales requieren un alto grado de administración, así como un pobre escalado de recursos. El enfoque tradicional para un usuario usualmente es utilizar al proveedor de Cloud para aprovisionar máquinas virtuales (VM) empleándolas de manera similar a una infraestructura local, con el consiguiente problema de la administración de recursos sumado a la degradación de la performance de las aplicaciones por la contextualización de los ambientes virtualizados. Serverless computing, permite a un usuario ejecutar código escrito en el lenguaje de programación de su elección, sin tener que aprovisionar primero una máquina virtual. Por otro lado, la elasticidad, disponibilidad, escalabilidad y la tolerancia a fallas son proporcionadas de manera transparente por el proveedor cloud. De esta manera es posible disminuir la complejidad de la administración de la infraestructura para el desarrollador, permitiéndole que se centre en la lógica de la aplicación. Y además surgen ventajas económicas, al pagar solo por el tiempo de uso. El trabajo se centra en el desafío de evaluar el costo, no solo monetario sino también de performance, de migrar aplicaciones de HPC a entornos serverless. Esta evaluación permitirá que se pueda tomar la decisión que infraestructura se usará con la finalidad que se obtenga el mejor beneficio de performance.

Keywords: Serverless Computing, HPC, Cloud Computing

1 Introducción

La popularización de IoT y la masificación de las infraestructuras cloud durante el último tiempo han abierto un mundo de posibilidades para las aplicaciones HPC. Esto se debe a que los dispositivos IoT generan una gran cantidad de datos, los cuales se hace impráctico tratarlos con paradigmas tradicionales. Para lograr el procesamiento adecuado de estos datos con características de velocidad y tamaño importantes, se hace necesario prescindir de los paradigmas de programación tradicionales [1]. Es por ello que es necesario aplicar algoritmos que permitan aprovechar la escalabilidad de recursos de cómputo y procesamiento de datos [2] [3]. En este sentido se plantea como solución al procesamiento de datos provenientes del IoT, técnicas de computación de alta prestaciones (HPC) con el fin de aumentar la performance de procesamiento.

Los entornos HPC son ideales para resolver aplicaciones científicas, computacionalmente costosas con manejo de grandes cantidades de datos, a fin de lograr resultados en menor tiempo. Dadas estas características, estas arquitecturas son las mejores candidatas para procesar datos provenientes del IoT. Si bien, las arquitecturas HPC han evolucionado en pos de obtener mejores tiempos de respuesta para las aplicaciones, presentan el inconveniente del escalado, de recursos de cómputo. Es por ello que una alternativa es migrar al cloud [4].

Cloud Computing se ha caracterizado por ser una tecnología centrada en ofrecer cómputo bajo demanda como cualquier otro servicio. Esto es una ventaja para montar aplicaciones donde es necesario el procesamiento intensivo, tales como aquellas aplicaciones que procesen y extraigan información de datos provenientes de dispositivos de IoT [5].

Los proveedores cloud alegan muchas ventajas en la migración de aplicaciones de HPC, como el acceso rápido a los recursos, costos más bajos y flexibilidad en la contratación y el aprovisionamiento de recursos. Un punto adicional es la seguridad, la cual, afirman es de alto nivel y en muchos casos muy difícil de implementar en la mayoría de los laboratorios, ya que tener personal de TI especializado en seguridad no es común [6].

Un aspecto más que lleva a la adopción del cloud como plataforma de despliegue de aplicaciones HPC es, mejorar la colaboración científica, es decir, facilitar la investigación colaborativa y la innovación; este aspecto ha sido el foco de este proyecto desde hace varios años al incorporar investigadores de otras universidades del país.

Otro tema es el potencial ahorro de tiempo que ofrece el cloud a los usuarios finales. Estos usuarios no necesitan preocuparse por actualizaciones de software, compatibilidad o parches de seguridad, pues todo esto es aprovisionado de forma transparente.

Sin embargo, el cloud tiene dos grandes desventajas, la primera es la degradación de la performance de las aplicaciones al montarlas sobre arquitectura virtualizada, debido a que genera overhead en la contextualización de las máquinas virtuales; la segunda desventaja cuando se despliegan aplicaciones en el cloud, es que es responsabilidad de la organización mantener funcionando de forma correcta la infraestructura que se necesite para el despliegue de las aplicaciones, lo cual lleva a cargar costos sobre el presupuesto para su mantenimiento y soporte [7].

En este sentido, la aparición del Serverless Computing [8] logra que los desarrolladores no tengan que preocupar por el aprovisionamiento y escalado de la infraestructura, por lo que se pueden centrar en la lógica de sus aplicaciones. De esta forma es posible lograr la abstracción de la gestión de servidores (aprovisionamiento, configuración, escalado, etc.) para que los usuarios, en este caso desarrolladores, puedan enfocarse en la lógica de sus aplicaciones.

2 Trabajos relacionados

Se ha mencionado en párrafos anteriores las ventajas y desventaja de migrar aplicaciones HPC al cloud. Sin embargo, el tiempo y el esfuerzo necesarios para configurar los recursos virtuales pueden ser mayores que el tiempo y el esfuerzo reales dedicados a hacer los cálculos. En contraposición, si se usa el paradigma serverless,

será posible tener control más granular sobre el servicio prestado al dejar en manos del proveedor cloud la administración de la infraestructura.
En [9], se ha realizado un mapeo sistemático de 89 casos de uso donde se aplicó el paradigma serverless, para resolver problemas en su mayoría que se encuadran en HPC. Pero hay escasa información sobre una comparativa de performance entre las aplicaciones ejecutándose sobre paradigma serverless frente a las mismas aplicaciones ejecutándose sobre una infraestructura cloud tradicional, en la cual se pueda hacer un análisis de comportamiento para posteriormente decidir cuál es la mejor solución para ejecutar aplicaciones HPC.
En este trabajo se usa como punto de partida [10] y profundiza las tareas de investigación en base a [11], [12], [13], entre otras de los últimos años, en las cuales se han explorado y evaluado el rendimiento del uso de serverless en aplicaciones HPC. Si bien estos estudios demuestran que serverless es fácil de usar y económico, no se ha cuantificado su efectividad sobre el enfoque convencional de aplicaciones corriendo en cloud.

3 Definición de la infraestructura

Para definir la infraestructura cloud que se utilizará, se han analizado los tres principales proveedores cloud públicos del mercado: AWS, Microsoft Azure y Google Cloud Platform y sus respectivas opciones para serverless: AWS Lambda, Azure Functions y Cloud Functions. Además se ha usado el estudio "Performance evaluation of heterogeneous cloud functions" (Figiela et al., 2018) donde se evalúa el desempeño del recurso FaaS que provee cada uno de los proveedores basándose en 7 características: rendimiento computacional, rendimiento de red, transferencia de datos entre la función y el storage del proveedor, sobrecargas de la función, duración de instancias, costos, heterogeneidad de la infraestructura. En este estudio se evalúan estas características en base a 5 hipótesis: el rendimiento de una función es proporcional al tamaño, el rendimiento de la red es proporcional al tamaño de la función, las sobrecargas no son relativas al tamaño de la función y son consistentes, las instancias se reutilizan entre llamadas y se reciclan a intervalos regulares, las funciones se ejecutan en hardware heterogéneo. Del análisis realizado, se concluye que el funcionamiento de AWS Lambda y Cloud Functions es consistente con la hipótesis de que el rendimiento de una función está relacionado con el tamaño de la misma. En el caso de Lambda, esta relación es más fuerte que en Cloud Functions ya que ésta, en más de un aspecto estudiado, presenta una mejora en el rendimiento en un 5% del total de ejecuciones realizadas. Por otro lado, Azure Functions presenta diferencias en el rendimiento con las dos primeras ya que el tamaño de la función no es configurable. En este trabajo se usa como infraestructura cloud Google Cloud Platform.

4 Escenarios de Trabajo

Se plantean tres escenarios en los cuales se evaluarán: tiempo de ejecución, tiempo de respuesta, performance, precio, resultados, facilidad de uso, tiempo de contextualización, aspectos particulares de cada escenario que lo destacan.

El primer escenario de evaluación es cloud “tradicional”. Este escenario se hace sobre un recurso que funciona como PaaS para el cual se utilizará Datproc, el cual permite poner a funcionar un cluster con Spark de n nodos, n-1 workers y 1 master.
El segundo escenario es serverless y el tercer escenario, se trata de una extensión del primero y la filosofía de serverless del segundo escenario, busca combinar ambos componentes. Para estos dos escenarios se usará Cloud Functions. La configuración de este escenario consta de n nodos, n-1 workers y 1 master para hacer una equivalencia con un cluster de n-1 workers y 1 master de la configuración de Dataproc. El nodo maestro será el encargado de desencadenar la ejecución de la función worker n-1 veces para arrancar n-1 instancias en “simultáneo” y luego recolectar los resultados que obtenga cada una de esas instancias para resolver un resultado único que será el resultado final del problema a resolver. Para resolver cada una de las partes del problema, en la función worker se utilizará pandas para trabajar el archivo de datos como un dataframe.

5 Resultados obtenidos

El objetivo del estudio es evaluar el comportamiento de FaaS para problemas de HPC con datos provenientes de IoT, y como el rendimiento de las funciones es variable respecto del tamaño de la misma, se evaluará el mismo problema con diferentes tamaños de función.
El problema a resolver es determinar, a partir de datos históricos obtenidos por sensores desde 2009 hasta 2022, la hora del día en que la cantidad de CO (monóxido de carbono) es más alta. El dataset en cuestión cuenta inicialmente con una cantidad aproximada de 110 mil registros, es decir, un registro por cada hora del día desde 2009 hasta el 2022 en las ciudades de Centenario, Córdoba y La Boca. Con el objetivo de dar resultados precisos, como el dataset contaba con alrededor del 40% de registros sin datos, lo que se hizo en un comienzo es una limpieza de los mismos Tras esto se obtuvo un dataset con 70 mil registros que será utilizado para comprobar el comportamiento de las dos configuraciones anteriormente descritas.
Para hacer una comparativa válida, se resolvió el problema con dos configuraciones equivalentes. La configuración de recursos mínima para cada nodo en Dataproc es de 1 vCPU y 3.5GB de RAM.
Si bien es interesante ver el comportamiento de la memoria, la comparativa principal es sobre el procesador. Por ello, para hacer una comparativa equivalente, se comenzará evaluando con 1 vCPU (3 workers, 1 master)
Para la configuración PaaS con Dataproc se configuran los recursos del cluster desde una interfaz gráfica y se trabaja sobre de Jupyter
Actualmente se está trabajando en la ejecución de los escenarios para FaaS con el objeto de poder realizar el análisis de performance de ambas soluciones. También se pretende lograr recabar información sobre los aspectos positivos y negativos del uso de FaaS para resolver problemas de HPC, así como la viabilidad de esta implementación.

6 Conclusiones y Futuros trabajos

Aun no se cuenta con información acabada sobre el comportamiento de Serverless para resolver problemas HPC. Solamente se han realizado las pruebas sobre una infraestructura tradicional. Las configuraciones de FaaS ya están listas, pero aún falta la ejecución de los escenarios con lo que es aun imposible realizar el análisis de performance que se fijó como objetivo.

Referencias

[1] Farhan, L., Kharel, R., Kaiwartya, O., Quiroz-Castellanos, M., Alissa, A., & Abdulsalam, M. (2018, July). A concise review on Internet of Things (IoT)-problems, challenges and opportunities. In 2018 11Th International Symposium On Communication Systems, Networks & Digital Signal Processing (CSNDSP) (pp. 1-6). IEEE.

[2] Medel, D., Murazzo, M. A., Molina, A. L., Sánchez, F., Cornejo, M., Rodríguez, N. R., ... & Piccoli, M. F. (2019). La Computación de Alta Performance como soporte a los sistemas altamente distribuidos. In XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan).

[3] Barrionuevo, M., Escalante, J., Lopresti, M., Lucero, M., Miranda, N. C., Murazzo, M. A., & Piccoli, M. F. (2020). Solución de grandes problemas aplicando HPC multitecnología. In XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz).

[4] de Souza Cimino, L., de Resende, J. E. E., Silva, L. H. M., Rocha, S. Q. S., de Oliveira Correia, M., Monteiro, G. S., ... & de Castro Lima, J. (2017, November). IoT and HPC integration: revision and perspectives. In 2017 VII Brazilian Symposium on Computing Systems Engineering (SBESC) (pp. 132-139). IEEE.

[5] Biswas, A. R., & Giaffreda, R. (2014, March). IoT and cloud convergence: Opportunities and challenges. In 2014 IEEE World Forum on Internet of Things (WF-IoT) (pp. 375-376). IEEE.

[6] Añel, J. A., Añel, J. A., Montes, D. P., Iglesias, J. R., & Romano. (2020). Cloud and Serverless Computing for Scientists. Springer International Publishing.

[7] Malla, S., & Christensen, K. (2020). HPC in the cloud: Performance comparison of function as a service (FaaS) vs infrastructure as a service (IaaS). Internet Technology Letters, 3(1), e137.

[8] Baldini, I., Castro, P., Chang, K., Cheng, P., Fink, S., Ishakian, V., ... & Suter, P. (2017). Serverless computing: Current trends and open problems. In Research advances in cloud computing (pp. 1-20). Springer, Singapore.

[9] Eismann, S., Scheuner, J., Van Eyk, E., Schwinger, M., Grohmann, J., Herbst, N., ... & Iosup, A. (2020). A review of serverless use cases and their characteristics. arXiv preprint arXiv:2008.11110.

[10] Rodríguez, N. R., Murazzo, M. A., Medel, D., Parra, L., Molina, A. L., Sánchez, F., ... & Vargas, L. (2021). Procesamiento paralelo sobre arquitecturas serverless para tratamiento de datos provenientes del IoT. In XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja).

[11] Niu, X., Kumanov, D., Hung, L. H., Lloyd, W., & Yeung, K. Y. (2019, September). Leveraging serverless computing to improve performance for sequence comparison.

In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (pp. 683-687).
[12] Spillner, J., Mateos, C., & Monge, D. A. (2017, September). Faaster, better, cheaper: The prospect of serverless scientific computing and hpc. In Latin American High Performance Computing Conference (pp. 154-168). Springer, Cham.
[13] Chard, R., Skluzacek, T. J., Li, Z., Babuji, Y., Woodard, A., Blaiszik, B., ... & Chard, K. (2019). Serverless supercomputing: High performance function as a service for science. arXiv preprint arXiv:1908.04907.

XXI Workshop Tecnología Informática Aplicada en Educación **(WTIAE)**

Coordinadores

Sonia Rueda (UNS)

Verónica Artola (UNLP)

Claudia Russo (UNNOBA)

Estrategias de comunicación en escenarios educativos híbridos: implementación y mejoras al sistema de notificaciones push de la aplicación de Moodle para AulasWebColegios de la UNLP

Alejandro Héctor Gonzalez[1], Lucas Ungaro[2], Leandro Matías Romanut[2]

[1]III-LIDI Instituto de Investigación en Informática LIDI - Facultad de Informática - Universidad Nacional de la Plata
[2]Dirección General de Educación a Distancia y Tecnologías - Universidad Nacional de la Plata

agonzalez@lidi.info.unlp.edu.ar
{leandro.romanut, lucas.ungaro}@presi.unlp.edu.ar

Abstract. El trabajo presenta un análisis del módulo de notificaciones de la aplicación de Moodle para dispositivos móviles y la implementación de un sistema de notificaciones que busca mejorar y agilizar la comunicación docente-estudiante, a través del uso de diversos dispositivos, dentro del entorno virtual de enseñanza y aprendizaje AulasWebColegios de la Universidad Nacional de La Plata (UNLP). Se desarrolla un análisis previo de las estrategias de comunicación utilizadas en los colegios de manera de poder determinar las necesidades existentes en cada establecimiento y propender a agilizar las comunicaciones entre docentes y estudiantes y estudiantes entre sí. También se describe la forma de implementar sistema de notificaciones push compatible con la app móvil de Moodle y las mejoras realizadas a las funcionalidades de la app.

Keywords: app, dispositivo móvil, moodle, comunicación

1 Introducción

La Universidad Nacional de La Plata (UNLP) desarrolla formalmente propuestas de Educación a Distancia desde la década de los 90. En el año 2002 se comenzaron a desarrollar los primeros cursos en opción pedagógica a distancia utilizando un desarrollo propio denominado WebLIDI [1], que luego se transformó en WebINFO y posteriormente en WebUNLP.

En el año 2008 comienza a utilizarse a nivel de presidencia de la universidad, el entorno virtual de enseñanza y aprendizaje (EVEA) Moodle bajo la denominación de AVA (Ambiente Virtual de Aprendizaje) [2]. Este nuevo EVEA logra rápida aceptación y surge a demanda de poder lograr mayor personalización de las aulas virtuales; si bien el EVEA que teníamos previamente resulta sencillo para los que se

inician en el armado de aulas virtuales, al tiempo surge la necesidad de incorporar nuevas herramientas y de realizar una personalización más específica de los espacios, situación que no es posible de resolver en tiempo y forma ante la demanda de nuevos servicios.

Desde el año 2017 se implementa el Sistema Institucional de Educación a Distancia de la UNLP (SIED-UNLP), como el conjunto de acciones, normas, procesos, equipamiento, recursos humanos y didácticos que permiten el desarrollo de propuestas a distancia. En la actualidad la Dirección General de Educación a Distancia y Tecnologías (DGEaDyT) administra el SIED-UNLP y varios entornos virtuales para alojar las diversas propuestas educativas que emergen de los distintos niveles (pregrado, grado, posgrado, extensión, capacitación) [3]. Para la gestión de sus aulas virtuales, la DGEaDyT, adopta la plataforma e-learning Moodle a partir de 2014 bajo la denominación de AulasWeb. Esta Dirección administra los siguientes entornos basados en una personalización de Moodle [4]: AulasWebGrado, AulasWebPosgrado, AulasWebFormacion, AulasWebColegios, AulasWebOficios, AulaCavila y Cursos Externos. Pueden ser visitados desde: http://www.entornosvirtuales.unlp.edu.ar/entornos%20EAD.html

En las aulas virtuales de los EVEA de la DGEADyT se ofrecen distintas herramientas de comunicación entre docentes y alumnos, como ser foros de debate, chat, correo electrónico interno, videoconferencia, entre otras.

Los Colegios de la UNLP utilizan aulas virtuales y durante la pandemia del COVID-19, en el año 2020, se produjo un uso masivo de los EVEA que llevó a tener un espacio propio en Moodle bajo la versión 3.9.2, personalizado y denominado AulasWebColegios. A partir del contexto pandémico, se registró un aumento en la creación y uso de aulas virtuales dentro del SIED de la UNLP, creciendo la cantidad de usuarios en los entornos. En particular al revisar las formas de acceder al entorno AulasWebColegios, se registró una creciente demanda de acceso a la plataforma vía dispositivo móvil por parte de los estudiantes que utilizan frecuentemente el celular para acceder al recorrido de las aulas virtuales. Entre las consultas que se reciben habitualmente en el correo electrónico del Webmaster de AulasWebColegios se hace mención a qué formas de comunicación hay dentro de un curso y en particular con los docentes. Se nota también en términos de peticiones a nivel servidor que hay una fluída comunicación dentro de los diferentes cursos. Al finalizar cada semestre se realiza desde el SIED de la UNLP un sondeo mediante encuestas breves a los docentes y estudiantes de los EVEA y en el caso de AulasWebColegios se observa alguna problemática referida a los anuncios, notificaciones de tareas nuevas, de entregas, de subidas de nuevos materiales y de mensajes no leídos tanto en docentes como estudiantes.

Por este motivo surge una propuesta que tiene por objetivo agilizar la comunicación docente-estudiante y estudiantes entre sí dentro del entorno AulasWebColegios, a través de la implementación y análisis del módulo notificaciones en la aplicación Moodle para dispositivos móviles.
Este desarrollo se plantea dentro de las líneas de investigación del III LIDI en el proyecto "Metodologías, técnicas y herramientas de ingeniería de software en

escenarios híbridos. Mejora de proceso" y es parte de un trabajo de tesina de grado para la Licenciatura en Informática de la Facultad de Informática de la UNLP. [5]

Se plantean como objetivos específicos del trabajo:

- Realizar un análisis del sistema de comunicación actual en el entorno de AulasWebColegios.
- Implementar un sistema de notificaciones push compatible con la aplicación móvil de Moodle.
- Identificar funcionalidades a mejorar del módulo de comunicación.
- Desarrollar funcionalidades nuevas en el módulo notificaciones dentro de la app.

2 Marco de referencia para el trabajo

Como temas que estructuran el marco teórico del trabajo se investigan temas referidos a la comunicación educativa, entornos virtuales, desarrollo de apps para Moodle y cómo se insertan estos desarrollos dentro de los escenarios híbridos.

La incorporación de tecnologías en los procesos escolares requiere de planificación y trabajo conjunto de directivos, docentes y estudiantes. Eva Da Porta indica que si: *"...incorporar tecnología digital en las escuelas hace visible un conjunto de problemas educativos que ponen de manifiesto cierto desfasaje de la institución escolar respecto de los modos en que nuestra sociedad se comunica, de los modos en que se produce el conocimiento, las formas en que se desarrollan las relaciones sociales y los procesos de subjetivación contemporáneos."* Esta forma de planificar permite definir etapas de planificación con tecnología en la escuela, donde participen se involucren los directivos, docentes y estudiantes. [6]

Desde el año 2020 se comenzó a utilizar con mayor énfasis la palabra "híbrido" como respuesta a escenarios de enseñanza que intercalan presencialidad y virtualidad. Según Andreoli este concepto no es nuevo en educación y hace referencia a otros términos que describen situaciones similares como como blended learning o aprendizaje combinado. [7]

El b-learning procura tomar en el contexto de aplicación las mejores prácticas de la educación presencial y las mejores prácticas de educación a distancia y combinarlas. Entre esas prácticas podemos mencionar el uso del aula virtual como espacio de comunicación y desarrollo de tareas para los tiempos asincrónicos. Estas aulas son construidas a través de diversos EVEA. Los EVEA son un mecanismo de comunicación en los procesos educativos. La incorporación de esta tecnología digital en las escuelas es conflictiva porque requiere de un proceso de hibridación donde se mezclan matrices culturales distintas y se requiere de un apoyo institucional para llevarlo adelante.

Un estudio realizado en 2016 por Arancibia se basó en la aplicación de un cuestionario de manera presencial en la Universidad Tecnológica de Chile, incluyó 5.234 docentes y 123.047 estudiantes chilenos, indica que: *"los resultados sobre el tipo de entorno o ambiente de aprendizaje que estudiantes y docentes prefieren para aprender arrojaron que el 5% de los estudiantes elige estudiar en línea, frente al 3% de los docentes; el 47% de los estudiantes opta por clases con algunos componentes en línea, frente al 44% de sus docentes; y el 15% de los estudiantes que se inclina por clases sin componentes presenciales, frente al 23% de los docentes."* [8]. En este estudio los docentes, a diferencia de los estudiantes, reportan valoraciones más favorables sobre la importancia de las tecnologías para el éxito académico. Esto hace notar que los docentes no son, necesariamente, menos expertos en el uso de la tecnología que los estudiantes y que ellos tienen actitudes favorables para la incorporación de tecnología digital en el proceso educativo.

Este conjunto de problemas educativos no es ajeno a la incorporación de entornos virtuales en la UNLP y en particular en los colegios de la universidad. El desarrollo histórico de los EVEA puede entenderse como un proceso lento pero paulatino apuntando a la mediación con tecnologías digitales. En la UNLP se pasó de un modelo de desarrollo propio de EVEA a la utilización y personalización de Moodle [4]. Por ejemplo, la utilización de personajes virtuales para acompañar las decisiones de los docentes en el selector de actividades de Moodle. Se creó un personaje virtual a través del sitio "Pocoyó" mediante su aplicación Pocoyize, la cual permite crear avatares de caricaturas y descargarlos. La incorporación de un personaje intenta generar una estrategia que oriente a los docentes en la comprensión de las actividades de trabajo colaborativo. [9]

En otros estudios como el de Herrera Cantillo se desarrolla una propuesta para optimizar los procesos comunicativos y de formación de los estudiantes de ciencias básicas en modalidad virtual de la Corporación Universitaria Americana, a través del uso de las aplicaciones Moodle Mobile y herramientas de mensajería instantánea [10]. Implementa la app Moodle Mobile y se focaliza en el uso de herramientas de mensajería instantánea como WhatsApp y Remaind. En la experiencia participaron 15 personas en un estudio focal de tipo cualitativo con encuestas y entrevistas. Destaca que los estudiantes y maestros coinciden en que las apps móviles permiten acortar los tiempos de respuesta, y en consecuencia los procesos de aprendizaje se optimizan. Es importante el uso de la mensajería instantánea del tipo Whatsapp y promover ambientes colaborativos para el desarrollo de actividades de aprendizaje. Indica la importancia de una capacitación en el uso adecuado de las herramientas de mensajería instantánea con fines educativos y establecer a si mismo uso de etiquetas con reglas claras que regulen la utilización.

El desarrollo en Moodle permite ampliar la funcionalidad y compartir la misma a la comunidad de software libre para que otros puedan acceder y utilizar los desarrollos. Tal es el caso del App de Moodle Mobile que permite mantenerse actualizada respecto a todo lo que está sucediendo en sus cursos y en el sitio. Cada vez que se abre la App, los eventos son sincronizados con el sitio web. Moodle viene con un servicio web integrado diseñado para aplicaciones móviles que permite

administrar las notificaciones push. Esta tecnología es de interés para poder probar en el espacio de las escuelas a través de AulasWebColegios, teniendo en cuenta el contexto y la edad de los participantes y su cercanía al uso de los dispositivos móviles, de manera de mantenerlos actualizados en las actividades escolares.

3 Análisis del caso a trabajar

A partir de la pandemia, la DGEaDyT registró un aumento en la creación y uso de aulas virtuales, creciendo la cantidad de usuarios en los entornos. Principalmente en uno de los entornos administrados: AulasWebColegios. Este contiene aulas del Bachillerato de Bellas Artes "Francisco Américo de Santo", Colegio Nacional "Rafael Hernandez", Escuela Agraria "M.C y M.L Inchausti", Escuela Graduada "Joaquín V. González" y el Liceo "Víctor Mercante". [11]

En el caso de los colegios de la UNLP al revisar las formas de acceder al entorno Moodle, se registró una creciente demanda de acceso a la plataforma vía dispositivo móvil, los estudiantes utilizan frecuentemente el celular para acceder al recorrido de las aulas virtuales de los entornos de la UNLP [11]. Existe una aplicación para celular de Moodle, pero la misma cuenta con algunas limitaciones como:

- No tiene logotipo o colores propios (Full app branding)
- No existe un servidor propio separado de Moodle para el envío de las notificaciones (Separate notifications server)
- Existen pocas características en la app que pueden ser configuradas (Customisable app features)

En términos específicos, hoy en día los estudiantes necesitan entrar al entorno virtual, localizar el aula virtual correspondiente y allí buscar la notificación de la actividad/recurso particular. Contar con un sistema de notificaciones push, aportaría comodidad y seguridad a los estudiantes a la hora de estar al tanto, en tiempo real, de las novedades que surgen en los cursos. Por lo cual dado el uso intensivo de celulares por parte de los estudiantes se hace necesario mejorar la comunicación docentes-estudiantes a través de la aplicación móvil de Moodle.

Entre las hipótesis de trabajo se busca implementar un sistema de notificaciones móviles y nos preguntamos si ¿puede aportar a la mejora de la comunicación en la app y lograr una mejor interacción de los estudiantes con el aula virtual?

4 Desarrollo propuesto

Para este trabajo de tesina de grado se propone implementar y agregar funcionalidades al módulo notificaciones de la aplicación Moodle, vinculada con un sistema ya establecido, en este caso AulasWebColegios. Se trabajará buscando la compatibilidad con el mismo.

Se realizará un análisis de la comunicación actual en AulasWebColegios a través de encuestas a estudiantes y docentes de manera de poder establecer los inconvenientes actuales y generar un informe de consulta sobre el estado actual de las comunicaciones entre los participantes y buscar de manera articulada ampliar las mejoras a las funcionalidades.

4.1 Análisis de las encuestas previas

El objetivo principal de las encuestas previas es analizar la necesidad, por parte de los usuarios, de agregar el servicio de notificaciones push a la aplicación para celular de AulasWebColegios. Se busca tener conocimiento en el dominio del problema que permita tener la perspectiva de los usuarios sobre qué canales de comunicación utilizan, poder analizar las opiniones de los usuarios en cuanto a ventajas y desventajas a la hora de notificar, de manera de poder detectar diferencias y similitudes de uso entre docentes y estudiantes. Se realizaron dos formatos de encuestas, uno para docentes y otro para estudiantes de 12 preguntas cerradas y abiertas.

En la encuesta previa participaron 271 estudiantes y 53 docentes de los diferentes establecimientos.

Con referencia a la perspectiva sobre los medios/canales de comunicación actuales los estudiantes dicen no estar conformes con la eficiencia en los medios/canales actuales de comunicación. Más de la mitad considera difícil estar al tanto de las novedades del curso y casi el 80% alguna vez se perdió la posibilidad de completar una tarea por no haber ingresado a tiempo al entorno web. En cuanto a los docentes, el 75.5% no quiere más vías de comunicación para notificar, considera que con los medios actuales de comunicación es suficiente, siendo este el planteo que más se repite en las respuestas abiertas. Los docentes que si desean más vías de comunicación piensan en el beneficio para el estudiante y en las ventajas de notificar al celular.

De los estudiantes encuestados (67.2%) se entera de las novedades dentro del aula virtual a través de Whatsapp. En porcentajes muy similares, se enteran a través de correo electrónico, redes sociales, comunicación presencial o ingresando a la plataforma web. Solo el 9.6% ingresa a la aplicación para celular Moodle.

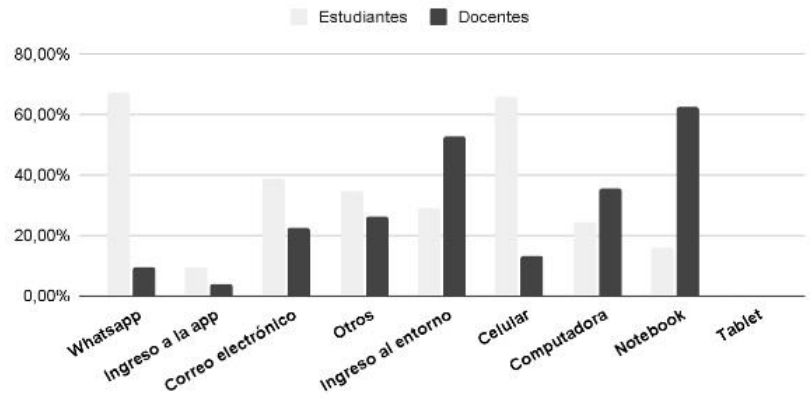


Fig1. Medios/canales de comunicación utilizados para acceder a los EVEA

La mitad de los estudiantes encuestados no conoce de la existencia de la app Moodle. Igualmente, el 65.7% de los estudiantes prioriza el celular para ingresar al aula virtual como se observa en al Fig 1. El resto ingresa por computadora de escritorio y/o notebook.

El canal más utilizado por los docentes para el envío de novedades es la plataforma web. En menor medida se utilizan las redes sociales, comunicación presencial o correo electrónico. Se advierte la similitud entre el porcentaje de los docentes que conocen la app Moodle (34%) y el porcentaje de docentes a favor de la implementación de notificaciones push (28.3%). Solo un 13% de los docentes ingresa a la plataforma a través de un dispositivo móvil, el resto lo hace con notebook o computadora de escritorio.

Se puede observar que la mayoría de los estudiantes perciben un problema en la forma y utilidad de las comunicaciones, a pesar de contar con varias opciones, y creen que se debería –de alguna forma- unificar/optimizar el proceso.

La opción más rápida sería concentrarse en el canal más utilizado, que es Whatsapp. Sin embargo, unificar las comunicaciones en este canal lograría el efecto contrario al buscado, dado que el Whatsapp es un recurso utilizado para varias actividades en diversos contextos. En la UNLP se busca tener canales de comunicación oficiales para dar contexto del proceso educativo. Por este motivo se decide avanzar en tener esa unificación a través de incorporar el uso de la aplicación Moodle con notificaciones push.

4.2 Implementación y mejoras

Según la documentación oficial de Moodle, si el administrador de un EVEA desea implementar un sistema de notificaciones móvil dentro del entorno debe seguir alguna de estas dos opciones [12]:

- Registrar el sitio, esto implica compartir información del EVEA con la empresa Moodle y estar sujeto a las restricciones de los diferentes planes para notificaciones propios.
- Instalar una infraestructura de notificaciones, pero utilizando las herramientas que sean compatibles con Moodle. Usando AirNotifier y una versión personalizada de la aplicación móvil de Moodle.

A fin de lograr el envío de notificaciones desde el sistema web de Moodle a los dispositivos móviles se pone en funcionamiento un servidor de aplicaciones, en este caso AirNotifier, que permite el envío de notificaciones en tiempo real. Además, se modifica la app de Moodle para su correcta comunicación con el servicio de notificaciones y se hará la conexión de este último con el entorno virtual de AulasWebColegios.

A partir del envío efectivo de las notificaciones, se desarrollan dos funcionalidades nuevas dentro del módulo notificaciones de la app. Lo que se busca es ir optimizando este módulo priorizando la usabilidad y la comunicación eficaz docente-estudiante.

Las nuevas funcionalidades desarrolladas son:

- Borrar de notificaciones.
- Destacar las notificaciones por curso.

5 Primeros resultados

Se logró la implementación y mejora del sistema de comunicación. Con referencia al envío de notificaciones a los dispositivos móviles se realizaron las siguientes tareas:

- Puesta en funcionamiento del servidor AirNotifier.
- Configuración del entorno virtual AulasWebColegios.
- Desarrollo mediante Firebase Cloud Messaging (FCM), la conexión entre la app Moodle y el servidor AirNotifier.

Se implementó la funcionalidad de borrar notificaciones se obtuvo:

- Correcta ejecución de la funcionalidad en el front-end (aplicación móvil).
- Correcto funcionamiento del nuevo servicio externo en el back-end (sistema web AulasWebColegios). El servicio atiende la solicitud del front-end y solicita el borrado a la base de datos.
- Verificación del borrado de las notificaciones en la base de datos.

Se desarrolló la función de destacar notificaciones por curso:

- Al igual que con el borrado de notificaciones, ver fig 2,. se verificó el correcto funcionamiento en el front-end, back-end y en la base de datos.

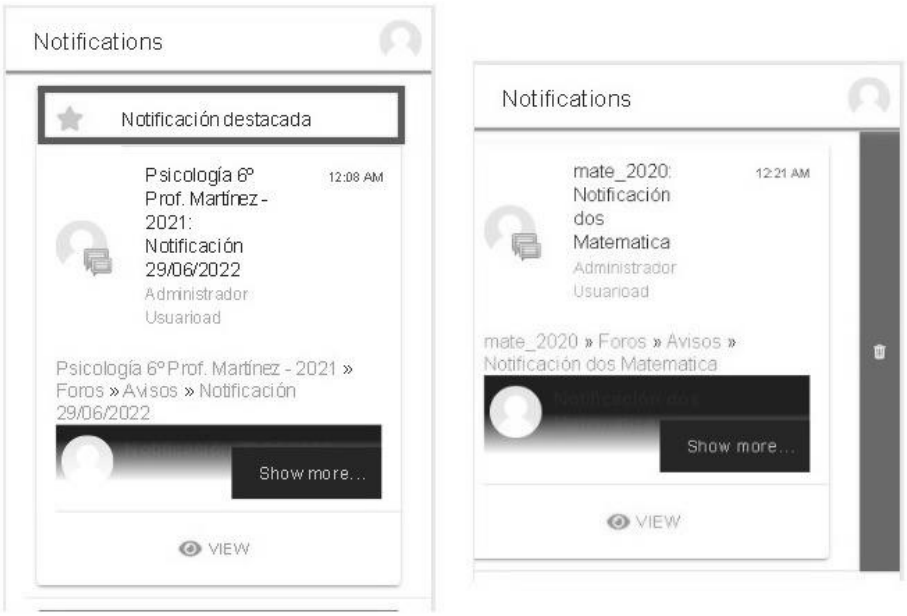


Fig 2. Implementación del destacado y del agregado del botón de trash para el borrado de notificaciones.

Se trabaja actualmente en la sensibilización entre estudiantes y docentes, del uso de la app de Moodle y ver el real alcance de la unificación de mensajes propuesta en este trabajo.

5 Conclusiones y trabajo futuro

Se trabaja en la puesta en producción de la implementación y mejora realizadas. Se realizará un sondeo luego del primer mes de uso de las nuevas funcionalidades.

Este trabajo de tesina nos permite reflexionar sobre el cambio en las dinámicas de comunicación a través de los entornos, en relación a lo ocurrido en la pandemia y a la pos pandemia del COVID19. Se pueden detectar nuevas formas y usos de los dispositivos y en particular los móviles. Nos permite analizar cómo se introduce el dispositivo móvil como medio de acceso a los EVEA , de una manera más sostenida, no sólo por el rango etario de los usuarios (al ser estudiantes de hasta 18 años de edad) sino también debido a que, en muchos casos, era la única herramienta que se disponía para interactuar durante la virtualidad así como en estos nuevos escenarios que se plantean después del retorno a la presencialidad.

Viendo este nuevo contexto se proyecta analizar otros EVEA de la UNLP que están requiriendo sistemas organizados de comunicación educativa como AulasWebOficios donde los participantes son de diferentes edades, lugares de residencia y diferentes niveles de estudio. En este espacio se desarrollan los cursos de educación alternativa para la comunidad, donde hay cursos a distancia de la Escuela de Oficios de la UNLP

Referencias

1. Sanz, C. V., Zangara, M. A., González, A. H., Ibañez, E., & De Giusti, A. E. (2003). WebLIDI: Desarrollo de un Entorno de Aprendizaje en la WEB. In V Workshop de Investigadores en Ciencias de la Computación.
2. Principe, A. G., Russo, C., Zangara, A., Esnaola, F., & Salvioli, A. P. S. (2009). La internacionalización de la educación a distancia: estrategias de abordaje. Presentación del Proyecto Aula Cavila UNLP. RIED. Revista iberoamericana de educación a distancia, 12(1), 95-111.
3. Gonzalez, A. H., & Martín, M. M. "Educación superior a distancia en Argentina: tensiones y oportunidades". Trayectorias universitarias, vol. 3, 2017.
4. Gonzalez A. H. "Tendencias en el desarrollo de plataformas educativas en el ambiente universitario. Presentación de casos" en LAS PLATAFORMAS VIRTUALES EN LA EDUCACIÓN SUPERIOR. Conferencias y comunicaciones de la Jornada de Plataformas Educativas en el Nivel Superior (JoPlEd). San Justo, 2019. Editorial UNLaM. Pags. 15 - 34. ISBN 978-987-4417-35-0
5. Sanz, C. V., Madoz, M. C., Gorga, G. M., González, A. H., Zangara, M. A., Iglesias, L., & Pesado, P. M. "Diseño y desarrollo de herramientas y entornos digitales para escenarios educativos híbridos". En XXIII Workshop de Investigadores en Ciencias de la Computación (WICC), Chilecito, La Rioja, 2021.
6. Da Porta, E. "El acceso a las tecnologías de la comunicación. Debates y perspectivas en América Latina" en Las significaciones de las TIC en educación. Córdoba, 2015. Ferreyra Editor.
7. Andreoli, S. "Modelos híbridos en escenarios educativos en transición". Citep. 2021.
8. Arancibia Muñoz, M. L.; Cabero Almenara, J., & Valdivia Zamorano, I. "Estudio comparativo entre docentes y estudiantes sobre aceptación y uso de tecnologías con fines educativos en el contexto chileno". Apertura, 2019. Guadalajara, Jal. 11(1), 104-119. Disponible: http://www.scielo.org.mx/pdf/apertura/v11n1/2007-1094-apertura-11-01-104.pdf
9. Romanut, L. M., González, A. H., & Madoz, M. C. "Asistente virtual para la utilización de herramientas de trabajo colaborativo en entornos educativos en línea". En XI Congreso de Tecnología en Educación y Educación en Tecnología (TE&ET). 2016.
10. Herrera Cantillo, C. P. "Optimización de los procesos comunicativos y de formación a través del uso de aplicaciones Moodle Mobile y mensajería instantánea en la formación virtual en la Corporación Universitaria Americana". Tesis de Magister, Universidad del Norte, 2018.
11. Jaime, C. J., González, A. H., & Barletta, C. M. (2021). Informe técnico. Dirección General de Educación a Distancia y Tecnologías.
12. Documentación oficial de Moodle. "Mobile app notificaciones". https://docs.moodle.org/all/es/Mobile_app_notificaciones (accedido 28 de Junio 2022)

Hope Project: Development of mobile applications with augmented reality to teach dance to children with ASD.

Mónica. R. Romero[1], Ivana Harari[1], Javier Diaz[1], Estela Macas[2]

[1]National University of La Plata, Faculty of Informatics, Research Laboratory in New Information Technologies (LINTI).
Calle 50 y 120, 1900 La Plata. Buenos Aires.

[2] International Ibero-American University - UNINI MX. Mexico

[1]monica.romerop@info.unlp.edu.ar, [1]iharari@info.unlp.edu.ar, [1]jdiaz@info.unlp.edu.ar, [2]estela.macas@doctorado.unini.edu.mxc

Orcid ID:https://orcid.org/0000-0002-6099-7039
Orcid ID: https://orcid.org/0000-0001-6350-7739
Orcid ID:https://orcid.org/0000-0002-4225-3829
Orcid ID:https://orcid.org/0000-0002-1237-1154

Abstract. New ICT information and communication technologies, and specifically augmented reality (AR), are providing educators with effective strategies that allow them to be more effective in teaching-learning processes in children with ASD. This research aimed to develop an intervention plan to improve certain skills in children with ASD. Through a qualitative-quantitative, descriptive, experimental study, the Hoope software used as an innovative resource. At the end of this research, we can conclude that through augmented reality (AR) we reinforce and innovate certain teaching-learning processes, showing favorable results that show that through an intervention plan through NICT there may be positive impacts on children with ASD.

Keywords: ASD, *Teaching, Hoope, Augmented Reality, NTIC*.

1 Introduction

One of the great challenges facing developing countries is the search for equity to educate in diversity[1], currently there are a large number of children who are diagnosed with developmental disorders, these children generally need particular attention and the implementation of strategies to improve the education they receive[4]–[5], since it is directly proportional to benefit their environment and therefore the quality of life[6].

Autism spectrum disorder, henceforth ASD, can be defined as a complex neurodevelopmental disorder[8]–[10], which is detected in the early years and lasts a lifetime [12]. The NICT, considering it as a facilitator of the decoding of information,

is logical, concrete, located in a space, not the verbal language that is invisible, temporary and abstract , [17], [18. Thus, the research first proposes an intervention plan that uses NTIC that can be used by educators, psych pedagogues, therapists, parents who work daily with ASD children.[15], [16], making use of augmented reality[17], for this purpose the software called Hope (Hoope) is used, in order that children can develop certain skills and abilities.

The study structured as follows: Section 2 explains the methodology for the investigation. Section 3 presents the results of the study, Section 4 presents a reflection of the results found in the application of the Hoope software in the experimental study, Section 5 presents the conclusions, recommendations, limitations, and future lines of research.

2 Material and method

The research addresses a mixed approach supported by the qualitative and quantitative method, additionally the study is of the type: descriptive, exploratory, because it seeks to know in a detailed way the relationship between pedagogical practice through technological innovation mediated using new emerging technologies and the benefits of the application of Hoope software in the teaching-learning processes of children with ASD.

The modality used is documentary and field research[18-19]. This is because the experimentation is conducted on a specific software called Hoope created in the Laboratory for Research of New Computer Technologies LINTI, of the National University of La Plata in Argentina, and documentary because the process and the results are supported in a methodology and in the theoretical support of previously conducted research[20-21].

The field work conducted in Ecuador, specifically in the city of Quito at the Ludic Place Therapeutic Center, which welcomes ASD children offers support for the prevention and addressing of specific learning needs associated with the presence and risk of spectrum disorders. In this area, various methodologies, programs, techniques, and instruments used to be able to support the children who attend consultations. The procedure for data collection will be through scheduled sessions where a multidisciplinary team intervenes through an interview, deep observation.

2.1 Population

The Director of the Ludic Place Center, destined for three professionals (teacher, psychologist, and psych pedagogue) participated in this process. These professionals were receptive to new and innovative strategies that include the use of new information and communication technologies. Additionally, parents, supported the proposal, and signed the informed consent for their children to participate in the intervention.

Children with ASD, Matias, who from now on will be identified with the letter M, is 4 years old and has high-functioning ASD; while Eidan, who from now on will be identified with the letter E, is 5 years old and has medium-functioning ASD; have been evaluated and diagnosed in the Ecuadorian Institute of Social Security of Ecuador (IESS).

2.3 Work plan.

The work plan was developed for a period of six months, from February to June 2021, where several activities were planned: the conceptualization of the project, the bibliographic review, the viability of the Project, validation of the current situation of infants, contextualization, needs analysis, development of the intervention plan: diagnostic phase, intervention phase, evaluation phase. For the intervention plan, the sessions were designed to work for 20 to 25 minutes, twice a week, for a period of several months.

2.4 Phases of the intervention

Phase I: Diagnostic or initial evaluation. Diagnostic and detailed evaluation of the student, before starting the intervention we conduct a complete and in-depth evaluation of M and E, to approach the intervention process in an individualized way. It is necessary to emphasize that the center has the diagnosis of children.

Phase II: During the Intervention. For the intervention phase, strategies were proposed to conduct a playful activity mediated through technology using the Hoope system. This system allowed the child to interact alone or with the help of the professional who guides the session. The activities that were conducted have a defined order, each session seeks a purpose and previously some aspects considered essential have been considered, such as the organization of spaces, the time of the sessions, the necessary materials, and the collaboration of the center team is counted on therapeutic and with parents.

Phase III - Final Evaluation - psych pedagogical. The purpose is to contrast the results obtained in the diagnostic evaluation with those that will be obtained after the process. Using the interviews, it is possible to obtain the necessary information to capture the results of the intervention of children with ASD with the Hoope Software.

2.5 Resources used in the intervention plan

To conduct this research, some resources were used, which are indicated below.

Human Resources: Multidisciplinary group made up of teachers, psychologists, educational psychologists, doctoral students, systems engineers, parents, and children with ASD.

ICT Resource: In relation to technological resources, the Hoope System created in the Research Laboratory of New Computer Technologies LINTI of the National

University of La Plata - Argentina was used, a Kinect device, and a laptop. The Hoope System is a system that is based on augmented reality, it is focused on children with ASD from 3 to 14 years old. This software allows the participant to choose options that allow reinforcing teaching-learning areas. Next, Figure 1 shows the resources used for the process.

Fig. 1. Main menu of the software Hoope main menu. Capture made of the software used for pedagogical intervention.

Fig.2. Resources used in pedagogical intervention

2.5 Activities designed to reinforce teaching-learning processes.

Next, the activities planned for the teaching-learning processes presented, for the intervention plan, we choose to reinforce several processes perception, imitation, fine motor skills, gross motor skills, visual motor skills, the same ones that are presented below in Table 4 planning temporary activities.

Table 1. Curricular content planning intervention project

Area: Education	Directed to: Children with autistic disorder ASD Time: 25 minutes per session
Theme:	Learn by dancing-Playful Activity. Use of Hoope system
	Objectives of the intervention plan
Perception:	Recognition, awareness and playful experimentation of the body. Visual perception (fundamental to the basis of cognitive processing and reasoning), is the ability to recognize and interpret different visual material correctly and transform this information into an adapted motor response. Therefore, it is an important skill, indispensable for school success.
Imitation.	Recognition, awareness and playful experimentation of the body as an expressive medium with the elements that make up the language of dance, space, time and energy.
Fine motor	Recognition, awareness, and playful experimentation of the body.
Gross Motricity	Recognition, awareness, and playful experimentation of the body.
Visio coordination Driving	Recognition, awareness and playful experimentation of the body.
	Name of the activity
Contact points/ touch point	They are interactive zones that appear randomly around the upper part of the user and are activated by being touched with the hands.
Kick points / kick points	They are interactive zones that appear randomly around the bottom of the user and are activated when touched with the feet.
Route tracking / tracking match:	They are a set of strokes that appear randomly around the top of the user, they are activated by touching the starting point and dragging the hand along the entire path to the end point.
Avatar pose / match poses:	They are a set of poses that appear randomly at each end of the user and are activated when he manages to imitate the pose by more than 80%.
Mix of poses and exercises	They are a set of poses that appear randomly
Skills with performance criteria	Learning Activities: perception, imitation, fine motor skills, gross motor skills

Shown in Figure 4 the activities proposed in the Software called Hoope to work the space of perception, imitation, fine and gross motor skills and visual motor coordination in children with ASD.

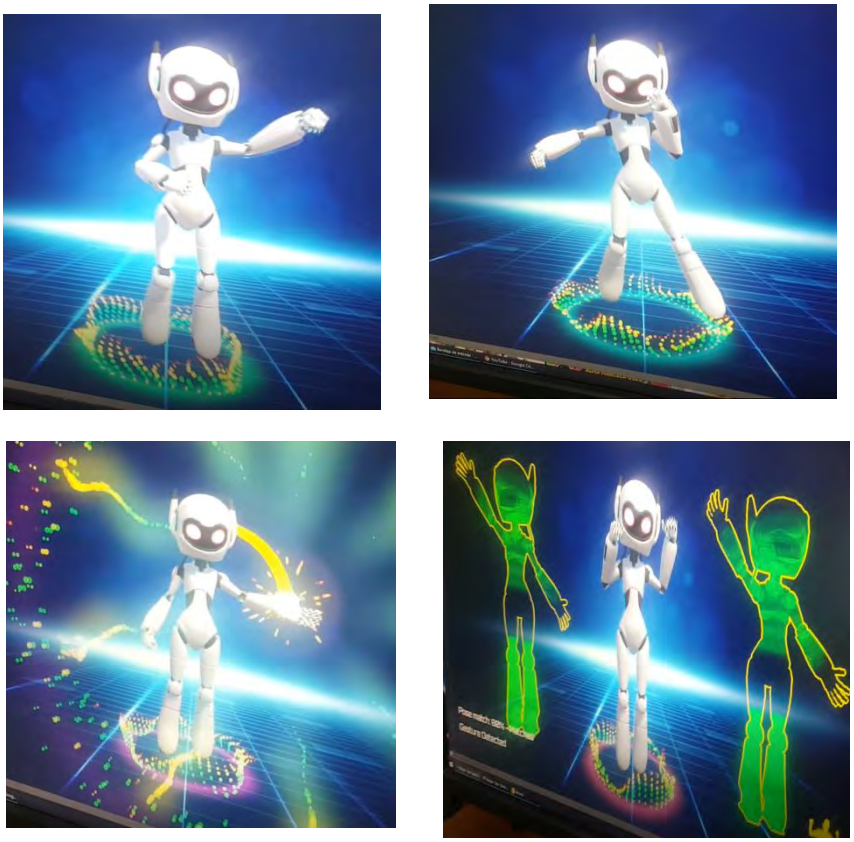


Fig. 3. Activity proposed to work imitation, perception, fine and gross motor skills and visual motor coordination. Software Hoope - playful game for children with ASD.

3 Results

Once the application of the Intervention Project is concluded, the purpose is to contrast the results obtained in the diagnostic evaluation with the results after the intervention process using the Hoope application. We focus on determining if the intervention plan generated favorable results and if there is evidence of any progress in the teaching-learning processes of children M and E.

For the evaluation of the proposed curricular activities, a scale of three possible options is used. The multidisciplinary team that accompanied the development of the pedagogical intervention plan was asked, if the child with ASD is this M or E carried out the proposed activity, it is classified as passed and it is scored as 3, if the child tries to carry out the activity it is determined that the activity is emergent and is scored with

2, if on the contrary the child fails in the process, the activity is scored with 1.The table of results of the proposed curricular activities is shown below.

Table 2. Results of the proposed curricular activities processes before and after AR

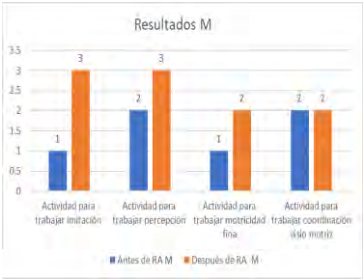
Results of the proposed curricular activities post-AR use				
Actions	**Before RA**		**After RA**	
Child with ASD	**M**	**AND**	**M**	**AND**
Activity to work imitation	1	1	3	3
Activity to work perception	2	2	3	2
Activity to work fine motor skills	1	2	2	3
Activity to work fine motor skills	1	2	2	3
Activity to work visual motor coordination	2	1	2	2

In the following image we can observe E using the Hoope Software during a scheduled session. In the Fig. 4 describes the activity proposed to work imitation: Software Hoope - playful game for children with ASD.

Fig. 4. Activity proposed to work imitation, Software Hoope - playful game for children with ASD

Next, the results presented to show the progress of each participant after complying with the proposed work schedule, through the designed phases and after the proposed sessions. Table 5 shows the comparison of M results, an analysis of the activities is conducted at the beginning or diagnostic phase and after the use of the Hope System that includes several activities. The interview conducted in the third phase of this intervention plan, it conducted in the educational center, they were informed in advance of the day and time where the meeting was to take place.

Table 3. Representative graphs of data analysis.



Results of M after the intervention with AR

Results of M after the intervention with AR

Proposed activity Interview results after the intervention process Imitation process

Perception process interview results

Results of the fine motor process interview

Gross motor process results.

Imitation activities. Question: Do you consider that the child's ability to imitate has improved after the use of augmented reality applications, specifically through the Hoope software. Analysis: When asking the multidisciplinary team (psychologist, teacher, and psych pedagogue), they totally agreed that the children reinforced the imitation process, developing the proposed exercises in an easier and more intuitive way with the application of augmented reality, using the option 1 of the proposed Hoope game.

Perception activities: Question: Do you consider that children's perception has improved after using the Hope software application that includes a natural interface

with augmented reality? Analysis: The child's perception has improved, after the use of the Hoope software, people from the multidisciplinary team indicated that they fully agree, and others agree.

Fine motor activities: Question: Do you think the child's fine motor skills have improved after using the Hoope software? Analysis: The fine motor skills of the child has improved after the use of applications with augmented reality, some people from the multidisciplinary team indicated that they were in complete agreement and others indicated that they agreed, as shown in the figure.

Gross motor activities: Question: Do you consider that gross motor skills on the part of the child have improved after the use of applications with augmented reality? Analysis: The gross motor skills of the child has improved after the use of applications with augmented reality, imitations of movements of the robot from the Hoope game, were carried out by the children during the sessions.

Visual motor coordination activities: Question: Do you think that the child's ability to associate animals with colors has improved after the use of augmented reality applications? Analysis: The visual motor coordination capacity of the children has improved after the use of applications with augmented reality, all the people on the team indicated that they were in complete agreement.

4 Discussion

It is necessary to review the fulfillment of the proposed objectives of the intervention project and determine to what extent these developed and fulfilled.

In relation to: Review and analyze updated bibliography in relation to the teaching-learning process of children with ASD that favor the approach, concretion and deepening of the proposal, it conducted considering different theories and research that exist in this regard and the subsequent selection of academic research to prepare the literature review.

As for carrying out an analysis of educational needs that allows knowing the incidence of the difficulty of certain teaching-learning processes of children with ASD who attend the Ludic Place Therapeutic Center, it was achieved by conducting interviews with the treating psychologist, Who can I obtain relevant information, taking into account that they are the ones who share directly with children with ASD and know what the needs of each one of them are.

About: Designing a plan for intervention mediated by information and communication technology, in particular augmented reality, was carried out taking into account the needs analysis, since the idea is precisely to cover the deficiencies that exist, once this aspect has been analyzed It helped a lot to take into account the general issues that were wanted to be addressed, and then to determine which were the areas that

would need to be worked to obtain the desired results and the time in which those changes are expected to be seen.

To select those activities that are the most appropriate for learning such as the processes of imitation, perception, fine and gross motor skills, and visual motor coordination, taking into consideration the context first, after that, the bibliographic review taken into consideration, to finally plan those activities that could be more suitable according to the augmented reality software application called Hoope.

5 Conclusions

The intervention plan developed a method that allowed to include emerging technologies in our case the use of Software Hoope, in it, intelligent objectives proposed in a time defined, helping the children who participated in reinforcing teaching-learning processes as perception. imitation, fine, and gross motor skills and visual motor coordination that are essential to reduce the existing gap and inequality to which they exposed daily. Regarding the work in the field, direct contact with the community established through the Ludic Place therapeutic center, where they worked with children with ASD, parents, and support professionals (psychologist, educational psychologists, teachers, information, and communication technology professionals).

These studies, which include experimentation as a fundamental basis, are important since they not only come to verify theories, concepts, and information from similar works, but also serve to develop new teaching processes, and hence the importance of being able to identify to personalize teaching. The incorporation of models, methodologies, and strategies especially with children with autism is a fundamental requirement, understanding that everyone has their own learning process. For the development of the intervention program, as well as for its monitoring, evaluation, and joint decision-making, it is necessary to work in a comprehensive and multidisciplinary way to obtain better results, given the multitude of professionals involved, the intervention approached from an interdisciplinary approach unifying goals, objectives and methodology used with the child.

Using innovative resources in the classroom, it seeks to stimulate the teaching-learning processes, it is essential to offer children with ASD during their school stage an adequate teaching-learning process that allows them to strengthen their skills, this intervention as an alternative in the educational process, to work in coordination towards joint goals and priorities (parents, professionals who accompany the child with ASD, psych pedagogues among others). New ICT information and communication technologies, and specifically augmented reality, are providing teachers with new and effective strategies that allow them to be more effective in education, generating significant interest in learning in children. The intervention proposal included two children with ASD diagnosed with moderate ASD (requires notable help) and severe ASD (they require a lot of help), however, it would be opportune to carry out the intervention in children with mild ASD, it is possible that this plan and its results are

better received, and that this intervention based on information and communication technology is of relevant help in these cases.

It is important to note that this intervention plan can be applied and reinforced if applicable, these adaptations related to content have been proposed as an orientation and exemplification, however, it will be the teacher in collaboration with his team of treating professionals, who will specify the elaboration of the individualized plan based on this intervention program, as well as on the orientations offered by the pedagogical counselor. Autism is a complex disorder that has characteristics of one child with another, therefore, the interventions must be different. Individuality is precisely the factor that should never lost sight of when planning an intervention, and what works for one case may not receive in the same way with another child; however, with the help and patience of the professionals in charge, adaptations of the plans can be made to individualize them and achieve better results.

In the case of children with ASD their development is not stable or predictable, therefore this plan must evaluate regularly, it should modify and perfected as many times as necessary. Computer applications in the field of education provide important advantages since they are means that tend to generate intrinsic motivation, being attractive and stimulating. We look at M and E, who like games as well as the music and sound effects provided by the Hoope Software, as well as animated characters. Regarding future lines of research, the application of this Hoope program would be of great interest not only in the therapeutic center but also in the home of children with ASD, or during schooling to reinforce the processes. We are grateful to the LINTI New Computer Technologies Research Laboratory of the National University of La Plata -Argentina, the National Secretary of Higher Education, Science and Technology SENESCYT- Ecuador, as well as the Ludic Place therapeutic center where this project conducted.

References

[1] I. Troya, A. D. R. Lalama, M. Pacheco, and M. Yépez, “Los retos de la docencia, frente a la educación inclusiva en el Ecuador,” *Espirales Rev. Multidiscip. Investig.*, vol. 2, no. 14, pp. 61–70, 2018, [Online]. Available: http://www.revistaespirales.com/index.php/es/article/view/190/131.

[2] A. Pantoja, “El modelo tecnológico de intervención psicopedagógica,” *REOP - Rev. Española Orientación y Psicopedag.*, vol. 13, no. 2, p. 189, 2014, doi: 10.5944/reop.vol.13.num.2.2002.11595.

[3] C. R. Tipo and C. Quir, “Propuesta de intervención psicopedagógica para el refuerzo de la lectura en el tercer año de primaria,” 2019.

[4] E. Guerrero Barona and L. J. Blanco Nieto, “Diseño de un programa psicopedagógico para la intervención en los trastornos emocionales en la enseñanza y aprendizaje de las matemáticas,” *Rev. Iberoam. Educ.*, vol. 34, no. 2, pp. 1–14, 2004, doi: 10.35362/rie3422990.

[5] M. Romero and I. Harari, “Uso de nuevas tecnologías TICS -realidad

aumentada para tratamiento de niños TEA un diagnóstico inicial," *CienciAmérica Rev. Divulg. científica la Univ. Tecnológica Indoamérica*, vol. 6, no. 1, pp. 131–137, 2017, [Online]. Available: https://dialnet.unirioja.es/descarga/articulo/6163694.pdf.

[6] F. A. Marín, Y. A. Esteban, and S. M. Iturralde, "Prevalence of autism spectrum disorders: Data review," *Siglo Cero*, vol. 47, no. 4, pp. 7–26, 2016, doi: 10.14201/scero2016474726.

[7] P. García-Primo, "La detección precoz de trastornos del espectro autista (TEA). El programa de cribado con M-CHAT en España y revisión de otros programas en Europa," p. 252, 2014.

[8] I. Málaga, R. B. Lago, A. Hedrera-Fernández, N. Álvarez-álvarez, V. A. Oreña-Ansonera, and M. Baeza-Velasco, "Prevalence of autism spectrum disorders in USA, Europe and Spain: Coincidences and discrepancies.," *Medicina (B. Aires).*, vol. 79, no. 1, pp. 4–9, 2019.

[9] C. Diagn, *American psychiatric association*, vol. 9, no. 5. 1923.

[10] J. Rodríguez Medina, "Mediacion entre iguales, competencia social y percepcion interpersonal de los ninos con TEA en el entorno escolar," 2019, doi: 10.35376/10324/39475.

[11] E. Reaño, "La Tríada de Wing y los vectores de la Electronalidad: hacia una nueva concepción sobre el Autismo," no. April 2015, pp. 0–13, 2016, [Online]. Available: https://www.researchgate.net/publication/274510152.

[12] R. Ventoso and Á. Brioso, "Ángel Rivière: La búsqueda del sentido en la clínica del autismo," *Infanc. y Aprendiz.*, vol. 30, no. 3, pp. 413–437, 2007, doi: 10.1174/021037007781787444.

[13] S. Corbellini, L. C. Real, and N. Silveira, "Intervenções Psicopedagógicas e Tecnologias Digitais na Contemporaneidade," *An. dos Work. do V Congr. Bras. Informática na Educ. (CBIE 2016)*, vol. 1, no. Cbie, p. 1394, 2016, doi: 10.5753/cbie.wcbie.2016.1394.

[14] L. Sobrado, C. Ceinos, and R. García, "Utilización de las TIC en orientación profesional: Experiencias innovadoras," *Rev. Mex. Orientación Educ.*, vol. 9, no. 23, pp. 2–10, 2012.

[15] M. Romero, E. Macas, I. Harari, and J. Diaz, "Eje integrador educativo de las TICS : Caso de Estudio Niños con trastorno del espectro autista .," *SAEI, Simp. Argentino Educ. en Informática Eje*, pp. 171–188, 2019.

[16] M. Romero, J. Díaz, and I. Harari, "Impact of information and communication technologies on teaching-learning processes in children with special needs autism spectrum disorder," *XXIII Congr. Argentino Ciencias la Comput.*, pp. 342–353, 2017, [Online]. Available: https://www.researchgate.net/publication/341282542.

[17] D. J. (2020) Romero MR, Macas E., Harari I., "Is it possible to improve the learning of children with ASD through augmented reality mobile applications ?," *Springer, Cham*, 2020.

[18] C. R. Oliva, "The use of ict in educational guidance: An exploratory study on the current situation of use and training among educational guidance professionals," *Rev. Esp. Orientac. y Psicopedag.*, vol. 26, no. 3, pp. 78–95, 2015, doi: 10.5944/reop.vol.26.num.3.2015.16402.

[19] M. Romero, I. Harari, J. Diaz, and E. Macas, "Hoope Project: User-centered

design techniques applied in the implementation of augmented reality for children with ASD.," *Int. Conf. Human-Computer Interact. (pp. 277-290).*, no. Springer, Cham., pp. 277–290, 2022.

[20] M. Romero, I. Harari, J. Diaz, and E. Macas, "Proyecto Esperanza: Desarrollo de software con realidad aumentada para enseñanza danza a niños con transtorno del espectro autista.," *Rev. Investig. Talent.*, vol. 9, no. 1, pp. 99–115, 2022.

[21] Monica, R., Ivana, H., Javier, D., & Jorge, R. (2020, June). Augmented reality for children with Autism Spectrum Disorder. A systematic review. In 2020 International Conference on Intelligent Systems and Computer Vision (ISCV) (pp. 1-7). IEEE Computer Society.

Asignación de Docentes a Establecimientos Educativos: Un Enfoque Multi-objetivo

Horacio Villalba-Martí[1], Fabio López-Pires[2]
y Eustaquio A Martínez[1].

[1] Facultad Politécnica, Universidad Nacional del Este,
Campus Km 8 Acaray, Ciudad del Este, Paraguay
{vimartih,amartinez}@fpune.edu.py
http://www.fpune.edu.py/
[2] Universidad Internacional Tres Fronteras,
Ciudad del Este, Paraguay
fabio.lopez@uninter.edu.py
http://uninter.edu.py/

Resumen Contar con una adecuada planificación logística contribuye a mejorar el funcionamiento del sistema educativo, impactando positivamente las condiciones asociadas al aprendizaje. Este trabajo propone una nueva formulación matemática del problema de Asignación de Docentes a Establecimientos Educativos (ADEE), con un enfoque multi-objetivo para: (1) minimizar la distancia entre la residencia del docente y el establecimiento educativo, (2) maximizar la cantidad de docentes asignados al mismo establecimiento educativo y (3) maximizar la cantidad de clases dictadas por un docente en diferentes turnos. Para resolver la formulación propuesta se presenta un Algoritmo Evolutivo Multi-Objetivo (MOEA) basado en el NSGA-II. Resultados experimentales con datos reales del Departamento de Alto Paraná del Ministerio de Educación y Ciencias (MEC) de Paraguay con 457 establecimientos educativos, 2995 clases y 1808 docentes, indican mejoras significativas en la asignación.

Keywords: Optimización Multi-Objetivo, Asignación de Docentes, Computación Evolutiva, Logística Educativa, Mejora de Condiciones de Aprendizaje.

1. Introducción

Un creciente número de investigaciones confirma la importancia de las políticas dirigidas a proteger la nutrición, la salud, el desarrollo cognitivo y socio-emocional de los niños en los primeros años de vida. Sin embargo, nuevas evidencias reunidas en investigaciones recientes indican que, una vez que los niños ingresan a la escuela, ningún otro factor es tan importante como la calidad de los docentes [1].

En un análisis del sistema educativo de Paraguay como parte del proyecto "Diseño de la Estrategia de Transformación Educativa del Paraguay 2030" [2], se indica que las condiciones laborales de los docentes en las instituciones educativas distan de ser óptimas, ya que un gran número de ellos trabaja en múltiples turnos, instituciones, jornadas y niveles. Además, los mismos cumplen con un gran número de tareas fuera de lo instruccional sin tener horas de contrato para cumplirlas, lo que imposibilita el trabajo colegiado y el aprendizaje organizacional al interior de las escuelas [3].

En ese contexto, el problema abordado en este trabajo es la Asignación de Docentes a Establecimientos Educativos (ADEE). Cabe mencionar que en el sistema educativo de Paraguay, la educación formal se divide en: la Educación Inicial (EI), la Educación Escolar Básica (EEB) en tres ciclos, la Educación Media (EM) y la Educación Superior (ES) [3].

Este trabajo se enfoca en la EEB del primer y segundo ciclo (1° Grado a 6° Grado). En estos niveles el docente imparte todas las materias de la clase, es decir, una clase (e.g. 1° Grado, Sección A) cuenta con un único docente asignado.

Algunos de los inconvenientes mencionados anteriormente pueden ser reducidos con un esquema de asignación o reasignación. Por ejemplo, minimizar la cantidad de instituciones donde el docente se desenvuelva profesionalmente, puede impactar positivamente en el trabajo de colegiado y el aprendizaje organizacional.

Otro punto clave en línea con los inconvenientes mencionados anteriormente, es el tiempo que el docente dedica al traslado. Minimizar la distancia de residencia del docente con el establecimiento educativo impactaría positivamente en los siguientes puntos:

- Ahorro en gastos de transporte, con impacto financiero positivo y cuidado del ambiente al reducir las emisiones de CO_2.
- Ahorro en tiempo, con el volumen automotor en crecimiento se puede perder hasta varias horas en el tráfico. En zonas rurales, reducir la distancia podría tener un impacto aún mayor con respecto al tiempo.

Las propuestas de nuevas formulaciones matemáticas del problema ADEE que contemplen diversos aspectos mencionados para mejorar la logística en el sistema educativo representa un tema de investigación con alto impacto para países en desarrollo, donde la educación es una de las principales alternativas para lograr los objetivos clave de crecimiento. Por lo tanto, este trabajo se enfoca en proponer una nueva formulación matemática que permita mejorar algunos de los puntos citados, y de esta manera ser una herramienta potencial para mejorar la logística del sistema educativo, mejorando las condiciones asociadas al aprendizaje.

Formalmente, se puede definir el problema ADEE como:

"*Dado un conjunto de docentes con sus respectivos datos asociados y un conjunto de establecimientos educativos con sus respectivos datos asociados, asignar a los docentes en los establecimientos educativos, considerando las restricciones operativas y de recursos y optimizando simultáneamente las funciones objetivo definidas.*"

El resto de este trabajo se encuentra estructurado de la siguiente manera: En la Sección 2 se resume una revisión sistemática de la literatura asociada al problema ADEE. En la Sección 3 se presenta la nueva formulación matemática propuesta para la optimización simultánea de 3 funciones objetivo, mientras que la Sección 4 presenta el Algoritmo Evolutivo Multi-objetivo (Multi-Objective Evolutionary Algorithm, MOEA) propuesto para la resolución de la formulación propuesta. Los resultados experimentales son resumidos en la Sección 5. Las principales conclusiones se detallan en la Sección 6.

2. Revisión de la Literatura

En [4] se propone la utilización de dos algoritmos, Simulated Annealing (SA) y Tabu Search (TS), para resolver el problema de asignación de profesores. En la formulación del problema se busca minimizar la varianza total de la carga docente en una primera fase, luego se toma como entrada para una segunda fase donde se busca minimizar la varianza ponderada total de la carga de docentes. Para evaluar el rendimiento de los algoritmos propuestos, han realizado experimentos con dos conjuntos de datos reales obtenidos de una Universidad de Indonesia y algunos conjuntos de datos generados aleatoriamente. Los resultados experimentales muestran que los algoritmos considerados encuentran mejores soluciones en comparación con asignaciones manuales y un trabajo anterior que utilizó un Algoritmo Genético (AG).

En [5] se aborda el problema de la asignación de clases a profesores en universidades mediante el método Beam Search (BS) y se añaden las preferencias del profesor. Además, se desarrolla una herramienta para realizar simulaciones. Se define una función objetivo para minimizar el costo total de las asignaciones según: las preferencias individuales de los docentes y la similitud entre diferentes disciplinas. Para resolver el problema, desarrollan una herramienta que utiliza una heurística mediante búsqueda en árboles y un algoritmo voraz. Se mostró un desempeño satisfactorio para una prueba con datos reales, donde fue necesario realizar la asignación de 63 clases a 11 profesores, mejorando así la asignación manual.

En [6] se proponen varios algoritmos de aproximación para el problema de asignación de profesores en formación a las escuelas, utilizando como punto de partida el sistema educativo eslovaco y checo, donde cada profesor en formación se especializa en dos materias. En la formulación del problema tienen en cuenta la capacidad de la escuela en cada materia, donde cada profesor indica la materia y la lista de escuela aceptables. Demuestran que es relativamente sencillo proponer algoritmos de aproximación para el problema de encontrar una coincidencia con la cardinalidad máxima para una instancia determinada del problema de asignación de profesores. Concluyen con las siguientes dos preguntas abiertas: (1) ¿podrían los algoritmos refinarse para obtener un mejor límite de aproximación? y (2) ¿es posible que el problema sea APX-Completo?.

En [7] se propone utilizar AGs y Asynchronous Cooperative Parallel Search (ACoPGA) para la resolución del problema de asignación de profesores con el fin de mejorar el rendimiento, la escalabilidad y reducir el largo tiempo de ejecución asociados. Como función objetivo se normalizan y combinan 5 funciones, dando a cada una de ellas un peso para realizar las asignaciones de los docentes. Conforme una revisión de la literatura, concluyen que los AGs demostraron ser uno de los mejores algoritmos meta-heurísticos para resolver problemas de asignación de profesores. Para mejorar el tiempo de resolución realizan diversos experimentos, utilizando un esquema paralelo contra secuencial, luego paralelo asíncrono contra paralelo síncrono. Finalmente encuentran que el enfoque Asynchronous Elite CoPGA (AECoPGA) es muy eficiente para resolver casos de óptimos locales y eleva la precisión de la solución y su velocidad de búsqueda.

En [8] se desarrolla un modelo de programación lineal mixto para balancear la carga de profesores mientras se maximiza las preferencias de los profesores por clase. Para validar el modelo realizaron dos experimentos: (1) utilizando los datos del Departamento de Gestión en la Escuela de Ingeniería de Barcelona de la Universidad Politécnica de Cataluña, uno de los departamentos con mayor número de clases y profesores, y (2) con 750 instancias generadas aleatorias con patrones reales. Los resultados que obtuvieron muestran que el modelo puede resolver escenarios de hasta 40 profesores, logrando soluciones aceptables en un tiempo de cálculo reducido.

En [9] se desarrolla un sistema de recomendación basado en técnicas de minería de datos para la asignación de clases a profesores en la Educación Superior (ES). En la propuesta se difiere de los modelos de optimización

tradicionales y se basa en un sistema de recomendación basado en el histórico de asignaturas dictadas por los docentes, en la evaluación del desempeño docente y el perfil de este. Se utilizó una base de datos de una Universidad de Ecuador, donde se tomaron 133000 registros acerca del perfil docente y 3000 registros correspondientes a evaluaciones estudiantiles e históricos académicos de los últimos 10 ciclos (5 años). Las recomendaciones sugeridas por el sistema fueron valoradas por 5 expertos del ámbito académico, considerando los criterios de pertinencia, coherencia y rendimiento de la recomendación obteniendo, en una escala del 1 al 5, resultados de 4.2, 4.0 y 4.2 respectivamente en promedio en cada criterio.

En [10] se realiza una revisión de la literatura sobre problemas de gestión de competencias, en particular, el Problema de Asignación de Docentes (PAD). Determinan que normalmente el PAD se resuelve antes que el problema de organizar el horario del curso. Las investigaciones relacionadas con el PAD tienen un área sin desarrollar ya que sus enfoques no permiten la síntesis de la estructura de competencias. Adicionalmente, la falta de soluciones potenciales al problema y su naturaleza NP-completo requieren la búsqueda de condiciones suficientes, cuyo cumplimiento garantice la existencia del plan de asignación docente. La búsqueda de esas condiciones se vuelve importante: determina la finalidad del trabajo, como el tiempo que consumen estas búsquedas.

En [11] desarrollan una ontología educativa para modelar la semántica de los cursos y los perfiles académicos en las universidades, con el fin de asignar al docente más adecuado para impartir un curso específico. En esta propuesta, se busca realizar el emparejamiento de los cursos con los docentes mediante el grado de afinidad de los tópicos del curso con el perfil del profesor utilizando la ontología, una propuesta diferente a los algoritmos de asignación o a las heurísticas utilizadas en la mayoría de los trabajos estudiados. Se aplica a la Universidad King Abdulaziz (KAU) en Arabia Saudita y se concentra en la Facultad de Tecnologías de la Información y la Computación (FCIT), incluidos sus tres departamentos que siguen las reglas de la Comisión Nacional de Acreditación y Evaluación Académica (NCAAA) para documentar sus datos. Aunque en la prueba no se ha asignado un número significativo de cursos y profesores, el sistema da resultados precisos.

En [12] se propone un enfoque de dos pasos para predecir las preferencias de los profesores utilizando técnicas de minería de datos y realizar la asignación de los profesores utilizando una programación lineal entera. Se vuelve a formular el problema para optimizar la preferencia de los profesores y la varianza de la carga de trabajo. Debido a la dificultad de obtener las preferencias, se realiza una estimación de estas mediante datos históricos. El modelo propuesto se valida mediante experimentos y validación manual. Tienen una desviación menor de la carga de trabajo real de los profesores respecto de la carga de trabajo objetivo y las soluciones se generaron en un tiempo significativamente menor que el proceso manual. La mayoría de las asignaciones de la sección del curso del maestro y el número correspondiente asignado de secciones son satisfactorias.

En [13] se diseñan mecanismos sin las limitaciones de la versión modificada de la Aceptación Diferida (DA), manteniendo las buenas propiedades de incentivo de este, es decir, a prueba de estrategias (lo que significa que los profesores tienen incentivos directos para informar sus preferencias con sinceridad). En esta propuesta los maestros tienen una lista estricta de preferencias y de la misma manera las escuelas cuentan con una lista de preferencia de profesores. Estos mecanismos buscan emparejar de una forma más equitativa que los métodos tradicionales. Muestran que el mecanismo de la versión modificada de DA no es justo y eficiente tanto para los maestros como para las escuelas. Confirman el desempeño de estos mecanismos alternativos, cuando el tamaño crece, se desempeñan mucho mejor en términos de eficiencia utilitaria y equidad. Por último, utilizando un conjunto de datos sobre las solicitudes de transferencia de maestros en Francia para medir las ganancias relevantes, los mecanismos alternativos generan ganancias significativas en eficiencia y equidad.

Cabe destacar que ninguno de los trabajos estudiados incluye la optimización de la distancia entre la residencia del docente y el establecimiento educativo, uno de los principales aportes de este trabajo. Adicionalmente, se destaca que un aspecto relevante identificado por el MEC de Paraguay, representa la optimización de la cantidad de docentes asignados, lo que puede ser enfocado con la optimización de docentes asignados al mismo establecimiento educativo y la cantidad de clases dictadas por un docente en diferentes turnos, siendo también esto un relevante aporte de este trabajo.

3. Formulación Matemática del Problema

En esta sección se presenta la formulación matemática con un enfoque multi-objetivo propuesta para el problema ADEE. Primeramente, se presenta una introducción a la optimización multi-objetivo, continuando con definiciones conceptuales, para posteriormente presentar la formulación compuestas por los datos de entrada, los datos de salida, el conjunto de restricciones, las funciones objetivo propuestas y se termina con un ejemplo básico.

3.1. Optimización Multi-Objetivo

Un Problema de Optimización Multi-Objetivo (POM) puede ser definido como el problema de encontrar un vector de variables de decisión que satisface las restricciones y optimiza una función vectorial, cuyos elementos

representan las funciones objetivo. Estas funciones forman una descripción matemática de los criterios de rendimiento que suelen estar en conflicto entre sí. Por tanto, el término "optimizar" significa encontrar una solución que dé valores de todas las funciones objetivo aceptables para el tomador de decisiones [14].

En [14] se expresa un problema general de optimización multi-objetivo como:

Un conjunto q de objetivos a optimizar:

$$y = f(X) = [f_1(X), f_2(X), \ldots, f_q(X)] \tag{1}$$

Sujeto a r restricciones:

$$e(X) = [e_1(X), e_2(X), \ldots, e_r(X)] \geq 0 \tag{2}$$

Con un conjunto de p variables de decisión:

$$X = [X_1, X_2, \ldots, X_p] \in \textit{Espacio de decisión} \tag{3}$$

y el vector objetivo y:

$$y = [y_1, y_2, \ldots, y_q] \in \textit{Espacio objetivo} \tag{4}$$

Un POM por lo general no tiene una única solución óptima, sino un conjunto de soluciones óptimas comprometidas, a diferencia de un problema de optimización mono-objetivo [15]. Mediante la dominancia de Pareto se puede comparar dos soluciones y determinar si una solución es mejor a otra en un contexto multi-objetivo. Esto ocurre cuando una solución es mejor o igual en cada función objetivo y estrictamente mejor en al menos uno [14].

3.2. Definiciones Conceptuales

Para comprender la formulación matemática propuesta, se define cada uno de los conceptos considerados en ésta.

- **Establecimiento:** Lugar físico que alberga una o más instituciones.
- **Institución:** Entidad habilitada para desarrollar las clases. Podría tener sede en más de un establecimiento.
- **Grado:** Corresponde al grado de enseñanza, e.g. 1° Grado, 2° Grado, entre otros.
- **Turno:** Corresponde al turno donde se imparte la clase. e.g. Turno Mañana, Turno Tarde, Turno Noche.
- **Sección:** En caso de que la misma institución, en el mismo establecimiento, tenga más de un grupo para el mismo grado y turno, es utilizada la sección para diferenciar, e.g. 1° Grado - Turno Mañana - Sección A y 1° Grado - Turno Mañana - Sección B.
- **Clase:** Compuesto por un grado, turno, sección, institución y establecimiento. Es la unidad de enseñanza donde debe ser asignado un docente para enseñar a un grupo de alumnos, e.g. 1° Grado – Turno mañana – Sección A – Institución 1 – Establecimiento 1.
- **Docente:** Profesional habilitado para impartir una clase.

3.3. Datos de Entrada

El conjunto de grados disponibles que se representa como un vector G de dimensión n_g:

$$G = \left[\, 1\ 2 \ldots n_g \,\right] \tag{5}$$

donde:

n_g: es la cantidad de grados.

El conjunto de turnos se representa como un vector T de dimensión n_t:

$$T = \left[\, 1\ 2 \ldots n_t \,\right] \tag{6}$$

donde:

n_t: es la cantidad de turnos.

El conjunto de secciones se representa como un vector S de dimensión n_s:

$$S = \left[\, 1\ 2 \ldots n_s \,\right] \tag{7}$$

donde:

n_s: es la cantidad de secciones.

El conjunto de instituciones se representa como un vector I de dimensión n_i:

$$I = \begin{bmatrix} 1\ 2\ \dots\ n_i \end{bmatrix} \tag{8}$$

donde:
n_i: es la cantidad de instituciones.

El conjunto de establecimientos se representa como una matriz E de dimensión $(n_e \times 3)$:

$$E = \begin{bmatrix} N_1 & X_1 & Y_1 \\ \vdots & \vdots & \vdots \\ N_{n_e} & X_{n_e} & Y_{n_e} \end{bmatrix} \tag{9}$$

Cada E_k es representado por el número del establecimiento, su coordenada geográfica X_k y su coordenada geográfica Y_k:

$$E_k = \begin{bmatrix} N_k\ X_k\ Y_k \end{bmatrix} \forall k \in \{1, \dots, n_e\}$$

donde:
N_k: número del establecimiento E_k;
X_k: coordenada geográfica X de la ubicación del establecimiento E_k;
Y_k: coordenada geográfica Y de la ubicación del establecimiento E_k;
n_e: cantidad de establecimientos.

El conjunto de docentes se representa como una matriz D de dimensión $(n_d \times 3)$:

$$D = \begin{bmatrix} N_1 & X_1 & Y_1 \\ \vdots & \vdots & \vdots \\ N_{n_d} & X_{n_d} & Y_{n_d} \end{bmatrix} \tag{10}$$

Cada D_i es representado por el número del docente, su coordenada geográfica X_i y su coordenada geográfica Y_i:

$$D_i = \begin{bmatrix} N_i\ X_i\ Y_i \end{bmatrix} \forall i \in \{1, \dots, n_d\}$$

donde:
N_i: número del docente D_i;
X_i: coordenada geográfica X de la ubicación de la residencia del docente D_i;
Y_i: coordenada geográfica Y de la ubicación de la residencia del docente D_i;
n_d: cantidad de docentes.

El conjunto de clases activas se representa como una matriz C de dimensión $(n_c \times 5)$:

$$C = \begin{bmatrix} G_1 & T_1 & S_1 & I_1 & E_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ G_{n_c} & T_{n_c} & S_{n_c} & I_{n_c} & E_{n_c} \end{bmatrix} \tag{11}$$

Cada C_j es representado por el Grado, Turno, Sección, Institución y Establecimiento como:

$$C_j = \begin{bmatrix} G_j\ T_j\ S_j\ I_j\ E_j \end{bmatrix} \forall j \in \{1, \dots, n_c\}$$

donde:
G_j: Grado de la clase C_j, entonces $G_j \in G$;
T_j: Turno de la clase C_j, entonces $T_j \in T$;
S_j: Sección de la clase C_j, entonces $S_j \in S$;
I_j: Institución de la clase C_j, entonces $I_j \in I$;
E_j: Número de establecimiento de C_j, entonces $E_j \in E$;
n_c: cantidad de clases disponibles.

La matriz calculada U de distancias entre la residencia del docente y los establecimientos de dimensión $(n_d \times n_e)$ se representa como:

$$U = \begin{bmatrix} U_{1,1} & U_{1,2} & U_{1,3} & \dots & U_{1,n_e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{n_d,1} & U_{n_d,2} & U_{n_d,3} & \dots & U_{n_d,n_e} \end{bmatrix} \tag{12}$$

Cada posición $U_{i,k}$ representa la distancia entre la residencia de un docente i y un establecimiento k.

$$\forall i \in \{1, \dots, n_d\} \wedge \forall k \in \{1, \dots, n_e\}$$

donde:
$U_{i,k}$: Distancia entre la residencia del docente D_i y el establecimiento E_k.
La matriz calculada V de distancias entre establecimientos de dimensión $(ne \times ne)$ se representa como:

$$V = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} & \dots & V_{1,n_e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{n_e,1} & V_{n_e,2} & V_{n_e,3} & \dots & V_{n_e,n_e} \end{bmatrix} \tag{13}$$

Cada posición $V_{k,m}$ representa la distancia entre establecimiento k y el establecimiento m.

$$\forall k \in \{1, \dots, n_e\} \wedge \forall m \in \{1, \dots, n_e\}$$

donde:
$V_{k,m}$: Distancia entre el establecimiento E_k y el establecimiento E_m.

3.4. Datos de Salida

Una solución al problema se representa por la matriz X de dimensión $(n_d \times n_c)$:

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & X_{1,3} & \dots & X_{1,n_c} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n_d,1} & X_{n_d,2} & U_{n_d,3} & \dots & X_{n_d,n_c} \end{bmatrix} \tag{14}$$

Cada posición $X_{i,j}$ representa el estado de asignado o no del docente i en la clase j.

$$\forall i \in \{1, \dots, n_d\} \wedge \forall j \in \{1, \dots, n_c\}$$

donde: $X_{i,j}$: es una variable binaria, donde 1 indica que el docente i fue asignado a la clase j, $0(cero)$ en caso contrario;

3.5. Restricciones

En esta sección se definen las restricciones que deben cumplir las soluciones factibles.

(a) Un docente asignado hasta en dos clases:
Un docente debería a lo sumo estar asignado a dos clases, que es el tiempo máximo de jornada laboral posible. Esta restricción se expresa como:

$$\sum_{j=1}^{n_c} X_{i,j} \leq 2 \; \forall i \in \{1, \dots, n_d\} \tag{15}$$

(b) Cada clase debe tener asignado un único docente:
Cada clase solo debe tener un único docente asignado y no puede quedar sin docente, esta restricción se expresa como:

$$\sum_{i=1}^{n_d} X_{i,j} = 1 \; \forall j \in \{1, \dots, n_c\} \tag{16}$$

(c) Turnos diferentes para las asignaciones de un docente:
Para los docentes con dos asignaciones, los turnos de las clases asignadas deben ser diferentes para que el docente pueda cumplir con la impartición de la clase, esta restricción se expresa como:

$$C_{j,2} \neq C_{l,2} \tag{17}$$
$$\forall i \in \{1, \dots, n_d\} \wedge \forall j \in \{1, \dots, n_c\}$$
$$\wedge \forall l \in \{1, \dots, n_c\} : j \neq l \wedge X_{i,j} = 1 \wedge X_{i,l} = 1$$

donde:
$C_{j,2}$: Turno de la clase asignada al docente;
$C_{l,2}$: El otro turno de la clase asignada al docente;
$X_{i,j}$: asignación del docente i en la clase j;
$X_{i,l}$: asignación del docente i en la clase l.

(d) Distancia máxima entre establecimientos asignados a un docente:
Los establecimientos asignados no deben distar más allá de una distancia prudencial que pueda permitir el traslado del docente en el cambio de turno.

$$P_{i,j,l} \leq D_{max} \tag{18}$$
$$\forall i \in \{1,\ldots,n_d\} \wedge \forall j \in \{1,\ldots,n_c\} \wedge \forall l \in \{1,\ldots,n_c\}$$
$$: j \neq l \wedge X_{i,j} = 1 \wedge X_{i,l} = 1$$

donde:
$C_{j,5}$: Establecimiento de la asignación uno del docente i;
$C_{l,5}$: Establecimiento de la asignación dos del docente i;
$P_{i,j,l}$: distancia $V_{C_{j,5},C_{l,5}}$ entre establecimientos asignados al docente i;
D_{max}: distancia máxima permitida entre establecimientos;
$X_{i,j}$: asignación del docente i en la clase j;
$X_{i,l}$: asignación del docente i en la clase l.

3.6. Funciones Objetivo

(a) Minimizar la distancia promedio entre la residencia del docente y el establecimiento educativo de su clase:
Este objetivo permite evaluar la afinidad de las soluciones basándose en las distancias (en KMs) menores entre la residencia del docente y el establecimiento educativo, la misma se expresa como:

$$f_1(X) = \frac{\sum_{i=1}^{n_d} \sum_{j=1}^{n_c} U_{i,C_{j,5}} * X_{i,j}}{n_c} \tag{19}$$

donde:
$U_{i,k}$: Distancia entre la residencia del docente i y el establecimiento k.

(b) Maximizar la cantidad de docentes con dos turnos dentro del mismo establecimiento:
Este objetivo permite evaluar la afinidad de las soluciones basándose en asignaciones dentro de la misma institución para el docente, la función se expresa como:

$$f_2(X) = \frac{\sum_{i=1}^{n_d} Y_i}{\sum_{i=1}^{n_d} R_i} \tag{20}$$

$$Y_i = \begin{cases} 1, \ Si \ \exists X_{i,j} = 1 \wedge X_{i,l} = 1 : j \neq l \wedge C_{j,5} = C_{l,5} \\ 0, \ caso \ contrario \end{cases}$$
$$R_i = \begin{cases} 1, \ Si \ \exists X_{i,j} = 1 \\ 0, \ caso \ contrario \end{cases}$$
$$\forall j \in \{1,\ldots,n_c\} \wedge \forall l \in \{1,\ldots,n_c\} \wedge$$

donde:
Y_i: variable binaria que indica si el docente i tiene dos asignaciones en el mismo establecimiento;
R_i: variable binaria que indica si el docente i cuenta con al menos 1 asignación, 0 (*cero*) en caso contrario.

(c) Maximizar la cantidad de aulas asignados a un docente:
Este objetivo permite evaluar la afinidad de las soluciones basándose en las asignaciones para ambos turnos disponibles del docente.

$$f_3(X) = \frac{\sum_{i=1}^{n_d} Z_i}{\sum_{i=1}^{n_d} R_i} \tag{21}$$

$$Z_i = \begin{cases} 2, \ Si \ \exists X_{i,j} = 1 \wedge X_{i,l} = 1 : j \neq l \\ 1, \ Si \ \exists X_{i,j} = 1 \\ 0, \ caso \ contrario \end{cases}$$
$$R_i = \begin{cases} 1, \ Si \ \exists X_{i,j} = 1 \\ 0, \ caso \ contrario \end{cases}$$
$$\forall j \in \{1,\ldots,n_c\} \wedge \forall l \in \{1,\ldots,n_c\}$$

donde:
Z_i: variable que representa la cantidad de clases asignadas al docente i;
R_i: variable binaria que indica si el docente (i) cuenta con al menos 1 asignación, 0 (*cero*) en caso contrario.

3.7. Ejemplo Básico

En la Fig. 1 se presenta una instancia simple del problema ADEE ubicada en un plano cartesiano. La instancia esta compuesta por 4 docentes y 3 establecimientos educativos con 6 clases en total.

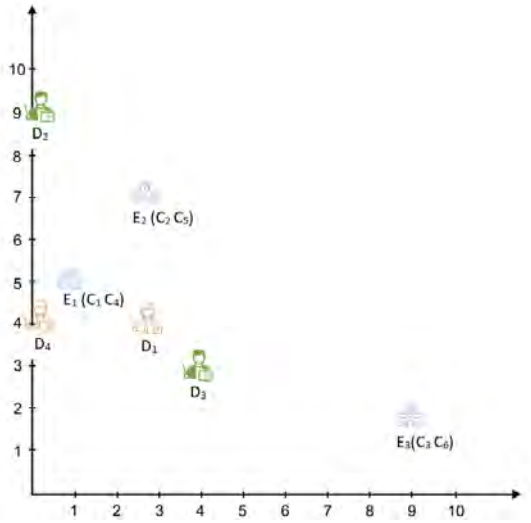


Figura 1. Establecimientos y Docentes ubicados en un plano cartesiano.

Los datos de entrada de la instancia de ejemplo se presentan a continuación G, T, S, I, E, D *y* C:

$$G = \begin{bmatrix} 1 & 2 \end{bmatrix} : 1 = Primer\ Grado \wedge 2 = Segundo\ Grado$$

$$T = \begin{bmatrix} 1 & 2 \end{bmatrix} : 1 = Turno\ Mañana \wedge 2 = Turno\ Tarde$$

$$S = \begin{bmatrix} 1 & 2 \end{bmatrix} : 1 = A \wedge 2 = B$$

$$I = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

$$E = \begin{bmatrix} 1 & 1 & 5 \\ 2 & 3 & 7 \\ 3 & 9 & 2 \end{bmatrix} \quad D = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 0 & 9 \\ 3 & 4 & 3 \\ 4 & 0 & 4 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 3 & 3 \\ 2 & 2 & 1 & 1 & 1 \\ 2 & 2 & 1 & 2 & 2 \\ 2 & 1 & 1 & 3 & 3 \end{bmatrix}$$

La Clase 1 C_1: 1° Grado, Turno Mañana, Sección A de la Institución A y del Establecimiento A.
Una solución factible X, que cumple con todas las restricciones, se muestra a continuación:

$$X = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

En esta solución factible X, al Docente 1 fue asignado a la Clase 2 del Establecimiento 2 y la Clase 5 del Establecimiento 2, mientras que al Docente 2 no se le ha asignado ninguna clase. Por otro lado, al Docente 3 se le ha asignado la Clase 3 del Establecimiento 3 y la Clase 6 del Establecimiento 3; y por último, al Docente 4 se le ha asignado la Clase 1 del Establecimiento 1 y la Clase 4 del Establecimiento 1.

4. Algoritmo Propuesto

En esta sección se presenta el Algoritmo Evolutivo Multi-Objetivo propuesto para la resolución de la formulación matemática del problema ADEE (ver Sección 3).

Los Algoritmos Evolutivos (AE) son métodos de búsqueda que se inspiran en la selección natural y la supervivencia del más apto en el mundo biológico. Los AE difieren de las técnicas de optimización más tradicionales porque involucran una búsqueda de una población de soluciones. Los AEs son particularmente adecuados para resolver problemas de optimización multi-objetivo, ya que consideran un conjunto de posibles soluciones (población) [16].

En este trabajo, se propone un MOEA basado en el NSGA-II propuesto por Kalyanmoy Deb en [15]. Para la utilización de un MOEA es necesario representar una solución al problema en una estructura de cadena denominada cromosoma. Para el problema ADEE, se propone una cadena de números enteros, donde cada gen (posición en la cadena) es una clase a ser asignada y cada alelo (valor que puede tomar el gen) representa al docente asignado a dicha clase.

En la Fig. 2 se muestra el cromosoma para la solución X del Ejemplo Básico (ver Sección 3).

Figura 2. Cromosoma propuesto para la solución X (ver Sección 3).

Cada cromosoma (representación de una solución) puede ser evaluado mediante las funciones objetivo definidas. Al valor de esta evaluación se le denomina fitness (aptitud) y es utilizado por los algoritmos evolutivos para determinar que individuos (soluciones) continúan en las siguientes generaciones.

A continuación, se presenta el Algoritmo 1 con el pseudo-código del MOEA propuesto para la resolución de la formulación matemática propuesta para el problema ADEE.

Algoritmo 1 - MOEA propuesto para resolver la formulación matemática para el problema ADEE (ver Sección 3).

Entrada: $G, T, S, I, E, C, D, U, V, N_{gen}$
Salida: Conjunto Pareto (Soluciones No Dominadas)
1: $P \leftarrow$ Población Aleatoria Factible
2: Evaluar P
3: **mientras** $N_{gen} \neq 0$ **hacer**
4: $\quad Q \leftarrow$ Seleccionar individuos de P según NSGA-II
5: $\quad Q \leftarrow$ Aplicar operador de cruzamiento
6: $\quad Q \leftarrow$ Aplicar operador de mutación
7: $\quad Q \leftarrow$ Reparar individuos
8: $\quad P \leftarrow$ Seleccionar individuos de $P + Q$ según NSGA-II
9: $\quad N_{gen} \leftarrow N_{gen} - 1$
10: **fin mientras**

Con la representación del cromosoma y la evaluación de su fitness es posible generar una población inicial de individuos (Paso 1-2 de Algoritmo 1) y mediante los operadores genéticos de selección, cruzamiento y mutación es posible ir mejorando dichas soluciones en cada generación.

El operador de selección es la estrategia utilizada para seleccionar que individuos sobreviven en la siguiente generación (Paso 4 y 8 de Algoritmo 1). El operador de cruzamiento (Paso 5 de Algoritmo 1) genera nuevos individuos denominados descendencia en base a combinaciones de los individuos actuales denominados padres.

Por último, el operador de mutación (Paso 6 de Algoritmo 1) suele intercambiar aleatoriamente un gen o más de uno en base a una probabilidad. Estos operadores podrían producir soluciones no factibles, por lo que normalmente se utiliza un operador de reparación (Paso 7 de Algoritmo 1) que ajusta la descendencia a soluciones factibles.

5. Solución Propuesta y Resultados Experimentales

En esta sección se presenta la solución propuesta para la aplicación del algoritmo propuesto (ver Sección 4) al caso del sistema educativo de Paraguay, el conjunto de datos considerado, detalles sobre la implementación de la solución y los resultados experimentales.

5.1. Diseño de la Solución Propuesta

Para la aplicación del algoritmo propuesto (ver Sección 4) al caso del sistema educativo de Paraguay, se propone la solución detallada en la Fig. 3. Para las fuentes de datos, se han obtenidos 3 orígenes heterogéneos con una variedad de formatos (característica de variedad de Big Data): (1) el Sistema de Gestión del MEC (SIGMEC), para datos **anonimizados** de los docentes, (2) archivos en formato *JSON* para las clases y (3) el Portal de Datos Abiertos del MEC [17] para establecimientos e instituciones. A continuación, estas fuentes de datos fueron consolidadas en un almacén de datos, mediante un proceso de extracción, transformación y carga (ETL). Por último, el algoritmo propuesto utilizó el almacén de datos para calcular y obtener los resultados de las asignaciones de los docentes a los establecimientos educativos en un Conjunto Pareto P, compuesto por soluciones no dominadas para soporte a la toma de decisión correspondiente.

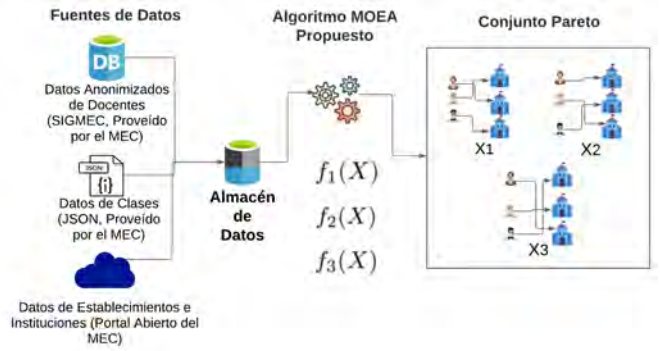


Figura 3. Diseño de la solución propuesta para Paraguay.

5.2. Conjunto de Datos Considerado

Se han utilizado datos de 2020 proveídos por el Ministerio de Educación y Ciencias del Paraguay (MEC). Si bien se ha diseñado una solución para el sistema educativo de Paraguay, los experimentos de este trabajo se han enfocado a un sub-conjunto de datos del Departamento de Alto Paraná del Paraguay. Características de los datos considerados en los experimentos están resumidas en la Tabla 1.

Tabla 1. Características del Conjunto de Datos de Alto Paraná.

Característica	Cantidad
Establecimientos	457
Docentes	1808
Clases	2995
Grados	6 (1° a 6°)
Turnos	2 (Mañana y Tarde)

Para calcular las coordenadas de la residencia de los docentes se ha utilizado el servicio de Geocoding API de Google. Se proveyó los datos en el siguiente formato: "*País, Departamento, Localidad, Distrito y Dirección*retornando las coordenadas en el siguiente formato: "*latitud,longitud*".

Para verificar la precisión del cálculo de las coordenadas se ha utilizado el servicio de Maps JavaScript API de Google. Se ubicaron por cada establecimiento y docente, un marcador dentro del mapa. Los docentes que no se encontraban dentro de los límites del departamento Alto Paraná fueron ajustados en un proceso manual. En la Fig. 4 se puede visualizar como los establecimientos y docentes quedaron dentro de los límites del Departamento Alto Paraná del Paraguay.

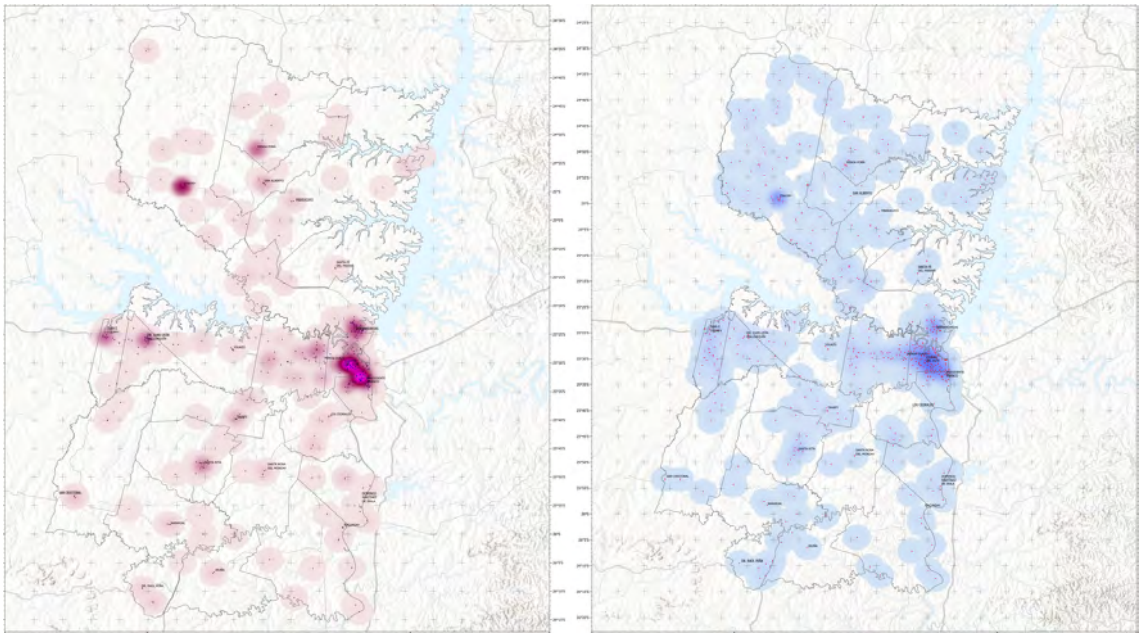


Figura 4. Docentes (izquierda) y Establecimientos (derecha) en Alto Paraná.

Para calcular la distancia entre establecimientos y la residencia de los docentes se utilizó el servicio geopy API, que utiliza la distancia geodésica entre dos coordenadas.

5.3. Implementación de la Solución

Para la implementación de la solución propuesta y para la ejecución de los experimentos se utilizó el siguiente conjunto de herramientas: Lenguaje de Programación Python[1] 3.9.11, Base de Datos PostgreSQL[2] 9.16, Framework de Optimización Multi-Objetivo pymoo[3] 0.5.0, librería geopy[4] 2.2.0 y el entorno de desarrollo Visual Studio Code[5] 1.66.2.

La selección de pymoo se basó en el estudio comparativo de frameworks resumido en la Tabla 2 presentado en [18].

Tabla 2. Comparativa de Frameworks de Optimización Multi-Objetivo presentada en [18].

Nombre	Licencia	Foco en Multi-Objetivo	Python Puro	Visualización	Toma de Decisión
jMetalPY	MIT	✓	✓	✓	–
PyGMO	GPL-3.0	✓	–	–	–
Platypus	GPL-3.0	✓	✓	–	–
DEAP	LGPL-3.0	–	✓	–	–
Inspyred	MIT	–	✓	–	–
pymoo	Apache 2.0	✓	✓	✓	✓

Con respecto al entorno utilizado en los experimentos, fue utilizado una computadora de escritorio con las siguientes características: Sistema Operativo Windows 10 Pro for Workstations, Procesador Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz de 8 Núcleos y Memoria RAM de 32GB.

Con el objetivo de dar reproducibilidad al trabajo, el código fuente, conjunto de datos y resultados experimentales se encuentran disponibles en línea[6].

[1] https://www.python.org/
[2] https://www.postgresql.org/
[3] https://pymoo.org/
[4] https://geopy.readthedocs.io/en/stable/
[5] https://code.visualstudio.com/
[6] https://github.com/horaciov/assign-teacher

5.4. Resultados Experimentales

Por la naturaleza del MOEA propuesto, se han realizado 10 ejecuciones del algoritmo para el conjunto de datos descrito (ver Sección 5.2) con una población inicial de 100 individuos, tamaño de población de 100 individuos y 100 generaciones.

Datos relacionados a la cantidad de soluciones no dominadas obtenidas en cada ejecución y el tiempo de ejecución del algoritmo son resumidos en la Tabla 3.

Tabla 3. Cantidad de Soluciones No Dominadas y Tiempos de Ejecución.

Corrida	**Cantidad de Soluciones**	**Tiempo de Ejecución (HH:mm)**
1	95	13:35
2	99	16:20
3	95	15:30
4	92	14:24
5	80	13:51
6	50	18:00
7	100	16:28
8	88	13:55
9	68	15:34
10	100	12:38

Considerando los resultados de las mencionadas ejecuciones del MOEA propuesto, las soluciones encontradas fueron combinadas en un único Conjunto Pareto P. De manera a tener un valor de solución de referencia del mencionado Conjunto Pareto P, se calculó el promedio de los valores de las funciones objetivo, representado por $y(\overline{X})$. Adicionalmente, y a modo de tener una comparativa con las solución actual del sistema educativo de Paraguay, se calculó el valor de las funciones objetivo de la solución proveída por el MEC, representado por $y(X_{MEC})$. Finalmente, para tener una idea del valor óptimo que las soluciones podrían llegar a tener, y solo a modo de referencia, se presenta también los valores de las funciones objetivo para una solución óptima teórica, representado por $y(X_{OPT})$. Los mencionados valores son resumidos a continuación en la Tabla 4.

Tabla 4. Vectores Objetivo: $y(X_{OPT})$, $y(X_{MEC})$ e $y(\overline{X})$

Vector Objetivo	**Min** $f_1(X)$	**Max** $f_2(X)$	**Max** $f_3(X)$
$y(X_{OPT})$	0.0000	1.0000	2.0000
$y(X_{MEC})$	7.6679	0.4850	1.5453
$y(\overline{X})$	6.3462	0.4533	1.9007

Como uno de los principales hallazgos de los experimentos realizados, se puede destacar que la asignación actual de docentes a establecimientos educativos en el sistema educativo de Paraguay X_{MEC} es una solución no dominada con respecto a las soluciones del Conjunto Pareto P. Sin embargo, considerando que una muy aceptada estrategia de selección entre soluciones no dominadas desde un Conjunto Pareto es utilizar el operador de preferencia, se destaca que las soluciones del Conjunto Pareto P, encontradas por el algoritmo propuesto, son preferidas en un 100 % a la solución X_{MEC}, es decir, siempre son mejores en 2 objetivos ($f_1(X)$ y $f_3(X)$. En ese contexto, es importante recordar que una solución no dominada A es preferida a otra B ($A \succ_p B$), si A es mejor en más objetivos B [19].

En la Fig. 5 se muestran los vectores objetivo de las soluciones no dominadas del Conjunto Pareto P, consolidadas de las 10 corridas del algoritmo propuesto. Además, se muestran los vectores objetivo de la solución actual del sistema educativo de Paraguay X_{MEC} y el vector objetivo $y(\overline{X})$. Se visualiza que todas las soluciones encontradas son mejores en las funciones objetivo $f_1(X)$ y $f_2(X)$, con respecto a la solución X_{MEC}.

Considerando los resultados experimentales, y con una comparativa de los valores de $y(X_{MEC})$ con $y(\overline{X})$, a continuación se presentan hallazgos relacionados a cada función objetivo propuesta (ver Sección 3):

- Con respecto a $f_1(X)$: se observa una mejora en promedio del 17.23 % de $y(\overline{X})$ sobre $y(X_{MEC})$, lo que implica una reducción significativa de la distancia promedio entre la residencia de los docentes y los establecimientos educativos.

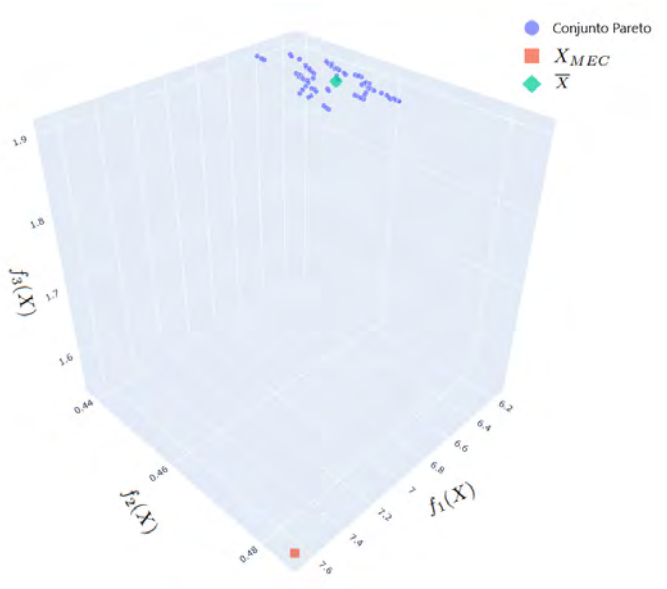


Figura 5. Resumen de Vectores Objetivo de Resultados Experimentales.

- Con respecto a $f_2(X)$: se observa que $y(X_{MEC})$ es mejor que $y(\overline{X})$ en promedio en un 6.53 %. Esto se encuentra relacionado con la mejora que se obtuvo en $f_3(X)$. Al aumentar la cantidad de docentes con ambos turnos, se ve una dificultad de encontrar o asignar docentes en el mismo establecimiento ($f_2(X)$).
- Con respecto a $f_3(X)$: se observa una mejora en promedio del 22.99 % de $y(\overline{X})$ sobre $y(X_{MEC})$, lo que permite que con una cantidad significativamente menor de docentes, se pueda atender la misma cantidad de clases ofertadas por los establecimientos educativos. Aproximadamente, se puede asignar con un 13 % menos de docentes la misma cantidad de clases, esto representa 235 docentes menos para nuestro caso. Esto resulta en un mejor aprovechamiento de los recursos.

6. Conclusiones y Trabajos Futuros

En este trabajo se ha propuesto una nueva formulación matemática para el problema ADEE, que permite estudiarlo en un contexto de optimización multi-objetivo (ver Sección 3). Se ha propuesto un Algoritmo Evolutivo Multi-Objetivo (MOEA) para validar la formulación propuesta y resolver el problema en tiempos razonables (ver Sección 4).

Resultados experimentales obtenidos con datos reales del sistema educativo de Paraguay, específicamente para el Departamento de Alto Paraná, muestran que que la propuesta representa una alternativa válida para resolver el problema ADEE, con mejoras significativas que podrían mejorar las condiciones de aprendizaje. En ese sentido, se redujo la distancia entre la residencia del docente y el establecimiento educativo, así como se mejoró la asignación para que el docente tenga asignados dos cursos, impactando así positivamente en su calidad de vida, reduciendo el tiempo de traslado y los gastos en el medio de transporte. Esto, por lo tanto, podrá contribuir a que puedan dar una clase de calidad a sus alumnos. Se consiguió optimizar la utilización de recursos, reduciendo la cantidad de docentes necesarios para atender la cantidad de clases ofertadas, en promedio en un 13 %.

En el contexto de este trabajo, se realizarán otros experimentos con los datos de los demás departamentos del Paraguay. Además, se proponen como trabajos futuros:

- Comparar experimentalmente el MOEA propuesto con otros Algoritmos Evolutivos Multi-Objetivo.
- Extender la formulación matemática propuesta para incluir no solo la distancia entre la residencia del docente al establecimiento educativo, sino también el tiempo de traslado como un criterio significativo.
- Proponer otras funciones objetivo que permitan mejorar la formulación matemática propuesta a otros escenarios de asignación de docentes, ya en un contexto de optimización de muchos objetivos (*Many-Objective Optimization*).

Referencias

[1] B. Bruns and J. Luque, "Foro sobre desarrollo de américa latina profesores excelentes cómo mejorar el aprendizaje en américa latina y el caribe," p. 76, 2014.

[2] "Diseño de la estrategia de transformación educativa del paraguay 2030," http://www.feei.gov.py/, accedido: 11/05/2022.

[3] G. N. del Paraguay, "Análisis del sistema educativo nacional," p. 64, 2021.

[4] A. Gunawan and K. Ng, "Solving the teacher assignment problem by two metaheuristics," *International Journal of Information and Management Sciences*, vol. 22, pp. 73–86, 03 2011.

[5] A. T. D. Azevedo, A. F. S. M. Ohata, J. A. Amorim, and P. M. Gustavsson, "Assigning classes to teachers in universities via mathematical modelling: Using Beam Search method and simulation in Java," pp. 577–586, 2013.

[6] K. Cechlárová, P. Eirinakis, T. Fleiner, D. Magos, I. Mourtos, E. Oce, and I. M. Preprint, "Approximation algorithms for the teachers assignment problem," 2014.

[7] E. Turki, "Solving teacher assignment problem by asynchronous cooperative parallel genetic algorithm," in *ICFCCS*, 05 2014.

[8] B. Domenech and A. Lusa, "A MILP model for the teacher assignment problem considering teachers' preferences," *European Journal of Operational Research*, vol. 249, no. 3, pp. 1153–1160, 2016.

[9] F. Pesántez-Avilés, D. Calle-López, V. Robles-Bykbaev, M. Rodas-Tobar, and C. Vásquez-Vásquez, "A recommender system based on data mining techniques to support the automatic assignment of courses to teachers in higher education," in *2017 International Conference on Information Systems and Computer Science (INCISCOS)*, 2017, pp. 231–236.

[10] E. Szwarc, I. Bach-Dąbrowska, and G. Bocewicz, "Competence management in teacher assignment planning," in *Information and Software Technologies*, R. Damaševičius and G. Vasiljevienė, Eds. Cham: Springer International Publishing, 2018, pp. 449–460.

[11] G. Ashour, A. Al-Dubai, and I. Romdhani, "Ontology-based course teacher assignment within universities," *International Journal of Advanced Computer Science and Applications*, vol. 11, 01 2020.

[12] R. Tejada and I. A. Martinez, "A two-step approach involving forecasting preferences integrating curriculum, rank, educational attainment and interest, and assignment to shorten teacher-course assignment process," in *2020 IEEE World Conference on Engineering Education (EDUNINE)*, 2020, pp. 1–6.

[13] J. Combe, O. Tercieux, and C. Terrier, "The Design of Teacher Assignment: Theory and Evidence," *The Review of Economic Studies*, 02 2022.

[14] C. A. C. Coello, G. B. Lamont, and D. A. V. Veldhuisen, *Evolutionary algorithms for solving multi-objective problems*. Springer, 2007.

[15] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[16] M. C. Bhuvaneswari, "Application of evolutionary algorithms for multi-objective optimization in vlsi and embedded systems," *Application of Evolutionary Algorithms for Multi-Objective Optimization in VLSI and Embedded Systems*, pp. 1–174, 1 2015.

[17] "Portal de datos abiertos del mec," https://datos.mec.gov.py/, accedido: 01/03/2022.

[18] J. Blank and K. Deb, "pymoo: Multi-objective optimization in Python," *IEEE Access*, vol. 8, pp. 89 497–89 509, 2020.

[19] F. Talavera, J. Crichigno, and B. Barán, "Policies for dynamical multiobjective environment of multicast traffic engineering," in *IEEE ICT*, 2005.

Tutores inteligentes en la enseñanza: Una revisión y análisis en la educación secundaria

María Cecilia Pezzini [1], Pablo Thomas [2]

[1] Alumna de la Especialización en tecnología informática aplicada a la educación. Facultad de Informática – Universidad Nacional de La Plata, La Plata, Argentina.
2 Instituto de Investigación en Informática LIDI (III-LIDI). Facultad de Informática – Universidad Nacional de La Plata, La Plata, Argentina. Centro Asociado a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)
c_pezzini@hotmail.com, pthomas@lidi.info.unlp.edu.ar

Resumen. Uno de los ámbitos más afectados por la crisis sanitaria ocasionada por la pandemia de COVID 19 fue la educación, debido a la interrupción de la enseñanza presencial durante los años 2020 y 2021, transformando las trayectorias de aprendizaje de más de 1.600 millones de estudiantes en todo el mundo [26].
En las pruebas Aprender 2019, el 72% de los alumnos terminó la secundaria con deficiencias en Matemática y anticipan que la pandemia agravó los resultados [3].
Es por esto, que la utilización de sistemas tutores inteligentes, puede resultar oportuno para acompañar el proceso de enseñanza y aprendizaje en el área de matemática, con el fin de ayudar a mejorar el nivel académico de los estudiantes.
En este trabajo se realiza un análisis comparativo de sistemas tutores inteligentes orientados a la educación secundaria para matemática.

Keywords: Sistemas tutores inteligentes, educación secundaria, enseñanza de matemática.

1 Introducción

El Ministerio de Educación de la Nación, a través de la Secretaría de Evaluación e Información Educativa (SEIE), realiza desde el año 2016, una evaluación nacional llamada Aprender [2] , que permite medir el nivel de desempeño de los estudiantes tanto de nivel primario como secundario, en áreas básicas de conocimiento, así como identificar distintos factores que inciden en los aprendizajes.

El 1 diciembre de 2021, se realizaron las pruebas Aprender 2021, de forma censal en sexto grado de las 23.000 escuelas primarias de Argentina. Los alumnos hicieron la evaluación de Lengua y Matemática [4][5].

Los resultados de la prueba Aprender 2021 permiten por primera vez tener una mirada sobre los niveles de desempeño de los estudiantes del último año de primaria luego de la interrupción de clases presenciales en 2020 y 2021, por la pandemia de COVID19.

En comparación con los datos previos, de las pruebas Aprender 2018, efectuada a alumnos de 6° año de primaria, los puntajes promedio disminuyeron tanto en Lengua como en Matemática.

En el año 2018, Argentina participó de las pruebas PISA (Programme for International Student Assessment, por sus siglas en inglés), que consiste en un operativo trienal, dirigido a alumnos de 15 años y que estén cursando 7° año o más.

Los alumnos argentinos, se ubicaron por debajo de la media de América Latina, (países latinoamericanos participantes Chile, Uruguay, Costa Rica, México, Brasil, Colombia, Perú, Panamá, República Dominicana) [24].

Mientras que en las pruebas Aprender 2019, dirigida a alumnos del nivel secundario, el 72% de los alumnos terminó la secundaria con deficiencias en Matemática y anticipan que la pandemia agravó los resultados [3].

La evolución de la tecnología, ha llevado a desarrollar sistemas de software que utilizan técnicas de inteligencia artificial, para favorecer la educación personalizada, actuando como un tutor personal para cada uno de los estudiantes, pudiendo discernir sus necesidades y los procesos meta cognitivos que requieren en el aprendizaje.

Tomando como premisa los resultados de las pruebas PISA 2018 y Aprender 2019, el presente trabajo tiene como objetivo, hacer una estudio bibliográfico del impacto de los sistemas tutores inteligentes, sobre investigaciones existentes, metodologías que se aplican; y qué trabajos se realizan en la educación secundaria, y cómo ayudan a mejorar el aprendizaje de los estudiantes.

Se pondrá especial atención en aquellos Sistemas Tutores Inteligentes (en adelante denominados STI) relacionados con la enseñanza de matemática.

La preferencia en su selección se debe a que el pensamiento matemático es aquel que mayor dificultad presenta para los alumnos; y la mejora en el mismo les brinda herramientas para desarrollar el pensamiento analítico, los predispone para el aprendizaje de otras disciplinas, pudiéndose considerar como una herramienta de aprendizaje bisagra.

Además, les permite desarrollar capacidades de razonamiento y abstracción, contribuyendo al análisis de estrategias tanto en la resolución de problemas como en situaciones concretas que se les presenten.

En la sección 2 se detalla la selección de sistemas tutores inteligentes. La sección 3 presenta el análisis de sistemas tutores inteligentes. En la sección 4 se presentan las conclusiones y trabajos futuros.

2 Selección de sistemas tutores inteligentes.

Los sistemas tutores inteligentes integran tres áreas básicas:

- La investigación educativa a través de herramientas que proporcionen una enseñanza personalizada asegurando el aprendizaje del estudiante.
- La inteligencia artificial, mediante la aplicación de técnicas de modelado de usuario, representación del conocimiento y razonamiento.
- La psicología cognitiva o educativa al aplicar la simulación cognitiva del comportamiento de un tutor: razonamiento, aprendizaje, conocimiento.

Los Sistemas Tutores Inteligentes, permiten la emulación de un tutor humano en el sentido de saber qué enseñar, cómo enseñar y a quién enseñar.

Por otra parte, se pueden concebir tutores que trabajen para solucionar paulatinamente los conceptos erróneos (misconceptions), contribuyendo a un cambio conceptual [22][23] ; en el modo de construir el conocimiento en forma significativa.

La búsqueda bibliográfica, se realizó en bases de datos de investigación académicas, a partir de cadenas específicas, como se indica en la tabla 1.

Se encontraron 183 trabajos que abordan la temática de interés; de los cuales se seleccionaron 10 sistemas tutores inteligentes, vinculados a la temática de STI en la enseñanza secundaria, con preferencia en el área de matemática, que pueden en algunos casos modelar la afectividad; cuyas fechas de publicación, corresponden a los últimos cinco años.

Como estrategias de selección se tuvo en cuenta:

1. El título de la publicación, se eliminaron las publicaciones con un título no relacionado con el objeto del trabajo.
2. Exclusión basada en resumen: se excluyeron las publicaciones que en el resumen o las palabras claves no estaban relacionadas con el enfoque de la revisión.
3. Exclusión basada en revisión rápida: se realizó una lectura rápida de las secciones y subsecciones, figuras, tablas y referencias para excluir las publicaciones que no estaban relacionados con el objetivo de la revisión.
4. Exclusión basada en el artículo completo: se realizó la lectura completa de la publicación y se eliminaron aquellos que no coincidían con los criterios de inclusión y exclusión.

La Tabla 2, muestra los sistemas tutores inteligentes seleccionados y la fuente de referencia.

Tabla 1. Motores de búsqueda y cadenas de referencia.

Biblioteca digital	Filtros de búsqueda
ACM Digital Library SEDICI IEEexplore Springer	Intelligent tutoring systems; year range; mathematics o stem; affective intelligent tutoring systems.

Table 2. Sistemas Tutores Inteligentes analizados en este trabajo.

Nombre del STI	**Descripción**
Al - Tutor	Generación de preguntas y respuestas correctivas personalizadas basadas en la evaluación de diagnóstico cognitivo[29]

Aleks	Evaluación y Aprendizaje en Espacios de Conocimiento [17]
Auto-Tutor	SKOPE-IT: superposición de tutoría de lenguaje natural en un sistema de aprendizaje adaptativo para matemática [14][19][20]
SPOKE IT	SKOPE-IT: superposición de tutoría de lenguaje natural en un sistema de aprendizaje adaptativo para matemática [21]
Lexue 100	Evaluación de un sistema de tutoría inteligente para la enseñanza personalizada de matemática [8]
An Intelligent Math E-Tutoring System for Students with Specific Learning Disabilities(SLDs)	Un sistema inteligente de tutoría electrónica de matemática para estudiantes con discapacidades específicas de aprendizaje [7]
SIMPLIFY ITS	Un sistema de tutoría inteligente basado en modelos de diagnóstico cognitivo y aprendizaje espaciado [18]
MathSpring	Avances de la Oficina de Investigación Naval STEM Grand Challenge: expandiendo los límites de los sistemas de tutoría inteligentes [12]
ASSITments	Avances de la Oficina de Investigación Naval STEM Grand Challenge: expandiendo los límites de los sistemas de tutoría inteligentes [16]
WAYANG OUTPOST	Avances de la Oficina de Investigación Naval STEM Grand Challenge: expandiendo los límites de los sistemas de tutoría inteligentes [6]

3 Análisis de Sistemas Tutores Inteligentes

De acuerdo a los avances que han tenido los sistemas tutores inteligentes en los últimos años, se propone como aporte del trabajo, un nuevo enfoque al estudio de los sistemas tutores inteligentes, mediante la evaluación de cinco aspectos: aspectos generales, aspectos relacionados con el feedback del STI, aspectos metodológico-educativos, elementos de evaluación de los sistemas tutores inteligentes y aspectos relacionados con la arquitectura del STI.

A continuación, se explican cada uno de los criterios propuestos, los cuales son utilizados en el análisis de los STI, cuyos resultados se pueden ver en la tabla 3.

- **Aspectos generales:** Los criterios que incluyen esta categoría están vinculados a contextualizar los Sistemas Tutores Inteligentes y dar una caracterización general de ellos. A partir de estos indicadores se puede determinar, el tipo de artículo; conocer el país y universidad de origen; y el nivel educativo al cual se dirigen.

— Tipo de Artículo: Este criterio busca determinar el formato de referencia en que fue publicado. Si en conferencia, en un journal, en un workshop, o si corresponde o pertenece al capítulo de un libro, chapter.

— País de investigación:Este criterio busca indagar los países en que se llevan adelante las experiencias con los STI seleccionados. Luego, se podrán resumir los países con mayor concentración de investigaciones encontradas y enfocadas en el desarrollo de sistemas tutores inteligentes. Los posibles valores del criterio serán los nombres de los países.

— Universidades de investigación: Este criterio busca indagar las universidades donde se investigan, prueban y/o desarrollan las experiencias con STI seleccionadas. Luego, se podrán resumir las universidades con mayor concentración de investigaciones encontradas y enfocadas en el estudio de los STI. Los posibles valores del criterio serán los nombres de las universidades.

— Nivel educativo: Los posibles valores del criterio son:

Educación Especial. Identifica las experiencias llevadas a cabo con personas de educación especial. En este caso, se detalla, si corresponde, las características particulares de la población destinataria.

Inicial. Esta etiqueta determina si los STI son aplicables a destinatarios comprendidos en edades entre los 3 y hasta los 5 o 6 años.

Primario. Esta etiqueta determina si los STI son aplicables a destinatarios, entre los 6 y hasta los 12 o 13 años.

Secundario. Esta etiqueta identifica los STI, donde los destinatarios son adolescentes y jóvenes, cuyas edades promedio se encuentran entre los 13 y 18 años.

Superior/universitario. Identifica los STI, donde los destinatarios se encuentran cursando en la universidad, una institución de educación superior o son investigadores de una institución educativa.

— Dominio: Este criterio indica el dominio de aplicación del sistema tutor inteligente (Matemática, lengua, ciencias, etc.).

- **Aspectos relacionados con el feedback del STI**. Los criterios incluidos en esta categoría buscan dar a conocer las estrategias y técnicas con las que se plantea desarrollar el STI, su posible vinculación con otros STI y su feedback con el alumno.
- **Aspectos metodológicos-educativos**. Al tratarse de STI orientados al proceso de enseñanza y aprendizaje, se busca conocer el tipo de proceso educativo que lleva adelante, y las metas que se propone alcanzar con su uso. Se pueden agrupar en funciones y herramientas del sistema tutor inteligente.

  Funciones del STI.

  — Determinar el nivel de diagnóstico inicial del estudiante. Se indicará el método utilizado para determinar el diagnóstico inicial.
  — Establecer una ruta personalizada de aprendizaje. Se indicará la metodología implementada, para guiar el aprendizaje del alumno en el STI:
  — Integrar al STI un agente pedagógico de lenguaje natural. Se indicará si el STI, tiene agente pedagógico de lenguaje natural, que interactúa con el estudiante. Es decir, si posee un interfaz con lenguaje natural.
  — Modelar el comportamiento de los estudiantes para determinar su estado emocional.

  Herramientas del STI.

  Un Sistema Tutor Inteligente debe adaptarse a las necesidades y preferencias del estudiante para que este obtenga mejores resultados.

  Es necesario contar con modelos computacionales que realicen el diagnóstico sobre el rendimiento de los estudiantes y que provean al STI de datos basados en predicción, para cambiar la estrategia de enseñanza cuando fuera necesario, o simplemente para recomendarle nuevos ejercicios y problemas [18], tales como redes neuronales, algoritmos genéticos, teoría de los espacios de conocimiento, etc.

- **Procesos de evaluación de los Sistemas Tutores Inteligentes**: El proceso de evaluación determina si los Sistemas Tutores Inteligentes seleccionados, han sido sometidos a un proceso de validación.

  La evaluación puede consistir en pruebas en laboratorios, verificación en ambientes controlados, y/o demostración en escenarios reales.
  A partir de este criterio se analizan las principales técnicas utilizadas y los resultados alcanzados.

  Los valores para este criterio son:
  — **Si / No**. Esta etiqueta se orienta a detallar si los STI seleccionadas fueron o no evaluados.
  — **Técnica de evaluación**. Esta etiqueta identifica el tipo de evaluación empleada en cada una de los STI; cualitativa, cuantitativa, cuasi-experimental.
  — **Enfoque de la evaluación**. Se indaga la finalidad de la evaluación.

Table 3. Características generales de los STI

Características de los STI					STI									
					Aleks	Ai- tutor	Assitments	Auto-Tutor	Lexue100	MathSpring	Simplify ITs	SLDs	Spoke IT	Wayang Outpost
Descripción general		Tipo de Artículo	Documento		X									X
			Conferencia			X					X	X		
			Simposio						X					
			Journal				X	X		X			X	
		País/Universidad	Japón	Instituto Nacional de Informática		X								
			China	Departamento de Tecnología Educativa					X					
			EEUU	Universidad de California	X									
				Universidad de Memphis				X					X	
				Universidad de Wisconsin								X		
				Universidad de Massachusetts						X				X
				Instituto Politécnico de Worcester y Universidad Carnegie Mellon			X							
			España	Centro de Investigación y Desarrollo de Galicia en Telecomunicaciones Avanzadas							X			
			Canadá	Universidad de Ottawa										X
		Dominio	Matemática		X	X	X		X	X		X	X	X
			Múltiples dominios					X						
			No especificado								X			
		Nivel educativo	Especial									X		
			Inicial											
			Primaria						X			X		
			Secundaria		X		X	X	X	X	X	X	X	X
			Superior/ Universitario											
Feedback con el STI		Feedback con otros STI			X		X	X		X	X		X	X
		Feedback con la Actividad	Sobre la tarea		X	X	X	X	X		X	X		X
			Procesamiento de la tarea		X	X	X	X	X		X	X		X
			Autoregulación			X		X	X		X	X		X
			Afectivo					X				X		X
Metodológico Educativo	Funciones del STI	Metodología para el diagnóstico inicial			X	X								
		Metodologia de personalización del aprendizaje			X	X			X					X
		Agente pedagógico de lenguaje natural				X						X		X
		Modelado del comportamiento de los estudiantes										X		X
	Herramientas del STI	Reglas de Clustering, árboles de desición, reglas de clasificación				X								
		Teoría del espacio de conocimiento			X									
		Algoritmos genéticos				X								
		Otras						X	X		X			X
Técnica de Evaluación			Cuantitativa		X		X						X	X
			Cualitativa		X	X	X	X		X	X	X		X
			Cuasiexperimental						X					

4 Conclusiones y trabajos futuros.

Del análisis de los trabajos seleccionados se desprenden las tendencias hacia las cuales avanzan los desarrollos de los STI. Estas tendencias involucran:

- La incorporación de agentes informáticos conversacionales, ayudando a la tutoría electrónica, a través del diálogo.
- El diagnóstico cognitivo, para determinar el nivel del estudiante en la materia que se está trabajando, generando una ruta de aprendizaje personalizada, de acuerdo a las necesidades del alumno.
- La integración de sistemas tutores inteligentes con fortalezas complementarias para potenciar el aprendizaje y permitir la complementariedad de los recursos de aprendizaje. Es el caso de SPOKE IT, y MathSpring. SKOPE-IT, integró AutoTutor [19] y ALEKS [13][17].

En términos de la taxonomía de Bloom, ALEKS (bucle externo) se enfoca principalmente en aplicar habilidades de matemática, mientras que las preguntas de AutoTutor (bucle interno) pueden ayudar a los estudiantes a comprender, analizar y evaluar conceptos matemáticos.

Al construir SKOPE-IT, se combinaron, ejemplos resueltos [21], autoexplicación [1], aprendizaje impulsado por un callejón sin salida [27] .

En el caso de MathSpring, integró Wayang Outpost y ASSISTments. Wayang Outpost [6][9], es un sistema de tutoría en línea que se enfoca en las habilidades de matemática para estudiantes de nivel secundario y ASSISTments es una plataforma que utilizan los profesores para asignar tareas digitales y actividades en el aula [16].

Es decir, Wayang Outpost, se ubica en el lazo externo, seleccionando los problemas apropiados a resolver por un estudiante. Mientras que ASSISTments, se ubica en el lazo interno, el tutor proporciona apoyo al estudiante dentro de la resolución de un problema, incluyendo la orientación paso a paso, la reflexión y revisión de la solución al final.

La utilización de sistemas híbridos reduce el esfuerzo en el desarrollo de los STI.

De la lectura de los textos se desprende que los STI tienen al estudiante como el centro del proceso educativo, siendo éste quien regula su aprendizaje.

Los hábitos de estudio autorregulados, se transforman entonces, en un elemento determinante para el éxito del proceso educativo.

También se observó que el uso del STI, como herramienta complementaria, reduce las diferencias entre alumnos de un mismo curso de matemática, al tiempo que da luces a los docentes sobre el estado de aprendizaje de cada uno de sus alumnos, permitiéndoles llevar un control detallado de sus dificultades y conocimientos alcanzados.

Las evaluaciones de los sistemas tutores inteligentes, llevadas a cabo por los autores de los trabajos muestran que los sistemas de tutoría, superan a los tutores no expertos e incluso podrían igualar a los tutores humanos expertos, en algunos temas [15][28].

Los resultados de este trabajo son un punto de partida para continuar con el desafío de mejorar el aprendizaje de los estudiantes en el área de matemática, donde se han

agravado aún más las deficiencias en el aprendizaje, como consecuencia de la pandemia del COVID 19.

Como trabajo futuro se plantea implementar el uso de alguno de los STI analizados, en alumnos de 1° año de educación secundaria , y estudiar su impacto en el aprendizaje.

Referencias

1. Aleven, V., McLaren, B., Roll, I., Koedinger, K. (2004). *Toward Tutoring Help Seeking.* In: Lester, J.C., Vicari, R.M., Paraguaçu, F. (eds) Intelligent Tutoring Systems. ITS 2004. Lecture Notes in Computer Science, vol 3220. Springer.
2. *Aprender*.https://www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/aprender
3. *Aprender 2019.* Disponible en: https://www.argentina.gob.ar/educacion/aprender2019. Accedido: Julio 2022.
4. *Aprender 2021*. Disponible en: https://www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/aprender/aprender-2021. Accedido: Julio 2022.
5. *Argentinos por la educación*. https://argentinosporlaeducacion.org/informes/ Accedido: Julio 2022.
6. Arroyo, I., Woolf, B.P., Burelson, W. et al. *A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition*, Metacognition and Affect. Int J Artif Intell Educ 24, 387–426 (2014). https://doi.org/10.1007/s40593-014-0023-y
7. ASSETS '21: The 23rd International ACM SIGACCESS Conference on Computers and Accessibility Virtual Event USA October 18 - 22, 2021
8. B. Zhang and J. Jia, "*Evaluating an Intelligent Tutoring System for Personalized Math Teaching*," 2017 International Symposium on Educational Technology (ISET), 2017, pp. 126-130, https://ieeexplore.ieee.org/abstract/document/8005404, Accedido: Julio 2022.
9. Beal CR, Walles, R, Arroyo, I, Woolf, BP. (2007). *On-line tutoring for math achievement testing: a controlled evaluation.* Journal of Interactive Online Learning, 6(1), 43–55.
10. Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-30139-4_22. Accedido: Julio 2022.
11. Cataldi Zulma, Fernando J. Lage."*Sistemas Tutores Inteligentes: Procedimientos, métodos, técnicas y herramientas para su creación*". (Este artículo es parte del PID Modelado del tutor basado en redes neuronales para un sistema tutor inteligente. SeCyT 2007-2008. UTN-FRBA EZINBA 639. Programa Incentivos código 25/C099).
12. Craig, S.D., Graesser, A.C. & Perez, R.S. *Advances from the Office of Naval Research STEM Grand Challenge: expanding the boundaries of intelligent tutoring systems.* IJ STEM Ed 5, 11 (2018). Accedido: Julio 2022

https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-018-0111-x

13. Falmagne, J. C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (Eds.). (2013). *Knowledge spaces: Applications in education*. Springer Science & Business Media.
14. Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). *AutoTutor: An intelligent tutoring system with mixed-initiative dialogue*. IEEE Transactions on Education, 48(4), 612-618.
15. Graesser, A. C., & D'Mello, S. (2012). *Emotions during the learning of difficult material*. In B. Ross (Ed.), The psychology of learning and motivation (Vol. 57, pp. 183–226). San Diego, CA: Elsevier.
16. Heffernan, Neil & Heffernan, Cristina. (2014). *The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching*. International Journal of Artificial Intelligence in Education. 24. 10.1007/s40593-014-0024-x.
17. N. L. Miller, J. E. Sanchez-Galan and B. E. Fernández, *"Use of an Intelligent Tutoring System for Mathematics by Students Who Aspire to Enter the Technological University of Panama,"* 2019 7th International Engineering, Sciences and Technology Conference (IESTEC), 2019, pp. 255-260, doi: 10.1109/IESTEC46403.2019.00-66. Accedido: Julio 2022
18. N. M. Villanueva, A. E. Costas, D. F. Hermida and A. C. Rodríguez, *"SIMPLIFY ITS: An intelligent tutoring system based on cognitive diagnosis models and spaced learning," 2018* IEEE Global Engineering Education Conference (EDUCON), 2018, pp. 1703-1712, https://ieeexplore.ieee.org/document/8363440. Accedido: Julio 2022
19. Nye, BD, Graesser, AC, Hu, X (2014). *AutoTutor and family: a review of 17 years of science and math tutoring. International* Journal of Artificial Intelligence in Education, 24(4), 427–469.
20. Nye, BD, Graesser, AC, Hu, X (2014a*). AutoTutor and family: a review of 17 years of science and math tutoring*. International Journal of Artificial Intelligence in Education, 24(4), 427–469.
21. Nye, B., Pavlik, P., Windsor, A. et al. *SKOPE-IT (Shareable Knowledge Objects as Portable Intelligent Tutors): overlaying natural language tutoring on an adaptive learning system for mathematics*. IJ STEM Ed 5, 12 (2018). https://doi.org/10.1186/s40594-018-0109-4. Accedido: Julio 2022.
22. Perkins, D. (1995) *La escuela inteligente*. Gedisa.
23. Pozo, J. I. (1998*). Aprendices y maestros*. Alianza
24. Programme for International Student Assessment (PISA).https://www.oecd.org/pisa/data/2018database/. Accedido: Julio 2022.
25. Schwonke R, Renkl, A, Krieg, C, Wittwer, J, Aleven, V, Salden, R (2009). *The worked-example effect: not an artefact of lousy control conditions*. Computers in Human Behavior, 25(2, SI), 258–266.
26. *UNESCO. 2022. Reimaginar juntos nuestros futuros*: Un nuevo contrato social para la educación. https://es.unesco.org/futuresofeducation/cumbre-sobre-la-transformacion-de-la-educacion. Accedido: Julio 2022.

27. VanLehn, K, Siler, S, Murray, C, Yamauchi, T, Baggett, WB (2003). *Why do only some events cause learning during human tutoring? Cognition and Instruction,* 21(3), 209–249.
28. VanLehn, K. (2011). *The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems*. Educational Psychologist, 46,97–221.
29. W. Gan, Y. Sun, S. Ye, Y. Fan and Y. Sun, *"AI-Tutor: Generating Tailored Remedial Questions and Answers Based on Cognitive Diagnostic Assessment,"* 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), 2019, pp. 1-6, doi: 10.1109/BESC48373.2019.8963236.

L.A.Z: Un Lenguaje Específico del Dominio para la Generación Automática de Sitios Web de Instituciones Escolares

Analía Zaldúa[1], Mario Berón[1], Julieta Gatica[1], and Mariano Luzza[1]

[1] Universidad Nacional de San Luis,
Ejército de los Andes 950, San Luis, Argentina
{amzaldua,mberon,mluzza}@unsl.edu.ar
jag81295@gmail.com

Abstract. Antes de la revolución tecnológica actual, los avisos institucionales se debían realizar con la suficiente antelación para que toda la comunidad educativa estuviera debidamente informada, aunque no se disponía de los medios ni del tiempo suficiente para hacerlo. En la actualidad esto cambió debido al avance de las Tecnologías de la Información y de la Comunicación. Se puede observar a nivel global que cada institución tiene presencia en internet a través de un sitio web que la representa y donde realiza la comunicación de los eventos más importantes. La realidad mencionada no sucede con todas las instituciones educativas dado que no cuentan con los recursos necesarios para construir su propio sitio web. Con el fin de solucionar este inconveniente se desarrolló LAZ, un lenguaje específico del dominio que permite especificar un espacio institucional utilizando el vocabulario empleado por los miembros de la institución y cuyo procesador genera automáticamente el entorno web institucional.

1 Introduction

Las instituciones educativas regionales de gestión pública brindan un servicio muy loable a la sociedad como lo es la educación gratuita y de calidad. Para poder llevar adelante dicha tarea, todo el personal realiza mucho esfuerzo para cumplir con los objetivos que propenden a que los estudiantes aprendan significativamente. Entre estas tareas se encuentra la de comunicar información relevante que la institución necesita notificar a su comunidad educativa para estar al tanto de los hechos que acontecen en la misma. Tradicionalmente, este tipo de actividades se lleva a cabo mediante notas en el cuaderno de comunicaciones que los estudiantes hacen firmar por sus tutores, de forma tal, de que conozcan las actividades institucionales. Si bien esta forma de llevar adelante la tarea ha funcionado, muchas noticias e información no llegan a conocerse hasta tanto el interesado se acerque a la institución o bien lea documentos asociados a la misma. A modo de ejemplo se pueden mencionar las noticias de último momento, las orientaciones disponibles en la institución, las fechas de inscripciones que se conocen cerca del inicio de clases, entre otras tantas posibilidades. Con el propósito de resolver este

inconveniente, las instituciones han evolucionado incorporando tecnologías de punta, las cuales han resuelto en gran medida los problemas antes mencionados. En la actualidad muchas instituciones tienen su propio sitio web en donde, cada tanto, van actualizando las noticias que el equipo de gestión desea que la comunidad educativa conozca. Esta aproximación ha resultado un adecuado método de comunicación, sin embargo, aún presenta un inconveniente de difícil solución: *las instituciones que poseen dicho avance tecnológico son, en general, instituciones privadas. Las mismas tienen sustento económico como para pagar a un profesional para que desarrolle el sitio y, si es necesario, lo actualice. Las instituciones públicas adolecen de ésta peculiaridad, con lo cual el diseño de un sitio web es una tarea que puede realizar, de buena voluntad, un docente de la misma, hecho poco probable dado que los mismos están dedicados tiempo completo a la tarea de educar*. Si, por el contrario, un integrante de la comunidad educativa desarrolla el sitio web, la tarea de actualización estaría a cargo de la misma persona dado que, por lo general, será un profesional o idóneo de la informática. Claramente, ninguno de los enfoques mencionados en los párrafos precedentes es viable en las instituciones educativas públicas porque no se disponen de costos ni de idóneos en informática. Para resolver este inconveniente, se desarrolló LAZ (**L**enguaje de **A**nalía **Z**aldúa). LAZ es un lenguaje específico del dominio cuyo propósito es facilitar la creación y actualización de sitios web institucionales, utilizando terminología empleada en las instituciones educativas. Este lenguaje permite especificar la información que estará disponible en el sitio institucional, el cual será generado por P-LAZ (El **P**rocesador de **LAZ**) una herramienta que toma como entrada una especificación LAZ y produce como salida una página web.

El artículo está organizado como se describe a continuación. La sección 2 describe los trabajos relacionados y los proyectos vigentes pertinentes a la temática. La sección 3 presenta el lenguaje específico del dominio LAZ. En la sección 4 se exhibe el procesador de LAZ (P-LAZ) y su arquitectura. La sección 5 presenta un ejemplo de aplicación. Finalmente, en la sección 6 se exponen las conclusiones del artículo y el trabajo futuro.

2 Trabajos Relacionados

Existen numerosas investigaciones referidas al uso, desarrollo e implementación de lenguajes específicos del dominio. En esta sección se describen los principales trabajos relacionados con la temática abordada en el artículo.

L. Cardelli y R. Davies [6] en su investigación presentan un lenguaje de programación para el procesamiento de documentos web llamado WebL. WebL es un lenguaje de scripting orientado a objetos de alto nivel que incorpora dos características novedosas: combinadores de servicios y un álgebra de marcado. Los combinadores de servicios son construcciones de lenguaje que brindan acceso confiable a los servicios web imitando el comportamiento de un internauta cuando se produce una falla al recuperar una página. Bergstra y P. Klint [5] describen cómo se puede usar un lenguaje basado en álgebra de procesos para la descripción

de la arquitectura de coordinación de sistemas de software heterogéneos y distribuidos. Wang, A. W. Appel, J. L. Korn [12] presentan el lenguaje de descripción de sintaxis abstracta Zephyr (ASDL). En su trabajo describen la sintaxis abstracta de las representaciones intermedias del compilador y otras estructuras de datos en forma de árbol. Así como las estructuras léxicas y sintácticas de los lenguajes de programación se describen con expresiones regulares y gramáticas libres de contexto. Giulianelli et. al. en [9] explican una propuesta que permite dentro del enfoque MDA (Model-Driven Architecture) utilizar UML y LEDs en distintos niveles de abstracción y generar mediante transformaciones el código fuente de una determinada aplicación. Esto se debe a que en algunos trabajos académicos surge la disyuntiva de utilizar UML (Unified Modeling Language) ó LED (Lenguaje Específico del Dominio) para modelar un determinado artefacto. Álvarez Herrero [4] en su trabajo sobre las páginas web de los centros educativos y su análisis de la situación en la Comunidad Valenciana refleja que las páginas web se han convertido en un requisito imprescindible de cualquier centro educativo. Más allá de ser plataformas para la información y la comunicación de toda la comunidad educativa, en la actualidad estas páginas web son también escaparates de los centros hacia la sociedad pero que presentan diferentes falencias y la mayoría de los problemas vienen motivados por la falta de una correcta gestión y mantenimiento de las páginas web. Luzza y su equipo de investigación en [11] presentaron PH-Asistido, una herramienta cuyo principal objetivo consiste en facilitar la enseñanza de la programación a través de la utilización de un dominio atrayente para este propósito: el Proyecto Hoshimi. PH-Asistido es una extensión del Lenguaje del Proyecto Hoshimi con acciones que simplifican la definición e implementación de un algoritmo para el Proyecto Hoshimi. Esto es llevado a cabo a través de un editor visual proactivo que asiste al usuario. La herramienta provee acciones propias del dominio, evitando que el estudiante disperse su atención en obstáculos poco relacionados con el problema a resolver, como la sintaxis del lenguaje. El editor también facilita la parametrización de dichas acciones, acotando las variables al tipo correcto, evitando así otro problema común a la hora de iniciarse en la programación. Gatica y su equipo de investigación [8], presentaron Vinculación 3.0, una herramienta que provee, a las distintas instituciones de nivel medio, terciarias o universitarias, un espacio para compartir su contenido. Esto lo hace a través de una biblioteca digital, una mesa de ayuda y de EDUTEC, un subsistema sencillo que genera páginas web con características básicas de forma automática para las instituciones secundarias. Un área de especial aplicación para los LED es en el ámbito de la enseñanza en programación, ya que posibilitan abstraerse de los problemas particulares de los lenguajes de Propósito General para centrarse en el tema particular que se desea enseñar [10] tal como lo intenta hacer PH-Asistido. En este sentido se pueden mencionar: Scratch [1,3], Alice [2,7], Logo [14,13] son lenguajes específicos del dominio que tienen como finalidad la enseñanza de la programación y son muy utilizados en la academia.

Como es posible percibir, a partir de los trabajos descritos en los párrafos precedentes, los LEDs tienen innumerables aplicaciones. Sin embargo hasta el

momento no fue posible encontrar herramientas que generen espacios institucionales siguiendo el enfoque presentado en este artículo el cual es: *Definir un lenguaje (y su porcesador correspondiente) estrechamente relacionado con el que se usa en las instituciones educativas para generar automáticamente sus páginas web. El enfoque además tiene como finalidad facilitar el mantenimiento del espacio web generado.*

3 LAZ: Lenguaje de Analía Zaldúa

LAZ (**L**enguaje de **A**nalía **Z**aldúa) es un lenguaje específico del dominio cuyo propósito es facilitar la creación y actualización de sitios web institucionales. Las palabras claves del mismo, están estrechamente vinculadas con el dominio del problema lo que facilita su aprendizaje y utilización. Esto se debe a que, al utilizar palabras que se emplean en el contexto institucional educativo, los usuarios (directivos, docentes, personal administrativo) se sentirán familiarizados con el lenguaje. LAZ es el punto de partida de diferentes tareas de procesamiento de lenguajes que traducen una especificación en código HTML que especifica el sitio web institucional. Básicamente, una página web institucional debe mostrar el nombre de la institución, información respecto de la misma y del personal. Esto puede ser especificado gramaticalmente de la siguiente manera:

$$\begin{aligned} \text{institución} \rightarrow\ & < COMIENZO - INSTITUCIÓN > \\ & \text{quienesSomos informacionInsitucional} \\ & \text{personal} \\ & < FIN - INSTITUCIÓN > \end{aligned}$$

donde el no terminal *quienesSomos* es una cadena de caracteres a través del cual la institución da a conocer sus objetivos principales, las bases de su formación, sus valores, entre otras características. El no terminal *informaciónInstitucional* sigue una lógica parecida al no terminal *quienesSomos*; es decir, es una cadena de caracteres compuesta por una o más letras, números y caracteres especiales.

El caso de *personal* es un tanto más complejo dado que, en una institución educativa, se cuenta con diferentes clases de personal. Por una parte, se tiene al equipo de conducción, el cual está formado por un director, un vicedirector y un regente, por otra parte están los docentes y el personal y administrativo. Estas características se especifican de la siguiente manera:

$$\text{personal} \rightarrow \text{directivos docentes administrativos}$$

La información asociada a cada miembro de la institución va a depender de lo que las instituciones consideren importante. Atento a este aspecto, se consultó a un grupo de directivos y se concluyó que para el caso de: i) *Directivos*, se registra el nombre, cargo y el correo electrónico o información de contacto; ii) *Profesores*, el nombre, materia, y el correo electrónico de contacto; iii) *Personal administrativo*, el nombre y el correo electrónico de contacto. La especificación gramatical de la información antes mencionada es la siguiente:

directivos→ <COMIENZO–DIRECTIVOS> director+ <FIN–DIRECTIVOS>
director→ <COMIENZO–DIRECTOR> nombre cargo contacto <FIN–DIRECTOR>
docentes→ <COMIENZO–DOCENTES> docente+ <FIN–DOCENTES>
docente→ <COMIENZO–DOCENTE> nombre materia contacto <FIN–DOCENTE>
administrativos→ <COMIENZO–ADMINISTRATIVOS>
administrativo+
<COMIENZO–ADMINISTRATIVOS>
administrativo→ <COMIENZO–ADMINISTRATIVO>
nombre contacto
<FIN–ADMINISTRATIVO>

Los no terminales *nombre*, *materia* y *contacto* generan un STRING. A continuación, se muestran todas las producciones que conforman el lenguaje LAZ:

institución→ $< COMIENZO - INSTITUCIÓN >$
quienesSomos informacionInsitucional
personal
$< FIN - INSTITUCIÓN >$
informaciánInstitucional→ STRING
personal→ directivos docentes administrativos
directivos→ <COMIENZO–DIRECTIVOS> director+ <FIN–DIRECTIVOS>
director→ <COMIENZO–DIRECTOR>
nombre cargo contacto
<FIN–DIRECTOR>
docentes→ <COMIENZO–DOCENTES> docente+ <FIN–DOCENTES>
docente→ <COMIENZO–DOCENTE>
nombre materia contacto
<FIN–DOCENTE>
administrativos→ <COMIENZO–ADMINISTRATIVOS>
administrativo+
<COMIENZO–ADMINISTRATIVOS>
administrativo→ <COMIENZO–ADMINISTRATIVO>
nombre contacto
<FIN–ADMINISTRATIVO>
quienesSomos→ STRING
nombre→ STRING
materia→ STRING
contacto→ STRING

4 PLAZ: Procesador de LAZ

La figura 1 muestra la arquitectura de P-LAZ. Como se puede observar la arquitectura consta de tres capas: la *Capa de Presentación*, la *Capa de Procesamiento de Lenguajes* y la *Capa de Generación Web*. Esta división se ha realizado con el

propósito de independizar los componentes de forma tal que una modificación en uno de ellos implique pocas o ninguna modificación en el otro.

La *Capa de Presentación* consta de una componente denominada *E-LAZ* (Editor de LAZ) la cual implementa un *Editor Específico del Dominio* que facilita el desarrollo de especificaciones LAZ. El editor posee facilidades para introducir patrones de especificaciones. Esta característica es útil porque reduce la cantidad de errores de sintaxis que el usuario puede introducir a medida que va construyendo la especificación de una institución.

La *Capa de Procesamiento* de Lenguajes está compuesta por los módulos *Lexer*, *Parser* y *Extractor de Información*. El *Lexer* recibe como entrada una especificación LAZ (generada por E-LAZ) y realiza el particionado de la misma en tokens, los cuales son pasados al *Parser* para que se lleve adelante el análisis sintáctico de la especificación. Si la especificación contiene un error sintáctico, el sistema indicará que el árbol de parse no pudo ser construido y le pedirá al usuario que corrija la especificación. En caso contrario, la salida del análisis sintáctico es el árbol de parse correspondiente a la entrada, el cual, a su vez, es la entrada al *Extractor de la Información*. Es importante mencionar que el procesador de lenguaje además de realizar el análisis sintáctico también realiza análisis semántico. Momento en el cual se pueden detectar errores tales como *demasiado directores*, *institución educativa sin docentes*, entre otros tantos errores. El *Extractor de Información* tiene definidos varios recorridos sobre el *Árbol de Parse*[1] los cuales tienen como finalidad extraer la información requerida por la *Capa de Generación Web*. La información recuperada por el *Extractor de Información* consta de la información institucional plasmada en la especificación la cual fue establecida en base a la experiencia de trabajo en el dominio del problema del equipo de docentes que colaboró en la definición de LAZ. A modo de síntesis se menciona que los recorridos recuperan y almacenan la siguiente información: *Nombre de la Institución*, *Breve Descripción de la Institución*, el *Personal* (directivos, docentes, administrativo), y una *Lista de Noticias*.

La *Capa de Generación Web* consta de varias componentes que colaboran entre si para generar el código html que define la página web. Tal como fue mencionado en secciones previas, estos módulos hacen uso de la información almacenada en estructuras de datos internas de P-LAZ. Nótese que la generación de código de estas componentes puede ser llevada a cabo utilizando cualquier técnica, considerando que mientras más simple sea más sencillo será comprender y modificar [4].

5 Ejemplo de Aplicación

A modo de prueba se aplicó LAZ y P-LAZ a un colegio estatal de la provincia de San Luis, obviamente se ha resumido la información que se pretende dar a conocer para que el artículo quede autocontenido. La institución seleccionada es el Colegio N° 13 Profesor Roberto Moyano que se encuentra ubicado en la ciudad

[1] Ésta estructura de datos la genera el parser cuando la especificación es correcta

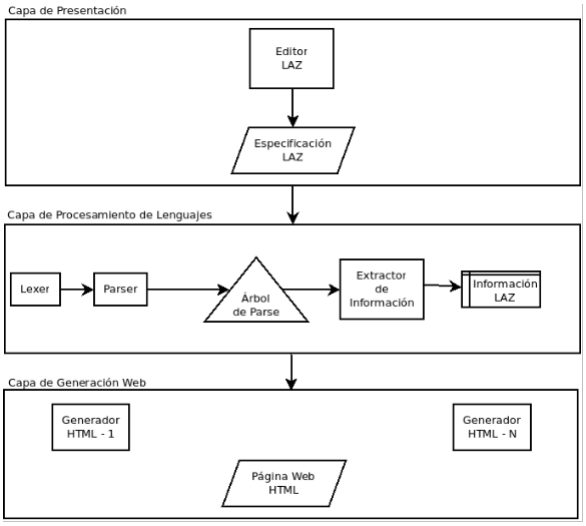


Fig. 1. Arquitectura de P-LAZ

de Juana Koslay en la Provincia de San Luis. El primer paso para comenzar a utilizar la herramienta es recopilar la información que una especificación LAZ necesita de forma obligatoria, a saber: el nombre de la institución, información respecto de quiénes somos, información institucional y personal.

5.1 Especificación LAZ de la Institución

Para comenzar a desarrollar una especificación LAZ el usuario hace uso de E-LAZ, con el cual se genera el patrón de código de una institución sin errores y el usuario solo debe completar con la información de la institución. Para el ingreso de información institucional se debe primero presionar el botón *Nueva Institución* ubicado en la barra de menú y luego completar los campos con la información institucional adquirida (ver figura 2).
Especificación de Personal: Para el ingreso del personal el usuario debe situarse en la sección personal[2], luego se debe ingresar el tipo de personal que se desea. En este caso se tienen tres tipos de personal: *directivo*, *docente* y *administrativo*. Para ingresar cada uno de esos tipos el usuario simplemente debe posicionarse en el lugar correcto es decir dentro de la sección personal y fuera de un personal específico y luego presionar el botón correspondiente al tipo de personal deseado y completar los datos requeridos. En la figura 2 se muestra la pantalla de P-LAZ con información institucional y un directivo especificado.
Ingreso de Noticias: Para el ingreso de noticias el usuario debe en primer lugar crear un bloque de noticias presionando el botón *Agregar Noticias* y

[2] Identificado por los terminales $< COMIENZO - PERSONAL >$ y $< FIN - PERSONAL >$

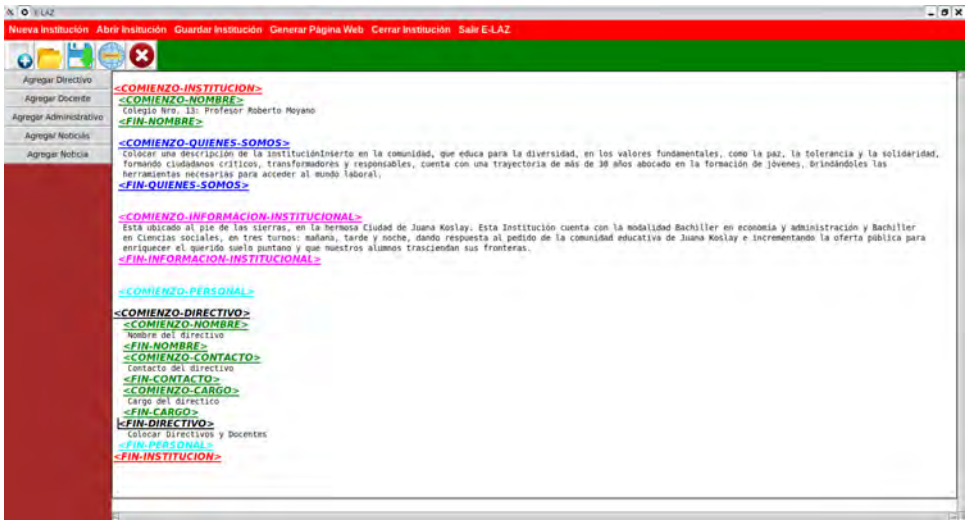


Fig. 2. Patrón LAZ generado con información

seguidamente colocarse dentro del bloque de noticias creado con anterioridad y agregar una noticia presionando el botón *Agregar Noticia*. En este bloque de noticias se pueden incorporar tantas noticias como el usuario desee (ver figura 3).

Fig. 3. Patrón LAZ generado con información

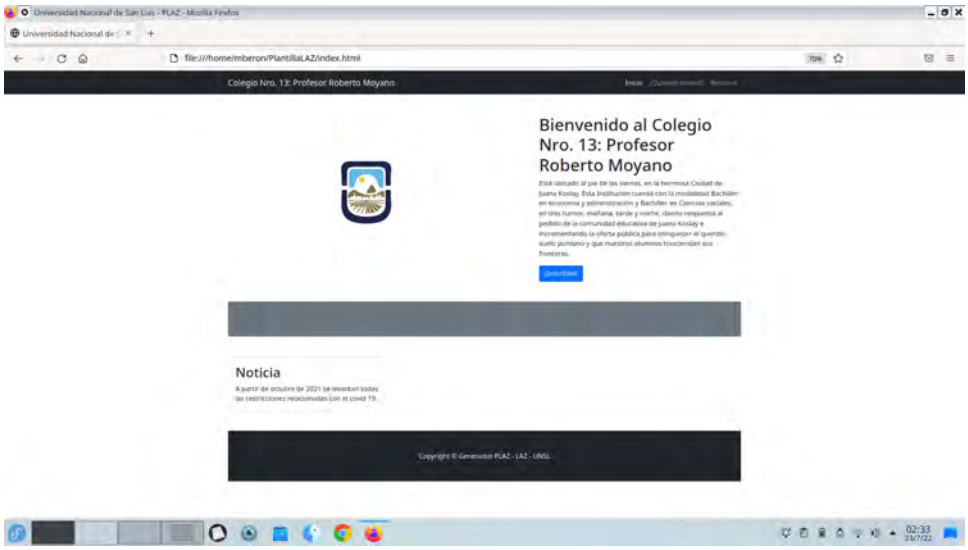


Fig. 4. Resultado del Generador Web

Generación de la Página Web: Una vez construida toda la especificación para la institución se está en condiciones de invocar al generador web. Esta tarea es simple de realizar y se lleva adelante presionando el botón destinado para tal fin (*Generar Página Web*). Luego de realizada esta actividad, si la especificación es correcta, el generador produce como resultado archivos html entre los cuales se encuentra *index.html*, el cual puede ser abierto por un navegador para observar su contenido de la página web especificada por el usuario. La figura 4 muestra el inicio de presentación de la institución.

6 Conclusiones

En este artículo se presentó una investigación que tiene como objetivo: Desarrollar una herramienta que genere espacios web institucionales de manera automática a partir de una especificación escrita en un lenguaje específico del dominio. El espacio web generado es básico e incluye: información de la institución en general, noticias, información académica y contacto con la misma. Para cumplir el objetivo mencionado, se llevaron a cabo las siguientes tareas: i) Se definió LAZ un lenguaje específico del dominio que utiliza terminología propia de las instituciones educativas y que posee construcciones sencillas que facilitan el uso del mismo por parte de los usuarios para especificar los espacios web institucionales; ii) Se implementó P-LAZ un procesador del LAZ el cual además de construir los componentes necesarios para el análisis sintáctico y semántico del lenguaje LAZ provee funcionalidades que facilitan, los miembros de la institución,

la construcción de especificaciones. Además posee una componente que permite generar automáticamente el espacio institucional.

La herramienta presentada en este artículo fortalece el proceso comunicacional entre la comunidad educativa y la institución escolar que es considerado un déficit teniendo en cuenta el contexto actual. El proceso de comunicación mostrará la institución educativa en un portal web de creación y mantenimiento sencillo y económico lo que justifica la importancia de automatizar los procesos.

Como trabajos futuros se pueden mencionar: i) Incrementar la expresividad de LAZ para que las instituciones puedan comunicar más información; ii) Autogestión del ambiente de trabajo, en cuanto a la elección de fuentes, formatos, colores del ambiente y que la herramienta sea adaptativa a los requerimientos y elecciones del usuario.

References

1. http://scratch.mit.edu.
2. https://www.alice.org/.
3. Achal. *Teach Yourself Animation Coding in Scratch 3: Programming for Kids and Beginners*.
4. Juan-Francisco Alvarez-Herrero and Rosabel Roig-Vila. Las pÁginas web de los centros educativos. anÁlisis de la situaciÓn actual en la comunidad valenciana - the websites of schools. analysis of the current situation in the valencian community. *Revista de Comunicación de la SEECI*, pages 129–147, 11 2019.
5. Jan A. Bergstra and Paul Klint. The discrete time toolbus—a software coordination architecture. *Science of Computer programming*, 31(2-3):205–229, 1998.
6. Luca Cardelli and Rowan Davies. Service combinators for web computing. *IEEE Transactions on Software Engineering*, 25(3):309–316, 1999.
7. Wanda P Dann, Don Slater, Laura Paoletti, and Dave Culyba. *Alice 3 to Java: Learning Creative Programming through Storytelling and Gaming*. Pearson, 2017.
8. Julieta Gatica and Camila Olguín. *Vinculación Educativa 3.0: Una herramienta para disminuir las brechas educativas regionales*. Proyecto final integrador de ingeniería en informática, 2021.
9. Daniel Alberto Giulianelli, Claudia Pons, Rocío Andrea Rodríguez, Pablo Martín Vera, and Víctor Fernandez. Integrando uml y dsl en el enfoque mda. In *XVI Congreso Argentino de Ciencias de la Computación*, 2010.
10. David A Ladd and J Christopher Ramming. Application languages in software production. In *USENIX 1994 Very High Level Languages Symposium (USENIX 1994 Very High Level Languages Symposium)*, 1994.
11. Mariano Luzza, Mario Marcelo Beron, and Pedro Rangel Henriques. Ph-helper-a syntax-directed editor for hoshimi programming language, hl. In *1st Symposium on Languages, Applications and Technologies*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.
12. Daniel C. Wang, Andrew W. Appel, Jeff L. Korn, and Christopher S. Serra. The zephyr abstract syntax description language. In *IN PROCEEDINGS OF THE CONFERENCE ON DOMAIN-SPECIFIC LANGUAGES*, pages 213–227, 1997.
13. Daniel Watt. *Learning with Logo*.
14. Molly Watt and Daniel Watt. *Teaching with Logo: Building blocks for learning*. Addison-Wesley Longman Publishing Co., Inc., 1985.

Aprendiendo a desarrollar un intérprete de un lenguaje de programación funcional

Lucas Spigariol[1234], Juan Bono[13], Francisco Sanchez Guijarro[1]
[1]Universidad Tecnológica Nacional
[2]Universidad Nacional de Gral San Martín
[3]Universidad de Buenos Aires
[4]Universidad Nacional de Hurlingham
{lspigariol, juanbono94, franleplant}@gmail.com

Abstract. Se presenta una herramienta de software, cuyo desarrollo está en progreso, que tiene como objetivo facilitar la comprensión del proceso de interpretación de un lenguaje de programación de alto nivel. Su elemento central es un intérprete propiamente dicho y está articulado con una serie de componentes gráficos que permiten visualizar las estructuras intermedias y hacer un seguimiento de los procesos internos que va realizando dicho intérprete, con un sentido pedagógico. Se apunta a un estudiante inicial que quiera aprender sobre el funcionamiento interno de los lenguajes y de la teoría que lo sustenta. Se permite también un uso avanzado, donde ya conociendo el intérprete, el estudiante pueda intervenir modificando o extendiendo la definición del mismo lenguaje.

Keywords: Intérpretes, Lenguajes de programación, Didáctica de la programación, Software educativo.

1 Presentación

El objetivo del software educativo en el que se centra el presente trabajo es facilitar la comprensión del proceso de interpretación de un lenguaje de programación de alto nivel. La principal motivación es la constatación de las dificultades de los estudiantes de comprender y poner en práctica los conceptos involucrados en el funcionamiento interno de los lenguajes de programación y la confianza en que una herramienta de software apropiada puede facilitar el proceso de aprendizaje.

Se trata de un trabajo en progreso, en el marco de un proyecto de investigación, que comienza con un diagnóstico de la situación de enseñanza y aprendizaje en relación a la temática en base al que se plantea el diseño y desarrollo de un software educativo.

Actualmente, se encuentra muy avanzado, al punto de contar con una versión funcionando del software que si bien presenta algunos aspectos pendientes se encuentra disponible para ser utilizada por estudiantes en situaciones concretas.

El diagnóstico realizado permitió identificar los aspectos principales a abordar pedagógicamente y definir el alcance y enfoque de la herramienta de software a desarrollar. Entre ellos, se destacan:

- Dificultades en la comprensión de cada uno de los procesos internos que se realizan sobre el código y del funcionamiento de manera integral.
- Necesidad de un soporte concreto para articular teoría y práctica.
- Importancia de la visualización de los procesos y estructuras de datos.

La etapa de diseño y desarrollo partió de elegir *Lisp* como lenguaje a interpretar, valorando que ya de por sí presenta un modelo relativamente simple y con la flexibilidad suficiente para seleccionar o adaptar diferentes aspectos.

Apuntando a un perfil de estudiante de las materias iniciales de carreras de informática, la herramienta consta como componente central de un intérprete propiamente dicho del lenguaje *Lisp*. La interacción básica se da mediante un editor en el que el estudiante ingresa su código fuente y una consola *REPL* en la que se evalúa dicho código y se obtiene el resultado. De manera simultánea, se visualizan gráficamente las principales representaciones intermedias que se generan y se permite ir avanzando paulatinamente en la ejecución de los procesos internos.

A su vez, previendo un estudiante con más experiencia o materias más avanzadas, se tuvieron en cuenta otros elementos. Se definió un subconjunto del lenguaje *Lisp* que se interpreta, dejando adrede funcionalidades sin implementar para permitir que un estudiante pueda completar o modificar el mismo intérprete y visualizar su funcionamiento con la misma interfaz. También se asumieron criterios de desarrollo y una arquitectura que facilita la modificación puntual de ciertos componentes del software sin necesidad de comprender toda su estructura.

Una decisión clave fue focalizar en los mecanismos de interpretación del código fuente, en particular superar las fases de análisis léxico y gramatical y llegar a abordar el proceso de evaluación, y no necesariamente lograr la generación de código ejecutable o plantear estrategias de optimización.

Por último, otro aspecto que se tuvo en cuenta como horizonte de trabajo, desde la convicción de la importancia estratégica del desarrollo de software de base y del rol de la universidad en el ambiente profesional de sistemas de información, fue dejar abierta la posibilidad de profundizar e ir un poco más allá de los contenidos mínimos que establecen los planes de estudio vigentes.

2. Diagnóstico

El contexto educativo en el que se realizó el estudio es la carrera de Ingeniería en Sistemas de Información de la Facultad Regional Delta de la Universidad Tecnológica Nacional. A la propia experiencia personal de docentes y estudiantes que conforman el equipo de investigación se sumó un relevamiento realizado entre los estudiantes de la carrera como así también a docentes de la misma facultad y de otras casas de estudios.

Dentro del plan de estudios, si bien no hay ninguna asignatura específica, ya sea obligatoria u optativa, sobre creación de lenguajes o que tenga como tema central el

estudio de intérpretes o compiladores, existe una materia en el segundo año de la carrera denominada "Sintaxis y semántica de Lenguajes", que es la que más se aproxima a la temática abordada y donde se gesta la motivación por llevar adelante la presente investigación.

2.1. Intuiciones preliminares

Analizar el funcionamiento de un intérprete o compilador de un lenguaje de programación de alto nivel es una tarea que se puede abordar desde un punto de vista estrictamente teórico, pero como tantos otros aspectos de la informática, la posibilidad de experimentarlo en la práctica, de probar cómo funciona, ayuda a la comprensión de los conceptos.

Una primera aproximación práctica puede realizarse con ejemplos simples en los típicos recursos áulicos de "papel" o "pizarrón". Allí, la ventaja es que el lenguaje a analizar no tiene por qué ser real sino que es posible utilizar una definición de lenguaje hecha especialmente con un fin didáctico, con un conjunto de reglas sintácticas y gramaticales propias, como así también tomar un lenguaje existente y adaptarlo o acotarlo adecuadamente. También permite abordar todo el proceso de compilación o sólo alguna de sus etapas, o hacerlo con diferentes niveles de profundidad.

Un camino diferente, tratándose de código, consiste en utilizar un lenguaje de programación de uso profesional, con su correspondiente compilador o intérprete, y ver en acción los procesos que permiten que finalmente se ejecute un programa escrito en dicho lenguaje, sobre todo al avanzar en casos de mayor complejidad. Lo que sucede, es que no sólo hay que asumir el lenguaje completo tal cual es, sino que generalmente sus compiladores están pensados para realizar eficientemente el proceso y lograr el resultado, pero no dan cuenta de qué manera lo obtuvieron. A su vez, analizar la documentación o su propio código fuente para entender su funcionamiento interno no siempre es posible, no es tarea sencilla y es discutible su sentido educativo.

Una de las intuiciones que fundamenta este trabajo es la posibilidad de hacer converger las ventajas de ambas perspectivas mediante la creación de un intérprete propio de un lenguaje de programación de alto nivel, que ejecute realmente el código fuente como lo hacen los intérpretes profesionales y que a la vez permita focalizar en aquellos aspectos conceptuales que se consideren más significativos y de esta manera sostener un proceso de enseñanza y aprendizaje articulando teoría y práctica.

2.2 Relevamiento

Un relevamiento realizado entre estudiantes y docentes permitió descubrir algunas tendencias en cuanto a dificultades o limitaciones y constatar algunas ideas previas. Los estudiantes consultados son todos de la misma facultad, mientras que entre los docentes hay también quienes provienen de otras instituciones universitarias, en asignaturas con contenidos afines. El instrumento de recolección de datos utilizado fue el mismo para todos, aunque en el caso de los docentes que accedieron a

participar se lo complementó con otras preguntas abiertas que permitieron recavar información caulitativa de interés.

Respecto del proceso de enseñanza y aprendizaje, es contundente la percepción que se trata de un tema complejo y que presenta dificultades. Un primer grupo destaca haber aprendido las etapas de compilación y hacer algunos ejercicios sobre ellas, como por ejemplo graficar autómatas en papel, reconoce que se abordan los elementos y etapas que conforman un lenguaje, pero de manera superficial. Otros señalan que se enseñan los conceptos y partes de un compilador y que se aprende a construir un analizador léxico y un analizador sintáctico mediante un lenguaje de programación. En ese sentido, valoran haber tenido una experiencia de tener que realizar un trabajo práctico donde se implementa alguno de los algoritmos estudiados. Un recurso utilizado es el desarrollo de componentes en relación a la definición de pequeños lenguajes. “Hicimos proyectos interesantes que nos permitieron comprender, por lo menos los primeros pasos, de lo que es crear un lenguaje de programación desde cero”, señala uno de los estudiantes consultados.

En relación a la ubicación de la materias dentro del plan de estudios se evidencia que el grado de profundidad que se puede lograr está limitado en gran parte por la poca experiencia previa en programación que tienen los estudiantes. Esto hace complejo no sólo poder llegar a una implementación práctica, sino que también muestra una falta de contextualización que ayude a interpretar el sentido de los conceptos. Lo cuenta con elocuencia un graduado de la carrera: “vi lo básico de compiladores en un momento de mi vida que no sabía muy bien de qué se trataba programar”.

Consultados puntualmente por las dificultades, las respuestas se pueden organizar mayoritariamente en tres tópicos. Por un lado, en entender el para qué del tema, lo cual guarda relación con la mencionado anteriormente sobre la ubicación de la asignatura en la carrera. Otro aspecto, el más mencionado, se refiere a la complejidad y carácter abstracto de los algoritmos y las estructuras de datos utilizadas; haciendo mención de diversos conceptos como por ejemplo la recursividad, el *backtracking* o los árboles. Un tercer elemento destacable es que aún quienes lograban seguir el funcionamiento de cada componente de manera individual -generalmente los primeros pasos- manifestaban dificultad para comprender el proceso de manera integral. Por último, el alcance predominante de la asignatura incluye el análisis lexicográfico, hay numerosas afirmaciones del estilo "vemos un *parser*", pero que no se avanza a las siguientes etapas, en particular, aquellas que permitan ver un resultado final.

Entre los docentes consultados se manifiesta lo valioso de contar con herramientas pensadas para estudiantes en vez de las que son diseñadas para profesionales, en las que cada docente elige dónde poner el foco o incluso adaptarla a su necesidad.

Resumiendo, se constata un enfoque teórico al que le falta más implementación práctica sobre lenguajes reales y cierta dificultad para ver integralmente el proceso de compilación/interpretación. Recuperando estas impresiones que no pretenden ser exhaustivas o excluyentes de otras realidades, lo que se destaca es que la construcción de lenguajes es un campo en el que hay mucho por hacer y se considera de suma importancia proveer de recursos tecnológicos y pedagógicos para que más docentes,

estudiantes e investigadores den pasos significativos en esta dirección. A su vez, construir una herramienta de software apropiada reafirma la importancia de una interacción fecunda entre teoría y práctica, en particular desde las miradas pedagógicas que interpretan la relación práctica/teoría desde la acción/reflexión [1].

3. Diseño y desarrollo de la herramienta

En sí mismo, el hecho de desarrollar un nuevo intérprete de un lenguaje de alto nivel, por la variedad de componentes de software y conceptos teóricos que combina, puede verse como una experiencia de aprendizaje y de construcción de conocimiento [2]. Pero se buscó dar un paso más allá y lograr que dicho componente se enmarque en una herramienta con sentido educativo.

La herramienta tiene un fin didáctico, no busca optimizar la *performance*, sino que prioriza que sea fácil de usar y ayude a entender mecanismos y conceptos. Se caracteriza por incluir un intérprete de un lenguaje funcional de alto nivel y ofrecer funcionalidades interactivas de visualización de mecanismos, entradas y salidas en los procesos intermedios y brindar la posibilidad de poder modificar, crear y extender los componentes del intérprete y adaptarlo.

Se propone para ser utilizada en asignaturas de la misma carrera universitaria de Sistemas de Información y carreras afines, pero también pueden ser de utilidad en cursos que abordan otro tipo de temáticas relacionadas con el diseño y análisis de lenguajes de programación. No está pensada como primer acercamiento a la programación, sino que puede hacer un mejor aprovechamiento de sus funcionalidades quienes ya tuvieron alguna experiencia en el uso de lenguajes de programación. En otras palabras, se asume como estrategia didáctica general que tiene más sentido profundizar en cómo funciona internamente algo -en este caso un lenguaje de programación- de lo que ya se conoce su utilidad.

De todas maneras, habilita a formas de uso que requieren diferentes niveles de conocimientos previos, por lo que puede utilizarse en espacios académicos que buscan un primer acercamiento a la teoría de lenguajes, como en otros que propongan profundizar en la materia, diseñar intérpretes y compiladores o crear nuevos lenguajes de programación.

Habiendo asumido una metodología de desarrollo incremental, se cuenta con una implementación que ya permite interpretar código y mostrar los resultados, que realiza las tareas principales y se encuentra productiva. En nuevas iteraciones se seguirá completando hasta contar con la herramienta terminada.

3.1. Lenguaje funcional

El lenguaje elegido para analizar y ejecutar es *Lisp*, lenguaje emblemático del paradigma funcional. En particular, se eligió un dialecto denominado *Racket Lisp* [3] que permite realizar las tareas más comunes. En el proceso de selección también se tuvo en cuenta la Construcción de *Lisp* en *Python* [4], por Peter Norvig [5], y un caso

más complejo, “Crafting Interpreters” [6] que toma un lenguaje similar a *Javascript* y muestra todo el camino hasta lograr un intérprete medianamente completo.

Entre las razones principales se encuentra la simplicidad constructiva, sintáctica y semántica, que facilita que los estudiantes y docentes puedan tener una primera experiencia accesible en su camino de aprender sobre compiladores. También, por su alto grado de expresividad. Se trata de una familia de lenguajes con más de 60 años de vigencia: Scheme, CommonLisp, Racket, Closure, entre otros son lenguajes con comunidades vibrantes y una búsqueda por la mejora continua.

Por otra parte, en los últimos años han crecido en el ambiente profesional del desarrollo de software los lenguajes de programación funcionales y muchos de sus conceptos característicos presentes en lenguajes de otros paradigmas. A su vez, en las carreras universitarias de sistemas, no solo de la UTN sino también de otras casas de altos estudios, se enseña la programación funcional. Esto hace que la elección de un lenguaje funcional como Lisp resulte familiar para los estudiantes y no represente una dificultad adicional.

3.2. Diseño preliminar

El diseño de la herramienta consiste en dos componentes principales: el intérprete propiamente dicho, que realiza todos los procesos internos, y una aplicación que proporciona la interfaz de usuario para mostrar y poder seguir de una manera didáctica su funcionamiento.

Las definiciones más importantes se orientaron a establecer los alcances de cada uno de los procesos sucesivos que se aplican sobre el código fuente, la formulación de las representaciones intermedias y la especificación de entradas y salidas de información, de manera de focalizar en los aspectos más significativo para el aprendizaje.

De esta manera, se prioriza el análisis lexicográfico, el análisis sintáctico y el mecanismo de evaluación y como elementos destacadas a mostrar de los resultados intermedios se seleccionó la lista de tokens y el árbol de sintaxis abstracta, siguiendo los criterios clásicos en la materia [7] [8]. Como datos de entrada, se contempla el ingreso del código fuente de la definición de un programa formado por un conjunto de funciones y las consultas que se realizan utilizando dichas funciones. Como datos de salida final, se considera el resultado de la evaluación, en caso que el proceso se complete, o información sobre los errores, si es que se produce alguno.

Por otra parte, en el diseño se dejó de lado el proceso posterior de generar un ejecutable o algún tipo de *bytecode* de más bajo nivel que luego evalúe una máquina virtual, que es un aspecto en el cual difieren los lenguajes actuales de uso industrial. Se trata de una discusión con mucha vigencia en las comunidades que se aglutinan alrededor de cada lenguaje y una decisión clave a la hora de modificar o crear de nuevos lenguajes de programación, y cada vez se encuentran más matices y combinaciones entre los conceptos tradiciones de "compilar" e "interpretar". En todo caso, constituye un posible futuro trabajo pensando en contextos educativos más avanzados o específicos en estas temáticas.

3.3. Uso principal

La utilidad básica consiste en ubicarse como usuario de la herramienta: Se escribe una porción de código que al hacerla evaluar desencadena una serie de procesos, cuyos pasos y estructuras intermedias son mostrados mediante gráficos adecuados, hasta que se obtiene el resultado final. En caso que todo funcione adecuadamente, el énfasis está puesto por un lado en mostrar el paralelismo entre cada porción del código fuente y sus representaciones correspondientes, como también poder visualizar la secuencia interna de evaluación.

Más en detalle, se comienza con el ingreso del código fuente de un programa sencillo en *Lisp*. Por ejemplo, en la figura 1 se detalla cómo se define una función.

```
(define (siguiente valor) (+ valor 1))
```

Fig 1. Definición de una función

Los *tokens* que genera el análisis lexicográfico se visualizan como una lista, en la que se refleja si se trata de identificadores, de palabras reservadas, literales o símbolos. La figura 2 muestra la lista de *tokens* correspondiente al código de la figura anterior.

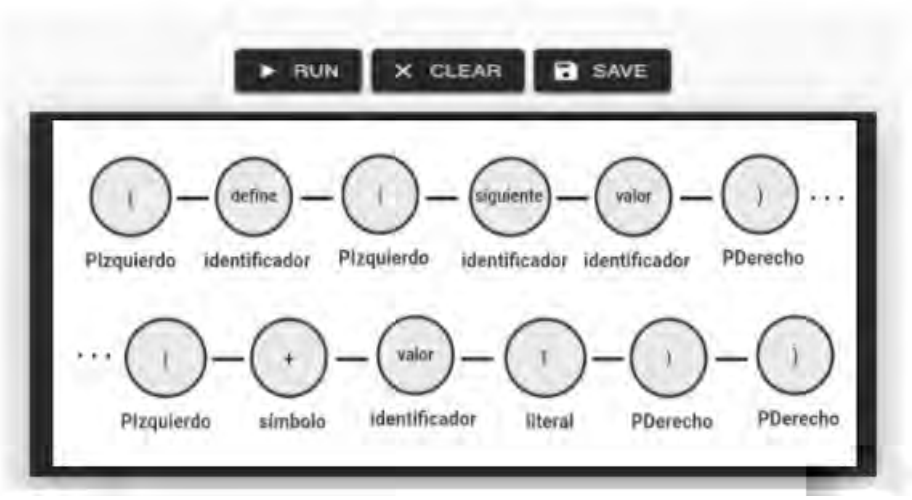


Fig 2: Lista de *tokens*

Luego, el análisis sintáctico “parsea” los *tokens* y construye el árbol de sintaxis abstracta, mejor conocido como AST, por su sigla en idioma inglés. En la figura 3 se muestra el árbol correspondiente al ejemplo anterior. Puede observarse que ya no

están presentes elementos como los paréntesis que son innecesarios en esta etapa y los diferentes tipos de nodos con sus respectivos colores.

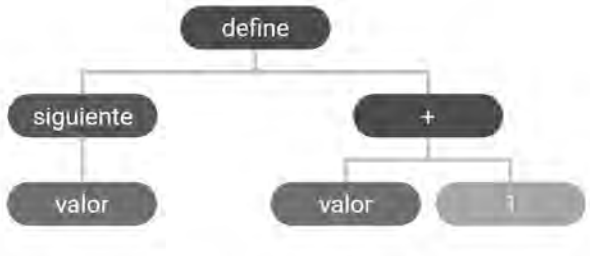


Fig 3: *Abstract Syntax Tree*

La instancia de evaluación implica analizar la expresión ingresada en la consola *REPL* e incorporarla al árbol ya construido, para luego recorrerlo realizando las sustituciones correspondientes, de manera de generar un resultado final que se vuelve a mostrar en dicha consola (Ver figura 4).

Fig 4: Consola REPL con el resultado de la evaluación.

3.4. Forma de uso alternativa

Una utilidad avanzada, que sin dudas requiere de un acompañamiento docente más intenso y un perfil de estudiantes con mayor experiencia, consiste en modificar el componente del intérprete mismo, y utilizando la misma interfaz y demás elementos

de la aplicación poder ver cómo se comporta diferente. Para ello, además de haber cuidado criterios de expresividad en el código y una adecuada modularización para que sea más sencillo de modificar, se dejaron adrede "huecos" en la formulación del lenguaje para habilitar a la realización de consigna de trabajo acotadas y precisas (por ejemplo, completar un tipo de dato faltante, agregar operaciones sobre un tipo existente, modificar una palabra reservada, etc.).

3.5. Decisiones técnicas

Se decidió utilizar *Rust* [9] para la construcción del intérprete propiamente dicho, teniendo en cuenta que es un lenguaje compilado, estáticamente tipado, moderno y con un enfoque en corrección que se destaca en el panorama actual de la industria de desarrollo de software. A la vez, comparte ciertos principios de funcionamiento con *Lisp* lo que le da mayor coherencia y robustez al desarrollo.

El código escrito en *Rust* se compila a un paquete de *WebAssembly*, que es un lenguaje de código binario que se puede ejecutar en el navegador. *React* consume el paquete de *WebAssembly*, lo que permite acceder a las funciones que expone el intérprete en un navegador.

4. Resultados y futuros trabajos

En primer lugar, en cuanto a la construcción de la herramienta como tal, los resultados alcanzados hasta ahora son satisfactorios, teniendo en cuenta que las funcionalidades desarrolladas funcionan de acuerdo a lo previsto y lo que se encuentra pendiente no impide el funcionamiento del resto.

El subconjunto de definiciones contempladas válidas de *Lisp* se evalúa satisfactoriamente, se obtiene el resultado y se visualizan los gráficos cuando el código es correcto. Las expresiones que producen un error en el *Lisp* original también dan un error en esta implementación y despliegan una leyenda similar, aunque no se exprese aún gráficamente.

En forma experimental, simulando un posible uso avanzado por parte de estudiantes y tomando como punto de partida una versión ya bastante avanzada de la herramienta, se hizo el ejercicio de puntualmente agregar un nuevo tipo de dato e implementar ciertas operaciones nativas sin alterar el resto del código. Reconociendo la distancia entre quien es miembro del equipo de desarrollo y quien no, aún tratándose potencialmente en ambos casos de estudiantes avanzados, se trata de un resultado provisorio pero prometedor.

Recuperando lo dicho anteriormente de que se trata de un trabajo en progreso, en cuanto al desarrollo mismo queda pendiente una mejor y más precisa forma de comunicar y mostrar los errores.

En la versión actual se informa que se produjo un error, indicando el motivo mediante una descripción textual e interrumpiendo el proceso de interpretación. (Ver figura 5). Entendiendo la importancia que tiene para un estudiante enfrentarse a un

error como parte de su proceso de aprendizaje, lo que se espera es poder representarlo gráficamente integrando la visualización de dicho error en los diagramas que se despliegan cuando el proceso termina correctamente. A su vez, dependiendo del motivo, ubicación o gravedad del error, se busca habilitar total o parcialmente la continuidad del proceso o habilitar la evaluación de las porciones del código fuente que no presentan errores.

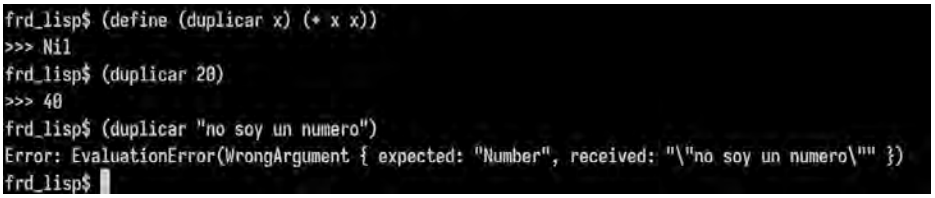


Fig 5: Visualización de un error de evaluación

Aún no se cuenta con los resultados reales de ver el impacto del uso de herramienta en estudiantes, lo que dará elementos más confiables acerca del logro de los objetivos planteados. Previendo esa instancia, el relevamiento realizado inicialmente establece un punto de referencia a la hora de realizar un nuevo estudio con estudiantes que hayan utilizado la presente herramienta y resultará oportuno utilizar como base el mismo instrumento de recolección de datos, sumándole alguna pregunta adicional. Cabe reconocer que los tiempos del recorrido académico de los estudiantes y de las decisiones de los docentes no son necesariamente los mismos que los que se dan en la dinámica de la investigación, por lo que la obtención de resultados requerirá como mínimo de un año. Asimismo, teniendo en cuenta que el período tomado para el relevamiento inicial abarca varios años de cursada de la materia para tener una muestra más sólida, sería consistente contemplar también más de una cursada con esta nueva herramienta.

Referencias

1. Freire, P. (1970) Pedagogía del Oprimido. Ed. Tierra Nueva. Montevideo
2. Moreno-Seco, Forcada. (2001). Learning compiler design as a research activity. Departament de Llenguatges i Sistemes Informatics, Universitat d'Alacant
3. Racket Lisp. 2019 Sitio web https://racket-lang.org
4. Lispy. 2019 Sitio web http://norvig.com/lispy.html
5. Norvig, Peter. 2019 Sitio web https://en.wikipedia.org/wiki/Peter_Norvig
6. Bob Nystrom.Crafting Interpreters. 2019 Sitio web http://craftinginterpreters.com
7. Aho, Lam, Sethi, Ullman. (1986). Compilers: Principles, Techniques, and Tools. Addison Wesley.
8. Cooper, Torczon. (2011). Engineering a Compiler. Morgan Kaufmann
9. Rust Programming Language. 2019 Sitio web https://www.rust-lang.org

Estudio exploratorio sobre el impacto causado por la pandemia en la docencia de la Universidad de Morón

Iris Sattolo[1], Marisa Panizzi[1], Vanesa Contreras[1]

[1]Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias Universidad de Morón. Cabildo 134 (B1708JPD), Partido de Morón, Argentina.
iris.sattolo@gmail.com, marisapanizzi@outlook.com , contreras_vane@yahoo.com.ar

Resumen. En este trabajo se presentan los resultados de un estudio exploratorio realizado mediante una encuesta con el propósito de obtener evidencia sobre el impacto causado por el cambio de modalidad que debieron afrontar los docentes de la UM (Universidad de Morón). Con los resultados del estudio se pretendía incrementar la calidad del proceso de enseñanza a distancia y conocer cuáles fueron los patrones en el comportamiento, las tasas de aceptación y la satisfacción. En esta encuesta participaron 77 docentes de las diferentes Escuelas Superiores de la UM. Los resultados reportan que los docentes no tuvieron dificultades para dictar clases en e-learning y refieren haber llevado a cabo sin inconvenientes la mayor parte de los contenidos propuestos para la materia. Si bien es evidente que el número de participantes es escaso, esto determina que nuestro trabajo futuro inmediato se centrará en ampliar la muestra.

Palabras claves: docencia en pandemia, virtualidad, Universidad de Morón, estudio exploratorio, encuesta.

1 Introducción

El surgimiento del COVID-19 trajo consigo cambios en las estructuras socioeconómicas a nivel global, como parte de estas, las instituciones de Educación Superior no fueron una excepción. Las medidas impartidas de "distanciamiento social" produjeron un contexto de digitalización forzada que, en el caso de las universidades, coaccionó los mecanismos de pedagogía a los de la teleeducación, para garantizar su funcionamiento y sostenibilidad [1]. La práctica pedagógica de los docentes universitarios sufrió cambios y debió adaptarse al nuevo contexto, los docentes debieron reinventar e insertar nuevas formas de enseñar en su proceso de trabajo apoyando el aprendizaje en herramientas web.

Nuestra institución había iniciado una transición a la digitalización antes de la pandemia y contaba ya con una infraestructura tecnológica adecuada permitiendo que algunos docentes tuviesen cierta experiencia en el desarrollo de una cultura digital. Desde el comienzo en el e-learning, la Universidad de Morón, contó con un sistema de gestión del conocimiento (en inglés, *Learning Management System* o LMS) a través del cual se ofrecían carreras a distancia, como también diversos cursos de capacitación. La plataforma que se utilizaba para su gestión era *Moodle*. En el año 2019 incorporó, como apoyo a la presencialidad, la plataforma *Blackboard* ofreciendo a sus docentes

cursos de capacitación en la misma. A comienzos del año 2020, como consecuencia de la llegada del COVID-19, las clases presenciales migraron a la plataforma *Blackboard* [2]. Surgieron entonces todos los problemas que la mayoría de las instituciones debieron solucionar al afrontar un cambio acelerado en la adopción de tecnologías digitales.

Los docentes que aún no habían incursionado en clases a distancia debieron adaptarse a la nueva modalidad abruptamente. Diseñar un curso a distancia presenta distintos desafíos a los de la modalidad presencial, más aún si se debe adaptar sobre la marcha la organización realizada con anterioridad. Esto incluye planificación de horarios, escritura y preparación de materiales para distancia, como también determinar tareas de evaluación. Para modificar un curso preparado para la presencialidad, el docente debió actualizar el material como también las estrategias abordadas en el curso.

Con el propósito de recolectar evidencia del impacto que produjo este cambio disruptivo, sobre los docentes en la Universidad de Morón e identificar las estrategias que pudieron utilizar en el primer cuatrimestre del año 2020 en la plataforma *Blackboard*, se decidió realizar un estudio exploratorio a través de una encuesta realizada según las directrices de Molléri *et al.*[3].

Este artículo se estructura de la siguiente manera: en la Sección 2 se describe la planificación de la encuesta, en la Sección 3 se describe su ejecución. Los resultados se presentan en la Sección 4. En la Sección 5 se presenta un análisis de las amenazas a la validez y finalmente, en la Sección 6 se exponen las conclusiones y trabajos futuros.

2 Planificación de la encuesta

Objetivos y preguntas de investigación. El uso de la plantilla de GQM "Goal-Question-Metric" [4] nos permitió formular el objetivo de la encuesta de la siguiente manera: "***Analizar*** *las prácticas que debieron afrontar los docentes ante el cambio de modalidad* ***con el propósito de*** *conocer el impacto generado que produjo este cambio disruptivo* ***con respecto a*** *obtener una percepción sobre las estrategias que pudieron utilizar en el primer cuatrimestre del año 2020 en la plataforma Blackboard* ***desde el punto de vista de*** *docentes de la Universidad de Morón*".

Las preguntas de investigación (PI) que guiaron este estudio son las siguientes:

- PI1: ¿Cuáles fueron las dificultades o problemas más comunes encontrados por los docentes al cambiar de modalidad abruptamente al e-learning?
- PI2: ¿Cómo impacta la formación docente en el dictado de clases especialmente para el cambio a e-learning?
- PI3: ¿Qué tipo de diseño utilizaron los docentes en sus primeras intervenciones en el e-learning?
- PI4: ¿Utilizaron algún tipo de seguimiento con el alumno que se corresponda con el *Learning Analytics*?

Proceso de ejecución. Este proceso consta de: 1) Diseño de un formulario para llevar un registro sistemático de la ejecución de la encuesta. Este se compone de dos partes: a) Proceso de envío de la encuesta: cantidad de envíos y a quienes. y b) Proceso de seguimiento del envío: mails erróneos, cantidad de cuestionarios respondidos y fecha de respuesta. 2) Envío de la encuesta por correo electrónico. Se diseño un texto de

presentación en el cual se menciona el propósito de la investigación, quiénes participan, tiempo estimado para responder la encuesta, agradecimiento por la colaboración y una invitación a que el encuestado difunda la encuesta entre sus contactos. 3) Revisión diaria de encuestas respondidas. 4) Extracción de las respuestas. 5) Revisión de si hay preguntas sin responder. 6) Extracción de los datos.

Población. La población a la cual se decidió enviar la encuesta son docentes pertenecientes a la Universidad de Morón que dictan clases en cualquiera de las Carreras y Escuelas Superiores de la Universidad, sin distinción de cargos relacionados al puesto de trabajo. Se utilizaron diferentes estrategias para seleccionar la muestra, todas a través de correo electrónico. La distribución de la encuesta se realizó vía correo electrónico, enviando un correo a cada uno de los docentes de la lista de contactos, explicando brevemente el motivo de la encuesta, el tiempo aproximado de llenado y el enlace (en inglés link) a la encuesta propiamente dicha para poder completarla.

Diseño de la encuesta. Se diseñó un cuestionario autoadministrado que se envió por correo electrónico a los encuestados y se utilizó la herramienta *Google Forms.* Para el diseño del cuestionario se definieron cuatro dimensiones de preguntas que, junto con la variable, la nomenclatura propuesta para cada dimensión y los indicadores que se presentan en la Tabla 1. El cuestionario se compone de 22 preguntas, una única pregunta opcional correspondiente al nombre de la materia. En el caso que el encuestado haya respondido afirmativamente la pregunta "En el primer cuatrimestre ¿dictó clases en otra materia?", se repite el bloque de preguntas de la dimensión académica, es decir que cada persona ha respondido un mínimo de 21 preguntas.

El cuestionario completo se encuentra disponible en el siguiente enlace: https://forms.gle/SWvpK61NkH1vSzmK6.

Tabla 1. Variable, sus dimensiones, la nomenclatura propuesta para cada dimensión y los indicadores.

Variable	Dimensión	Nomenclatura de la Dimensión	Indicador
Problemática de la docencia de la UM en pandemia.	Personal	PER	Rango de edad, si es profesor universitario.
	Herramienta	HER	Experticia del docente con respecto a la herramienta.
	Técnica	TEC	Problemas de conexión.
	Académica	ACA	Actividad docente, dedicación en la docencia.

A continuación, se describe cada una de las dimensiones:

1) Personal: se refiere a la información sobre el docente encuestado, como ser: el rango de edad, si tiene título de profesor universitario o no, si realizó cursos de e-learning. Esta dimensión contiene 3 preguntas para poder responder parte de la PI 1 y PI2.

2) Herramienta: se refiere a la información relacionada con la utilización de la herramienta Blackboard por parte del docente. Esta dimensión consta de 3 preguntas, las cuales están relacionada a la PI 3.

3) Técnica: se refiere a la información relacionada con el acceso a internet, con los dispositivos utilizados para dar las clases, sobre el conocimiento de la herramienta Blackboard, sobre la disponibilidad de la herramienta, si tuvo alguna capacitación sobre la herramienta, si utilizó alguna de las opciones brindadas por la plataforma para dar seguimiento a sus alumnos. Esta dimensión consta de 7 preguntas relacionadas a la PI 4.
4) Académica: corresponde a la información relacionada a la materia que dicta el docente, como ser: nombre (es una pregunta opcional), escuela superior a la que pertenece, año de la materia, antigüedad en lo que respecta al dictado de la materia. En esta dimensión se realizaron preguntas relacionadas a la aplicación utilizada para comunicarse con los alumnos, material didáctico utilizado y si pudo adaptar su estrategia de enseñanza la nueva modalidad o no fue necesario. Esta dimensión cuenta con 9 preguntas las cuales responden en parte a las PI 1, PI 3 y PI4.
Validación. Antes de comenzar con el envío de la encuesta, se realizó una prueba piloto inicial con un grupo reducido de 3 docentes que permitió verificar el tiempo de respuesta inicial estimado de 15 minutos, quedando reducido a 10 minutos. Esta prueba permitió evaluar el lenguaje y la redacción utilizada en la encuesta. Como resultado se realizaron algunos cambios en preguntas relacionadas al ámbito personal y también académico, obteniendo una segunda versión. Se realizó luego una comprobación enviando nuevamente al mismo grupo, donde los participantes de la prueba confirmaron su claridad y legibilidad de la encuesta.

3 Ejecución de la encuesta

Reclutamiento de los participantes. El envío de la encuesta ha sido directo dado que se envió a docentes de la Universidad de Morón.
Gestión de las respuestas. La gestión de la ejecución de la encuesta se realizó de acuerdo con el procedimiento definido en la sección 2. Se realizó un seguimiento diario de las respuestas con el objeto de comprobar que las respuestas estén completas.
Análisis de los datos. Para asegurar la calidad de los datos obtenidos de la encuesta se revisaron estos con el propósito de encontrar errores (completitud y errores de tipeo). Se realizó un análisis de contenido de las respuestas de texto libre [5]; el análisis de los datos se basó en un análisis cuantitativo centrado principalmente en estadísticas descriptivas y porcentajes de la información recopilada.

4 Análisis e interpretación de los resultados obtenidos

En la encuesta participaron 77 docentes de diferentes Escuelas Superiores de la Universidad de Morón, a continuación, se presentan, los resultados que permitieron dar respuesta a cada PI.

PI1: ¿Cuáles fueron las dificultades o problemas más comunes con las que se presentaron los docentes que comienzan a dictar clases en e-learning?

El 22 % de los encuestados respondió que su conexión a internet fue regular y el 9 % de estos utilizó datos móviles para poder conectarse. Si bien el 85,71 % de los

encuestados tiene una edad mayor a 50 años, el 57,14 % de estos pertenece a ESIICA (Escuela Superior de Ingeniería Informática y Ciencias Agroalimentarias). El 16 % de los encuestados hace menos de 5 años que dicta la materia. El 27 % de los encuestados dicta materias correspondientes a 1er año y el 21 % a 2do año, se sabe que los primeros años de las carreras tienen mucha cantidad de alumnado.

En el análisis de la encuesta observamos que la gran mayoría de los encuestados tienen una edad mayor a los 50 años (66 docentes; 85,71 %), esto puede ser un indicador de cómo ha sido su adaptación a la nueva modalidad de enseñanza a distancia. Siendo conscientes de las dificultades tecnológicas que se les presentan a la gente de mayor edad. 7 docentes (9,09 %) entre 40 y 50 años y sólo 4 docentes (5,19 %) entre 30 y 40 años.

En la Figura 1, se muestra la cantidad de docentes encuestados por rango de edad.

Fig. 1. Encuestados por rango de edad.

PI 2: ¿Cuán importante es el nivel académico para poder dictar clases en e-learning?

El 45 % de los encuestados posee título de profesor universitario, el 75 % de los encuestados realizó cursos de e-learning. El 62 % de los encuestados hace más de 10 años que dicta la materia, es decir cuenta con cierta experiencia en el dictado de la materia. Respecto a si cuenta con título de profesor universitario o posgrado relacionado con la educación, observamos que únicamente 35 de los docentes encuestados respondieron de manera afirmativa a esta pregunta. Podríamos decir que la mitad de los encuestados no poseen título de profesor universitario o posgrado y determinar si esto condiciona a la adaptación de la estrategia de enseñanza de la materia o materias que dicta.

En la Figura 2, se muestra la cantidad de docentes que poseen título de profesor universitario o posgrado relacionado con la educación. Se puede observar que el (45 %; 35 docentes) tiene título de profesor, mientras que el (55 %; 42 docentes) responde no tenerlo.

Con relación a si el docente realizó cursos de capacitación en e-learning cualquiera fuese el momento, observamos que el 75% de los encuestados tuvo alguna capacitación en e-learning. Con lo cual nos da la pauta de que un alto porcentaje de los docentes tiene conocimientos previos a través de los cursos realizados, y posiblemente haya podido aplicarlos al dictado de sus clases.

Fig. 2. Encuestados con formación en docencia.

En la Figura 3, se muestra el porcentaje de docentes que realizó alguna capacitación en e-learning. Se puede observar que el (25 %), 19 de los docentes no tomó capacitaciones en e-learning, mientras que el (75 %), 58 docentes si lo hizo.

Fig. 3. Encuestados con formación en e-learning.

Además del análisis de la encuesta podemos observar que del total de encuestados 48 de los docentes, (62%) dicta la materia hace más de 10 años en la Universidad. Con estos valores podríamos determinar que contamos con un alto grado de experiencia en el dictado de clases por parte de los docentes encuestados. Luego tenemos 17 docentes, (22%) que dictan la materia entre 5 a 10 años, y el resto correspondiente a 12 docentes, (16%) que dictan la materia hace menos de 5 años.

Como podemos ver en la Figura 4, se muestra el porcentaje según el rango de años que el docente lleva dictando la materia.

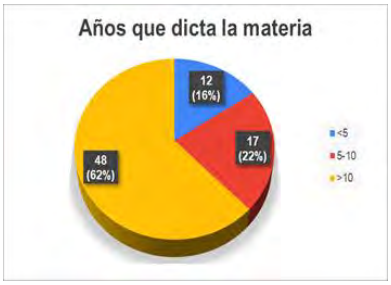


Fig. 4. Encuestados según el rango de años de dictado de la materia.

PI 3: ¿Qué tipo de diseño utilizaron los docentes en sus primeras intervenciones en el e-learning?

El 96,10 % de los encuestados utilizó material didáctico PDF, el 79,22 % PPT y Video, el 46,75 % utilizaron Programas específicos de la materia, el 28,57 % utilizó Biblioteca digital, el 20,78 % utilizó Simulaciones, los encuestados que utilizaron Scorm y Programas de diseño suman el 12,98 % y finalmente el 27,27 % responde haber utilizado "Otro" material didáctico.

PI 4: ¿Utilizaron algún tipo de seguimiento con el alumno que se corresponda con el Learning Analytics?

El 57,14 % de los encuestados utilizó la herramienta de "Libro de calificaciones" y el 19,48 % la utilizó pocas veces. El 51,95 % utilizó "Debates y actividades", mientras que el 29,87 % la utilizó pocas veces. El 35,06 % de los encuestados utilizó la opción de "Tomó asistencia" registrando la asistencia de los alumnos a las clases, mientras que el 28,57 % la utilizó pocas veces. Y el 22,08 % utilizó la herramienta de "Miró las estadísticas" mientras que el 28,57 % la utilizó pocas veces.

Observamos además que, para dar seguimiento a los alumnos más del 50 % de los docentes utilizó las herramientas que brinda la plataforma Blackboard como ser, "Libro de calificaciones" 44 docentes (57,14 %) y "Debates y Actividades" 40 docentes (51,95 %). Observamos que un porcentaje bajo utilizó la opción de "Tomar asistencia", solo 27 docentes (35,06 %) mientras que 22 docentes (28,57%) dice haber tomado asistencia muy pocas veces, y 24 docentes (31,17 %) directamente no tomó asistencia. Notamos también que el 32 % de los docentes en promedio, no utilizó ninguna de las herramientas disponibles en la plataforma Blackboard.

Como podemos ver en la Figura 5, se muestra cada una de las herramientas disponibles en la plataforma *Blackboard* y la utilización efectuada por los docentes.

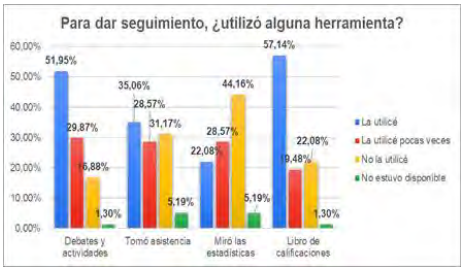


Fig. 5. Encuestados según el uso de herramientas disponibles en la plataforma *Blackboard*.

5 Amenazas a la validez

Para analizar la validez de la encuesta se tuvieron en cuenta los tipos de validez propuestos por Wohlin *et al.*[6]:

- Validez de conclusiones. El tamaño de la muestra (77 respuestas) se considera pequeño para considerar los resultados estadísticamente aceptables. Si bien se reconoce que es aconsejable ampliar la muestra, pero por tratarse de un estudio exploratorio sobre el impacto causado por el cambio de modalidad que debieron

afrontar los docentes de la UM permitió cumplir con el propósito definido para esta encuesta. Además, se aplicó el proceso de manera sistemática y rigurosa para permitir que el proceso sea reproducible.

- Validez interna. Los principales problemas que afectan la validez interna de nuestro estudio se refieren al encuadre y al muestreo de los participantes. Nuestra estrategia de reclutamiento de participantes podría haber sido sesgada por profesionales sin experiencia. Si bien hay una variedad en los roles que participaron en la encuesta, la mayoría de los encuestados cuentan con más de 10 años de experiencia en el dictado de la materia. Otro factor negativo podría haber sido la dificultad para comprender las preguntas (por ejemplo, ambiguas, poco claras, mal formuladas), esto quedó resuelto con la prueba piloto realizada con 10 encuestados. Las motivaciones de los encuestados también podrían haber afectado las respuestas y, por lo tanto, los resultados de la encuesta, esto quedó resuelto porque en el mail que se envió junto con la encuesta quedó explicito el compromiso de compartir los resultados de la encuesta con los participantes.
- Validez externa. Se seleccionaron a los participantes de la encuesta de forma de que sean docentes pertenecientes a la Universidad de Morón que dictan clases en cualquiera de las Carreras y Escuelas Superiores de la universidad, sin distinción de cargos relacionados al puesto de trabajo. Esto permitió realizar observaciones sobre el impacto causado por el cambio de modalidad que debieron afrontar los docentes de la UM.
- Validez de constructo. Se definieron las preguntas de investigación de manera cuidadosa, así como también el esquema de las categorías de preguntas y los posibles valores de las respuestas. Todo el proceso ha sido consensuado entre los investigadores de manera que no sea sesgado el objeto de estudio. Otros aspectos que permitieron disminuir esta amenaza han sido que, en el mail de la invitación a los participantes, se les explicó claramente el propósito del estudio y del propio cuestionario se visualiza su anonimato salvo que el participante estuviese interesado en que lo podamos contactar para profundizar el cuestionario.

6 Conclusiones y trabajos futuros

En este trabajo se han presentado los resultados de un estudio exploratorio mediante una encuesta cuyo objetivo consistió en analizar el impacto generado por el cambio de modalidad del dictado de clases que debieron afrontar los docentes de la UM en el año de la pandemia por COVID-19.

Se ha mencionado la aplicación de técnicas para asegurar la calidad de la encuesta: validez, confiabilidad y objetividad.

En el análisis se consideraron 77 respuestas válidas y completas provenientes de los docentes de las distintas Escuelas Superiores de la UM. Se reconoce que la muestra utilizada es reducida, pero sin embargo se considera que las respuestas a las preguntas de investigación brindan información interesante a los investigadores de este estudio.

En relación con la PI1, el mayor conflicto manifestado por los docentes fue la conectividad.

Con respecto a la PI2, no se observa diferencia entre los docentes con título docente, y sin título docente, El problema del abordaje a esta nueva modalidad, es evidente que se debió a otras causas no consultadas en la encuesta.

Se observó que más del 50 % de los docentes encuestados no utilizaron herramientas que brinda la plataforma para realizar un seguimiento sobre los estudiantes, posiblemente debido al desconocimiento de las bondades que ofrece la plataforma al utilizar estas herramientas, o a la falta de tiempo de los docentes para investigar e implementarlas.

En relación con la PI3, no hay evidencias que se utilizaran metodologías propias sobre Diseños de aprendizaje relacionados al e-learning.

Con respecto a la PI4, se observó que si bien existen herramientas de seguimiento de aprendizaje en la plataforma, los docentes sólo utilizaron las más conocidas, tales como "el libro de calificaciones".

Nuestro trabajo futuro inmediato, será centrará en incrementar la muestra para lograr un análisis más profundo y en segundo lugar extender la encuesta a otras universidades para luego poder contrastar los resultados.

Referencias

1. Bank, IDB inter-Américan Developen (2020). "La educación superior en tiempos de Covid-19" Whashintong.
2. Plataforma Blackboard. Disponible en: https://www.unimoron.edu.ar/area/blackboard
3. Molléri J., Petersen K., & Mendes E. (2020). An empirically evaluated checklist for surveys in software engineering. Information and Software Technology. Vol 119 (106240).
4. Basili V.; Rombach D. (1988). The TAME project: towards improvement-oriented software environments. IEEE Transactions on Software Engineering, 14(6), pp. 758-773.
5. Krippendorff K. (2012). Content analysis: An introduction to its methodology, 3rd edn. Sage Publications, Thousand Oaks.
6. Wohlin C., Runeson P., Höst M., Ohlsson MC., Regnell B., Wesslén A. (2012). Experimentation in software engineering: an introduction. Springer.

Asignación de Estudiantes a Establecimientos Educativos: Un Enfoque Multi-objetivo

Maria Cecilia Casco[1], Fabio López-Pires[2], Benjamín Barán[1], and Eustaquio A Martínez[1]

[1]Facultad Politécnica, Universidad Nacional del Este
Ciudad del Este, Paraguay
{ceciliacasco,bbaran, amartinez}@fpune.edu.py
[2] Universidad Internacional Tres Fronteras
Ciudad del Este, Paraguay
fabio.lopez@uninter.edu.py

Resumen En este trabajo se aborda el problema de Asignación de Estudiantes a Establecimientos Educativos (AEEE). Dicha problemática afecta la logística del sistema educativo, ya que se ve influenciada por variantes como: disponibilidad, distancia, infraestructura, entre otros. Se propone una nueva formulación matemática al problema de AEEE con un enfoque multi-objetivo para: (1) minimizar la diferencia entre la cantidad de estudiantes asignados y la cantidad óptima de estudiantes por clase, (2) minimizar la distancia promedio entre la vivienda del estudiante y el establecimiento y (3) maximizar la utilización de establecimientos con mejor infraestructura. Para resolver la formulación propuesta se plantea un Algoritmo Evolutivo Multi-Objetivo (MOEA) basado en el NSGA-II. Para la validación de esta propuesta se consideraron los datos provistos por el Ministerio de Educación y Ciencias del Paraguay (MEC) correspondientes a Ciudad del Este - Alto Paraná, del primer al tercer grado, de 90 establecimientos y 15.763 estudiantes, los resultados arrojan mejoras significativas en la cantidad de alumnos asignados por clases.

Keywords: optimización multi-objetivo, asignación de estudiantes, computación evolutiva, algoritmos genéticos

1. Introducción

Problemas relacionados a cuestiones logísticas como optimizar el tiempo de transporte del estudiante, la asignación de cursos u otros factores relacionados con el correcto aprovechamiento de los recursos educativos pueden ser abordados con una técnica de resolución y una formulación matemática adecuada. Esto permitiría aprovechar al máximo los recursos físicos del sistema educativo, e.g. infraestructura física de establecimientos.

La mala distribución de recursos en el sistema educativo afecta directamente la formación de los estudiantes, impidiendo la consecución de los objetivos pedagógicos. Para el caso de Paraguay, se contrapone a los derechos fundamentales de todos los ciudadanos establecido en la Ley 1264/98 General de Educación [1], la cual en su artículo número 3 menciona: *."El Estado garantizará el derecho de aprender y la igualdad de oportunidades de acceder a los conocimientos y a los beneficios de la cultura humanística, de la ciencia y de la tecnología, sin discriminación alguna."*

En este contexto, resolver el problema de Asignación de Estudiantes a Establecimientos Educativos (AEEE) podría contribuir a mejorar la logística de los sistemas educativos. Con una correcta asignación de los recursos a los establecimientos, se mejora el aprovechamiento de los recursos ya existentes y se podría lograr una mejor distribución de las aulas, evitando sub o sobre asignación. Además, se podrían prevenir casos de deserción escolar por problemas relacionados al traslados de los estudiantes a la institución asignada.

La implementación de una alternativa tecnológica para la Asignación de Estudiantes a Establecimientos Educativos, podría tener una incidencia directa en lo social y económico, considerando los siguientes aspectos:

- Reducción de gastos de operación y transporte.
- Mejor distribución de recursos físicos escolares.
- Garantía de plazas escolares para los alumnos.
- Disminución de los tiempos en los procesos de inscripción.
- Optimización en la cantidad de alumnos asignados por aulas, evitando sub o sobre asignación.

Teniendo en cuenta los puntos citados, este trabajo se enfoca en proponer una formulación matemática que permita mejorar algunos de estos puntos, para de esta manera ser una herramienta potencial para mejorar la logística del sistema educativo, impactando positivamente en la vida de los alumnos y sus familias.

Formalmente, se puede definir el problema AEEE como:

"Dado un conjunto de estudiantes A y un conjunto de establecimientos educativos E, asignar los estudiantes A a los establecimientos educativos de E, considerando las restricciones de los recursos y optimizando las funciones objetivo definidas."

El resto del trabajo se encuentra estructurado en las siguientes secciones: en la Sección 2, se revisan trabajos relacionados de los últimos 10 años; la Sección 3 presenta la formulación matemática; la sección 4, muestra el conjunto de datos utilizado y los resultados experimentales, y por último, la sección 5 presenta las conclusiones y propone trabajos futuros.

2. Revisión de la Literatura

Afacan et al. en [2] buscan maximizar el número de estudiantes asignados a instituciones educativas, comparando las técnicas de Gale-Shapley Deferred Acceptance (GDA), Boston Mechanism (BM), Top-Trading Cycles (TTC), y Serial Dictatorship (SD) con la técnica que ellos denominaron Efficient Assignment Maximizing Mechanisms (EAMs). Esta técnica hace asignaciones combinando las demás técnicas en ciclos que tienen como objetivo lograr una mayor cantidad de asignaciones en cada iteración. Las pruebas fueron realizadas considerando 400 estudiantes y 20 escuelas. Utilizando la técnica de EAM se redujo a 21 la cantidad de estudiantes sin asingación, en comparación a las demás técnicas que dejaban hasta 61 estudiantes sin escuela, lo que implica un aumento de la eficiencia del 65 % en el proceso de asignación.

El proceso de asignar estudiantes a grupos de laboratorios fue abordado por Agustín-Bla et al. en [3], donde se plantean las restricciones de espacio en los laboratorios y de los recursos con los que se disponen, además de preferencias en cuanto a horarios o grupos en el momento de la inscripción. Utilizaron algoritmos genéticos para el proceso de asignación, buscando maximizar la eficiencia en asignaciones escolares y la satisfacción considerando preferencias, lograron que el 90 % de los estudiantes hayan sido asignados a los grupos por los cuales demostraron preferencia.

De acuerdo a lo expuesto por Baker y Powell en [4] uno de los objetivos principales en los procesos de asignación de estudiantes a cursos o escuelas, es maximizar la diversidad de estudiantes en los grupos formados y maximizar las diferencias entre los grupos. En este caso se aplicó la técnica de búsqueda de vecinos. El método fue aplicado en la Escuela de Negocios de Tuck - Estados Unidos para la asignación de 200 estudiantes a 4 secciones de un mismo curso. Para la evaluación de los resultados se realizó una media de las características de cada estudiante, tomando en cuenta la nacionalidad, la clase social y el nivel de formación académica de los mismos. Luego de las pruebas se obtuvieron grupos heterogéneos el 90 % de las veces, comparando, finalmente las asignaciones realizadas por el algoritmo y las realizadas manualmente.

La eficiencia del recorrido del transporte escolar es otro objetivo muy estudiado en la literatura. En general se busca minimizar el tiempo de viaje del autobus.Surgen así distintas formulaciones Multi-Objetivo, en el caso de Bouzarth et al. en [5] añaden el objetivo de minimizar las diferencias socioeconómicas de los estudiantes entre las instituciones de un mismo distrito. Los resultados demostraron que ambos objetivos entran en conflicto al momento de optimizarlos, por lo que se manejaron pesos para las distintas pruebas, de modo a que cada distrito que desea aplicar el método escoja el grado de priorización de cada objetivo.

Caceres et al. en [6] presentan una formulación Multi-Objetivo, considerando minimizar el tiempo de recorrido del transporte y la cantidad de buses escolares utilizados. En las pruebas realizadas redujeron la cantidad de buses utilizados de 86 a 77 en el distrito escolar.

En el caso de Schittekat et al. en [7], plantean como funciones objetivo minimizar la distancia recorrida por los estudiantes desde sus hogares hasta la parada del autobus y minimizar la cantidad de paradas del autobus escolar. Las pruebas fueron realizadas tomando 200 estudiantes y 8 paradas del bus. La solución óptima definida previamente consideraba la distribución de estos estudiantes y la ruta recorrida por el bus.

Budish y Castillon en [8] realizaron una formulación Mono-Objetivo considerando maximizar la eficiencia en las asignaciones escolares, tomando como caso de estudios la escuela de negocios de Harvard. Aplicaron el método Random Serial-Dictatorship (RSD), considerando las preferencias sobre las cursos, manifestadas por los estudiantes en el proceso de inscripción. Budish y Castillon llegaron a la conclusión que los procesos manuales aplicados por la institución arrojaban mejores resultados, pero automatizar el proceso garantizaba que las asignaciones serian justas y no podrian ser sesgadas por la intervención humana.

El plan de asignación de estudiantes a escuelas públicas en los Estados Unidos, tiene como objetivo mantener la diversidad entre sus estudiantes, poniendo como meta que entre el 15 % a 50 % de los alumnos de una institución sean originarios de barrios de una clase social media-baja. T.H. Rao et al. en [9] realizaron una formulación Multi-Objetivo, utilizando programación matemática, considerando maximizar la diversidad, minimizar el recorrido del transporte escolar y maximizar la satisfacción de los padres, considerando las instituciones que estos han escogido para sus hijos. Las pruebas demostraron que es posible minimizar la distancia recorrida y maximizar la diversidad, siempre que se de mayor flexibilidad a las preferencias de los padres.

Sönmez y Ünver en [10] realizaron una investigación sobre los procesos de asignación de recursos y su aplicación a casos práticos. En el área escolar analizaron tres situaciones (1) proceso de admisión a las universidades, (2) asignación de los estudiantes a las escuelas, y (3) elección de la escuela; en todos los casos la formulación es Mono-Objetivo, planteando maximizar la eficiencia en la asignación de los recursos escolares.

Cabe destacar que ninguno de los trabajos estudiados incluye la optimización de asignaciones tomando en consideración la calidad de la infraestructura de los establecimientos, uno de los principales aportes de este trabajo. En Paraguay, este es un aspecto relevante considerando que el Ministerio de Educación y Ciencias (MEC) enfoca muchos recursos en los procesos de priorización de establecimientos considerando aspectos de aulas, mobiliarios e infraestructura edilicia, para la asignación de recursos [11].

3. Formulación Matemática Propuesta

En esta sección se presenta la formulación matemática propuesta del problema AEEE con un enfoque multi-objetivo. Se inicia con una introducción a la optimización multi-objetivo, las definiciones conceptuales, seguido de los datos de entrada, los datos de salida, el conjunto de restricciones y las funciones objetivo propuestas.

3.1. Conceptos de Optimización Multi-objetivo

Un problema general de optimización multiobjetivo (PMO) puro se compone de un conjunto de p variables de decisión, de q funciones objetivo y de r restricciones. Las funciones objetivo y las restricciones son funciones de las variables de decisión. En una formulación de PMO, x representa el vector de decisión, mientras que y representa el vector objetivo. El espacio de decisión se denota por X y el espacio objetivo como Y. Estos se pueden expresar como [12]:

Optimizar

$$y = f(x) = [f_1(x), f_2(x), ..., f_q(x)] \tag{1}$$

Sujeto a

$$e(x) = [e_1(x), e_2(x), ..., e_r(x)] \geqslant 0 \tag{2}$$

donde

$$x = [x_1, x_2, ..., x_p] \in X \tag{3}$$

$$y = [y_1, y_2, ..., y_q] \in Y \tag{4}$$

Cabe señalar que optimizar, en un contexto particular, puede significar maximizar o minimizar. El conjunto de restricciones $e(x) \geqslant 0$ define el conjunto de soluciones factibles $X_f \subset X$ y su correspondiente conjunto de vectores objetivo factibles $Y_f \subset Y$. El espacio de decisión factible X_f es el conjunto de todos los vectores de decisión x en el espacio de decisiones X que satisfacen la restricción $e(x)$, y se define como:

$$X_f = \{x | x \in X \wedge e(x) \geqslant 0\} \tag{5}$$

El espacio objetivo factible Y_f es el conjunto de vectores objetivo que representa la imagen de X_f sobre Y y se denota por:

$$Y_f = \{y | y = f(x) \ \ \forall x \in X_f\} \tag{6}$$

Para comparar dos soluciones en un contexto multiobjetivo, se utiliza el concepto de dominancia de Pareto. Dadas dos soluciones factibles $u, v \in X_f$, u domina v, denotado como $u \succ v$, si $f(u)$ es mejor o igual a $f(v)$ en cada función objetivo y estrictamente mejor en al menos una función objetivo. Si ni u domina a v, ni v domina a u, se dice que u y v no son comparables (denotados como $u \backsim v$).

Un vector de decisión x no está dominado con respecto a un conjunto U, si no hay ningún elemento de U que domine a x. El conjunto de soluciones no dominadas del conjunto de soluciones factibles X_f, se conoce como conjunto Pareto óptimo $P*$. El conjunto correspondiente de vectores objetivo constituye el frente de Pareto óptimo $PF*$.

3.2. Definiciones Conceptuales de la Formulación

Para comprender la formulación matemática se definen, seguidamente, cada uno de los conceptos considerados.

- **Establecimiento:** Lugar físico que alberga una o más instituciones.
- **Institución:** Entidad habilitada para desarrollar las clases, podría tener sede en más de un establecimiento.
- **Grado:** Corresponde al grado de enseñanza. Ejemplos: Preescolar, primer grado, etc.
- **Turno:** Corresponde al turno donde se imparte la clase. Ejemplos: Turno mañana, turno tarde, etc.
- **Sección:** En caso de que la misma institución, en el mismo establecimiento, tenga más de un grupo para el mismo grado y turno, es utilizada la sección para diferenciar. Ejemplo: 1er grado - Turno Mañana - sección A y 1er grado - Turno Mañana - sección B.
- **Clase:** Compuesto por un grado, turno, sección, institución y establecimiento. Es la unidad de enseñanza donde debe ser asignado un docente para enseñar a un grupo de alumnos. Ejemplo: 1er grado – Turno mañana – sección A – Institución 1 – Establecimiento 1.
- **Estudiante:** Alumno con necesidad de acceder a una oferta académica específica, de acuerdo con el grado al cual pertenece.

3.3. Datos de Entrada

En este apartado se presentan los datos de entrada. Primeramente, se define el conjunto de grados disponibles que se representa como un vector G de dimensión n_g:

$$G = \{1, 2, \ldots, n_g\} \tag{7}$$

donde:
n_g: representa la cantidad de grados.

El conjunto de turnos se representa como un vector T de dimensión n_t:

$$T = \{1, 2, \ldots, n_t\} \tag{8}$$

donde:
n_t: representa la cantidad de turnos.

El conjunto de secciones se representa como un vector S de dimensión n_s:

$$S = \{1, 2, \ldots, n_s\} \tag{9}$$

donde:
n_s: representa la cantidad de secciones.

El conjunto de instituciones se representa como un vector I de dimensión n_i:

$$I = \{1, 2, \ldots, n_i\} \tag{10}$$

donde:
n_i: representa la cantidad de instituciones.

El conjunto de establecimientos se representa como una matriz E de dimensión $(n_e \times 6)$:

$$E = \begin{bmatrix} N_1 & Lat_1 & Lng_1 & V_{i_1} & V_{s_1} & V_{m_1} \\ \vdots & \vdots & \vdots & & & \\ N_{n_e} & Lat_{n_e} & Lng_{n_e} & V_{i_{n_e}} & V_{s_{n_e}} & V_{m_{n_e}} \end{bmatrix} \tag{11}$$

Cada E_k es representado por el número del establecimiento, su latitud coordenada geográfica Lat, su longitud geográfica Lng y priorizaciones:

$$E_k = \begin{bmatrix} N_k & Lat_k & Lng_k & V_{i_k} & V_{s_k} & V_{m_k} \end{bmatrix} \forall k \in \{1, \ldots, n_e\}$$

donde:
N_k: número del establecimiento E_k;
Lat_k: latitud geográfica de la ubicación del establecimiento E_k;
Lng_k: longitud geográfica de la ubicación del establecimiento E_k;
V_{i_k}: priorización del establecimiento E_k en el área de infraestructura;
V_{s_k}: priorización del establecimiento E_k en el área de sanitarios;
V_{m_k}: priorización del establecimiento E_k en el área de mobiliarios;
n_e: cantidad de establecimientos.

El conjunto de estudiantes se representa como una matriz A de dimensión $(n_a \times 4)$:

$$A = \begin{bmatrix} N_1 & Lat_1 & Lng_1 & G_1 \\ \vdots & \vdots & \vdots & \\ N_{n_a} & Lat_{n_a} & Lng_{n_a} & G_{n_a} \end{bmatrix} \tag{12}$$

Cada A_i es representado por el número de estudiante, su Latitud geográfica Lat, su Longitud geográfica Lng y el grado que debe cursar::

$$A_i = \begin{bmatrix} N_i & Lat_i & Lng_i & G_i \end{bmatrix} \forall i \in \{1, \ldots, n_a\}$$

donde:
N_i: número del estudiante A_i;
Lat_i: latitud geográfica de la ubicación de la residencia del estudiante A_i;
Lng_i: longitud geográfica de la ubicación de la residencia del estudiante A_i;
G_i: grado que debe cursar el estudiante A_i;
n_a: cantidad de estudiantes.

El conjunto de clases activas se representa como una matriz C de dimensión $(n_c \times 5)$:

$$C = \begin{bmatrix} G_1 & T_1 & S_1 & I_1 & E_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ G_{n_c} & T_{n_c} & S_{n_c} & I_{n_c} & E_{n_c} \end{bmatrix} \tag{13}$$

Cada C_j es representado por el Grado, Turno, Sección, Institución , Establecimiento y Capacidad como:

$$C_j = \begin{bmatrix} G_j & T_j & S_j & I_j & E_j \end{bmatrix} \forall j \in \{1, \ldots, n_c\}$$

donde:
G_j: Grado de la clase C_j, donde $G_j \in G$;
T_j: Turno de la clase C_j, donde $T_j \in T$;
S_j: Sección de la clase C_j, donde $S_j \in S$;
I_j: Institución de la clase C_j, donde $I_j \in I$;
E_j: Número de establecimiento de C_j, donde $E_j \in E$;
n_c: cantidad de clases disponibles.

La matriz calculada U de distancias entre la residencia del estudiante y los establecimientos de dimensión $(n_p \times n_e)$ se representa como:

$$U = \begin{bmatrix} U_{1,1} & U_{1,2} & U_{1,3} & \ldots & U_{1,n_e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{n_a,1} & U_{n_a,2} & U_{n_a,3} & \ldots & U_{n_a,n_e} \end{bmatrix} \tag{14}$$

Cada $U_{i,k}$ representa la distancia entre la residencia del estudiante y un establecimiento.

$$U_{i,k} = [U_{i,k}] \forall i \in \{1, \ldots, n_a\} \wedge \forall k \in \{1, \ldots, n_e\}$$

donde:
$U_{i,k}$: Distancia entre el estudiante P_i y el establecimiento E_k.

3.4. Datos de Salida

Una solución al problema se representa por $S = S_{i,j,k}$

$$S_{i,j,k} = [S_{i,j,k}] \tag{15}$$
$$\forall i \in \{1, \ldots, n_a\} \wedge \forall j \in \{1, \ldots, n_c\} \wedge \forall k \in \{1, \ldots, n_e\}$$

donde:
$s_{i,j,k}$: es una variable binaria, donde 1 indica que el estudiante i fue asignado a la clase j del establecimiento k, 0($cero$) en caso contrario.

3.5. Restricciones

En esta sección se definen las restricciones que deben cumplir las soluciones factibles.

1. El establecimiento debe ser el asignado a la clase.

$$C_{j,4} = k \tag{16}$$
$$\forall i \in \{1, \ldots, n_p\} \wedge \forall j \in \{1, \ldots, n_c\} \wedge \forall k \in \{1, \ldots, n_e\}$$
$$: X_{i,j,k} = 1$$

2. Un alumno solo puede ser asignado a un aula.
   Esta restricción se expresa como:

$$\sum_{j=1}^{n_c} \sum_{k=1}^{n_e} x_{i,j,k} = 1 \ \forall i \in \{1, \ldots, n_a\} \tag{17}$$

3.6. Funciones Objetivo

En esta sección se presentan las 3 funciones objetivo propuestas. Las mismas se enumeran a continuación.

1. Minimizar la diferencia entre la cantidad de alumnos asignados y la cantidad óptima de alumnos por clase.
   Este objetivo mide el criterio de elección, comparando la cantidad de estudiantes asignados y la cantidad recomendada de estudiantes pedagógicamente por aula:

$$f_1(X) = \frac{\sum_{j=1}^{n_c} \sum_{i=1}^{n_a} |Q_{opt} - \sum_{k=1}^{n_e} X_{i,j,k}|}{n_c} \tag{18}$$

   donde:
   Q_{opt}: Cantidad de estudiantes recomendados pedagógicamente por aula.

2. Minimizar la Distancia promedio entre la vivienda del estudiante y el establecimiento.
   Este objetivo mide el criterio de elección, promediando las distancias entre la residencia del estudiante y el establecimiento educativo, la misma se expresa como:

$$f_2(X) = \frac{\sum_{j=1}^{n_c} \sum_{i=1}^{n_a} \sum_{k=1}^{n_e} U_{i,k} * X_{i,j,k}}{n_a} \tag{19}$$

   donde:
   $U_{i,k}$: Distancia entre el estudiante i y el establecimiento k.

3. Maximizar la utilización de establecimientos con mejor nivel de infraestructura.
   Este objetivo mide el criterio de elección, calculando la media del nivel de infraestructura y mobiliario de las aulas asignadas.

$$f_3(X) = \sum_{j=1}^{n_c} \sum_{i=1}^{n_a} \sum_{k=1}^{n_e} (E_{k,3} + E_{k,4} + E_{k,5}) * X_{i,j,k} \tag{20}$$

4. Resultados Experimentales

4.1. Implementación

Según [13], los Algoritmos Evolutivos (Evolutionary Algorithms - EAs) han demostrado ser especialmente adecuados para la optimización multiobjetivo. En este trabajo se propone un MOEA basado en NSGA-II que de acuerdo con [14] es apropiado para este tipo de problemas.

A continuación, se presenta el Algoritmo 1 con el pseudo-código del MOEA propuesto para la resolución de la formulación matemática propuesta para el problema AEEE.

Algoritmo 1 - MOEA propuesto para resolver la formulación matemática para el problema AEEE.

Entrada: $G, T, S, I, E, C, A, U, V, N_{gen}$
Salida: Conjunto Pareto (Soluciones No Dominadas)
1: Inicializar conjunto de soluciones P_0
2: $P' \leftarrow$ Reparar soluciones del conjunto P_0
3: Evaluar P'
4: **mientras** $N_{gen} \neq 0$ **hacer**
5: $Q \leftarrow$ Seleccionar individuos de P' según NSGA-II
6: $Q \leftarrow$ Aplicar operador de cruzamiento
7: $Q \leftarrow$ Aplicar operador de mutación
8: $Q \leftarrow$ Reparar individuos
9: $P \leftarrow$ Seleccionar individuos de $P \cup Q$ según NSGA-II
10: $N_{gen} \leftarrow N_{gen} - 1$
11: **fin mientras**
12: **retorna** Conjunto Pareto

Con el objetivo de dar reproducibilidad al trabajo, el código fuente, conjunto de datos y resultados experimentales se encuentran disponibles en línea[1].

4.2. Conjunto de Datos

Para los experimentos se han utilizado datos del Ministerio de Educación y Ciencias del Paraguay (MEC). Los datos se encuentran disponibles en el portal de datos abiertos de Paraguay [15]. También se utilizaron datos no públicos (anonimizados) de los estudiantes proveídos en el contexto de este trabajo. Los datos corresponde al año 2020.

En los experimentos considerados eb este trabajo se han utilizado un subconjunto de datos pertenecientes a Ciudad del Este, departamento Alto Paraná del Paraguay. Estos datos son resumidos en la Tabla 1.

Tabla 1. Resumen del conjunto de Datos de Ciudad del Este

Conjunto de datos	Cantidad
Establecimientos	90
Estudiantes	15763
Clases	501
Grados	3 (1° a 3°)
Turnos	2 (Mañana y Tarde)
Sección	6 (A, B, C, D, E y F)

[1] https://github.com/cecicasco/assign-student

De acuerdo a los datos provistos por el MEC, la distribución por grado de los establecimientos, las clases habilitadas y de los alumnos, son detallados en la Tabla 2.

Tabla 2. Distribución de establecimientos, clases y alumnos por grado

Grado	Establecimientos	Clases habilitadas	Estudiantes
1	90	169	5382
2	90	170	5393
3	90	162	4988

Como no se contaba con las coordenadas de las viviendas de los estudiantes, se ha utilizado un georreferenciador de ESRI [2], para la ubicación de las viviendas de los estudiantes de quienes se tenía los datos en el siguiente formato "País, Departamento, localidad, distrito y dirección" y el algoritmo retornó las coordenadas en el padrón "latitud, longitud". Considerando que estos registros representaban el 40 % del total, para los demás registros se generaron coordenadas de forma aleatoria dentro de los límites definidos.

Para verificar la precisión del cálculo de las coordenadas se utilizó la herramienta ArGis y se ubicó por cada establecimiento y estudiante un marcador dentro del mapa.

En la Fig. 1 se puede visualizar como los establecimientos y las viviendas quedaron dentro de los límites de Ciudad del Este, departamento Alto Paraná del Paraguay.

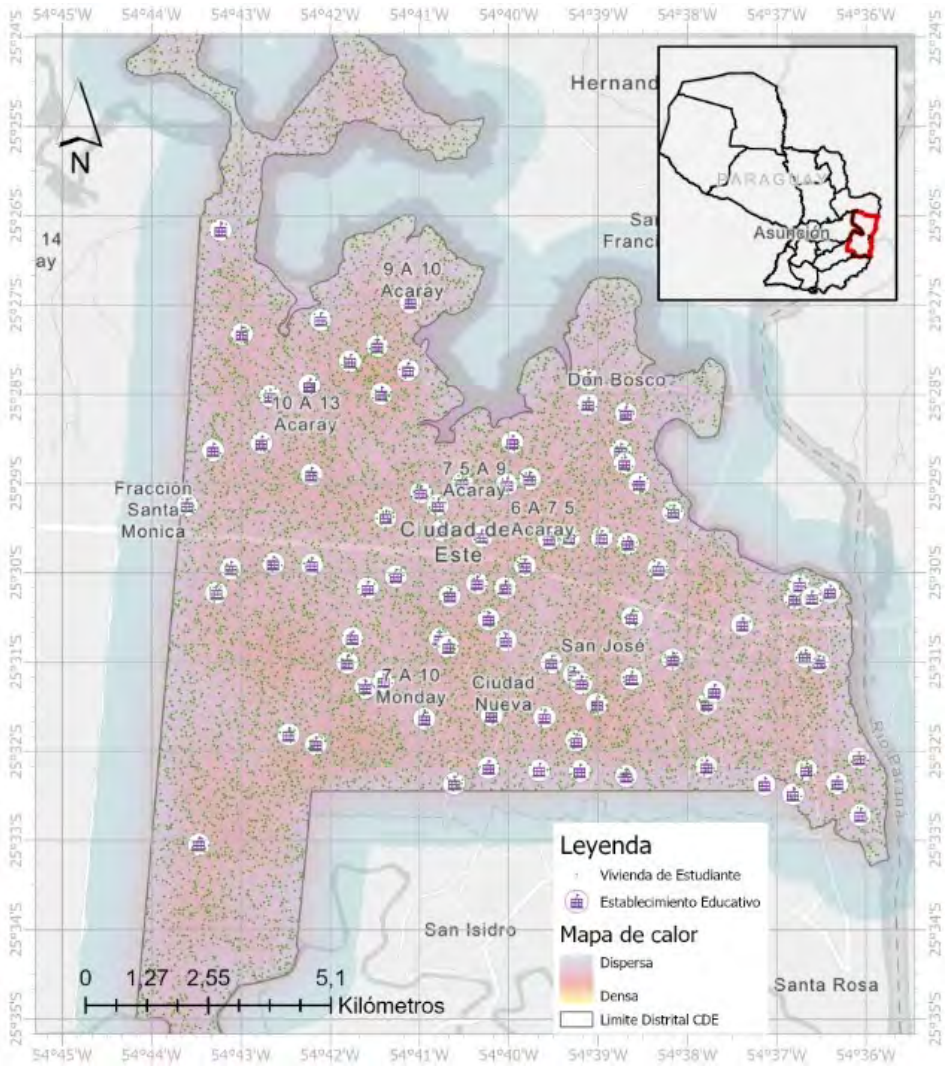


Figura 1. Ubicación de viviendas y establecimientos

[2] https://www.esri.com/en-us/home

4.3. Resultados

Dada la naturaleza del MOEA propuesto, por cada grado se han realizado 10 ejecuciones del algoritmo para el conjunto de datos descrito con una población de 100 individuos y por 100 generaciones.

Se halló un valor promedio del conjunto Pareto para cada una de las funciones objetivo, y se comparó con el valor de las mismas funciones considerando las asignaciones realizadas actualmente por el MEC. El resultado de esta comparación se puede ver en la Tabla 3.

Tabla 3. Vectores Objetivo: $y(X_{MEC})$ e $y(\overline{X})$

Grado	**Vector Objetivo**	**Min** $f_1(X)$	**Min** $f_2(X)$	**Max** $f_3(X)$
Primer	$y(X_{MEC})$	8.165	6.4812	51.028
	$y(\overline{X})$	3.301	6.458	51.921
Segundo	$y(X_{MEC})$	7.7	6.726	51.293
	$y(\overline{X})$	3.517	6.734	51.904
Tercer	$y(X_{MEC})$	6.951	6.695	51.084
	$y(\overline{X})$	2.957	6.629	52.569

Analizando los resultados obtenidos se puede destacar:

- En la función objetivo $f_1(X)$ se mejora en un 40 % el promedio de alumnos asignados por clase, considerando un óptimo teórico de 30. Esto equivale a un mayor aprovechamiento de las clases puesto que se evita la sub y sobre asignación. En las asignaciones actuales se pudo constatar que existen clases con hasta 68 o inclusive con solamente 6 alumnos por clase.
- En la $f_2(X)$ los resultados obtenidos por el algoritmo y las asignaciones actuales realizadas por el MEC no difieren significativamete, pero esta función se considera prácticamnete como no confiable puesto que el 60 % de los datos de ubicación de la vivienda de los alumnos fueron generados de manera aleatoria y no representan necesariamente la realidad.
- En $f_3(X)$ se busca maximar las asignaciones en clases con mejor nivel de infraestructura. En este sentido se mejoró las asignaciones actuales en 1 %. Esta pequeña variación se debe a que los niveles de infraestructura están distribuidos de manera bastante desigual y los establecimientos con mejor infraestructura cuentan con el total de disponibilidad ya asignada.

5. Conclusión y Líneas de trabajos Futuros

En este trabajo se ha propuesto una formulación matemática para resolver el problema de Asignación de Estudiantes a Establecimientos Educativos, implementando un nuevo algoritmo evolutivo multi-obejtivo basado en el NSGA-II para el problema AEEE.

Aplicando el algoritmo a datos de establecimientos educativos y estudiantes del primer a tercer grado, de Ciudad del Este - Paraguay, se comprobó que es posible mejorar el promedio de alumnos asignados por clase sin que esto implique un cambio en la infraestructura, simplemente que aprovechando los recursos ya disponibles y buscando aproximar las asignaciones al número de alumnos por clase recomendado pedagógicamente.

Considerando el alcance del trabajo, y las posibilidades de ampliación identificadas en el proceso de desarrollo, se propone como líneas de investigación futuras:

- Sustituir la función de minimización de distancias por minimización de tiempo de traslado, considerando medio de transporte, tipo de pavimento, entre otros.
- Incluir criterios de valoración de preferencias de los estudiantes o de los tutores para la asignación de establecimientos.
- Incorporar funciones objetivos que evaluen la diversidad étnica y socio-económica en las clases.
- Comparar experimentalmente el MOEA propuesto con otros Algoritmos Multi-Objetivo.

Referencias

[1] C. de la Nación, "Ley 1264 general de educación," vol. 9, no. 2, pp. 25–30, 2007.

[2] M. Afacan, I. Bó, and B. Turhan, "Assignment maximization," *SSRN Electronic Journal*, 10 2017.

[3] L. E. Agustín-Blas, S. Salcedo-Sanz, E. G. Ortiz-García, A. Portilla-Figueras, and Ángel M. Pérez-Bellido, "A hybrid grouping genetic algorithm for assigning students to preferred laboratory groups," *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 7234 – 7241, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417408006635

[4] K. Baker and S. Powell, "Methods for assigning students to groups: A study of alternative objective functions," *Journal of the Operational Research Society*, vol. 53, 04 2002.

[5] E. L. Bouzarth, R. Forrester, K. R. Hutson, and L. Reddoch, "Assigning students to schools to minimize both transportation costs and socioeconomic variation between schools," *Socio-Economic Planning Sciences*, vol. 64, pp. 1 – 8, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0038012116301756

[6] H. Caceres, R. Batta, and Q. He, "School bus routing with stochastic demand and duration constraints," *Transportation Science*, vol. 51, no. 4, pp. 1349–1364, 2017. [Online]. Available: https://doi.org/10.1287/trsc.2016.0721

[7] P. Schittekat, J. Kinable, K. Sörensen, M. Sevaux, F. Spieksma, and J. Springael, "A metaheuristic for the school bus routing problem with bus stop selection," *European Journal of Operational Research*, vol. 229, no. 2, pp. 518 – 528, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0377221713001586

[8] E. Budish and E. Cantillon, "The multi-unit assignment problem: Theory and evidence from course allocation at harvard," *American Economic Review*, vol. 102, no. 5, pp. 2237–71, May 2012. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/aer.102.5.2237

[9] T. Rao, A. Paleshi, G. DePuy, and B. Erenay, "A mathematical programming approach for assigning students to schools," *61st Annual IIE Conference and Expo Proceedings*, 01 2011.

[10] T. Sönmez and M. Utku Ünver, "Chapter 17 - matching, allocation, and exchange of discrete resources," ser. Handbook of Social Economics, J. Benhabib, A. Bisin, and M. O. Jackson, Eds. North-Holland, 2011, vol. 1, pp. 781 – 852. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780444531872000176

[11] "Aprendamos sobre: Priorización en fonacide." https://bit.ly/3cRl62w, accedido: 01/03/2022.

[12] C. Coello, D. Veldhuizen, and G. Lamont, *Evolutionary Algorithms for Solving Multi-Objective Problems Second Edition*, 01 2007.

[13] C. Von Lücken, A. Hermosilla, and B. Barán, "Algoritmos evolutivos para optimización multiobjetivo: Un estudio comparativo en un ambiente paralelo asíncrono," in *X Congreso Argentino de Ciencias de la Computación*, 2004.

[14] C. A. C. Flórez, R. A. B. Ocampo, and A. M. Cabrera, "Algoritmo multiobjetivo nsga ii aplicado al problema de la mochila." *Scientia et technica*, vol. 2, no. 39, pp. 206–211, 2008.

[15] "Portal de datos abiertos del mec," https://datos.mec.gov.py/, accedido: 01/03/2022.

Accesibilidad Web centrada en discapacidades visuales. Estudio empírico longitudinal de un portal de formación docente

Verónica K. Pagnoni[1], Sonia I. Mariño [2]

[1] Dirección General de Nivel Superior, Ministerio de Educación, Provincia de Corrientes, vero_pagnoni@hotmail.com

[2] Departamento de Informática. Facultad de Ciencias Exactas y Naturales y Agrimensura, Universidad Nacional del Nordeste, Corrientes, Argentina, simarinio@yahoo.com

Resumen. En contextos de educación asegurar la Accesibilidad Web (AW), es relevante para contribuir a consolidar una sociedad más justa y equitativa. El artículo analiza la Accesibilidad Web a través de un estudio descriptivo longitudinal desarrollado sobre dos versiones de un portal de formación docente continua disponible en los años 2018 y 2019, enfocado en los usuarios de una localidad del Nordeste Argentino. Para caracterizar un aspecto de las páginas validadas: la AW, se aplicaron criterios expuestos por el WCAG 2.0 y WAI-ARIA 1.1. Las evidencias de sendos estudios son objeto de comparación y análisis con la finalidad de conocer el cumplimiento de la AW, en particular se consideró su repercusión en los usuarios con discapacidad visual. Los hallazgos dan cuenta de que la incorporación de cuestiones definidas por el W3C en la actualización del portal posibilitó percepciones más positivas en los usuarios, centradas en el uso del contraste, el tamaño de fuente, el titulado de páginas, la señal visual del foco y el cambio de idioma, elementos que contribuyen a mejorar la accesibilidad web visual. Por ello, se afirma que el uso de estándares favorece efectivamente la AW, lo que da cuenta de la relevancia de fomentar su uso en todo el ciclo de vida de las soluciones informáticas, para conseguir contenidos web más accesibles para todos los usuarios.

Palabras clave: Accesibilidad Web, Educación Superior, Portales educativos, Estándares de Accesibilidad Web, Discapacidad Visual.

1 Introducción

La importancia de asegurar la accesibilidad a la educación superior y formación continua, está dada por el hecho de que estas permiten la obtención de mejores trabajos, logar una participación activa en la sociedad, por ende, a la igualdad de oportunidades. Y en lo que respecta a las personas con discapacidad les permite obtener mayor autonomía e independencia [1].

Este trabajo se enmarca en la Tesis de Maestría [2] donde se partió del planteo de la hipótesis: La aplicación de estándares de Accesibilidad Web en portales educativos mejorará la equidad e inclusión educativa de personas con discapacidades.

El artículo presenta un abordaje empírico de la Accesibilidad Web (AW) materializado en la medición de la AW de un portal, en su versión 2018 y 2019, destinado a la formación continua para el Nivel Superior, y el impacto que ésta tiene en el acceso efectivo de los usuarios al contenido Web.

1.1 Accesibilidad Web

En un ámbito en donde es relevante la igualdad de oportunidades, asegurar la Accesibilidad Web, es un elemento relevante para contribuir a consolidar una sociedad más justa y equitativa. Se entiende a la AW como la capacidad de acceso y contenidos digitales por todas las personas, sin importar las discapacidades que puedan poseer y de las características de su contexto [3].

Existen numerosas organizaciones internacionales dedicadas a esta temática, entre las que se mencionan: el W3C [4], la ISO [5] [6] [7] [8] [9], la Fundación Sidar [10], el Centro de Investigación y Desarrollo de Adaptaciones Tiflotécnicas (CIDAT), promovido por ONCE [11]. Uno de ellos es ampliamente reconocido como el WCAG enmarcado en el W3C [12].

El Consorcio World Wide Web (W3C) definió las Pautas de Accesibilidad para el Contenido Web (WCAG) [13]. Las WCAG se componen de principios, pautas y criterios de conformidad que permiten clasificar la accesibilidad de un contenido web en tres niveles: A, AA y AAA [14].

1.2 Contexto del estudio / Descripción del portal evaluado

Los institutos de educación superior no universitaria constituyen numerosas posibilidades de capacitación continua para los docentes del nivel. Un instituto de formación nacional ofrece un portal a través del cual se concretan numerosas capacitaciones en temas relacionados con áreas específicas de conocimiento, así como también concernientes a práctica docente.

2 Metodología

Se realizó una investigación cuantitativa descriptiva longitudinal, se busca caracterizar un aspecto de las páginas validadas: la Accesibilidad Web. Para ello, se aplicaron criterios y procedimientos sistemáticos [15] como los expuestos por el WCAG 2.0 [16] y WAI-ARIA 1.1 [17]. Constó de las siguientes fases:

Fase 1: Profundización de los aspectos teóricos referentes a Accesibilidad Web:

- Definición de destinatarios: el análisis de la Accesibilidad Web se centró en la discapacidad visual.
- Relevamiento de proyectos similares desarrollados en el dominio de la AW en la educación superior.

Fase 2: Definición y aplicación de una metodología para el abordaje empírico del tema:

- Estudio y elección de estándares referentes a la Accesibilidad Web: Se seleccionó la norma WC3 en su versión WCAG 2.0.
- Revisión de la Accesibilidad Web, se utilizaron:
    - Guías de revisión manual para el experto y herramienta WAVE.
    - Guías de revisión manual para el usuario.
- Selección de las páginas web a evaluar.
- Evaluación de las páginas seleccionadas.
- Procesamiento de los datos. Se utilizó una planilla de cálculo para sistematizar la información.

Fase 3: Análisis de resultados.

- Comparación de los resultados de las validaciones.
- Elaboración de conclusiones.

3 Resultados

El estudio se centró en analizar la discapacidad visual. Se realizó utilizando revisiones manuales aplicadas a las páginas seleccionadas en dos períodos diferentes: 2018 y 2019.

La elección de las páginas a evaluar se basó teniendo en cuenta las más representativas, debido a que el docente que realiza una formación, debe acceder al contenido de cada una de éstas para conocer, inscribirse, e ingresar a las clases virtuales y realizar las actividades propuestas en la formación.

Este estudio longitudinal consideró los hallazgos resultantes de evaluar el portal en el año 2018 valiéndose de la intervención de expertos. En la primera validación, se consideraron las páginas denominadas: Página 1, Página 2, Página 3. Se utilizaron una "Guía de revisión manual para el experto" [18] y la herramienta WAVE [19]. Luego un grupo de usuarios realizaron sus apreciaciones respeto de la AW en el portal utilizando la "Guía de revisión manual para el usuario" detallada en [20].

La segunda validación llevado a cabo en 2019, permitió analizar la existencia de avances en el cumplimiento de la AW, se realizó utilizando otras tres páginas web identificadas como Página 1A, Página 2A, Página 3A, con finalidades similares a las primeras, y forman parte de la actualización del portal objeto del estudio.

En la Tabla 1 se definen las páginas a analizar.

Tabla 1. Páginas seleccionadas**.**

Función en el sitio web	Denominación de Páginas iniciales	Denominación de Páginas actuales
Página de ingreso a un curso	Página 1	Página 1A
Página de ingreso a un curso	Página 2	Página 2A
Página de ingreso a un curso	Página 3	Página 3A

En las siguientes Tablas 2, 3 y 4 se detallan los resultados en la revisión del experto en ambos portales. Se sombrearon los aspectos medianamente o no cumplimentados, y se subrayaron en verde los que se modificaron en el segundo conjunto de páginas.

Tabla 1-Aplicación de grilla de revisión manual del experto a la Página 1 y 1A

Criterio	Página 1			Página 1A		
	Cumple (C)	Cumple Medianamente (CM)	No cumple (NC)	Cumple (C)	Cumple Medianamente (CM)	No cumple (NC)
Principio Perceptible						
1-Imágenes 1.1.1 A	X			X		
3-Etiquetas 1.3.1 A		X		X		
4-Color 1.4.1 A	X			X		
6-Contraste 1.4.3 AA			X	X		
7-Tamaño del texto 1.4.4 AA	X			X		
Principio Operable						
8-Accesible con el teclado 2.1.2 A	X			X		
12-Título de página 2.4.2 A		X			X	
14-Foco visible 2.4.7 AA		X		X		
15-Enlaces 2.4.9 A	X					
Principio Comprensible						
16-Idioma de la página 3.1.1 A		X		X		
Principio Robusto						
20-Hojas de estilo 4.1.1 A	X			X		
21-Maquetación 4.1.1 A	X			X		

Tabla 2-Aplicación de grilla de revisión manual a la Página 2 y 2A.

Criterio	Página 2			Página 2A		
	Cumple (C)	Cumple Medianamente (CM)	No cumple (NC)	Cumple (C)	Cumple Medianamente (CM)	No cumple (NC)
Principio Perceptible						
1-Imágenes 1.1.1 A	X			X		
3-Etiquetas 1.3.1 A	X			X		
4-Color 1.4.1 A	X			X		
6-Contraste 1.4.3 AA			X	X		
7-Tamaño del texto 1.4.4 AA	X			X		
Principio Operable						
8-Accesible con el teclado 2.1.2 A	X			X		
12-Título de página 2.4.2 A		X		X		
14-Foco visible 2.4.7 AA		X		X		
15-Enlaces 2.4.9 A	X			X		
Principio Comprensible						
16-Idioma de la página 3.1.1 A		X		X		
17-Formularios 3.2.2 A	X			X		
18Asociación de etiquetas y controles 3.3.2 A	X			X		
19 Sugerencias ante errores 3.3.3 A		X		X		
Principio Robusto						
20-Hojas de estilo 4.1.1 A	X			X		
21-Maquetación 4.1.1 A	X			X		

Tabla 3-Aplicación de grilla de revisión manual a la Página 3 y 3A.

Criterio	Página 3			Página 3A		
	Cumple (C)	Cumple Medianamente (CM)	No cumple (NC)	Cumple (C)	Cumple Medianamente (CM)	No cumple (NC)
Principio Perceptible						
1-Imágenes 1.1.1 A	X			X		
3-Etiquetas 1.3.1 A		X		X		
4-Color 1.4.1 A	X			X		
6-Contraste 1.4.3 AA			X	X		
7-Tamaño del texto 1.4.4 AA	X			X		
Principio Operable						
8-Accesible con el teclado 2.1.2 A		X		X		
12-Título de página 2.4.2 A		X		X		
14-Foco visible 2.4.7 AA		X		X		
15-Enlaces 2.4.9 A	X			X		
Principio Comprensible						
16-Idioma de la página 3.1.1 A		X		X		
17-Formularios 3.2.2 A						
18Asociación de etiquetas y controles 3.3.2 A	X			X		
19 Sugerencias ante errores 3.3.3 A	X			X		
Principio Robusto						
20-Hojas de estilo 4.1.1 A		X		X		
21-Maquetación 4.1.1 A	X			X		

El análisis de las evidencias recuperadas en las Tablas 2 a 4, permiten apreciar que:

- En la Página 1: al deshabilitar los estilos en esta página no se pierde información; existen 30 errores de contraste; el título de página, podría ser más descriptivo de su contenido; el foco se muestra con una señalización visual, aunque es difícil de notar y el idioma es únicamente el español.
- En la Página 1A: al deshabilitar los estilos en esta página no se pierde información, el contraste es aceptable considerando la relación de color de fondo y primer plano, el título de página es descriptivo del contenido, el foco se muestra con una señalización visual intensa, el idioma de la página se puede cambiar.
- En la Página 2: deshabilitando los estilos en esta página no se pierde información; posee 9 errores de contraste; para el envío de los datos se usa el botón de envío convencional; en el formulario que posee la página cada campo cuenta con etiquetas representativas; si bien se dan sugerencias cuando el usuario

comete errores al ingresar datos, el mensaje brindado podría ser más claro; el foco se muestra con una señalización visual, esta no se visualiza fácilmente; el título de la página se considera adecuado y descriptivo de su contenido y el idioma de la página no se puede modificar.

- En la Página 2A: deshabilitando los estilos en esta página no se pierde información; el contraste es aceptable considerando la relación de color de fondo y primer plano; para el envío de los datos se usa el botón que indica fehacientemente la acción a realizar; en el formulario que posee la página cada campo cuenta con etiquetas representativas; las sugerencias cuando el usuario comete errores al ingresar datos son claras; el foco se muestra con una señalización visual intensa; el título de la página se considera adecuado y descriptivo de su contenido y el idioma de la página se pude modificar.
- En la Página 3: al deshabilitar los estilos en esta página no se pierde información; posee 17 errores de contraste; el título de página, podría ser más descriptivo de su contenido; en cuanto al acceso por teclado, existen dos elementos que no pueden ser accedidos utilizando la tecla TAB; el foco se muestra con una señalización visual, pero ésta no es fácilmente identificable y el idioma de esta página es español y no hay puede ser modificado.
- En la Página 3A: al deshabilitar los estilos en esta página no se pierde información; el contraste es aceptable considerando la relación de color de fondo y primer plano; el título de página es descriptivo de su contenido; todos los elementos que pueden ser accedidos utilizando la tecla TAB; el foco se muestra con una señalización visual intensa y el idioma se puede ser modificar.

Comparando los datos obtenidos, se evidencia la existencia de mejoras de cumplimiento de la AW en la versión 2019 de portal analizado, en los principios Perceptible, Operable y Comprensible. Particularmente, es notorio un mejor uso de: etiquetas, contraste, accesibilidad con teclado, foco visible, idioma de la página, sugerencias de error y título de página.

Respecto de las pruebas manuales, las ejecutaron se llevaron adelante por 15 docentes que realizaron formaciones académicas utilizando el portal. Los implicados recibieron electrónicamente un formulario a completar según se especificó en la descripción del instrumento. Las respuestas al cumplimiento de los criterios planteados en las páginas seleccionadas siguen el siguiente formato: Si (S), Medianamente (M) o No (N). La caracterización de los evaluadores de los aspectos generales indagados, se detalla en la Tabla 5. Para ejemplificar los resultados de la evaluación, se incluyen las Tablas 6 y 7. Se sombrearon los aspectos medianamente o no cumplimentados, y se subrayaron en verde los modificados en el segundo conjunto de páginas.

Tabla 4 - Caracterización de los evaluadores

Datos	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
Edad	51	66	28	27	28	65	45	38	28	35	41	46	38	38	26
Género	F	M	M	F	F	M	F	F	F	M	M	M	M	F	F
Tipo de dispositivo utilizado	PC, Celular	PC	PC, Tablet	PC, Celular	PC, Celular	PC	PC, Celular	PC, Celular	PC, Celular	PC, Celular	PC, Celular	PC, Celular	PC, Celular	PC, Celular	PC, Celular
Problema visual	Si	Si	No	Si	No	Si	Si	Si	No	No	Si	Si	Si	No	No

Tabla 5 -Respuestas de usuarios a la grilla de revisión manual- Página 1

Criterios	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1.Imágenes	Si	Si	Si	No	Si	No	Si	Si	Si	Si	Si	Si	Si	Si	Si
3.Etiquetas	Si	Si	Si	Si	S	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
4.Color	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
6. Contraste	M	No	Si	No	Si	No	No	No	Si	M	No	No	M	Si	Si
7.Tamaño de fuente	M	No	Si	No	Si	No	M	No	Si	Si	M	M	M	Si	Si
8.Acceso por teclado	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
12.Título de página	Si	No	M	No	No	M	M	No	No	M	M	M	No	No	Si
14.Señal visual de foco	M	No	Si	No	Si	No	M	No	Si	Si	M	M	Si	Si	Si
15.Propósito de link	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
16.Cambio de idioma	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

Tabla 6-Respuestas de usuarios a la grilla de revisión manual- Página 1ª

Criterio	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1.Imágenes	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
3.Etiquetas	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
4.Color	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
6. Contraste	M	Si	Si	Si	Si	Si	Si	Si	Si	M	M	M	M	Si	Si
7.Tamaño de fuente	M	No	Si	Si	Si	No	No	M	Si	Si	M	M	M	Si	Si
8.Acceso por teclado	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
12.Título de página	Si	M	M	No	No	M	M	No	No	M	No	M	No	No	Si
14.Señal visual de foco	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
15.Propósito de link	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
16.Cambio de idioma	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si

Tabla 7-Respuestas de usuarios a la grilla de revisión manual- Página 2

Criterio	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1.Imágenes	Si	Si	Si	No	Si	No	Si	Si	Si	Si	Si	Si	Si	Si	Si
3.Etiquetas	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
4.Color	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
6. Contraste	Si	Si	Si	No	Si	No	Si	Si	Si	Si	Si	Si	Si	Si	Si
7.Tamaño de fuente	M	No	Si	No	Si	No	M	No	Si	Si	M	M	Si	Si	Si
8.Acceso por teclado	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
12.Título de página	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
14.Señal visual de foco	M	No	Si	Si	Si	No	M	No	Si	Si	M	M	Si	Si	Si
15.Propósito de link	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
16.Cambio de idioma	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
17.Envío de datos	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
18.Etiquetas en formularios	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
19.Mensajes de error	M	No	Si	No	Si	No	No	No	Si	Si	M	No	Si	Si	Si

Tabla 8-Respuestas de usuarios a la grilla de revisión manual- Página 2A

Criterio	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1.Imágenes	Si	Si	Si	Si	Si	No	Si	Si	Si	Si	Si	Si	Si	Si	Si
3.Etiquetas	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
4.Color	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
6. Contraste	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
7.Tamaño de fuente	M	No	Si	Si	Si	Si	Si	Si	Si	Si	M	M	Si	Si	Si
8.Acceso por teclado	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
12.Título de página	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
14.Señal visual de foco	Si	Si	Si	Si	Si	M	M	Si	Si	Si	M	Si	Si	Si	Si
15.Propósito de link	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
16.Cambio de idioma	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
17.Envío de datos	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
18.Etiquetas en formularios	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
19.Mensajes de error	M	Si	Si	Si	Si	Si	Si	Si	Si	Si	M	Si	Si	Si	Si

Tabla 9-Respuestas de usuarios a la grilla de revisión manual- Página 3

Criterio	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1.Imágenes	Si	Si	Si	No	Si	No	Si	Si	Si	Si	Si	Si	Si	Si	Si
3.Etiquetas	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
4.Color	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
6.Contraste	M	No	Si	No	Si	No	No	No	Si	M	No	No	M	Si	Si
7.Tamaño de fuente	M	No	Si	No	Si	No	M	No	Si	Si	M	M	M	Si	Si
8.Acceso por teclado	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
12.Título de página	No	No	M	No	No	M	M	No	No	M	M	M	No	No	M
14.Señal visual de foco	M	No	Si	No	Si	No	M	No	Si	Si	M	M	Si	Si	Si
15.Propósito de link	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
16.Cambio de idioma	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No

Tabla 10-Respuestas de usuarios a la grilla de revisión manual- Página 3A

Criterio	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15
1.Imágenes	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
3.Etiquetas	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
4.Color	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
6.Contraste	M	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	M	Si	Si
7.Tamaño de fuente	M	No	No	No	No	No	M	No	No	No	M	M	M	No	No
8.Acceso por teclado	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
12.Título de página	M	M	M	No	No	M	M	No	No	M	M	M	No	No	M
14.Señal visual de foco	Si	Si	Si	Si	Si	Si	M	M	Si	Si	Si	Si	Si	Si	Si
15.Propósito de link	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si
16.Cambio de idioma	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si	Si

Comparando los datos de las Tablas 6 y 9, 7 y 10, 8 y 11, se determina que existe una percepción más positiva por parte de los usuarios al validar la segunda versión del portal. Según sus apreciaciones, en las Página 1A, Página 2A y Página 3A se mejoró el uso del contraste, el tamaño de fuente, el titulado de páginas, la señal visual del foco y el cambio de idioma, elementos que contribuyen a mejorar la accesibilidad web visual.

4. Conclusiones

La validación realizada por el experto corrobora la idea que la segunda versión del portal cumplimenta en mayor medida los estándares de AW. Además, al realizar la comparativa entre las percepciones de los usuarios a las dos versiones del portal, se evidenció su mejor aceptación al segundo conjunto de páginas web. Esta notable mejoría de la percepción en diversos aspectos, permiten acceder a la información de manera más rápida y clara, propiciando el desarrollo de las formaciones académicas en igualdad de condiciones que los restantes.

Estas mejorías redundan en un mayor acceso a la información. Por ello, el estudio reafirma la necesidad de fomentar el uso de estándares en el ciclo de vida de las soluciones informáticas.

Por último, se concluye que, este estudio realizado permite afirmar que la aplicación de estándares de Accesibilidad Web, como los definidos por el W3C, en estos portales mejora la equidad e inclusión educativa de personas con discapacidades, en este caso visuales. Por lo expuesto, se comprueba la hipótesis propuesta inicialmente: "La aplicación de estándares de AW en plataformas y portales educativos mejorará la equidad e inclusión educativa de personas con discapacidades".

Referencias

1. B. S. Misischia, "Derecho a la educación universitaria de personas con Discapacidad", 2013. *Revista Latinoamericana de Educación Inclusiva* Vol. 8, Nº 1, marzo - agosto 2014, pp. 25 - 33
2. V. K. Pagnoni, "Aportes a la inclusión educativa. Indagación en torno a la Accesibilidad Web de un portal educativo nacional según el estándar WCAG 2.0", Tesis de la Maestría en Educación en Entornos Virtuales, Dir. S. I. Mariño, Universidad de la Patagonia Austral, 2021.
3. Y. Stable Rodríguez & C. A. Sam Anlas, "National Libraries and Web Accessibility. Situation in Latin America", *Revista Interamericana de Bibliotecología*, 41(3), 253-265. 2018.
4. W3C (2021). ACERCA DEL W3C [Online]. Available: https://www-w3-org.translate.goog/Consortium/
5. ISO. Organización Internacional para la Estandarización. [Online]. Available: http://www.iso.org/iso/home.html
6. ISO (2012). ISO/IEC 40500:2012. Information technology - W3C Web Content Accessibility Guidelines (WCAG) 2.0. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=58625
7. ISO (2008). ISO/IEC 9241-151:2008 Ergonomics of human-system interaction – Part 151: Guidance on World Wide Web user interfaces.
8. ISO (2008). ISO/IEC 9241-171:2008. Ergonomics of human-system interaction – Part 171: Guidance on software accessibility.

9. ISO (2008). ISO/IEC 9241-20:2008. Ergonomics of human-system interaction – Part 20: Accessibility guidelines or information/communication technology (ICT) equipment and services.
10. Fundación Sidar (2015). Fundación Sidar - Acceso Universal. [Online]. Available: http://www.sidar.org/
11. ONCE (2016). Centro de Investigación. Desarrollo y Aplicación Tiflotécnica. [Online]. Available: http://cidat.once.es
12. W3C (2018). Introducción a las Pautas de Accesibilidad para el Contenido Web (WCAG). [Online]. Available: https://www.w3.org/WAI/standards-guidelines/wcag/es
13. L. F. Londoño Rojas, V. Tabares Morales, M. R. Bez & N. D. Duque Mendez, "Análisis comparativo de guías para el desarrollo web accesible", *Ciencia e Ingeniería Neogranadina*, 28(1), 101-115. 2017.
14. N. Duque, J. Flores, & N. Castaño, "Accesibilidad en sitios web colombianos", *Ingeniería E Innovación, 2*(1), 34-41. 2014.
15. G. P.Guevara Alban, A. E. Verdesoto Arguello& N. E. Castro Molina, "Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción)", *RECIMUNDO*, 4(3), 163-173. 2020.
16. W3C (2018). Web Content Accessibility Guidelines (WCAG) 2.0. [Online]. Available: https://www.w3.org/TR/WCAG20/
17. W3C (2017). Accessible Rich Internet Applications (WAI-ARIA) 1.1. [Online]. Available: https://www.w3.org/TR/wai-aria/
18. V. K. Pagnoni & S. I. Mariño, "Accesibilidad Web centradas en revisiones manuales. Estudio de un EVA de formación docente continua", (inédito).
19. WAVE. Web Accessibility Evaluation Tool. [Online]. Available: https://wave.webaim.org/
20. V. K. Pagnoni, S. I. Mariño. "Una guía de Accesibilidad Web para portales educativos. La revisión de usuarios", Congreso Argentino de Ciencias de la Computación, p. 133- 141, 2021.

Tecnologías de Reconocimiento Automático de Voz en Contextos Educativos. Una Revisión Sistemática de Literatura.

Marcelo Zampar, Susana Herrera

Instituto de Investigación en Informática y Sistemas de Información, Universidad Nacional de Santiago del Estero, 1912 Av. Belgrano (S), Santiago del Estero, Argentina
{mzampar,sherrera}@unse.edu.ar

Resumen. Este artículo se vincula con las tecnologías educativas en la educación superior de las personas con discapacidad auditiva, quienes tienen distintas formas de comunicarse (lengua de señas, lectura de labios, escritura); siendo la escritura la forma más aceptada. Constantemente, se van generando alternativas de uso del lenguaje escrito como dispositivos de educación inclusiva para optimizar la comprensión lingüística en los hipoacúsicos. De allí la necesidad de que la clase del profesor pueda ser comunicada en lenguaje escrito. Y, para ello, las instituciones y los docentes requieren tener acceso a sistemas de reconocimiento automático de voz, que convierten el audio en texto. En este contexto, se realizó una revisión sistemática de literatura con el objetivo de analizar cómo estos sistemas mejoran la comprensión lingüística del sordo. La misma incluyó 171 artículos que fueron analizados teniendo en cuenta aspectos tecnológicos, pedagógicos y auditivos. Los resultados obtenidos muestran un recorrido incipiente en esta área y abre caminos para que los investigadores trabajen con el lenguaje escrito como medio para que la comunidad sorda pueda optimizar su competencia lingüística.

Keywords: reconocimiento automático de voz, tecnologías educativas, educación inclusiva, educación superior de personas con discapacidad auditiva.

1 Introducción

La UNESCO y las Naciones Unidas proclaman una serie de principios relacionados a las personas con discapacidad. Especialmente, definen que estas personas deben tener las mismas oportunidades de desarrollo que las demás. En consecuencia, la educación inclusiva es un objetivo global importantísimo en la política, la investigación y la práctica educativa [3]. Varias discapacidades deben ser tenidas en cuenta para que haya una verdadera inclusión educativa. Una de ellas es la deficiencia auditiva, la cual se define como la pérdida de la función anatómica o fisiológica del sistema auditivo. Tiene su consecuencia inmediata en un déficit del acceso al lenguaje oral, afecta el desarrollo lingüístico y comunicativo; y la posterior integración escolar, social y laboral del sordo [4]. Es así, que ya hace varios años el nivel superior en integración con instituciones especiales aceptan intérpretes de lengua de señas en el aula de clase. Sin embargo, estos profesionales no son docentes o no son capaces de comprender correctamente las lecciones, con implicaciones negativas en el aprendizaje [5]. Por ello, estos establecimientos especiales comienzan a combinar profesores con intérpretes para acompañar al sordo en la clase del profesor de la asignatura.

Fen & Cheng, [5] afirman que todos los métodos de enseñanza tradicionales como el lenguaje de señas, la lectura de labios y la escritura en gran medida tienen deficiencias. Aun así, aseveran que la escritura es la forma más clara, precisa y fácil de

aceptar para las personas sordas y con problemas de audición. Aseguran que en el aula, este lenguaje escrito puede expresar con mayor eficacia el contenido de la enseñanza, atraer el interés y mejorar la calidad de aprendizaje.

Continuamente, se generan alternativas de cómo utilizar el lenguaje escrito y se conciben técnicas para optimizar la comprensión lingüística escrita en el hipoacúsico. Un caso concreto es la Logogenia. Este método considera que la adquisición de una lengua, desde la mirada de la Gramática Generativa Transformacional [6], es una facultad biológica innata, que utiliza como entrada la lengua oral. Esta, activa los mecanismos de adquisición para comprender la lengua materna. En base a ello, se creó el método denominado Logogenia; en la cual se sustituye la entrada de la lengua oral en el oyente por la lengua escrita en el sordo [7, 8 y 9]. Pero para aplicar este método, se requiere que los alumnos hipoacúsicos dispongan de los textos escritos.

Una clase contiene actividades, evaluaciones, apuntes propios del docente, bibliografía y demás dispositivos didácticos y pedagógicos. Es muy difícil llevar todos estos elementos al lenguaje escrito. Y es aquí donde los Sistemas de Reconocimiento de Voz (SRA del inglés Speech Recognition Automatic) tienen un rol importante como tecnología que facilita la conversión de lenguaje oral a lenguaje escrito. Conceptualmente, un SRA es una herramienta computacional capaz de procesar la señal de voz emitida por el ser humano y reconocer la información contenida en esta, convirtiéndola en texto o emitiendo órdenes [10]. Un SRA intenta resolver el problema de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido [11].

Los SRA convierten, de forma automática, la voz en texto. Y, contar con lenguaje escrito permite mejorar la competencia lingüística del hipoacúsico, mediante la utilización de métodos como la Logogenia. Es por ello que los autores consideraron relevante llevar adelante una RSL sobre los SRA y su uso en contextos educativos, cuyos resultados se presentan en este artículo.

El artículo está organizado de la siguiente manera. En la Sección 2, se describe el método utilizado para la revisión. En la Sección 3, se presentan el análisis y resultados de la RSL, incluyendo los hallazgos relacionados a las preguntas de investigación. Finalmente, en la Sección 4 se sintetizan las conclusiones del estudio.

2 Método

La RSL sigue los lineamientos propuestos por Petticrew & Roberts [12] para este tipo de investigación científica, renovados a la luz de los aportes de Lavallée [13]. En consecuencia, el protocolo de búsqueda, selección, y análisis de la evidencia empírica se ajusta a las siete etapas sugeridas por los autores para su desarrollo. Estas etapas, esquematizadas en la Fig. 1, son: (1) definición de las preguntas de investigación o de las hipótesis; (2) especificación de los tipos de estudios que deben ser considerados; (3) realización de una búsqueda exhaustiva de la literatura; (4) evaluación de los resultados de la búsqueda y selección de artículos; (5) análisis de los estudios incluidos; (6) síntesis; y (7) difusión de los hallazgos de la revisión.

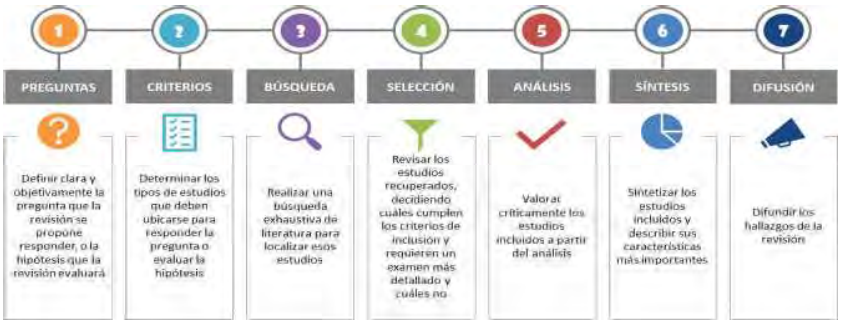


Figura 1. Fases de RSL según Petticrew & Roberts. Elaboración Dieser [15].

En relación a la etapa 1, la pregunta que ha guiado la revisión es: ¿Cuáles tecnologías SRA se usan en el proceso de aprendizaje de personas hipoacúsicas o sordas? ¿Contribuyen estas tecnologías a mejorar la comprensión escrita de las personas sordas?

Asimismo, se definieron aspectos para analizar cuantitativamente los resultados:

- año de publicación
- contexto educativo: virtual, presencial, bimodal o no formal
- escenario de uso: educación o campo general
- tipo de tecnología: asistiva o de uso general
- idioma de la tecnología
- tipo de institución: ordinaria, especial, no formal
- vinculación con entornos virtuales de enseñanza-aprendizaje (EVEA)

Respecto a la etapa 2, se definió realizar una búsqueda de artículos científicos, en inglés o en español, publicados en los últimos doce años (2010 de 2022). La búsqueda en inglés se hizo en las bibliotecas digitales IEEE Xplore (http://ieeexplore.ieee.org), ACM (http://dl.acm.org) y WOS (https://www.webofscience.com). La búsqueda en español se realizó en SeDiCI (http://sedici.unlp.edu.ar) y Redalyc (http://www.redalyc.org). Las cadenas de búsqueda utilizadas fueron las siguientes:

- En inglés, “speech recognition” AND (education OR learning) AND ("hearing loss" OR "hearing impairment" OR deaf)
- En español, “reconocimiento de voz" AND (educación OR aprendizaje) AND (hipoacusia OR "discapacidad auditiva" OR sordo)

Además, para el proceso de filtrado se definieron criterios de exclusión (CE) e inclusión (CI). Los CE fueron:

- CE 1: artículos no escritos en inglés o español.
- CE 2: documentos correspondientes a actas de congresos o revistas, donde la palabra clave correspondan a diferentes artículos del mismo documento.
- CE 3: artículos que incluyen las palabras clave pero que no tratan del tema en sí mismo, es decir, experiencias que no tienen que ver con educación o aprendizaje, experiencias informales, etc.
- CE 4: artículos orientados a mejorar la comunicación de los sordos mediante el uso de la lengua de signos o gestos.
- CE 5: artículos repetidos

Los criterios de inclusión fueron:

CI 1: Estudios empíricos sobre la conversión de voz a texto, reconocimiento de voz o subtitulado para colaborar con la discapacidad auditiva en contextos educativos.

CI 2: En caso de artículos de un mismo proyecto, se consideró el más completo.

Luego, se desarrolló la etapa 3 del método, ejecutándose las búsquedas. Se obtuvo 171 artículos de todas las bases de datos seleccionadas. Posteriormente, en la etapa 4, se aplicaron los CE y CI, mediante la lectura de títulos y resúmenes de dichos artículos. Luego del filtrado quedaron seleccionados 12 artículos. El CE 4 ha sido el criterio que más artículos excluyó. La tabla 1 muestra el detalle de la cantidad de artículos encontrados y seleccionados en cada biblioteca.

Tabla 1. Resultado de las etapas 3 y 4. Artículos encontrados y seleccionados.

Biblioteca	**Artículos encontrados**	**Artículos seleccionados**
ACM	7	2
IEEE	31	5
Redalyc	97	1
Sedici	9	1
WoS	27	4
Total	171	13

3 Resultados de la revisión

En este apartado se describe la etapa 5 de la RSL. Es decir, se presenta un análisis cuantitativo y cualitativo de los 12 artículos seleccionados. En la tabla 2 se presentan dichos artículos, referenciados desde la letra A hasta la K.

3.1 Análisis cuantitativo

Siguiendo los aspectos mencionados previamente, se analizaron los artículos seleccionados y se obtuvieron los resultados expuestos a continuación.

Se observa que la mayoría de los artículos se publicaron desde el año 2015 en adelante, excepto tres. Esto muestra que los SRA evolucionaron rápidamente en los últimos años. Según el contexto educativo, seis de los artículos se utilizan en la educación virtual [A, G,H,I,J,K], ocho en la educación presencial [A,C,D,G,H,I,J,K], seis en la educación bimodal [A,G,H,I,J,K] y cuatro no son del ambiente educativo formal [B,E,L], como se muestra en la tabla 2.

Tabla 2: Año de publicación de los artículos y contexto educativo de tecnologías RSA.

			Contexto educativo			
			Virtual	Presencial	Bimodal	No Formal
Ref	**Artículo**	**Año**	6	8	6	4
A	Kushalnagar & Cols .	2012	1	1	1	
B	Glasser	2019				1
C	Jun & Cheng	2011		1		
D	Fen & Cheng	2010		1		
E	Kosuke & Cols.	2017				1
F	Aye, & Cols.	2020				1
G	Batista	2016	1	1	1	

H	Kheir & Way	2015	1	1	1	
I	Kuldeep & Cols.	2021	1	1	1	
J	Alvarez & Rufrancos	2016	1	1	1	
K	Shashidhar & Cols.	2021	1	1	1	
L	Le & Cols.	2011				1

En cuanto al escenario de uso, como se ve en la Fig. 2, diez artículos están orientados a educación y solo dos artículos al campo general [B,L]. En cuanto a las tecnologías involucradas, en la Fig. 3 se aprecia que ocho son asistivas y cuatro de uso general [B,G,I,L].

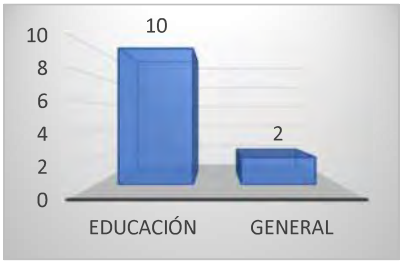


Figura 2. Escenario de Uso.

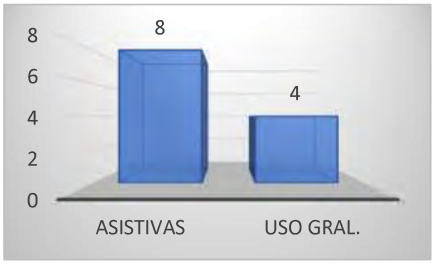


Figura 3. Tecnologías involucradas.

Según el idioma en el que se han desarrollado las tecnologías, la Fig. 4 muestra que tres son en inglés [A,B,H], uno en Myanmar ex Birmania [F], dos en español [G,J], tres en chino [C,D,L], uno en japonés [E] y dos en indiano [I,K]. Como se observa en la Fig. 5, en cuanto a la vinculación con un EVEA once artículos no están vinculados con EVEA, mientras que un artículo propone el uso de chatbots como apoyo para la comunicación en el aula [G].

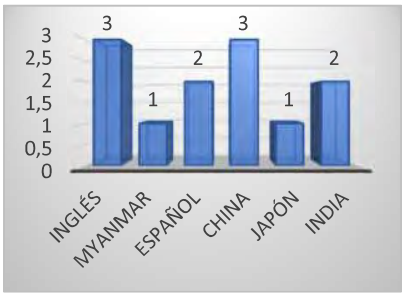


Figura 4. Idioma.

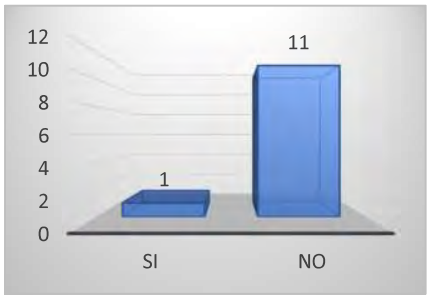


Figura 5. Vinculación a un EVEA.

En la Fig. 6 se observa en qué tipo de instituciones se utilizó la tecnología. Cinco artículos se implementan en instituciones educativas ordinarias, ocho en especiales y cinco de manera no formal. Es decir, hay tecnologías implementadas en más de una institución. Los SRA están mayormente disponibles en el idioma local. Las excepciones son los grandes SRA como el de Google. Esto puede inferir que cada lengua tiene

fonemas diferentes en calidad y en cantidad. Incluso, dentro de un mismo idioma, las diferencias regionales también involucran diferentes fonemas. Finalmente, la Fig. 7 hace referencia a la manera en que fueron validadas las tecnologías propuestas. Ocho de los SRA [C,D,E,G,H,J,K,L] fueron evaluados por usuarios finales en contextos reales, ya sea en instituciones de educación ordinaria o especial. El resto de las contribuciones se evaluaron en el laboratorio, es decir, no en su contexto real de uso. Al mencionar la realización de evaluaciones indicaron que los resultados obtenidos tuvieron aspectos positivos para el aprendizaje. Por supuesto, en general, aprecian que todavía hay puntos en los que necesitan seguir investigando.

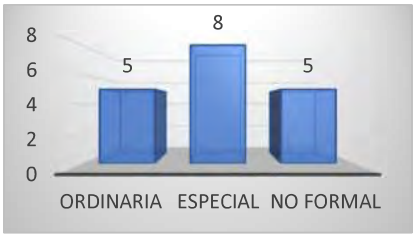


Figura 6. Tipo de instituciones educativas.

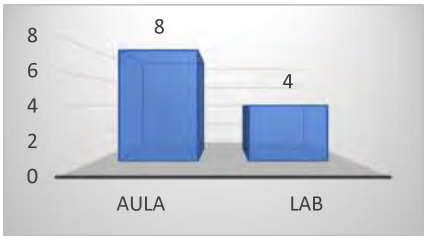


Figura 7. Evaluación.

3.2 Hallazgos relacionados a las preguntas de investigación

A continuación, se presentan los principales hallazgos de la RSL, en relación a los SRA para personas con discapacidad auditiva.

- Además de los enfoques tradicionales de subtituladores profesionales y los SRA en un artículo se ofrece un nuevo enfoque colaborativo de subtítulos entre compañeros de clase [A]. Permite la transcripción en tiempo real de múltiples personas no expertas, en el que se pueden utilizar mecanismos de acuerdo colectivo para evaluar la calidad de la transcripción. Posibilita la conversión de voz a texto de una manera más económica que contratar un subtitulador profesional y señala que existen casos de uso actualmente que están más allá del alcance de los SRA. Los subtituladores no expertos no necesitan una capacitación extensa para adquirir habilidades específicas y pueden obtenerse de una variedad de fuentes: compañeros de clase, miembros de la audiencia, voluntarios, etc.
- En [B] se evalúan RSA con voces de hablantes sordos y con problemas de audición (DHH del inglés Deaf and Hard-of-Hearing). RSA ha mejorado a lo largo de los años y es capaz de alcanzar tasas de errores tan bajos como 5-6 % con la ayuda de algoritmos de computación en la nube y aprendizaje automático que toman en cuenta modelos de vocabulario. Se utilizaron dos motores de RSA muy populares y disponibles gratuitamente para uso público, como lo son Microsoft Translator Speech API y el servicio IBM Watson Speech to Text. Es probable que obtengan resultados impredecibles de los ASR, ya que los patrones en el habla son muy diversos dentro de la población DHH. En consecuencia, la tasa de error aumenta en comparación a la población que no es DHH.
- [C] afirma que la elaboración de texto escrito para comunicarse en la educación de estudiantes sordos siempre ha sido difícil. Con la maduración de los SRA la conversión de voz a texto puede hacer posible que llegue más rápido al aula

y profesores. Se prueba con los modelos de SRA ViaVoice y ViaScribe para cubrir en tiempo real esta situación. Mejorar la tasa de reconocimiento es la mayor preocupación del docente.

- Otro estudio [D] también prueba los modelos de SRA ViaVoice y ViaScribe; ahora, desde el punto de vista del estudiante sordo y del profesor. Aseguran los autores que el SRA en el aula promueve el desarrollo de la competencia profesional del docente, expresa con eficacia el contenido de la enseñanza y compensa la insuficiencia de los métodos de enseñanza tradicionales; como el lenguaje de señas, la lectura de labios y la escritura en gran medida. Específicamente para el alumno sordo propicia el interés y la calidad del aprendizaje; y puede revisar oportunamente después de clase los contenidos de la asignatura.
- En [E] se describe un RSA que aprende la voz humana circundante y los sonidos emitidos por un objeto. Transmite visualmente en el dispositivo que reside la aplicación a las personas con discapacidad auditiva y sus cuidadores. A partir del resultado del aprendizaje, el sonido se analiza mediante la comparación de patrones para determinar qué tipo de sonido es. El RSA reduce la carga de los cuidadores y propicia para que estos puedan comunicarse mejor con los discapacitados auditivos.
- La aplicación móvil VOIS para niños sordos utiliza un SRA birmano basado en una red neuronal convolucional (CNN), en la que los datos de voz son entrenados para reconocer cada palabra en su pronunciación correcta [F]. Los modelos entrenados y el motor RSA está integrado en la aplicación para poder reconocer el habla sin conexión. Cuando se conecta a Internet, los modelos de redes neuronales se pueden actualizar para reconocer nuevas palabras. Puede ayudar a los niños con discapacidad auditiva a entrenar el idioma a su propio ritmo y a comprender los conceptos básicos. Proporciona palabras estructuradas de una y dos sílabas, recopiladas en materiales educativos y de comunicación de la vida real.
- [G] asevera que para optimizar la comprensión escrita del alumnado se puede apelar a mecanismos interactivos y eficientes que operan con cierta autonomía en una amplia disponibilidad de acceso, como los programas robot conversacionales, o chatbots. Es adecuado para sistematizar respuestas a dudas o consultas de índole operativa que suelen repetirse de manera constante entre los participantes de los diferentes cursos. Asegura que un individuo al interesarse en el desarrollo de la lógica de su programación, se puede animar a incorporar conversaciones de conceptos o temáticas de la asignatura correspondiente en formato texto.
- Las clases en el nivel superior con frecuencia contienen terminología específica de un dominio y representan un desafío para los SRA, sistemas que generalmente se basan en un diccionario de palabras comunes para guiar el reconocimiento. [H] consideran también al SRA ViaScribe, el cual cuenta detección de pausas en el habla, inserción de oraciones y saltos de párrafo, ortografía fonética cuando el SR no está seguro y un modo independiente del hablante para adaptarse a varios de ellos. Esto mejora la capacidad del SRA para ayudar a los estudiantes sordos y con problemas de audición a tomar notas en el aula y optimizar su comprensión escrita
- En un trabajo se utiliza SRA que se puede categorizar en acústico-fonético, detección de patrones y otros enfoques matemáticos[I]. Aquí, el SRA separa los argumentos pronunciados en el procesamiento de señales que transforma el audio

hablado en texto en un formato legible. Los autores afirman que el SRA permite autenticar la individualidad del usuario utilizando su voz, mejorando la eficiencia y precisión de la comprensión escrita en diferentes lugares de trabajo.

- En otro trabajo declaran que el mercado actual es muy limitado en cuanto a aplicaciones dedicadas a sordos, e incluso las que hay poseen escasas funciones. Si nos enfocamos en herramientas asistivas para oyentes en español, la oferta de software tiende a ser aún menor [J]. Por tal motivo, desarrollaron la aplicación "TalkLouder!" con interfaz gráfica para Windows Phone y Android. Afirman que puede ayudar al hipoacúsico en la comprensión del lenguaje, con el objetivo de favorecer su inclusión en la sociedad.
- En [K] se presenta un SRA VGG16, arquitectura de modelos de visión de red neuronal de convolución. Se observa que incluso con un enfoque simplificado se pueden obtener altos rendimientos. Se afirma que, con la adición de algoritmos de extracción de características, como los puntos de referencia faciales, se puede mejorar aún más el rendimiento del modelo. La sincronización de labios con la voz puede aumentar el costo del sistema.
- [L] desarrolla un SRA que convierte la pronunciación china en forma de boca. Favorece al sordo en la comprensión del lenguaje chino. También es una forma importante de y en la visualización de la pronunciación. Se desarrolló el software sobre tecnología multimedia y orientada a objetos. El diseño se basa en la solicitud de código interno de caracteres chinos, en el SDK de voz de Microsoft SAPI5 y en la tecnología RSA. Puede alcanzar una precisión relativamente alta solo después de que los usuarios realicen entrenamiento de pronunciación.

En cuanto a las preguntas que guiaron la investigación, las tecnologías que SRA que se usan en el proceso de aprendizaje de las personas sordas son: uno sobre subtituladores profesionales [A], Microsoft Translator Speech API, IBM Watson Speech to Text [B], ViaVoice y ViaScribe en [C,D y H], VOIS basado en una red neuronal convolucional [F], un artículo sobre Chatbots [G], otro sobre diversos enfoques matemáticos [I], TalkLouder [J], SRA VGG16 con arquitectura de modelos de visión también de red neuronal de convolucional [K] y finalmente SDK de voz de Microsoft SAPI5 [L]. En solo un artículo no se ha proporcionado información técnica de los SRA utilizados.

En cuanto a la segunda pregunta, evidentemente estas tecnologías contribuyen a mejorar la comprensión escrita de las personas sordas. La maduración de esta tecnología propició los hallazgos encontrados. Todos ellos consideran el lenguaje escrito como la vía para optimizar el aprendizaje del lenguaje. Además, en un trabajo [B] se tuvo en cuenta que el DHH también puede generar texto a partir de su voz, aunque con una tasa de error mayor. En otra aplicación [E] se tuvo en cuanta el sonido ambiental además del habla como estrategia de aprendizaje para el sordo. Otra vía alternativa al habla se consideró en [G] con chatbots, en donde la comunicación docente y estudiante es todo sobre lenguaje escrito, sin necesidad de conversión, pero con mucho potencial para el aprendizaje del sordo en el trayecto educativo. En otro artículo se ha combinado el habla con algoritmos de extracción de características faciales o la sincronización de los labios con la voz [K]

4 Conclusiones y trabajos futuros

En este artículo, se presentó una revisión de los RSA orientados al uso en la educación por personas con discapacidad auditiva. La revisión abarcó un conjunto de artículos científicos. Se describieron los métodos utilizados para la revisión y los aspectos tomados en cuenta para el análisis. Los resultados de la búsqueda se presentaron y discutieron sistemáticamente. Después de la discusión, se identificaron algunas características generales en los sistemas revisados, que se presentan a continuación.

Se corrobora que los SRA ayudan a la comprensión lingüística escrita, mejora la calidad de la enseñanza, tiene grandes beneficios para la revisión después de clase, aumenta el interés de los estudiantes en el aprendizaje y colabora con los distintos actores del trayecto educativo. La conversión de audio a texto que se logra con el SRA es el lenguaje escrito y este la base fundamental para la implementación de técnicas que favorecen la comprensión lingüística escrita.

Se evidenciaron tres enfoques distintos de subtítulos: subtítulos profesionales, subtítulos colectivos y los SRA.

Las personas DHH no podrían lograr las mismas tasas de errores que la población que no es DHH en SRA.

La evolución de los SRA ha permitido que el texto escrito tenga más presencia en el aula y en los distintos dispositivos de la asignatura.

Existen SRA que mejoran la eficiencia y precisión mediante detección de pausas en el habla, inserción de oraciones, saltos de párrafo y en algunas aplicaciones se adaptan a varios hablantes, logrando la individualidad del usuario. Esto, ayuda a superar la terminología especifica de un dominio en la educación superior.

Los chatbots son herramientas interactivas efectivas que operan con cierta autonomía para optimizar la comprensión lingüística escrita del sordo.

Aplicaciones que propician el aprendizaje del sonido ambiental es otra alternativa para la comunidad sorda.

El habla puede ser integrado a otros recursos como extracción de características faciales o sincronización de los labios.

El presente artículo abre las puertas para que los investigadores trabajen con el lenguaje escrito como medio para que la comunidad sorda pueda optimizar su comprensión lingüística.

Apéndice: artículos incluidos en la RSL

A Kushalnagar, Laseckit & Bigham A readability evaluation of real-time crowd captions in the classroom. 2012.

B Glaser. Automatic Speech Recognition_Services Deaf and Hard-of-Hearing Usability. 2019.

C Jun & Cheng The Exploration of the Strategies and Skills of Effective Use of Voice Recognition Software in the Classroom for Deaf Students. 2010.

D Fen & Cheng. Using Speech Recognition Technology To Support Education For Deaf Students. 2010.

E Kosuke, Seiichi & Yuhki. Voice Recognition and Information Transmission-System for Hearing Impaired People. 2017.

F Aye, Nway & Sheinn. VOIS: The First Speech Therapy App Specifically Designed for Myanmar Hearing-Impaired Children. 2020.

G Batista, Alejandro. Uso de chatbots como apoyo para la comunicación en el Aula. Un asistente virtual 24x7x365 colaborando con el curso. Facultad de Ciencias Jurídicas y Sociales. Universidad Nacional de La Plata. 2016.
H Kheir & Way. Improving speech recognition to assist real time classroom note taking. 2015.
I Kuldeep, Anurag, Gangadhar, Vijay, Ravindra & Sanjiv. Speech Recognition Classification with ANN Implementation Using Machine Learning Algorithm. 2021.
J Alvarez y Rufrancos Talk-Louder! 2016.
K Shashidhar, Patilkulkarni & Nishanth. Visual Speech Recognition using Convolutional Net Neuronal. 2021.
L Le, Pan & Ding The design and development of a software to convert Chinese pronunciation into mouth. 2011.

Referencias

1. UNESCO. Directrices para la inclusión: garantizar el acceso a la educación para todos. París. 2006.
2. Naciones Unidas. Convención sobre los derechos de las personas con discapacidad y su Protocolo facultativo. Res 61/106. Nueva York. 2006.
3. Herrera, H; Manresa Yee, C y Sanz, C. Aprendizaje móvil para niños con discapacidad auditiva: revisión y análisis. 2021
4. FIAPAS. Manual Básico de Formación Especializada sobre Discapacidad Auditiva (4ª ed.). Madrid. 2010
5. Fen & Cheng. Using Speech Recognition Technology To Support Education For Deaf Students. 2010.
6. Chomsky, Noam. La arquitectura del lenguaje. Barcelona, Ed Kairós. 2003.
7. Fernández Botero, Eliana. Logogenia: desde la gramática generativa, una nueva opción para los sordos: estudio de caso. 2004.
8. Radelli, Bruna. "Una aplicación de la lingüística: la logogenia", en Dimensión Antropológica, vol. 23, septiembre-diciembre, pp. 51-72. Disponible en: http://www.dimensionantropologica.inah.gob.mx/?p=652. 2001.
9. Sarmiento, O y Valdeblanquez, D. "AIUTA: Software de apoyo a las terapias de logogenia en niños sordos de 8 a 12 años", Tesis de Pregrado, Dpto. Ingeniería de Sistemas, Pontificia Universidad Javeriana, Colombia. 2010.
10. Moreno, A. La Lengua española y las nuevas tecnologías. Inteligencia Artificial y lengua española. Congreso de la Lengua Española, Sevilla. 1992.
11. Casacuberta Nolla, F. La Lengua española y las nuevas tecnologías. Análisis y síntesis de la señal acústica. Congreso de la Lengua Española, Sevilla. 1992.
12. Petticrew, M., & Roberts, H. Systematic reviews in the Social Sciences: A practical guide. Oxford, UK: Blackwell Publishing. 2006
13. Lavallée, M., Robillard, P. N., & Mirsalari, R. Performing systematic literature reviews with novices: An iterative approach. IEEE Transactions on Education, 57(3), 175–181. 2014.
14. Dieser Paula. Estrategias de autorregulación del aprendizaje y rendimiento académico en escenarios educativos mediados por TIC. 2019

Análisis de indicadores de presencias en cursos mediados por tecnología digital en tecnicaturas de Educación Superior

Omar Spandre[1], Paula Dieser[1], Cecilia Sanz[1,2,3]

[1] *Maestría en Tecnología Informática Aplicada en Educación. Facultad de Informática, UNLP*
[2]*Instituto de Investigación en Informática LIDI – CIC. Facultad de Informática, UNLP*
[3]*Comisión de Investigaciones Científicas de la Provincia de Buenos Aires*
spandreomar@gmail.com, pauladieser@gmail.com, csanz@lidi.info.unlp.edu.ar

Resumen. El modelo de Comunidad de Indagación viene siendo utilizado como marco teórico para analizar el aprendizaje en línea reconociendo la importancia de contar con tres tipos de presencias: cognitiva, docente y social. En particular, el empleo de este modelo ganó visibilidad a partir de la interrupción de las clases presenciales, a causa de la pandemia provocada por COVID-19. Como consecuencia de ello, las Instituciones de educación superior adaptaron sus cursos a un formato virtual para atender a los estudiantes durante la contingencia. Este trabajo aborda un estudio de caso en el que se analizan indicadores de las tres presencias en cursos de tecnicaturas de Educación Superior durante los años 2020 y 2021 en el Instituto Superior de Formación Docente y Técnica N° 93 (ISFDyT 93) de la ciudad de San Vicente. Con este fin se implementó una encuesta, respondida por 119 estudiantes. Los principales resultados dan cuenta de una alta valoración de las Presencias Docente y Cognitiva y, en menor medida, de la Presencia Social, y de los indicadores que hacen referencia al diálogo entre los estudiantes.

Keywords: Comunidad de Indagación, educación superior, educación mediada por tecnología digital, educación remota de emergencia, tipos de presencias.

1 Introducción

Este trabajo analiza las interacciones entre los miembros de una comunidad educativa, cuyos estudiantes de Educación Superior desarrollaron sus actividades pedagógicas entre los años 2020 y 2021 en el período denominado Aislamiento Social Preventivo y Obligatorio (ASPO). Este período especial motiva el análisis de los componentes del diálogo, mediados por entornos virtuales de enseñanza y de aprendizaje. Estos componentes, y sus relaciones, en un período de cursada virtual obligatoria, debido al aislamiento producido por la pandemia, están influenciados por el uso de las nuevas

tecnologías, y, por lo tanto, por la brecha entre los que disponen de recursos para acceder a ellas y los que han tenido que apropiarse en forma rápida e improvisada de los conocimientos necesarios para aprovecharlas. En este sentido, un gran número de profesores y estudiantes desarrollaron sus actividades en aulas virtuales interactuando entre sí, mediados por diferentes entornos, produciendo conocimiento, utilizando diferentes estrategias didácticas, adaptando sus planificaciones al nuevo contexto y desarrollando diálogos en modos sincrónicos y asincrónicos que pueden estudiarse desde varias dimensiones. Según [1], estas dimensiones o presencias y la interrelación entre ellas son necesarias para que el aprendizaje en línea sea posible. En este artículo se analizan las relaciones entre las presencias y cuál es la percepción de los estudiantes, a partir de las dimensiones e indicadores del Modelo de CoI [1]. Cabe aclarar que más allá de la situación de pandemia, esta temática resulta de interés y actualidad, ya que permite echar luz sobre procesos educativos mediados por tecnologías digitales, por lo que los resultados aquí encontrados serán un aporte para el diseño de futuros cursos en modalidades híbridas. De aquí en más este artículo se organiza de la siguiente manera: en la sección 2 se incluye el marco teórico y la encuesta a la luz del Modelo de CoI. En la sección 3 se sintetizan algunos antecedentes, en la sección 4 se detallan los aspectos metodológicos utilizados para el análisis de presencias del Modelo de CoI que se aplicarán al estudio de caso que aquí se aborda, en la sección 5 se presenta un análisis y discusión de los resultados alcanzados. Finalmente, en la sección 6 se enuncian conclusiones y trabajos futuros.

2 Marco teórico

2.1 El Modelo de Comunidad de Indagación

El modelo de CoI (*Community of Inquiry*) es un marco teórico desarrollado por Garrison y Anderson [1]. Según este modelo, para que el aprendizaje en línea sea posible, es necesaria la interrelación de tres dimensiones o presencias: cognitiva, docente y social [2]. A continuación, se describen estos tres tipos de presencia:

Presencia Cognitiva. Indica hasta qué punto los estudiantes son capaces de construir significado a través de la reflexión continua en una comunidad de investigación crítica [1,3], a través de una comunicación sostenida [4]. El modelo propuesto identifica cuatro fases no secuenciales en la Presencia Cognitiva: activación, exploración, integración y resolución [1,3].

Presencia Social. Es la capacidad de los participantes para proyectarse social y emocionalmente como personas, para promover la comunicación directa entre individuos y para hacer la representación personal explícita. La Presencia Social marca una diferencia cualitativa entre una comunidad de investigación/acción colaborativa y el proceso de meramente descargar información [3].

Presencia Docente. Se define en el modelo CoI como el acto de diseñar, facilitar y orientar los procesos de enseñanza y aprendizaje para obtener los resultados previstos de acuerdo con las necesidades y capacidades de los estudiantes [5].

2.2 La encuesta de la Comunidad de Indagación

El marco de CoI se ha adoptado en todo el mundo como una guía para la investigación y la práctica en el aprendizaje en línea, sin embargo, en 2006, dos cuestiones estaban desafiando la investigación de CoI en particular. La primera fue la falta de medidas comunes en los estudios que investigaban las presencias individuales, lo que dificultaba las generalizaciones entre los estudios. La segunda cuestión fue que pocos estudios antes de esa fecha exploraron las tres presencias y, lo que es más importante, las interacciones entre ellas, cuando éstas son fundamentales para el modelo en sí [6].
En 2006, varios investigadores de referencia en relación al Modelo de CoI comenzaron a trabajar en la creación de un instrumento de encuesta para medir las tres presencias y coincidieron que una encuesta de autoinforme es totalmente apropiada para medir las presencias, ya que se basan en las percepciones de los estudiantes evaluados, mediante una escala tipo Likert de 5 puntos (1: muy en desacuerdo; 5: totalmente de acuerdo). Se conciliaron los puntos en común entre los elementos existentes, para capturar cada una de las presencias [6] y lo resultante es un instrumento de 34 ítems [7], que aquí se presenta más adelante a través de un enlace, ya que es el que se utilizará para el estudio de caso de este trabajo.

3 Análisis de antecedentes

En un trabajo previo realizado por los autores [8], se analizan las estrategias utilizadas durante el proceso de cuarentena en varios países de la región a la luz del marco teórico del Modelo de CoI, mediante una revisión sistemática (RS) de 24 artículos publicados entre los años 2020 y 2021. Estos trabajos abordan el problema en la Educación Superior, en comunidades que cursaron durante esos años en la modalidad virtual, a causa de la suspensión de las actividades presenciales.
En el artículo precedente, se analizan las implicancias de reorganizar los procesos de enseñanza y aprendizaje, y los resultados muestran que un alto porcentaje de las investigaciones ponen el foco en la presencia docente, particularmente en el diseño educativo y de organización, con mucho menor énfasis se estudia la presencia cognitiva. Al mismo tiempo, hay escasa producción sobre la dimensión Presencia Social y el análisis de indicadores que refuercen el aprendizaje [8].
Los resultados ponen de manifiesto entonces la escasa producción sobre la dimensión Presencia Social, y el análisis que haga foco en el afecto, la cohesión del grupo y la comunicación abierta, los cuales son indicadores que refuerzan el aprendizaje y mantienen una dinámica de relaciones sociales positiva. Este aspecto solamente es tratado en algunos artículos desde la perspectiva de cohesión entre los estudiantes, en particular en un trabajo se aborda con estudiantes de primer año en las instituciones de educación superior, y como una variable de deserción, tomando en cuenta la alta escolarización de los estudiantes que han transitado el nivel secundario [8].
Los aspectos antes mencionados motivaron este nuevo trabajo, en el que se realiza una encuesta, basada en la propuesta de los investigadores del modelo CoI. La encuesta es aplicada en un estudio de caso con estudiantes de una comunidad educativa de Educación Superior, que cursaron en modalidad virtual en los años 2020 y 2021. Así el

objetivo de este trabajo es analizar las relaciones entre las presencias del modelo de CoI, utilizando el instrumento de encuesta indicado en la subsección 2.2, a la luz de dicho marco teórico.

En este sentido, para este análisis de antecedentes se seleccionan artículos de investigación publicados, cuyos títulos advierten el estudio específico de la Presencia Social en el marco del modelo Comunidad de Indagación.

Como primera observación, es posible afirmar que la mayoría de los autores referenciados acuerdan en definir la estrecha relación que existe entre la Presencia Social, la Presencia Cognitiva y la Presencia Docente en el modelo Comunidad de Indagación [9-13].

Al respecto, en [10] postulan una serie de afirmaciones que vale la pena destacar: la articulación entre las tres presencias se evidencia en el rol mediador que la Presencia Social adquiere en relación a las otras dos presencias, ya que la Presencia Social es condición para alcanzar el pensamiento crítico (Presencia Cognitiva) y es, a su vez, responsabilidad del docente a partir de las propuestas de actividades que realice.

En la práctica educativa en línea, la Presencia Social se pone de manifiesto mediante la interacción y la colaboración que se originan, desarrollan y potencian por la comunicación afectiva, comunicación abierta y cohesión de los miembros que integran la comunidad [9].

Los hallazgos de algunos de estos estudios [11-12] parecen indicar que la Presencia Social es relevante como predictora de procesos de aprendizaje exitosos, dada su estrecha relación con la Presencia Cognitiva en determinadas prácticas en entornos virtuales, como ser los foros, donde se forman con mayor facilidad ambientes amistosos y relajados que contribuyen a la cohesión de grupo y al trabajo colaborativo [9].

4 Aspectos metodológicos para el estudio de caso sobre análisis de presencias del modelo de Comunidad de Indagación

Los resultados del artículo anterior [8] y los antecedentes mencionados motivaron este trabajo y la necesidad de indagar en el estudio de caso aquí presentado, la valoración de los estudiantes en relación a las presencias del Modelo de CoI en los procesos educativos mediados por tecnologías digitales en los que han participado. En particular, se plantean nuevas preguntas de investigación:

P1. ¿Cuáles son las percepciones de los estudiantes participantes de la encuesta en relación a las presencias del modelo de CoI?

P2: ¿Existen diferencias entre las presencias según dichas percepciones?

P3: ¿Se encuentran diferencias en dichas presencias según carrera y el rango etario de los estudiantes?

Para encontrar respuestas a estas preguntas de investigación se utiliza la encuesta del modelo de CoI, el cual es un instrumento de 34 ítems que se presenta en el siguiente enlace:

https://drive.google.com/file/d/1MZlcTBJpEnVEOUEail6gBYSbKtDSFxEB/view?usp=sharing

Para el estudio de caso, la comunidad está integrada por estudiantes de 1°, 2° y 3° año de Educación Superior que cursaron tecnicaturas superiores durante los años 2020 y 2021 en el período de ASPO en modalidad virtual.

Se realizó la encuesta presentada en la subsección 2.2 entre estudiantes del Instituto Superior de Formación Docente y Técnica N° 93 (ISFDyT 93) de la ciudad de San Vicente. Las tecnicaturas superiores involucradas son: Administración Contable, Administración contable con orientación en marketing, Administración Pública, Análisis desarrollo y programación de aplicaciones, Enfermería y Guía de turismo. Además, se incluyeron variables para su posterior análisis, como el género y la edad de los participantes.

La metodología utilizada para que los estudiantes completaran la encuesta fue por medio de un formulario de Google y difundido, en primera instancia, a través de los estudiantes referentes de cada carrera y profesores que tenían conformados grupos de WhatsApp en las materias seleccionadas, mediante este camino se recibieron 50 respuestas. Luego, se reforzó la difusión mediante el correo electrónico institucional a todos los estudiantes de las carreras mencionadas, recibiendo 70 respuestas más. Se aplicó un filtro para eliminar duplicados, quedando un total de 119 respuestas, las cuales se analizan en este trabajo.

Analizando la muestra (ver Fig.1), contestaron 5 estudiantes de la Tecnicatura en Administración Contable, 11 de la Tecnicatura en Administración Contable con orientación en Marketing, 7 de Administración Pública, 26 de Análisis, Desarrollo y Programación de Aplicaciones, 60 estudiantes de la carrera de Enfermería y 10 de la Tecnicatura Superior en Guía de Turismo.

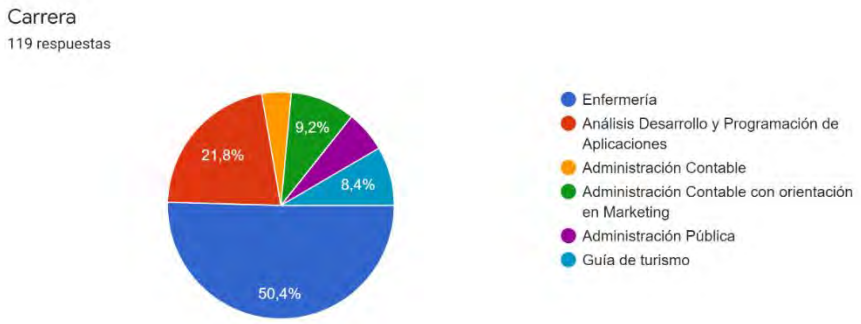


Fig. 1 - Cantidad de estudiantes por carrera del total de 119 que integran la muestra

En relación al perfil de los estudiantes que contestaron la encuesta, la mayoría pertenecen al género femenino (ver Fig. 2) y se concentra en el rango de edades que va desde los 26 a los 41 años.

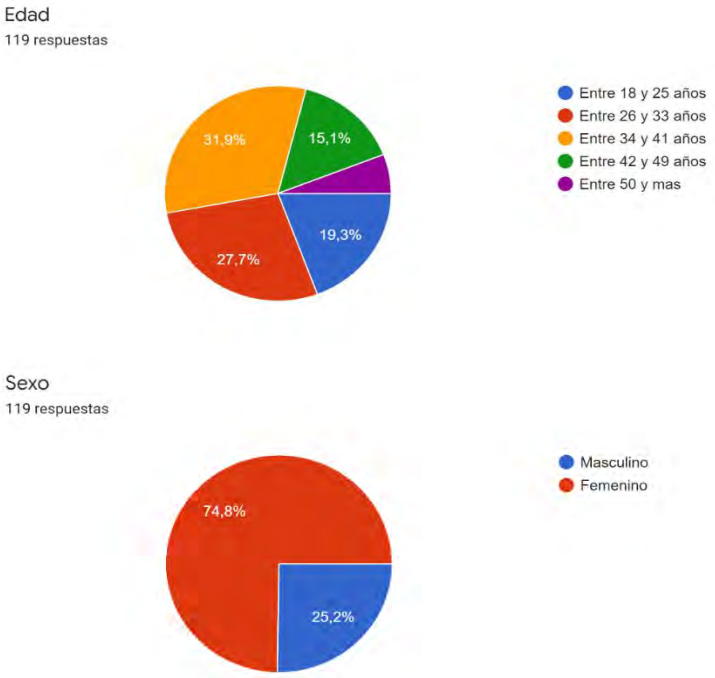


Fig. 2 - Superior: muestra la cantidad de estudiantes según rango etario; Inferior: detalla los estudiantes participantes según género.

Con respecto a los medios utilizados para la comunicación entre el profesor y el estudiante, éste podía optar por varios medios a la vez, siendo las herramientas más elegidas el Entorno Virtual de Enseñanza y Aprendizaje (EVEA) del instituto, WhatsApp y Google Meet, según se presenta en la Fig.3.

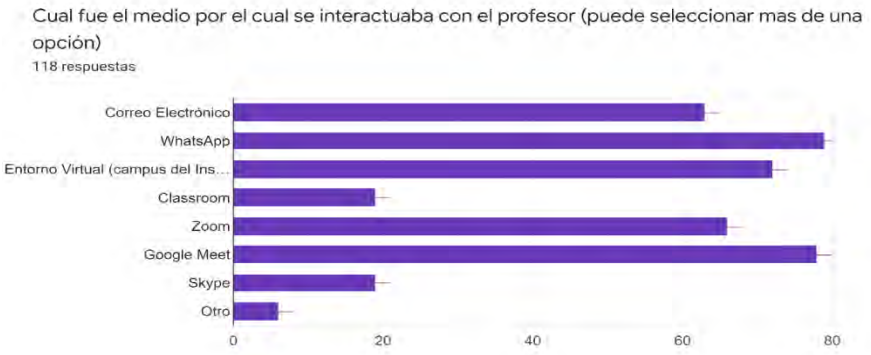


Fig. 3 - Herramientas digitales utilizadas para la comunicación de acuerdo a los respondido por los participantes

5 Resultados alcanzados

A continuación, se presentan los resultados en relación a cada pregunta de investigación planteada para este trabajo. Para procesar los resultados se utilizó una escala para evaluar cada respuesta, siendo 1 = Muy en desacuerdo; 2 = En desacuerdo; 3 = Neutral; 4 = De acuerdo; 5 = Totalmente de acuerdo, luego se calculó el promedio en cada una de las afirmaciones.
En relación a la pregunta de investigación P1. ¿Cuáles son las percepciones de los estudiantes participantes de la encuesta según las presencias del modelo de CoI? Se encuentran respuestas que confirman el desacuerdo con respecto a las afirmaciones que hacen referencia a la Presencia Social, ítems 14 a 22, ya que las respuestas están concentradas en el nivel 3.5 con énfasis en la neutralidad en las respuestas que responden a las afirmaciones 16, 17 y 18.
La percepción de los estudiantes en el caso de las afirmaciones de la encuesta que involucran la Presencia Cognitiva, ítems 23 a 34, muestra un nivel de acuerdo mayor, ya que están concentradas sobre el nivel 4 (De acuerdo).
La Presencia Docente, que incluye los ítems 1 a 13, aparece con el mayor nivel de satisfacción de las tres presencias, ya que están concentradas en el valor 4.5 de la escala.
Las valoraciones medias de los estudiantes participantes de la encuesta, en cada uno de sus 34 ítems, se presenta en la figura 4.

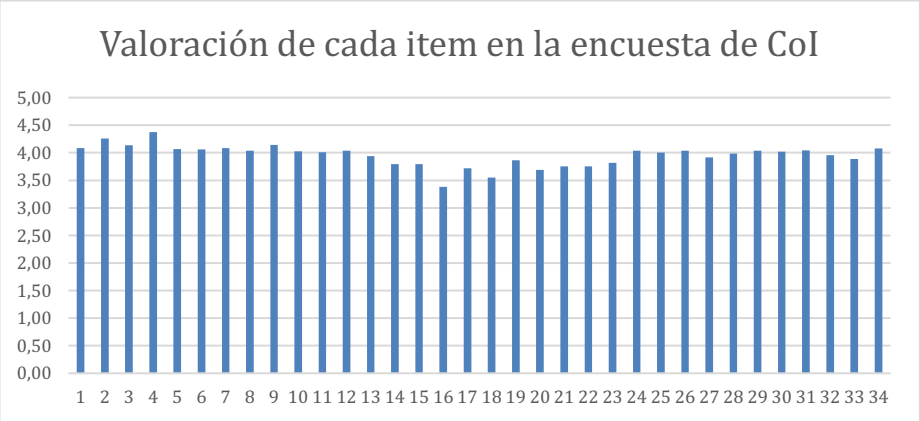


Fig. 4 – Valoraciones medias de cada uno de los ítems de acuerdo a lo respondido por los participantes mediante la encuesta de la CoI.

En relación a la pregunta 2, P2: ¿existen diferencias entre las presencias según dichas percepciones?
Se visibiliza, según los resultados descritos anteriormente, la mayor valoración con respecto a las declaraciones que refieren a la Presencia Docente y principalmente las que hacen foco en el Diseño y Organización de los cursos (1-4). Menor es el nivel de acuerdo respecto de las afirmaciones que involucran a la Presencia Cognitiva, y menor aún, las que hacen referencia a la Presencia Social.
Así se puede afirmar que a nivel general tomando en cuenta todos los participantes de la encuesta y de acuerdo con el diseño del instrumento, los ítems 1-13 (Presencia Docente) se valoraron más fuertemente en el nivel 4.5. Los ítems 14-22 (Presencia Social) más fuertemente en el nivel 3.5. Finalmente, los ítems 23-34 (Presencia Cognitiva) se valoraron con mayor intensidad en el nivel 4.
Con respecto a la pregunta 3, P3: ¿se encuentran diferencias en dichas presencias según carrera y el rango etario de los estudiantes? Para abordar esta pregunta se efectúa un filtro, analizando en forma separada los estudiantes de la carrera de enfermería, ya que dichos estudiantes conforman el 50% de la muestra. Se visibiliza una merma en el nivel de acuerdo con respecto a las respuestas que involucra a la Presencia Docente, estando más concentradas hacia el nivel 4. El resto de las respuestas conservan los mismos valores que la muestra principal con todos los estudiantes.
Con respecto a las edades de los estudiantes, se hace un corte entre la franja etaria que va de los 18 a 33 años y el resto de los estudiantes mayores a 33 años, los resultados arrojan menor valoración aun de los más jóvenes que el resto de la muestra, valores que reflejan descontento entre la población de menor edad, que se intensifican en la Presencia Social y los ítems 16, 17 y 18, que hacen referencia directamente a la comunicación mediada por la WEB, y la participación de los estudiantes en las conversaciones grupales.

5.1 Discusión

En línea con los trabajos citados a lo largo de este artículo, las tres presencias del modelo de CoI tienen estrecha relación entre sí [2, 6, 8-13], pudiéndose observar en este trabajo, la aceptación en general de las herramientas y estrategias didácticas implementadas por los profesores durante la pandemia.
Existe, según los resultados de este trabajo, una conformidad de los estudiantes por las decisiones tomadas por los profesores en lo que respecta al armado de los cursos, la facilitación y la instrucción directa, aspectos que demuestran la preocupación por resolver el problema de la virtualidad en un contexto de emergencia sanitaria, y la implementación de nuevos canales de comunicación sincrónicos y asincrónicos para las propuestas en línea.
Sin embargo, se observa que los estudiantes perciben dificultades para la participación en discusiones, en la interacción con otros compañeros y en el uso de herramientas digitales o basadas en la WEB. En este sentido, se coincide con resultados del trabajo anterior [8], en el cual, se manifiesta la escasa producción sobre la dimensión Presencia Social, y que haga foco en el afecto, la cohesión del grupo y la comunicación abierta [8]. Se considera, entonces, que la Presencia Social fue la menos valorada por los estudiantes en su percepción de las experiencias vividas durante el período de ASPO.

Como consecuencia podría pensarse que los estudiantes no han podido tener una interacción con sus pares y docentes totalmente positiva o que ésta ha resultado escasa, con pocos momentos de diálogo. Como hipótesis, es posible que los profesores hayan utilizado los formatos de clase presencial en la virtualidad, poniendo mayor énfasis en los materiales de estudio y en contar con herramientas, pero tal vez menor atención en promover la interacción. La falta de conectividad, el tiempo destinado a las clases virtuales, la escasa preparación en el manejo de las herramientas digitales, pueden ser las variables que intervinieron en los procesos de enseñanza y de aprendizaje para que la dimensión de Presencia Social sea menos valorada en la encuesta que se aplica en este trabajo. Por esto se refuerza la necesidad de implementar estrategias que fortalezcan los indicadores que mantienen una dinámica de relaciones sociales positiva, tal como se manifiesta en los trabajos citados previamente [11, 12].

6 Conclusiones y Trabajos Futuros

En este trabajo se presentó un estudio de caso en el que se analizan los tres tipos de presencia: social, cognitiva y docente, que se proponen como parte del modelo de Comunidad de Indagación. Estas presencias se consideran fundamentales para el desarrollo de propuestas educativas mediadas por tecnologías digitales que resulten efectivas. A partir de los resultados alcanzados se puede observar que se requiere fortalecer, en el diseño de los cursos mediados, el componente de diálogo vinculado con la Presencia Social. Al mismo tiempo, se perciben positivamente los esfuerzos de los docentes por diseñar y abordar propuestas educativas durante el período de ASPO. Algunos resultados también permitirán echar luz sobre el trabajo en carreras específicas de la muestra que se ha tomado, ya que se han encontrado algunas diferencias en este sentido, y también en el rango etario.

Las conclusiones de este artículo, serán consideradas en el proceso de avance de una tesis desarrollada en el marco de la Maestría en Tecnología Informática Aplicada en Educación, de la Facultad de Informática de la UNLP. Además, se abre el camino para profundizar en el diseño de propuestas educativas que tengan en cuenta las tres dimensiones, dada la incidencia que éstas tienen en el proceso educativo.

Referencias

1. Garrison, D. R., Anderson, T.: E–learning in the 21st century: A framework for research and practice. London: Routledge Falmer (2003)
2. Garrison, R., Cleveland-Innes, M., Fung, T. S.: Exploring causal relationships among teaching, cognitive and social presence: Student perceptions of the community of inquiry framework", The Internet and Higher Education, vol. 13, núm. 1-2, 31--36 (2010). DOI: 10.1016/j.iheduc.2009.10.002
3. Garrison, R., Anderson, T., Archer, W.: Critical inquiry in a textbased environment: computer conferencing in higher education, Internet and Higher Education, vol. 11 , núm. 2, 1--14 (2000). DOI: 10.1016/S1096-7516(00)00016-6
4. Gunawardena, C. N., Lowe, C. E., Anderson, T.: Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. J. Ed. Comp. Res., vol. 17, núm. 4, 397--431 (1997)

5. Gutiérrez-Santiuste, E., Rodríguez-Sabiote, C., Gallego-Arrufat, M-J.: Cognitive presence through social and teaching presence in communities of inquiry: A correlational–predictive study. Australasian J. of Educational Technology, vol. 31, núm. 3, 349--362. (2015)
6. Arbaugh, B., Cleveland-Innes, M., Diaz, S., Garrison, R., Ice, P., Richardson, J., Shea, P., Swan, K.: Community of Inquiry Framework: Validation and instrument development. The International Review of Research in Open and Distributed Learning. vol. 9, núm. 2 (2008). DOI: https://doi.org/10.19173/irrodl.v9i2.573.
7. Richardson, J.C., Arbaugh, J.B., Cleveland-Innes, M., Ice, P., Swan, K.P., & Garrison, D.R.: Uso del marco de la comunidad de investigación para informar un diseño instruccional eficaz. En: Moller L., Huett J. (eds) La próxima generación de educación a distancia. Springer, Boston, MA (2012). DOI: https://doi.org/10.1007/978-1-4614-1785-9_7
8. Spandre, O., Dieser, P., Sanz, C.: Revisión Sistemática de Metodologías Educativas Implementadas Durante la Pandemia del COVID-19 en la Educación Superior en Iberoamérica. En Congreso Argentino de Informática, pp. 49--63. Springer, Cham (2022)
9. Ferreyra, E. G., Strieder, S., Valenti, N. B.: La presencia social en la educación en línea según el Modelo Comunidad de Indagación. Abordajes. Revista de Ciencias Sociales y Humanas, vol. 6, núm. 12, (2022).
10. Gutiérrez-Santiuste, E., Gallego-Arrufat, M. J.: Presencia social en un ambiente colaborativo virtual de aprendizaje: análisis de una comunidad orientada a la indagación. Revista mexicana de investigación educativa, vol. 22, núm 75, 1169--1186 (2017) Recuperado http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662017000401169&lng=es&tlng=es.
11. Gutiérrez-Santiuste, E., Rodríguez-Sabiote, C. Gallego-Arrufat, M. J.: Cognitive presence through social and teaching presence in communities of inquiry: A correlational-predictive study, Australasian J. Educational Technology, vol.31, núm.3, 349--362. (2015). Disponible en: http://ajet.org.au/index.php/AJET/issue/view/112.
12. Costley, J., Lange, C.: The Relationship between Social Presence and Critical Thinking: Results from Learner Discourse in an Asynchronous Learning Environment. J.l Information Technology Education: Research, vol. 15, 89--108 (2016). DOI:10.28945/3418.
13. Yang, J. C., Quadir, B., Chen, N., Miao, Q.: Effects of online presence on learning performance in a blog-based online course. The Internet and Higher Education, vol. 30, 11--20 (2016). DOI:10.1016/j.iheduc.2016.04.002

Alfabetización digital con impacto en la lectoescritura

Viviana Harari[1] e Ivana Harari[1]

[1] Facultad de Informática de la Universidad Nacional de La Plata, La Plata, Buenos Aires. Argentina.
vharari@info.unlp.edu.ar, iharari@info.unlp.edu.ar

Resumen. La Facultad de Informática de la Universidad Nacional de La Plata viene trabajando desde hace más de 15 años en la alfabetización digital de niños/as y adolescentes de sectores vulnerables de la ciudad de La Plata y alrededores. Frente a la detección de grandes dificultades en la lectoescritura por parte de escolares de grados avanzados del nivel primario, desde el año 2019 se comenzó a pensar la enseñanza de informática orientada a fortalecer y estimular este proceso educativo. En este artículo se presenta el desarrollo y aplicación de un software para el aprendizaje de lectoescritura, basado en la metodología de enseñanza Dale!, en donde se disponen de cuadernillos con actividades y juegos. Se describe sus características, detalles de la capacitación donde se complementa la tecnología digital con el recurso didáctico físico y el impacto en el estudiantado, donde se demuestra la importancia del uso de la tecnología en los procesos de enseñanza aprendizaje.

Keywords: DALE, Alfabetización, Brecha digital, Capacitación digital.

1 Introducción

Desde el año 2007 la Facultad de Informática [1] de la Universidad Nacional de La Plata [2] viene trabajando en forma continua e ininterrumpida en la alfabetización informática de niños/as y jóvenes de sectores vulnerables de la ciudad de La Plata y alrededores. Este trabajo actualmente se desarrolla en el marco del proyecto denominado "El barrio va a la universidad" [3].

La interacción y comunicación permanente que se mantiene con estos sectores tan necesitados, ha permitido visualizar varias problemáticas que atraviesan a todos los grupos con los que se trabaja. La dificultad en la lectoescritura que presentan muchos niños/as que toman la capacitación informática y que cursan grados avanzados de la escuela primaria, es una problemática que se viene detectando desde hace mucho tiempo con todos los grupos con los que se trabaja. La pandemia del COVID'19 desatada en el año 2020 y la cuarentena estricta que provocó la no escolarización presencial, ha hecho que esta problemática se profundice aún más.

La preocupación permanente del equipo de trabajo sobre esta temática ha llevado a que indague sobre diversas maneras o formas de colaboración, que permitan ayudar a revertir esta situación. A través de esta búsqueda se llega a una propuesta denominada Dale! (Derecho a Aprender a Leer y Escribir) dirigido por la Dra. Beatriz Diuk [4].

Dicha propuesta presenta una metodología de enseñanza, manuales para implementarlo y, actividades que se ajustan a las condiciones de trabajo en que se desarrolla el proyecto de alfabetización informática. A raíz del descubrimiento del mismo, se comenzó a implementar, en forma de juegos digitales, las tareas lúdicas, con fichas de papel, que presenta parte de la propuesta Dale!.

Si bien en el año 2019 se comenzó a desarrollar un software con algunos de los juegos, a partir del segundo semestre del año 2021 y, tomando como base lo ya implementado, se comenzó con el desarrollo de todos los juegos del nivel 1de la propuesta, para poder utilizarlos en una capacitación informática específica, con niños/as que presenten serias dificultades en la lectoescritura.

Si bien actualmente se continúa con el desarrollo de la aplicación, en el primer semestre del año, se pudo comenzar a dar la capacitación informática a dos grupos de niños/as con dificultades serias de lectoescritura, de dos asociaciones civiles.

A través de este artículo se contará detalles de la capacitación y del software.

2 El barrio va a la Universidad

El proyecto "El barrio va a la universidad", que forma parte del Programa de Educación para la Inclusión de la UNLP, trabaja con sectores vulnerables de la cuidad de La Plata y alrededores, brindando capacitación informática a niños/as y jóvenes que concurren a diferentes asociaciones civiles sin fines de lucro como lo son los comedores barriales, fundaciones, entre otras. Nació con el solo objetivo de acortar la brecha digital en estos sectores pero luego se amplió, incorporando el objetivo de acercar a estos grupos a la universidad para que, desde edades tempranas, puedan incorporar en sus imaginarios la idea de poder estudiar, en el futuro, una carrera universitaria o, capacitarse en algunos de los oficios que ofrece la UNLP.

A lo largo de los años se trabajó en función de ambos objetivos, realizando actividades relacionadas con cada línea de trabajo. En lo que respecta a acortar la brecha digital, se realizan desde actividades básicas, que permiten que los/as niños/as y jóvenes puedan aprender a utilizar aplicaciones de uso cotidiano hasta, actividades más complejas, donde se aprende a programar haciendo uso de herramientas específicas [5]. En lo que respecta a acercar a la universidad se han realizado múltiples actividades, en conjunto con diferentes unidades académicas de la UNLP, destinadas a mostrar lo que se puede estudiar [6].

La problemática relacionada con la deficiencia en la lectoescritura, que presentan muchos/as de los/as niño/as que toman los cursos de capacitación, atraviesa a todas las asociaciones con las que se trabaja, por eso siempre se pensó en sumar una capacitación específica destinada a estos grupos.

3 Dale!

Dale! de la Dra Beatriz Diuk, es una propuesta pensada para ser aplicada en niños/as de sectores sociales de suma vulnerabilidad, escolarizados y con problemas de lectoescritura. Es una propuesta sistemática y se organiza en tres niveles. El nivel 1 es

para aquellos/as niños/as que no pueden escribir palabras sencillas con sílabas con estructuras CV (consonante vocal), el nivel 2 para aquellos/as que presenten problemas para escribir palabras con estructuras como CVC (consonante vocal consonante) y, el nivel 3 para aquellos/as que solo omiten letras cuando escriben palabras con estructuras como CCV (consonante consonante vocal).

Está organizada por sesiones, 40 en cada nivel, que se llevan a cabo con los materiales correspondientes: cuadernillo del alumno, de juegos y de tareas. La propuesta plantea el dictado de dos sesiones por semana con una duración de 20 minutos cada una.

Para cada sesión se plantean 3 momentos: conversar y escribir, escritura y lectura. En los dos primeros niveles, en el momento de escritura, se plantean juegos con recursos tradicionales como ser fichas de papel. En todas las sesiones los/as niños/as hacen uso de los cuadernillos del alumno, juegos y de tareas.

Para el momento "conversar y escribir" se propone mantener una conversación con el/la niño/a relacionada con su cotidianeidad, generar una oración y, ayudar a que el niño/a la escriba. En el momento de "escritura", en los niveles 1 y 2 se lleva a cabo un juego, donde se plantean diferentes actividades lúdicas que permiten al niño/a trabajar los objetivos planteados en esa sesión. Una vez finalizado el juego se pasa al cuadernillo para escribir las palabras relacionadas con lo trabajado.

En el momento de "lectura" se lleva a cabo actividades donde el/la docente cuenta un cuento o, el/la niño/a lee una serie de palabras, oraciones o historietas, de acuerdo al nivel que se esté trabajando.

El tercer cuadernillo de "tareas" se diseñó con actividades que los/as niños/as deben realizar una vez terminada la sesión Dale! [7].

4 Capacitación informática con impacto en la lectoescritura.

La capacitación informática con impacto en la lectoescritura se está aplicando actualmente con niños/as de las asociaciones civiles sin fines de lucro Padre Cajade Niños [8] y, Nam Qom [9] que, luego de una pequeña evaluación, fueron clasificados como de Nivel1.

Para la capacitación se hace uso de recursos tanto digitales como no digitales. Respecto a estos últimos se utilizan aquellos propuestos por Dale!, como ser: lápiz, cuadernillo, goma, lapicera, fichas de papel con imágenes, entre otros. Respecto a lo digital se utilizan aplicaciones que sirven para editar textos (editores de textos) y aquellas que sirven para dibujar (graficadores).

La metodología que se utiliza en cada sesión es la que propone Dale! pero, se le agregan actividades digitales con algunos adicionales.

Se llevan a cabo los tres momentos de Dale!: conversar y escribir; escritura y lectura, de la manera que se detalla a continuación:

- Conversar y escribir: se mantiene una conversación sobre el entorno del niño/a y se genera en forma consensuada una oración. Se utiliza el cuadernillo del/de la niño/a para que escriba la oración, haciendo uso de recursos tradicionales y, luego se pasa a escribir la misma oración en la computadora haciendo uso de un editor de textos.

De acuerdo al nivel de conocimiento que presentan los/as niños/as se los/as capacita desde el uso y funcionalidad básica del teclado, hasta cuestiones más avanzadas del editor como: cambio de tamaño y color de las letras, justificación de párrafos, viñetas, entre otros.

- Escritura: Se hace uso del software desarrollado para realizar el juego y, la escritura de las palabras que propone Dale! para esa sesión. Luego se vuelve al cuadernillo y se dicta al niño/a las mismas palabras con las que trabajó con el software.
- Lectura: Se le lee al niño/a el cuento, en el caso que la sesión así lo proponga o, se le pide que el/ella lea las palabras propuestas para esa sesión, haciendo uso de su cuadernillo. Como algo adicional, en el caso de lectura de los cuentos, se le pide al niño/a que realice con el graficador, un dibujo relacionado con el cuento. El mismo luego debe ser recortado, copiado y, pegado en el documento donde se encuentra la oración y las palabras tipeadas por el/la niño/a.

1. Software.

Para el desarrollo del software que acompaña la capacitación informática con impacto en la lectoescritura se tomó como base el desarrollo comenzado en el año 2019 [10]. En esa oportunidad se desarrollaron algunos juegos puntuales pensados para ser utilizados en las clases de informática, no asociadas a ninguna capacitación específica y, para cualquier persona que dictase la capacitación Dale!.

Uno de los principales cambios introducidos en esta versión fue la posibilidad de que el/la capacitador/a pueda acceder directamente a la sesión deseada (fig. 1).

Fig. 1. Pantalla inicial y de elección de sesión.

El cliquear en un link de sesión se abre la ventana con el juego correspondiente, pasando luego a la actividad de escritura. Como ejemplo, en la figura siguiente se puede ver la composición de la sesión 1 (fig.2).

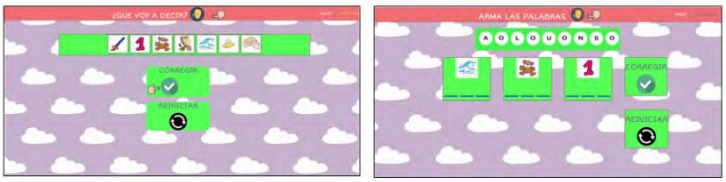


Fig. 2. Juego sesión 1 y escritura palabras de la misma sesión.

Si bien se continua con el desarrollo del software, el uso del mismo, por parte de los/as niños/as, ha permitido detectar errores, cambiar algunas interfaces, modificar algunos sonidos, agrupar ventanas, entre otras cosas.

5 Conclusiones.

La capacitación informática con impacto en la lectoescritura está demostrando una aceptación por parte de los niños/as muy alto. De por sí, la mayoría de ellos/as demuestran mucho interés en utilizar la computadora para escribir y para jugar. Las sesiones se pueden dar de la forma pensada, durando unos 40 minutos aproximadamente. Si bien duplica el tiempo propuesto por Dale! los/as niños/as, salvo el juego, repiten las actividades dos veces en formato tradicional y digital.

En lo que respecta al avance en lo digital, en la mayoría de los casos se ha tenido que enseñar cosas básicas del uso de la computadora, dado que los/as niños/as no sabían utilizar el recurso. A medida que fueron pasando las sesiones se fueron afianzando cada vez más, tanto en el uso del recurso como en el uso de las aplicaciones.

Respecto del avance en la lectoescritura, hubo muchos casos de niños/as a los cuales se los avanzó en las sesiones ya que comenzaron a mostrar facilidad al realizarlas. A dos de las niñas se las pasó a nivel 2.

Después de muchos años de identificar esta problemática e intentar buscar, desde la capacitación informática, alguna manera de colaborar con la mejora de lectoescritura de muchos niños/as, por los resultados vistos hasta ahora, la capacitación informática con impacto en la lectoescritura, parece ser eficaz.

Referencias

1. FI, http:// www.info.unlp.edu.ar, último acceso 2022/08/15.
2. UNLP, http:// www.info.unlp.edu.ar, último acceso 2022/08/15
3. Javier Díaz, Ivana Harari y Viviana Harari “University project: The marginalized neighborhood goes to University”. Congreso Internacional ICEER - International Conference on Engineering Education and Research. ISBN 978-9954-9091-2-6. (2013).
4. Dale!, http://propuestadale.com/, último acceso 2022/08/15.
5. Ivana Harari y Viviana Harari. “Teaching Programming to Children in Vulnerable Situations”. Twelfth Latin American Conference on Learning Technologies LACLO. ISBN: 978-1-5386-2377-0. (2017).
6. Ivana Harari, Mariela Stavale y Viviana Harari. “Extensión, Docencia e Investigación destinada a los Comedores Barriales: un caso testigo de cómo se combinan los pilares de la Universidad en forma multidisciplinaria”. XXIII Congreso Argentino de Ciencias de la Comunicación CACIC 2017. ISBN: 978-950-34-1539-9. (2017).
7. Diuk, B. G. Propuesta DALE!: Derecho a Aprender a Leer y Escribir. (2013).
8. Padre Cajade Niños, https://is.gd/9mY57p, último acceso 2022/08/15..
9. Oliveros, Nila Vigil. "Diagnóstico sociolingüístico participativo del barrio Toba/Qom las Malvinas de La Plata". Revista Lengua y migración, Universidad de Alcalá. (2018).
10. Viviana Harari, Ivana Harari y Luján Rojas. “Optimizando estrategias de enseñanza a través de las TICs”. Jornadas Argentinas de Informática JAIIO. ISSN:2451-7496. (2019).

Análisis de Impacto e Implementación de la Retroalimentación en la Plataforma H.E.R.A., Herramienta de Desarrollo y Administración de Material Pedagógico Multimedial

Vanina Cecilia Chiavetta[1] [0000-0002-2222-3333]; Luis Mariano Mongelo[1] [0000-0002-9656-7233]; Marcela Fabiana Dávila[1] [0000-0002-6346-8327]

[1] Universidad Nacional de La Matanza, Departamento de Ciencias Económicas, Florencio Varela 1903, Buenos Aires, Argentina.

vchiavetta@gmail.com, luis.mongelo@gmail.com, lic_marceladavila@yahoo.com.ar

Abstract. La presente investigación analiza a modo de experiencia de prueba y reestructuración de prototipos a la Herramienta H.E.R.A., una aplicación ejecutable desde internet para administrar el desarrollo de material pedagógico-didáctico, que brinda la posibilidad de adaptar también el mismo a una población de alumnos diversos funcionales, con la supervisión de asesores pedagógico-didácticos y desarrolladores multimediales de contenidos educacionales, en línea. Esta segunda etapa, elabora nuevos mecanismos de ponderación y calificación de estos últimos al momento de la selección y la implementación de un sistema de semáforos inteligentes, para el seguimiento del proceso de retroalimentación en varios ciclos, de un mismo proyecto de creación de material, debidamente supervisado por los asesores y super-usuarios administrativos en representación de las instituciones universitarias a la que pertenecen los profesores solicitantes del material multimedial. Asimismo, se amplían los alcances y contenidos de los bancos de recursos contenidos en la herramienta primigenia, agregando además de texto, video, audio e imágenes, componentes comunicativos para diversos funcionales como text to voice o impresión Braille.

Keywords: Diversidad Funcional, Banco Multimedial, Administración de Materiales Didácticos, Asesores Pedagógicos-didácticos

1 Introducción

A partir de la creación del software H.E.R.A. (Herramienta Educativa de Recursos Áulicos) desarrollada como un prototipo en una investigación anterior (bajo el nombre y código de proyecto: Sistema de comunicación multimedial para el desarrollo de material pedagógico para estudiantes regulares y diversos funcionales en la educación superior – PIDC-55-B-224), destinada a generar un sistema de comunicación multimedial tendiente a gestionar requerimientos de material pedagógico para estudiantes regulares

y diversos funcionales en la educación superior; se decidió emprender este segundo proyecto para dar continuación a una línea de investigación implícita.

En la misma, se intenta llevar una solución informática a la administración de proyectos de diseño de material pedagógico para estudiantes universitarios regulares y diversos funcionales (alumnos con impedimento de disminución visual, ciegos, sordos y disminuidos motrices). A tal efecto, un usuario profesor, accede a la aplicación en Internet y se identifica para dar comienzo a un nuevo proyecto de creación de material pedagógico. El sistema recibe las indicaciones del profesor y le asigna, de un banco de desarrolladores multimediales asociados, a la o las personas que desarrollarán el proyecto.

Estos recibirán un formulario de requerimientos diseñado por la herramienta y generarán un proceso circular retroalimentado (como explicaremos más adelante) para ir dando forma mediante un sistema de semáforos virtuales y ponderaciones (diseñado por este grupo de trabajo) al material terminado. En medio del proceso, un asesor pedagógico-didáctico (también suministrado por un banco de profesionales administrado por el sistema) realizará las correcciones pedagógicas del material, habilitando o deshabilitando las señales del semáforo de control, hasta llegar a un punto final, donde un super-usuario del sistema (miembro de la dirección académica de la casa de altos estudios, con permisos especiales de administrador dentro de H.E.R.A.) lo aprueba y libera el proyecto para su implementación y puesta en funciones.

2 Desarrollo del proyecto

Desde el punto de vista del aspecto y usabilidad de la interface, fue el sustento teórico de Tona Monjo Palau, y en particular de su obra Diseño de Interfaces Multimedia (2011), el que nos guió a lo largo de las tres etapas de diseño, creación y re-diseño de nuestra herramienta. En ella, la autora presenta una metodología de diseño paso a paso, basada en el análisis de tareas que se pretende, debería realizar la aplicación a desarrollar. Este método evalúa cómo consiguen plasmar las personas sus objetivos mediante el software. Mediante la observación y entrevistas con los usuarios, un analista de sistemas determina el conjunto de objetivos de los usuarios previstos. A continuación, se definen las tareas que permiten conseguirlos, y se ordenan de acuerdo con la importancia del objetivo y la frecuencia de ejecución de la tarea. Las prioritarias se descomponen en pasos individuales y el nivel de descomposición puede variar, dependiendo del sistema evaluado. A continuación, el análisis sugiere cómo puede realizarse la tarea más eficientemente, o propone nuevas que puedan alcanzar más efectivamente los objetivos. En base a este desglose de tareas, se van generando los módulos correspondientes a las funciones de administración de la información que deberá llevar adelante el software que encierra la interfaz. Esto se ve en la Fig. 1.

Una vez diseñada, nuestra herramienta se dividió en 3 módulos bien diferenciados:

- El módulo Gestor de Requerimientos.
- El módulo Gestor de Soluciones.
- El Sistema de Control.

Fig. 1. Diagrama de metodología para el desarrollo de la herramienta.

Desde el Gestor de Requerimientos, el usuario docente presentará un conjunto de archivos de base necesarios para la construcción del material pedagógico del proyecto (documentos de texto con la teoría o los textos a plasmar, archivos de imágenes, diagramas e ilustraciones y de tratarse de un proyecto multimedia, archivos de video o de sonido). La herramienta H.E.R.A. le brindará entonces el acceso a un banco de contenidos o base de datos de recursos de la misma, donde se van acumulando los materiales del tipo creative commons (acceso libre) aportados por los proyectos anteriores o agregados en forma periódica por los superusuarios del sistema. Asimismo, el banco posee enlaces a bancos de contenidos freeware de internet, como Freepik o FreeImage. El usuario podrá entonces apropiarse de más material para su proyecto, y terminará generando un formulario digital, con los contenidos y objetivos del material pedagógico a desarrollar.

3 Metodología de retroalimentación del proyecto

El formulario generado, se deberá ofrecer a un desarrollador multimedial, incluido en el banco de multimediales almacenado en la misma herramienta. El sistema posee un proceso de ponderación que califica las aptitudes de los desarrolladores, buscando a los tres más adecuados por sus habilidades y especialidades, para cada proyecto en particular. Desde la pantalla de acceso al sistema, se puede ingresar para postularse como desarrollador de proyectos, mediante la recolección de un curriculum y el completado de una pantalla de indicación de habilidades como la de la Fig. 2. En ella, se derivan las marcaciones de los postulantes a un sistema de calificación por ponderación que obedece a la teoría expuesta por Alfonso DePirenne en su trabajo “Administración de la Educación Virtual” (2008). Allí se le asigna un puntaje a cada una de las marcas de especialización (Experto - Medio – Básico – Sin Conocimientos) y se lo multiplica por un factor de ponderación aplicado a la especialidad requerida por el proyecto en su génesis original (versión primaria o de borrador).

El Gestor de Soluciones se divide pues en dos componentes lógicos, el Sistema de Análisis Pedagógico y el Gestor de Diseñadores Multimediales. A los primeros se los requiere en la secretaría o departamento de pedagogía universitaria del claustro solicitante y se los asigna también por sus aptitudes evaluativas, mediante otro sistema de ponderación gobernado por los superusuarios administradores del sistema. Estos tres actores ingresarán entonces en un proceso circular de retroalimentación, gobernado por un sistema de semáforos de nuestra creación. Este panel de semáforo está visible en la

mayoría de las pantallas del sistema, y presenta cuatro estadíos bien diferenciados: Borrador – Pedagógico – Multimedial – Retroalimentación.

Cuando uno de ellos se encuentra en verde, esto significa que el proyecto de creación del material se encuentra en esa etapa de desarrollo, por ejemplo, semáforo en verde marcando el estadío pedagógico, indica que en ese momento el material se encuentra bajo análisis pedagógico didáctico, por el asesor correspondiente, que generará un nuevo formulario con las correcciones y reformas sugeridas para indicar al siguiente actor en el círculo de actividades del semáforo (por ejemplo, del asesor pedagógico-didáctico puede pasar al multimedial y de él, regresar reformado al profesor solicitante, para solicitar su aprobación a las reformas).

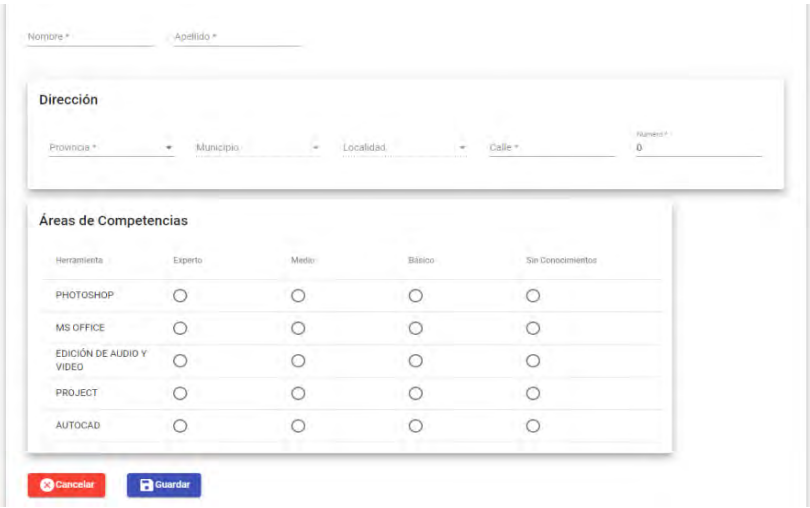


Fig. 2. Pantalla de carga e indicación de habilidades del Multimedial.

Finalmente, la herramienta permite una adaptación dinámica del material pedagógico o proyecto, para su utilización por una población de alumnos diversos funcionales. Al ingresar por la página de inicio, se le pide al usuario identificarse con un rol dentro del sistema (Profesor – Multimedial – Asesor Pedagógico-Didáctico – Superusuario), su clave de acceso, y la indicación del proyecto sobre el que se quiere trabajar en la presente sesión. Si el proyecto es nuevo, se deberá indicar si el mismo necesita adaptación para usuarios diversos funcionales, y se le ofrecerá al participante, diferentes facilidades de adaptación para su material. Allí contamos con indicaciones de accesibilidad para llevar el material a Braille, utilizando una interfaz de adaptación de código abierto facilitada por Lumi Industries, íntegramente desarrollada en lenguaje Delphi 7, que digitaliza el material final en archivo de texto, a una plancha de Braille fácilmente imprimible en impresora 3D. También permite traspaso de Text to Voice bajo el sistema FreeTTS disponible en Sourceforge bajo licencia GNU.

Finalmente, se le ofrece al usuario en línea una interfaz del sistema accesible, en consideración de que el mismo usuario (profesor o multimedial) sea una persona diverso

funcional. Allí accederá a un menú de adaptación de interfaz que contempla las opciones de Ayuda óptica – Interfaz audio-descriptiva – Interfaz táctil – Subtítulos.

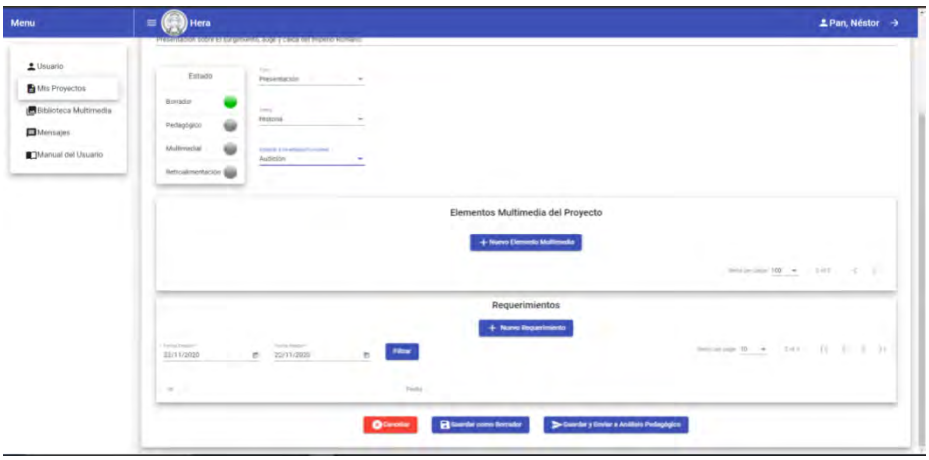


Fig. 3. Interfaz de administración de proyectos con sus semáforos.

4 Resultados

Nuestra herramienta busca dinamizar la comunicación entre los usuarios docentes y los desarrolladores multimediales, también ofreciendo una interfaz sencilla, de gran usabilidad y operación intuitiva, inspirada en otras interfaces comunes del área educativa y comercial de nuestra región, tales como la plataforma educativa Collaborate, de Blackboard Inc., las páginas Wiki de Pbworks Inc., o la interfaz del Image Bank de Gettyimages Latinoamérica. Para evaluar si se cumplen con las condiciones enunciadas, recurrimos a los métodos de evaluación de software propuestos por Eric Zabre e Islas, que permiten puntuar y analizar cuantitativamente y cualitativamente estos trabajos de implementación.

Bibliografía

1. Blackman, R.: Nuevos Desarrollos para el Nuevo Mundo Digital. Ediciones Orbe. Ciudad de México (2009)
2. DePirenne, A.: Administración de la Educación Virtual. Publicaciones Planeta Inteligente. Ciudad de México (2008)
3. Eric Zabre, B. e Islas, P.: Evaluación de herramientas de hardware y software para el desarrollo de aplicaciones. Wiley-Interscience. Barcelona (2011)
4. Monjo Palau, T.: Diseño de Interfaces Multimedia. Edicions Universitat Oberta de Catalunya. Cataluña (2011)
5. Suárez Turbón, I. y Sueiras Rodríguez, E.: Guía multimedia de recursos educativos para alumnado con necesidades educativas especiales. Centro del Profesorado y de Recursos de Gijón. Principado de Asturias (2017)

Diseño de una herramienta para la creación de actividades educativas basadas en Realidad Aumentada

Natali Salazar Mesia[1][0000-0003-3946-8383], Cecilia Sanz[1,2][0000-0002-9471-0008]

[1] Instituto de Investigación en Informática LIDI. Facultad de Informática – Universidad Nacional de La Plata
[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires
{nsalazar, csanz}@lidi.info.unlp.edu.ar

Abstract. En este trabajo se presenta el avance de una tesis de Maestría en Tecnología Informática Aplicada en Educación de la Facultad de Informática de la Universidad Nacional de La Plata. Se continua con el desarrollo de la herramienta de autor AuthorAR en el marco de la investigación que se lleva adelante en el Instituto de Investigación en Informática LIDI. Se detallan los avances en la investigación, y la propuesta de implementación de una plantilla de AuthorAR que permite crear actividades educativas basadas en Realidad Aumentada, con geolocalización como tipo de reconocimiento. Se detallan aspectos de la refactorización de la herramienta con nuevas funcionalidades de manejo de usuario para que cada uno acceda a sus proyectos. Se espera finalizar la aplicación móvil para realizar las evaluaciones con docentes y estudiantes.

Keywords: Realidad Aumentada, Herramienta de Autor, Actividades Educativas, Geolocalización

1 Motivación

La Realidad Aumentada (RA) es una tecnología que complementa la percepción e interacción con el mundo real y permite a la persona vivenciar un entorno real aumentado, con información digital generada por la computadora [1]. Además, posibilita el desarrollo de aplicaciones interactivas que combinan la realidad con información sintética, tal como imágenes 3D, sonidos, videos, textos, sensaciones táctiles, en tiempo real, y de acuerdo con el punto de vista de quien está observando la escena [2]. En este tipo de tecnología, la información virtual, tiene que estar vinculada especialmente al mundo real, es decir, un objeto virtual, siempre debe tener una ubicación relativa al objeto real. La visualización de la escena aumentada (mundo real + sintético) debe hacerse de manera coherente [3].

En particular, la RA es una tecnología que puede ayudar a mejorar el proceso de enseñanza y de aprendizaje. Su utilización, en algunos procesos educativos, puede aportar aspectos diferenciales respecto a los métodos tradicionales de enseñanza, entre los que se mencionan: realismo, interactividad, motivación e interés en aprender [4].

En los últimos años se observa un número creciente de experiencias educativas que involucran RA en el aprendizaje y se distinguen diferentes formas de reconocimiento de elementos de la escena. Por ejemplo, en [5] y [6] se aplica RA a la enseñanza y aprendizaje de temas de química para facilitar ciertos niveles de abstracción, utilizando aplicaciones ya disponibles en los diferentes *stores* como QuimicAR. En dicha experiencia se lleva un trabajo en el aula que propone un aprendizaje más activo por parte de los alumnos. En [7] se demuestra que el uso de RA mejora tanto el rendimiento espacial como el académico en pruebas realizadas con estudiantes de ingeniería. Los autores de [8] muestran el desarrollo de un material educativo para alumnos interesados

en aprender, conocer o reforzar los conocimientos referentes a las estructuras de control iterativas, recolectando objetos 3D aumentados de la escena que el alumno captura a partir de la elección de la estructura de control adecuada. En dicho trabajo se detalla una experiencia con 20 estudiantes de primer año de una carrera de Computación donde se prueba que la aplicación logra el objetivo de ser un apoyo a los estudiantes.

Estos ejemplos, en general, presentan experiencias en las que se han encontrado beneficios en el uso de la RA.

Las herramientas de autor de RA permiten generar contenidos de RA sin la necesidad de tener conocimientos avanzados de programación. Algunas de ellas se orientan a crear actividades educativas y están destinadas principalmente a docentes y estudiantes. Se puede afirmar que funcionan como puente para que los docentes puedan diseñar este tipo de propuestas con RA. Además, cuentan con facilidades para organizar actividades o interconectar diferentes componentes que permiten adecuar el contenido a los objetivos, los conocimientos y habilidades que se buscan desarrollar. Brindan la posibilidad de participar en decisiones de diseño, a partir de plantillas, lo que convierte a las herramientas de autor en instrumentos de uso cada vez más frecuente en el ámbito educativo [9].

En [10] se presenta una herramienta de autor con AR *nuggets*, aplicaciones de RA basadas en patrones. Cada AR *nugget* es un escenario que incluye objetos con parámetros y *targets* donde el contenido virtual se ancla en el mundo real. Por defecto, se incluye la posibilidad de crear tres escenarios con plantillas predefinidas, y se tiene la posibilidad de crear un propio escenario. Utiliza objetos 3D simples como cubos y cilindros para facilitar su creación. Se realizó una prueba con 48 voluntarios que tuvieron diferentes dificultades a la hora de configurar sus actividades. Se indican resultados positivos relacionados con la ayuda que brinda la herramienta para crear sus actividades. Si bien se nota un esfuerzo en la comunidad académico – educativa por ofrecer este tipo de herramientas a los docentes y estudiantes, se observa que pocas de estas iniciativas se encuentran disponibles para ser accedidas, o se trata de prototipos parciales, demos o productos comerciales.

Es por ello que en esta tesis de maestría se propone aportar al diseño de actividades educativas que incorporen la tecnología de RA con el objetivo de incrementar las posibilidades de los docentes de acercarse al diseño de este tipo de actividades, y ponerlas en juego en situaciones educativas concretas. En particular, se está diseñando una plantilla que servirá para enriquecer una herramienta de autor, cuyo desarrollo se encuentra en el marco de otra tesis de maestría [11].

2 Aporte del trabajo

Esta tesis retoma el trabajo presentado en [11], donde se describe la herramienta de autor AuthorAR y sus plantillas de actividades educativas. Esta herramienta permite crear dos tipos de actividades, de exploración y de estructuración de frases.

En el proceso de esta tesis se ha analizado el estado del arte en relación a herramientas de autor, y se han llevado a cabo diferentes publicaciones al respecto. Por ejemplo, en [12] se realiza un análisis de herramientas de autor con RA, y se describe el inicio del proceso para extender la herramienta AuthorAR con propuestas de nuevas plantillas de actividades educativas. En [13] se llevó a cabo un estudio de librerías de RA que aportan a los programadores al momento de crear aplicaciones con esta tecnología. Se hizo una comparativa, considerando criterios de análisis tales como documentación y versionado, tipos de reconocimiento, detección y seguimiento para RA, plataforma de ejecución: dispositivos móviles o para *desktop*, y tipos de licencias que posee: libres, comerciales o educativas.

Actualmente, se está realizando una refactorización de AuthorAR con la incorporación de una nueva plantilla de actividad educativa basada en geolocalización. Se desarrolla en Java 8 con una API rest y se mantiene la base de datos SQlite.

A su vez, se está desarrollando una aplicación móvil con Unity y Vuforia para que funcione como visor de las actividades educativas diseñadas en AuthorAR.

En la figura 1 se muestra la pantalla de inicio con la incorporación del manejo de usuarios, lo que es una nueva funcionalidad desarrollada en el marco de esta tesis para la herramienta AuthorAR. Cada usuario tiene acceso a todos sus proyectos en AuthorAR.

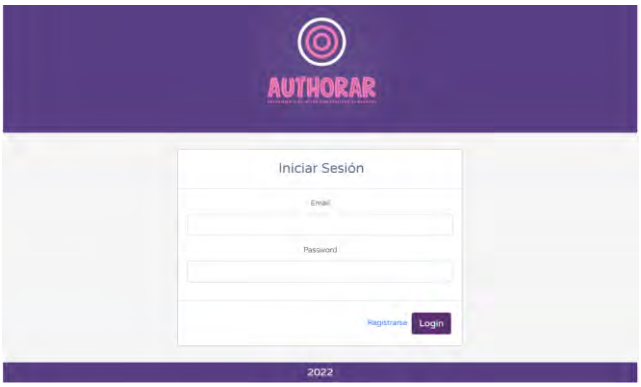


Fig. 1. AuthorAR inicio de sesión

Un proyecto cuenta con los campos de nombre, temática y etiquetas en su creación, y cada proyecto permite crear los diferentes tipos de actividades, a partir de plantillas que orientan su diseño.

En la figura 2 se muestra la página inicial con toda la información de los proyectos.

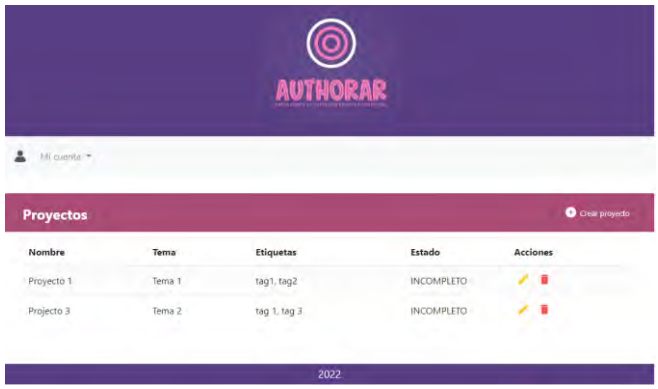


Fig. 2. Página de inicio de AuthorAR

De esta manera, el usuario que ingrese a AuthorAR va a tener acceso web para configurar sus proyectos.

Una vez que un proyecto tiene el estado completo, va a poder ser ejecutado desde la aplicación móvil. Cada usuario accede a la aplicación con un código único para realizar las actividades del proyecto.

Implementación de la plantilla de geolocalización

Entre los aportes de la tesis, se ha puesto el foco en la creación de una plantilla dentro de la herramienta AuthorAR, que posibilite diseñar una actividad educativa basada en geolocalización. De esta manera, se pueden generar recorridos a ser realizados por los

estudiantes, con preguntas que pueden estar relacionadas con un contexto específico, y/o que involucren el movimiento de los participantes. En este sentido se ha avanzado con su desarrollo. Actualmente, la implementación desarrollada permite configurar el nombre, el objetivo y el *feedback* de una actividad que contiene una serie de preguntas y respuestas para mostrar contenido de RA en diferentes ubicaciones.

Una pregunta contiene el texto de la pregunta y el *feedback* opcional sobre la respuesta que se espera. Se muestra la lista de preguntas creadas en la figura 3.

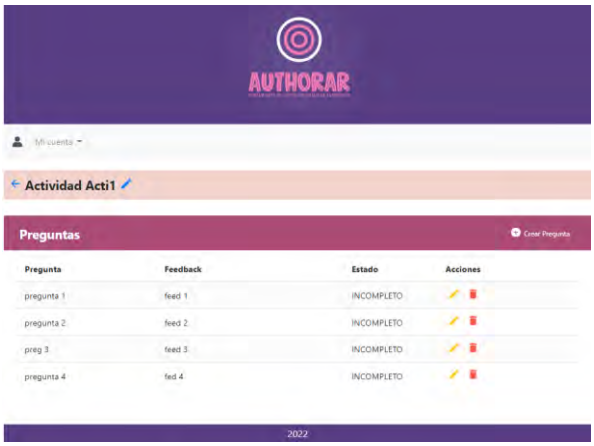


Fig. 3. Listado de preguntas de una actividad de Geolocalización

Al configurar una respuesta se describe su texto, el recurso a mostrar con RA, como una imagen, un objeto 3D o un video, su ubicación, que se selecciona utilizando un mapa, y además se debe indicar si es la respuesta correcta y un *feedback* opcional que explique su elección. En la figura 4 se muestra un ejemplo.

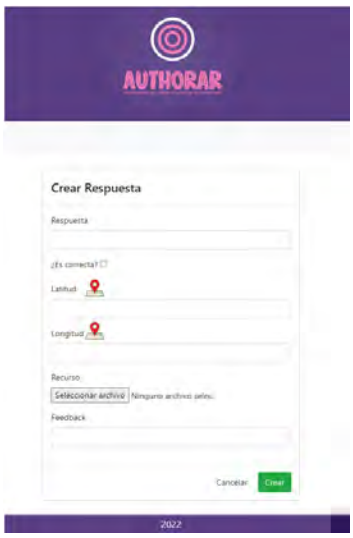


Fig. 4. Creación de una respuesta de la actividad de Geolocalización

Una pregunta tiene el estado completo si: hay más de una respuesta, y si solo una es correcta. La actividad tiene el estado completo si tiene al menos una pregunta con estado completo. Así se busca desarrollar un proyecto, cuyo estado sea completo, para poder ejecutarlo en la aplicación móvil.

Se está trabajando en completar este desarrollo y dejarlo disponible para la comunidad educativa, de manera tal de atender a la necesidad de ofrecer una herramienta de autor que acerque el diseño de actividades educativas con RA a docentes y estudiantes.

3 Líneas de investigación futura

El siguiente paso es finalizar la implementación de la aplicación móvil para ejecutar las actividades creadas en AuthorAR y realizar una evaluación con docentes y alumnos. Esta evaluación será de importancia para mejorar aspectos de usabilidad.

Como parte de la experiencia de evaluación, se trabajará con una actividad de geolocalización para que se muestren las preguntas y respuestas, y el usuario con la utilización del dispositivo móvil, que debe contar con GPS, pueda avanzar en la/s pregunta/s para completar la actividad. En cada caso, el usuario debe elegir una de las respuestas a la pregunta y reubicarse de acuerdo con su elección.

Al mismo tiempo, se espera continuar con el desarrollo de otras plantillas para incorporar a la herramienta de autor AuthorAR.

4 Referencias

1. Van Krevelen, D., y Poelman, R. (2010). A survey of augmented reality technologies, applications and limitations. International Journal of Virtual Reality, 9 (2), 1 - 20.
2. Azuma, R. (2001). Augmented reality: Approaches and technical challenges. Fundamentals of Wearable Computers and Augmented Reality, 27–63.
3. Milgram Kishino, P., Takemura, H., Utsumi, A., y Kishino, F. (1994). Augmented reality: A class of displays on the reality-virtuality continuum. En Telemanipulator and telepresence technologies (p. 282-292)
4. Ibáñez, M. B., y Kloos, C. (2018). Augmented reality for stem learning: A systematic review. Computers & Education(123), 109-123.
5. Carrizo, M. A., Barutti, M. E., y Soto, S. B. (2022). Incorporación de realidad aumentada como propuesta didáctica para la enseñanza y el aprendizaje de ciencias. Educación En La Química, 28(01), 63–73.
6. Bustillo López, M. F., Ferrer, L., Videla, S., Ohanian, G., y Vardaro, S. (2022). Realidad Aumentada como recurso disruptivo para explorar la Química Orgánica. Educación En La Química, 28(01), 74–83
7. Gómez Tone, H.; Martin-Gutierrez, J. y Valencia-Anci, B. (2022). Entrenamiento Basado en Realidad Aumentada para Mejorar las Habilidades Espaciales y la Consiguiente Mejora del Rendimiento Académico en Estudiantes de Ingeniería. Digital Education Review, 2022, Núm. 41, p. 306-322, https://doi.org/10.1344/der.2022.41.306-322.
8. Rojas Torres, J. M. (2021) Objeto de aprendizaje para la enseñanza de las estructuras de repetición utilizando dispositivos móviles y realidad aumentada. https://hdl.handle.net/20.500.12371/15031
9. Jesionkowska, J., Wild, F., y Deval, Y. (2020). Active Learning Augmented Reality for STEAM Education-A Case Study. Edu. Sci. 10, 198. doi:10.3390/educsci10080198
10. Rau L., Döring Dagny C. y Horst Robin, D. R. (2022) Pattern-Based Augmented Reality Authoring Using Different Degrees of Immersion: A Learning Nugget Approach. Frontiers in Virtual Reality. DOI:10.3389/frvir.2022.841066.
11. Moralejo, L., Sanz, C., Pesado, P., y Baldasarri, S. (2013). Authorar: Authoring tool for building educational activities based on augmented reality. En 2013 international conference on collaboration technologies and systems (cts) (p. 503-507).
12. Salazar Mesia, N.; Sanz, C. y Gorga, G. (2019). Posibilidades de las librerías de realidad aumentada en el desarrollo de actividades educativas. XIV Congreso Nacional de Tecnología en Educación y Educación en Tecnología (TE&ET 2019) (San Luis, 2019), 47-55.
13. Salazar Mesia, N. (2019). Análisis comparativo de librerías de realidad aumentada. sus posibilidades para la creación de actividades educativas. Trabajo Final Integrador Especialización en Tecnología Informática Aplicada en Educación. Universidad Nacional de La Plata. Descargado de http://sedici.unlp.edu.ar/handle/10915/76545

XX Workshop Computación Gráfica, Imágenes y Visualización **(WCGIV)**

Coordinadores

María Luján Ganuza (UNS)

Roberto Guerrero (UNSL)

Oscar Bría (UNLP)

Detección de signos de COVID-19 en radiografías de tórax a través del procesamiento digital de imágenes con redes neuronales convolucionales

Guido Sebastián Armoa[1], Nuria Isabel Vega Lencina[1] y Karina Beatriz Eckert[1]

[1]Universidad Gastón Dachary, Posadas, Misiones, Argentina.
{guidoarmoa777, nurivega.nv, karinaeck}@gmail.com

Resumen. El presente trabajo de se vio motivado por la histórica pandemia que afectó a todo el mundo desde fines del 2019. El diagnóstico temprano de la enfermedad del COVID-19 es crucial para el tratamiento y control de la enfermedad. En este contexto, la radiografía de tórax juega un papel importante; precisamente este trabajo tiene como objetivo el desarrollo y análisis de un prototipo de software para el reconocimiento de signos de COVID-19 en radiografías de tórax, a partir del procesamiento de imágenes utilizando redes neuronales convolucionales. Se propone un modelo de red neuronal convolucional para detectar signos de COVID-19 en imágenes de radiografías de tórax. La metodología propuesta experimenta y analiza el comportamiento de la misma, mediante el entrenamiento de la red utilizando distintos conjuntos de datos disponibles públicamente. Los resultados experimentales demuestran la efectividad y las limitaciones de la metodología propuesta, logrando un 79% de exactitud en la clasificación.

Palabras clave. Procesamiento digital de imágenes, Redes neuronales artificiales, Redes neuronales convolucionales, Radiografía de tórax.

1 Introducción

El coronavirus 2019 (COVID-19) es una enfermedad infecciosa que ha afectado a todo el mundo y causado millones de muertes [1].

Un importante obstáculo para controlar la propagación de esta enfermedad es la falta de experiencia o conocimiento por la novedad de la enfermedad, y la escasez de pruebas. Hasta el momento, para saber si un paciente está o no infectado las técnicas más comúnmente utilizadas por los médicos son la reacción en Cadena de la Polimerasa de Transcripción Inversa (PCR, por sus siglas en inglés de Polymerase Chain Reaction) y los test rápidos con anticuerpos. Las primeras son consideradas las más fiables hasta el momento, pero se necesitan de 4 a 6 horas para obtener resultados. Sumado a esto, dada la enorme demanda que tienen los laboratorios autorizados, estos resultados pueden demorarse varios días y además de esto los kits de prueba de PCR son muy escasos [2]. Por otra parte, los test rápidos permiten conocer en 10 o 15 minutos si una persona está o no infectada, pero tienen sensibilidad inferior al 30% y no son aconsejables para una rutina de diagnóstico [3].

El diagnóstico temprano de la enfermedad del nuevo COVID-19 es crucial para el tratamiento y control de la enfermedad. Esto motiva a estudiar formas alternativas de prueba, como las Radiografías de Tórax (RXT). Las pruebas de imagen tienen un papel importante en la detección y manejo de los pacientes, se han utilizado para apoyar el diagnóstico, determinar la gravedad de la enfermedad, guiar el tratamiento y valorar la respuesta terapéutica [4].

La RXT generalmente es el estudio por imágenes de primera línea en la evaluación de pacientes con sospecha de COVID-19 por su utilidad, disponibilidad y bajo coste junto con la evaluación clínica y los exámenes de laboratorio, colabora en la evaluación inicial y en el seguimiento de esta enfermedad. Puede ser empleada como método para la organización de la atención de las personas según los recursos existentes y las necesidades de los individuos en determinados escenarios, acelerando su proceso de clasificación, ingreso hospitalario y tratamiento [3].

Actualmente se está viviendo un enorme desarrollo en la tecnología asociada a la Inteligencia Artificial (IA), la parte de la ciencia que se ocupa del diseño de sistemas de computación inteligentes, dando lugar a nuevas herramientas y aplicaciones. Los sistemas basados en IA integran algoritmos como el aprendizaje automático y aprendizaje profundo, en entornos complejos que permiten la automatización [4], [5].

El aprendizaje automático (ML, Machine Learning), tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender. Esta trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. El aprendizaje profundo, es un subcampo dentro de ML, definido como un algoritmo automático estructurado o jerárquico que emula el aprendizaje humano utilizando distintas estructuras de redes neuronales con el fin de obtener ciertos conocimientos y lograr el aprendizaje. Destaca porque no requiere de reglas programadas previamente, sino que el propio sistema es capaz de “aprender” por sí mismo para efectuar una tarea a través de una fase previa de entrenamiento. La principal distinción del aprendizaje profundo se establece por su estructura y procesamiento de la información el cual imita las redes neuronales del cerebro humano [6], [7], [8].

Una red neuronal es un modelo de computación cuya estructura de capas se asemeja a la estructura interconectada de las neuronas en el cerebro, con capas de nodos conectados. Puede aprender de los datos, de manera que se puede entrenar para que reconozca patrones, clasifique datos y pronostique eventos futuros. Uno de los tipos más populares de Redes Neuronales Artificiales (RNA) son las conocidas como Redes Neuronales Convolucionales (CNN, por sus siglas en inglés Convolutional Neural Network). Las mismas, eliminan la necesidad de una extracción de características manual, por lo que no es necesario identificar las características utilizadas para clasificar las imágenes. Funciona mediante la extracción de características directamente de las imágenes. Las características relevantes no se entrenan previamente; se aprenden mientras la red se entrena con una colección de imágenes. La CNN dispone de decenas o cientos de capas ocultas que aprenden a detectar diferentes características de una imagen. Se aplican filtros a cada imagen de entrenamiento con distintas resoluciones, los filtros pueden variar desde características muy simples como el brillo y los bordes, hasta más complejas, como las características que definen el objeto de manera única.

Es decir, que las primeras capas pueden detectar líneas o curvas y se van especializando hasta llegar a capas más profundas que reconocen formas complejas como un rostro o una silueta. La salida de cada imagen convolucionada se emplea como entrada para la siguiente capa. Estas redes son particularmente útiles para encontrar patrones en imágenes para reconocer objetos, caras y escenas [9].

Es por esto que teniendo en cuenta la situación en la que se vive la pandemia en la Argentina, se propone el desarrollo y entrenamiento de una CNN para poder analizar en forma automatizada radiografías de tórax de pacientes con sospecha COVID-19. Para así, en pocos minutos, contribuir a saber si un paciente presenta algún signo típico de esta afección y ayudar a los médicos a tomar decisiones sobre la internación y el tratamiento inmediato, mientras se esperan los resultados de laboratorio.

El desarrollo del presente trabajo se estructura en secciones, en cada una de ellas se detallan las actividades llevadas a cabo para lograr los objetivos planteados. En la sección 2 se describe la metodología propuesta y su desarrollo. Posteriormente, en la sección 3, se exponen las configuraciones y desarrollo de cada una de las pruebas realizadas, como así también, la comparación y análisis de los resultados obtenidos en cada una de ellas. Finalmente, en la sección 4, se presentan las conclusiones obtenidas y los trabajos a futuro propuestos.

2 Propuesta

Para la implementación de la CNN, se siguieron una serie de pasos, comenzando por la carga y configuración del dataset, luego se siguió con la creación del modelo de red con cada una de sus capas, incluyendo las capas de convolución, max pooling, aplanamiento y capa de densa. Por último, se procedió al entrenamiento de la red y las pruebas de la misma.

2.1 Obtención del dataset

Se decidió utilizar el conjunto de imágenes de una base de datos abierta creada por un equipo de investigadores y médicos, pertenecientes a la Universidad de Qatar, Doha, y la Universidad de Dhaka, Bangladesh, junto con sus colaboradores de Pakistán y Malasia [9]. La misma está compuesta de radiografías de tórax con casos positivos de COVID-19, casos Normales (pulmones sanos), casos de neumonía viral y casos de opacidad pulmonar. Todas las imágenes se encuentran en formato Portable Network Graphics (PNG), y con una resolución de 299x299 píxeles.

Para la realización de las pruebas, se generaron distintas distribuciones de imágenes en diferentes datasets. El dataset número 1 (formado por 10972 imágenes para la clase COVID y 10192 para la clase no COVID) considera como COVID a las imágenes con opacidad pulmonar, COVID positivo y neumonía viral. El dataset número 2 (4960 imágenes por cada clase) considera como COVID a las imágenes con COVID positivo y neumonía viral. El dataset número 3 (3615 imágenes por clase) considera COVID solo a las imágenes clasificadas con COVID positivo. El dataset número 4 (2462 imágenes por clase) considera COVID solo a las imágenes con COVID positivo y

además se encuentra depurado, es decir, que se eliminaron previamente las imágenes con errores o de mala calidad que pudieran sesgar los resultados. En todos los casos, solo se consideran como no COVID las imágenes de pulmones normales.

El objetivo de formar distintos grupos de datos, es encontrar el conjunto de imágenes que demuestre el mejor rendimiento en el aprendizaje de la red. Para ello, se determinó la realización de 3 entrenamientos con cada conjunto de prueba para evaluar la capacidad de aprendizaje del modelo propuesto.

Se tomó el 80% del conjunto de imágenes seleccionadas para que formen parte del conjunto de entrenamiento, el 10% para el conjunto de validación y el 10% restante para el conjunto de prueba.

2.2 Desarrollo del prototipo

El prototipo desarrollado fue implementado en Python, dado que es un lenguaje multiparadigma y multinivel, de código abierto y gratuito. El mismo fue seleccionado ya que es ideal para la implementación de técnicas de IA. A su vez, se utilizó la plataforma de Anaconda ya que puede crear e implementar modelos de aprendizaje profundo que utilizan redes neuronales. Además, la misma se integra fácilmente con herramientas como TensorFlow y Keras para poder crear y entrenar modelos de redes neuronales, incluidas redes neuronales convolucionales [10]–[13].

Para la creación de la red neuronal se utilizó la función Sequential() que agrupa una pila lineal de capas en un archivo tf.keras.Model [14].

Luego se procedió a añadir las capas que forman parte de la red. Para agregar cada capa se utilizó la función add(). Los parámetros que recibe esta función son los siguientes: Número de filtros utilizados, tamaño del filtro, la variable padding referente al filtro en las esquinas, la altura y longitud que tienen las imágenes de entrada y la función de activación utilizada. A su vez, se añadió una capa de MaxPooling, para el cual se utilizó un filtro de tamaño 2x2.

Con ésta última capa finaliza la primera convolución, en la figura 1 se observa de manera gráfica la arquitectura de la misma.

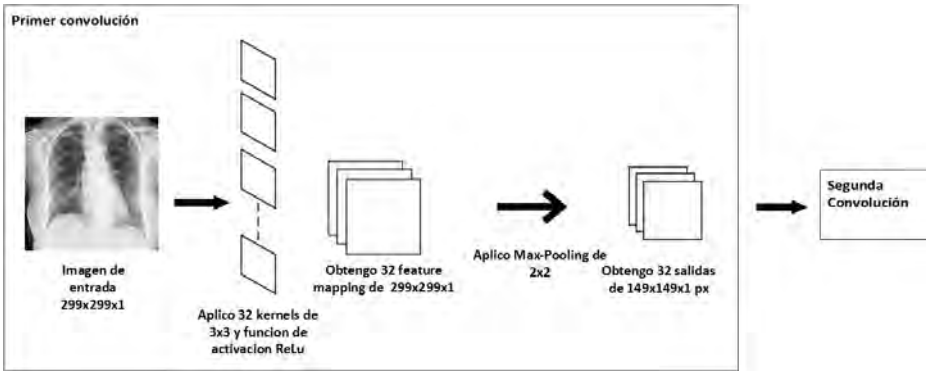


Figura 1. Arquitectura de la primera convolución.

A continuación, se agrega la segunda y tercera capa de convolución, en este caso, capas ocultas, por lo que no se especifica el tamaño de la imagen (inputShape). En ambos casos los parámetros que reciben son los siguientes: El número de filtros, el tamaño del filtro, la variable padding y la función de activación utilizada.

Al finalizar, se añade una capa de MaxPooling, para el cual se utilizó un filtro de tamaño 2x2, con esto finalizan las convoluciones.

Por último, se añade una capa de aplanamiento para transformar la red multidimensional a una dimensión para que luego se pase a una red neuronal tradicional. En la figura 2, se puede apreciar la arquitectura final de la red neuronal convolucional completa, formada por las 3 capas de convoluciones, el aplanamiento y el perceptrón multicapa. Las salidas de la MLP [15] (perceptrón multicapa, por sus siglas en inglés multilayer perceptrón) representan a las salidas de toda la red, dando como resultado un porcentaje de acierto para cada clase.

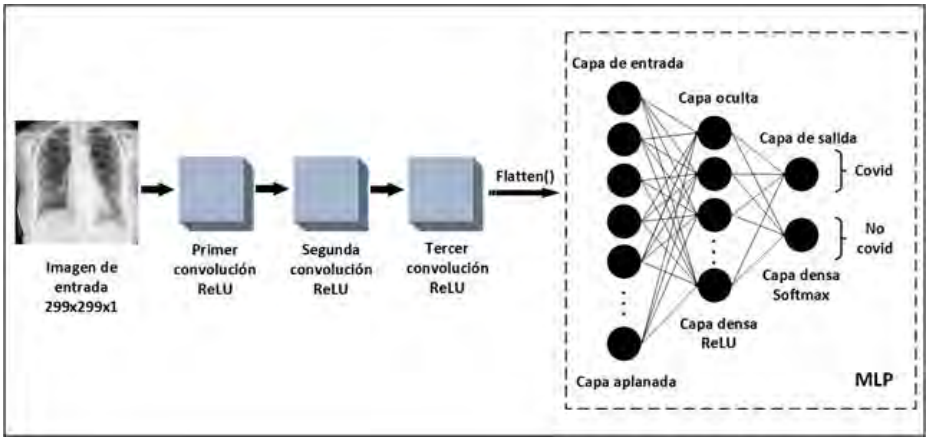


Figura 2. Arquitectura completa del prototipo con CNN.

2.3 Entrenamiento de la red

El siguiente paso es la especificación del entrenamiento y el dataset. Para ello se utilizó la función fit_generator(), que recibe los siguientes parámetros: conjunto de entrenamiento, número de pasos por épocas, número de épocas (en este caso 20), conjunto de validación y número de pasos de validación [16].

Una vez entrenada la red, el modelo se guarda en el directorio correspondiente, en este caso, en la carpeta específica llamada modelo, la estructura se guarda bajo el nombre "modelo.h5" y los pesos bajo el nombre de "pesos.h5".

Las cuatro métricas que devuelve el entrenamiento de la red tanto para el conjunto de entrenamiento como para el conjunto de validación son la exactitud del aprendizaje y el error.

3 Pruebas y Resultados

Con el fin de validar y analizar la funcionalidad del prototipo desarrollado con la metodología propuesta detallada en la sección anterior, se han planeado las siguientes pruebas:

Teniendo en cuenta el cuantioso número de imágenes de prueba, se recurrió al uso de una matriz de confusión, una herramienta útil para medir el desempeño en la clasificación de las imágenes de manera eficiente [17], [18].

Se realizaron tres pruebas por cada dataset, obteniendo un total de 12 pruebas. Es decir, cada prueba consistió en entrenar el modelo con su respectivo conjunto de datos (tabla 1). Una vez finalizado el entrenamiento, se utiliza la matriz de confusión, para poder evaluar el desempeño de la red ante el conjunto de prueba y analizar los valores de cada una de sus métricas (tabla 2). Probando automáticamente todas las imágenes del conjunto de prueba y comparando sus etiquetas reales con las predicciones hechas por la red.

A continuación, en la tabla 1 se presenta un cuadro comparativo con los resultados de error y exactitud obtenidos durante el entrenamiento de cada red. A su vez, en la tabla 2 se presentan los valores de las métricas obtenidas a partir de las matrices de confusión generadas en cada prueba.

Tabla 1. Resumen resultados de entrenamientos.

Dataset	Prueba	Error entrenamiento	Error validación	Exactitud entrenamiento	Exactitud validación
	1	0,3840	0,3319	0,8292	0,8552
1	2	0,3575	0,3030	0,8478	0,8725
	3	0,3478	0,2951	0,8511	0,8730
	4	0,2956	0,6591	0,8715	0,6099
2	5	0,2928	0,6354	0,8806	0,6069
	6	0,2939	0,5788	0,8752	0,6633
	7	0,3993	0,2759	0,8009	0,8920
3	8	0,3993	0,2759	0,8009	0,8920
	9	0,3602	0,2588	0,8387	0,9034
	10	0,3510	0,3900	0,8410	0,8208
4	11	0,3971	0,3815	0,8122	0,8562
	12	0,4212	0,3944	0,7909	0,8375

En la tabla 1 se puede observar que la gran mayoría de los entrenamientos tuvieron resultados que en primera instancia parecen prometedores. Se destaca el entrenamiento de la prueba 9 por lograr alcanzar una exactitud de 90% y un error de tan solo 0,25 sobre el conjunto validación. Por otra parte, la prueba 5 fue la prueba con los resultados menos favorables, contando con un error de 0,63 y una exactitud de 60% sobre el conjunto de validación.

Tabla 2. Resumen de las pruebas.

Dataset	Prueba	Precisión		Sensibilidad		F1-score		Exactitud
		Covid	No Covid	Covid	No Covid	Covid	No ovid	
1	1	0,6667	0,7757	0,8310	0,5845	0,7398	0,6667	70,78%
	2	0,5804	0,8560	0,9470	0,3153	0,7197	0,4609	63,11%
	3	0,5142	0,9118	0,9941	0,0609	0,6778	0,1142	52,75%
2	4	0,4995	0,5000	0,9455	0,0544	0,6536	0,0982	49,95%
	5	0,4945	0,4400	0,9152	0,0665	0,6421	0,1156	49,04%
	6	0,4995	0,0000	1,0000	0,0000	0,6662	0,0000	49,95%
3	7	0,5142	0,9545	0,9972	0,0580	0,6786	0,1094	52,76%
	8	0,5385	0,8378	0,9669	0,1713	0,6917	0,2844	56,91%
	9	0,4794	0,4727	0,5470	0,4061	0,5110	0,4368	47,65%
4	10	0,7912	0,7984	0,8008	0,7886	0,7960	0,7935	79,47%
	11	0,6579	0,6858	0,7114	0,6301	0,6836	0,6568	67,07%
	12	0,6667	0,6549	0,6423	0,6789	0,6542	0,6667	66,06%

Realizando un breve análisis de los resultados obtenidos con cada conjunto de datos de prueba, presentes en la tabla 2 se puede observar que todas las pruebas realizadas con un mismo dataset dieron resultados similares entre ellas.

El dataset 1, es el conjunto que presenta mayor variación en sus resultados, alcanzando una exactitud del 71% en la prueba 1 (tabla 2).

Las pruebas realizadas con el dataset número 2 se destacan por obtener los resultados menos favorables, en ningún caso logró superar el 50% de exactitud. El hecho de tener la misma cantidad de imágenes de cada clase, un 50% de exactitud indica que la red clasifica prácticamente a todas las imágenes como una misma clase.

Para el caso del dataset número 3, se ve un comportamiento similar a las pruebas realizadas con el dataset número 2, con resultados poco destacables. Es conveniente resaltar que la prueba 9 realizada con dicho conjunto de datos, presenta el mejor valor de exactitud alcanzando un 90% sobre el conjunto de validación durante el entrenamiento (tabla 1). Hay que tener en cuenta que, si bien el valor de exactitud en el entrenamiento es alto, no asegura un buen rendimiento en la clasificación de las imágenes, tal como se puede apreciar en los resultados de clasificación de la tabla 2.

Para finalizar, el dataset número 4 se destaca por obtener los mejores resultados, superando en todos los casos el 66% de exactitud sobre el conjunto de prueba. Resaltando la prueba 10 donde se obtuvo una exactitud del 79% en la clasificación de las imágenes del conjunto de prueba, en la figura 3 se observa la matriz de confusión generada.

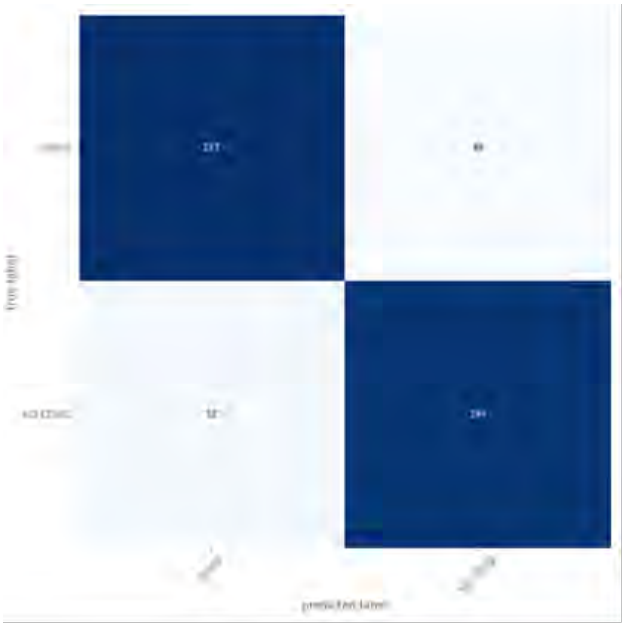


Figura 3. Matriz de confusion de la prueba 10

3.1 Validación de la red

Finalizadas las pruebas anteriores, se seleccionó la red que mejor desempeño demostró en la clasificación de imágenes, dicha red fue la obtenida en la prueba 10 con el dataset 4 (obteniendo un 79% de exactitud) y se procedió a la validación de la misma.

Ésta actividad fue realizada con la ayuda de un profesional en el área de diagnóstico por imágenes del Hospital Escuela de Agudos Dr. Ramón Madariaga.

Se suministró al profesional un dataset con 52 imágenes sin etiquetas provenientes del conjunto de prueba del dataset utilizado para el entrenamiento de la red, para las cuales se solicitó al mismo que las clasifique según su criterio. De las 52 imágenes de radiografías de tórax, 37 de ellas fueron clasificadas por el profesional de la misma manera que la red, 3 no pudieron ser definidas bajo ninguna clasificación por lo cual se decidió no tomarlas en cuenta, y las 12 imágenes restantes fueron clasificadas contradiciendo a las salidas obtenidas por la red.

Se obtuvo un 75% de coincidencia con la clasificación realizada por el médico. Es conveniente destacar, que el criterio de clasificación varía dependiendo de cada persona, dado que diferentes factores como la experiencia o el error humano pueden afectar a la decisión tomada. Por consiguiente, se consideran favorables los resultados obtenidos en la validación.

4 Conclusiones

Con el fin de cumplir con los objetivos de este trabajo fue necesario un exhaustivo proceso de investigación y análisis sobre los signos del COVID-19 en las radiografías de pulmón, profundizar y comprender la aplicación de las redes neuronales convolucionales en la clasificación de imágenes, diseñar la arquitectura de la red y los diferentes datasets, implementar el prototipo, entrenar y probar la red con cada conjunto de datos, y validar el funcionamiento del prototipo junto a un profesional. Para finalmente, analizar los resultados obtenidos y determinar la viabilidad de la propuesta.

Tras el análisis, se puede observar que las tablas de resultados y las matrices de confusión demuestran que el dataset número 4 presentó el mejor desempeño en la clasificación de imágenes, alcanzando un 79% de exactitud. En base a esto, se puede deducir que la depuración realizada sobre el mismo fue el factor clave para lograr esta mejora. En este sentido, las pruebas indican que la eliminación de imágenes de mala calidad produce una mejora exponencial en los resultados de la clasificación, demostrando la sensibilidad que poseen las CNN ante este tipo de imágenes y la importancia de tener un conjunto de datos representativo.

Dentro del análisis expuesto es posible observar que la red entrenada posee dos grandes limitaciones, la técnica utilizada para realizar la radiografía, lo cual puede afectar a la calidad de la imagen, y las afecciones crónicas que pueda presentar el paciente. Por otra parte, es importante destacar el gran costo computacional que demandan las CNN para su desarrollo.

Tal y como se ha podido comprobar, el prototipo obtenido resultó prometedor para el médico especialista en diagnóstico por imágenes, ya que, mediante el mismo se logró la clasificación de radiografías de pulmón con signos de COVID-19 de manera eficaz, en cuestión de segundos una vez obtenida la radiografía. De esta manera, se podría generar una alternativa para cooperar con los organismos de salud en el diagnóstico de COVID-19.

De acuerdo con lo expresado anteriormente, se demostró que con la aplicación de redes neuronales convolucionales se logró construir un prototipo capaz de reconocer eficazmente los signos de COVID-19 en radiografías de pulmón. De esta manera, queda en evidencia que el presente estudio logró cumplir satisfactoriamente los objetivos propuestos al inicio de la investigación.

El prototipo desarrollado puede presentar mejoras a futuro, una de ellas es implementar una red que, en vez de realizar una clasificación binaria como en este trabajo, realice una clasificación de múltiples clases. Además, se propone utilizar transferencia de aprendizaje y probar la eficiencia entre los distintos modelos disponibles (ResNet, VGG, etc.).

5 Referencias

[1] "COVID-19 Visualizer." https://www.covidvisualizer.com/ (accessed Aug. 12, 2020).

[2] E. M. Chamorro, A. D. Tascón, L. I. Sanz, S. O. Vélez, and S. B. Nacenta, "Diagnóstico

radiológico del paciente con COVID-19," *Radiologia*, vol. 63, no. 1, p. 56, Jan. 2021, doi: 10.1016/J.RX.2020.11.001.

[3] E. Martínez Chamorro, A. Díez Tascón, L. Ibáñez Sanz, S. Ossaba Vélez, and S. Borruel Nacenta, "Radiologic diagnosis of patients with COVID-19," *Radiologia*, vol. 63, no. 1, pp. 56–73, Jan. 2021, doi: 10.1016/j.rx.2020.11.001.

[4] "¿Qué es la inteligencia artificial (IA)? - MATLAB & Simulink." https://la.mathworks.com/discovery/artificial-intelligence.html (accessed Sep. 24, 2020).

[5] *The Handbook of Artificial Intelligence*. Elsevier, 1981.

[6] "Introducción al Aprendizaje Automático - Fernando Sancho Caparrini." http://www.cs.us.es/~fsancho/?e=75 (accessed Sep. 07, 2020).

[7] "¿Qué es el Deep Learning? | SmartPanel." https://www.smartpanel.com/que-es-deep-learning/ (accessed Sep. 07, 2020).

[8] "Deep Learning - Libro online de IAAR." https://iaarbook.github.io/deeplearning/ (accessed Nov. 17, 2021).

[9] "COVID-19 Radiography Database | Kaggle." https://www.kaggle.com/tawsifurrahman/covid19-radiography-database (accessed May 21, 2021).

[10] "Anaconda | Use Cases." https://www.anaconda.com/use-cases (accessed Mar. 09, 2022).

[11] "Librerías ML: TensorFlow, Scikit-learn, Pytorch y Keras - Platzi." https://platzi.com/blog/librerias-de-machine-learning-tensorflow-scikit-learn-pythorch-y-keras/?gclid=CjwKCAiA4KaRBhBdEiwAZi1zzrm5QrcLNP_R6BqpM9DZj0H6v9yvzsHEXltymGQzgu3FfBQaImjc_hoCmHMQAvD_BwE&gclsrc=aw.ds (accessed Mar. 10, 2022).

[12] "Keras: the Python deep learning API." https://keras.io/ (accessed Jul. 04, 2022).

[13] "TensorFlow." https://www.tensorflow.org/?hl=es-419 (accessed Jul. 04, 2022).

[14] "The Sequential class." https://keras.io/api/models/sequential/ (accessed Mar. 02, 2022).

[15] J. Hilera and V. Martínez, "Redes neuronales artificiales: fundamentos, modelos y aplicaciones," *Madrid: Ra-ma*, no. January, p. 9, 1995, [Online]. Available: http://en.scientificcommons.org/7007722.

[16] "fit_generator function - RDocumentation." https://www.rdocumentation.org/packages/keras/versions/2.4.0/topics/fit_generator (accessed Jul. 04, 2022).

[17] "La matriz de confusión y sus métricas – Inteligencia Artificial –." https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/ (accessed Mar. 15, 2022).

[18] "Evaluando los modelos de Clasificación en Aprendizaje Automático: La matriz de confusión.| profesorDATA.com." https://profesordata.com/2020/08/07/evaluando-los-modelos-de-clasificacion-en-aprendizaje-automatico-la-matriz-de-confusion-claramente-explicada/ (accessed Apr. 04, 2022).

RSL Sobre Diagnóstico de COVID-19 Utilizando Redes Neuronales Artificiales Convolucionales

Eduardo Hugo Bennesch[1], Rocio Klan[2], Juan Claudio Mousquere[3]
Postgrado, Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones
Posadas, Misiones, República Argentina.

1 Ingbennesch@gmail.com, 2 Rocioklan@gmail.com, 3 J.C.Mousquere@gmail.com

Resumen. La rápida propagación del COVID-19 a nivel mundial obligó a desarrollar sistemas y métodos para predecir el comportamiento del virus o detectar la infección. Una de las formas de detectar al COVID-19 es a través del análisis de Rayos X o de Tomografías Computarizadas del tórax, por lo que resulta relevante desarrollar modelos de IA que puedan asistir en la toma de decisiones. El objetivo de este informe es presentar una Revisión Sistemática de la Literatura (RSL) para evidenciar los avances en el desarrollo de soluciones software utilizando Redes Neuronales Artificiales (RNA) enfocadas en la detección del COVID-19. La búsqueda de artículos se realizó en seis fuentes diferentes. Como resultado, se obtuvieron 18 estudios clasificados en 5 dimensiones: Tipos de Propuestas, Tipos de datos, Validación, Características y Tipo de Soporte. Este trabajo evidencia que existen en simultáneo una gran cantidad de investigaciones relacionadas, que apuntan a la necesidad de encontrar soluciones prácticas, de bajo costo y de rápida evolución. La mayoría de los trabajos estudiados hacen hincapié en el dinamismo de los métodos de entrenamiento y la precisión de las respuestas.

Palabras Clave: redes neuronales convolucionales, machine learning, detección por imágenes, COVID-19, diagnóstico temprano.

1 Introducción

El diagnóstico precoz de COVID-19 permite a los profesionales de la salud conjuntamente con las autoridades gubernamentales romper la cadena de transmisión y aplanar la curva epidémica [1]. En marzo de 2020, el crecimiento de casos de COVID-19 escaló a aproximadamente 35000 diagnósticos en solo 15 días en todo el mundo [2].

El COVID-19 causa, desde síntomas leves a casos de estados graves o críticos [3]. El número total de casos de COVID-19 al momento de realizar la RSL es de aproximadamente 336.068, de los cuales 223.819 se consideran activos y 112.249 son catalogados como casos cerrados.

De los casos activos, el 95% de los pacientes, es decir 213.179 personas presentan síntomas leves y el 5% restante es, decir: 10.640 presentan cuadros severos o críticos.

Mientras que el 87% de los casos cerrados, en otras palabras, 97.636 personas se recuperaron y el 13%, 14.613 personas, fallecieron.

Las Redes Neuronales Convolucionales también conocidas como RNC o CNN por sus siglas en inglés [4] representan una técnica de aprendizaje profundo que consta de varias capas apiladas juntas que utilizan conexiones locales conocidas como campo receptivo local y distribución de peso para un mejor rendimiento y eficiencia. La arquitectura profunda ayuda a estas redes a obtener muchas características diferentes y complejas que una red neuronal simple no puede aprender. Las RNC han demostrado un excelente rendimiento en varias aplicaciones, como la clasificación de imágenes, detección de objetos, reconocimiento de voz, procesamiento del lenguaje natural y análisis de imágenes médicas. En la **Imagen 1** se puede observar la visión general de esta metodología.

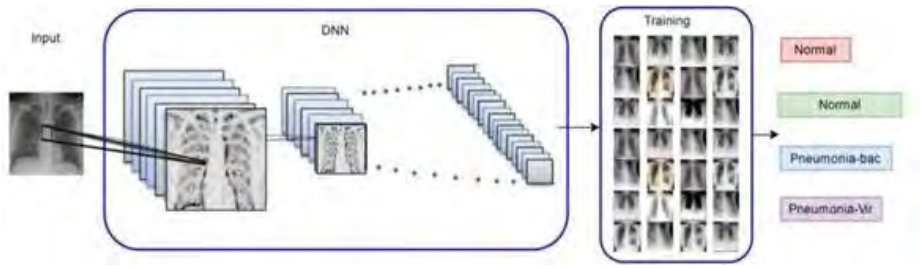


Imagen 1*: Funcionamiento de una RNC [5].*

El presente trabajo está organizado de la siguiente manera, en la sección 2 son presentadas las características de la RSL, objetivos, preguntas de investigación, cadenas de búsqueda y criterios de inclusión y exclusión. La sección 3 contiene la ejecución de la RSL. La sección 4 se presenta el reporte de resultados obtenidos y finalmente, en la sección 5 se muestran las conclusiones y futuras líneas de trabajo.

Trabajos Relacionados. Mediante el trabajo de investigación ejecutado, podemos afirmar que al momento de la realización del mismo no existe ninguna RSL realizada sobre el tema elegido.

2 Planificación de la RSL

El objetivo de esta etapa es definir el protocolo de la revisión. Para ello serán elaborados el objetivo de la revisión, las preguntas de investigación que se pretenden responder con esta revisión, la cadena de búsqueda, las fuentes donde se realizarán las búsquedas, los criterios de inclusión y exclusión de artículos y las dimensiones a utilizar para clasificar los estudios primarios obtenidos.

2.1 Objetivos y Preguntas de Investigación

2.1.1 Objetivos

Esta RSL se realiza con el objetivo de *sintetizar* la literatura existente acerca del Diagnóstico de COVID-19 Utilizando Redes Neuronales Artificiales Convolucionales. El conocimiento extraído de esta RSL será la base para un mejor entendimiento de los adelantos científicos en materia de investigación acerca de la utilización del Diagnóstico por Imágenes en la detección temprana de síntomas compatibles con COVID-19.

2.1.2 Preguntas de Investigación

- ¿Qué tipo de Metodologías existen para detectar COVID-19 con RNC?
- ¿Cómo fue el tratamiento de los orígenes de datos?
- ¿Cómo puede ayudar las RNC a mejorar la velocidad de detección de casos de COVI- 19?
- ¿Cuáles son las ventajas de las redes neuronales convolucionales sobre otras redes neuronales en el manejo de imágenes?
- ¿Existen formas de automatizar los resultados?
- ¿Cómo se validan las metodologías?

A continuación, en la **Tabla 1** se detallan las preguntas realizadas y las dimensiones correspondientes a cada pregunta.

Tabla 1: *Preguntas y dimensiones.*

Preguntas	Dimensiones
¿Qué tipo de Metodologías existen para detectar COVID-19 con RNC?	Tipos de Propuestas: método, conocimiento, herramienta, características.
¿Cómo fue el tratamiento de los orígenes de datos?	Tipo de datos.
¿Cómo puede ayudar las RNC a mejorar la velocidad de detección de casos de COVID-19?	Validación: propuesta, ejemplo, conclusión.
¿Cuáles son las ventajas de las redes neuronales convolucionales sobre otras redes neuronales en el manejo de imágenes?	Característica: definiciones.
¿Existen formas de automatizar los resultados?	Tipo de Soporte: manual, herramienta.
¿Cómo se validan las metodologías?	Validación

2.2 Cadena de Búsqueda

En la **Tabla 2** se enumeran las palabras clave y las relacionadas seleccionadas, para conformar la cadena de búsqueda.

Tabla 2: *Palabras clave y relacionadas*

Palabras Clave	Palabras Relacionadas
covid-19	Coronavirus, nova covid-19
Convolutional neural Network	

Utilizando los operadores lógicos AND y OR se obtuvo la siguiente cadena de búsqueda:

"(covid-19 OR nova covid-19 OR coronavirus) AND (convolutional) AND (neural) AND (network)"

2.3 Fuentes de Búsqueda

Las búsquedas fueron realizadas en los siguientes repositorios digitales:

- https://scholar.google.es
- https://www.sciencedirect.com
- https://ieeexplore.ieee.org/Xplore/home.jsp
- http://dl.acm.org/
- http://www.biblioteca.mincyt.gob.ar/recursos/ver?id=scielo
- https://www.elsevier.com/solutions/scopus

2.4 Criterios de Exclusión e Inclusión

Criterio de exclusión:

- Fecha de publicación del artículo, fue omitido todo lo anterior a la fecha de aparición del "paciente cero" de la enfermedad.

Criterios de Inclusión:

- Artículos sobre diagnóstico de COVID-19 utilizando redes neuronales.
- Diagnóstico automatizado de COVID-19.
- Aplicación de técnicas de Deep Learning en la detección de COVID-19.

3 Ejecución de la RSL

- Se realizó la búsqueda según la cadena de búsqueda en el título (artículos encontrados), teniendo en cuenta las facilidades que proporciona cada fuente, para filtrar los artículos.
- Se filtraron los artículos de acuerdo al título y al abstract.
- Se obtuvieron estudios primarios leyendo el texto completo.
- Se clasificaron los estudios primarios según las dimensiones definidas.

En la **Tabla 3** se presenta una descripción cuantitativa de los resultados, donde se indica para cada fuente, la cantidad de artículos encontrados, los artículos restantes luego de leer el título y el abstract, y finalmente, los artículos obtenidos luego de leer el texto completo. También se observa la gran cantidad de resultados en bruto en el google académico, debido a la carencia de filtros más específicos que si brindan las otras fuentes de búsqueda.

Tabla 3: *Síntesis de las búsquedas.*

Fuente	Artículos encontrados	Después del primer filtrado	Después del texto completo
Science Direct	20	3	3
Google Académico	1.290	15	13
ACM	0	0	0
MINCYT	0	0	0
IEEE	7	0	0
SCOPUS	0	0	0
Totales	1.317	18	16

4 Reporte de Resultados

4.1 ¿Qué tipo de Metodologías existen para detectar COVID-19 con RNC?

4.1.1 Redes Neuronales Concatenadas

En el estudio [6] se entrenaron redes ResNet50V2, Xception y una concatenación de redes neuronales Xception y ResNet50V2 utilizando la biblioteca Keras.

La red neuronal concatenada está diseñada concatenando las características extraídas de Xception y ResNet50V2 y luego conectando las características concatenadas a una capa convolucional que está conectada al clasificador. Esta red neuronal ha mostrado una mayor precisión en comparación con las demás.

4.1.2 CoroNet

En la investigación [1] se implementó CoroNet que es una arquitectura RNC diseñada para la detección de infección por Covid-19 a partir de imágenes de rayos X de tórax. Se basa en la arquitectura Xception RNC. Xception, que significa la versión Extreme de Inception (su modelo predecesor) es una arquitectura RNC de 71 capas de profundidad previamente entrenada en el conjunto de datos ImageNet. Xception utiliza capas de convolución separables en profundidad con conexiones residuales en lugar de convoluciones clásicas.

4.1.3 Deep Feature Extraction

En [7] se usó un modo de sintonización superficial durante la adaptación y el entrenamiento de un modelo RNC pre-entrenado de ImageNet usando el conjunto de datos de imágenes de rayos X de tórax recopilados. Se utilizaron las características de RNC disponibles en el mercado de modelos pre-entrenados en ImageNet (donde el entrenamiento se realiza solo en la capa de clasificación final) para construir el espacio de características de la imagen.

Sin embargo, debido a la alta dimensionalidad asociada con las imágenes, se aplicó en [7] Análisis de Componentes Principales (por sus siglas en inglés PCA) para proyectar el espacio de características de alta dimensión en una dimensión inferior, donde se ignoraron las características altamente correlacionadas. Este paso es importante para que la descomposición de la clase produzca clases más homogéneas, reduzca los requisitos de memoria y mejore la eficiencia del marco.

4.1.4 Arquitectura DeTraC

El modelo DeTraC consta de tres fases. En la primera fase, se entrenó el modelo RNC pre-entrenado de backbone de DeTraC para extraer características locales profundas de cada imagen. Luego se aplicó la capa de descomposición de clase de DeTraC para simplificar la estructura local de la distribución de datos. En la segunda fase, el entrenamiento se lleva a cabo utilizando un sofisticado método de optimización de descenso de gradiente. Finalmente, se usó la capa de composición de clase de DeTraC para refinar la clasificación final de las imágenes. Los componentes de descomposición de clase y composición se agregan respectivamente antes y después de la transformación del conocimiento de un modelo RNC pre-entrenado de ImageNet. El componente de descomposición de clase que apunta a dividir cada clase dentro del conjunto de datos de imagen en k subclases, donde cada subclase se trata de forma independiente. Luego, esas subclases se ensamblan nuevamente utilizando el componente de composición de clase para producir la clasificación final del conjunto de datos de imagen original [7].

4.1.5 SegNet

En [8] se presentó una red unificada de alta precisión para la segmentación de la infección por COVID-19 a partir de imágenes de TC de tórax. Esta red consta de dos partes: codificador y decodificador. El codificador con 4 capas (es decir, E1, E2, E3, E4) obtiene información sólida a través del extractor de características y PASPP (Pooling Progresivo de la Pirámide Espacial de Atrous). Cada capa emplea bloques residuales y FV (Feature Variation) como operaciones básicas para extractores de características, excepto la capa E4. El bloque residual suma las características de entrada y los resultados después de dos capas convolucionales, lo que alivia efectivamente el gradiente de fuga.

4.1.6 Enfoque de Aprendizaje Desequilibrado

En [9] se utilizó un conjunto de datos altamente desequilibrado, lo que podría conducir a un aprendizaje sesgado del modelo, sin embargo, el número de imágenes de CXR infectadas con coronavirus es muy inferior en comparación con otras clases, por lo tanto, las técnicas de equilibrio de clase deben asegurarse para suavizar el proceso de aprendizaje. En este trabajo se discuten dos enfoques para manejar el problema de desequilibrio de clase: enfoque de clase de peso y sobre-muestreo aleatorio.

4.1.7 Redes Neuronales Convolucionales Basadas en la Evolución Diferencial Multiobjetivo

En [10] se plantea la utilización de redes neuronales convolucionales basadas en la evolución multiobjetivo-diferencial (MODE) para la clasificación de pacientes infectados por COVID-19 a partir de imágenes de TAC del tórax. El trabajo refleja la buena performance que propone el modelo y hace referencia al problema que presenta en los ajustes de hiperparámetros.

En la **Imagen 2** se muestra el funcionamiento del proceso de entrenamiento y testing de éste modelo de clasificación.

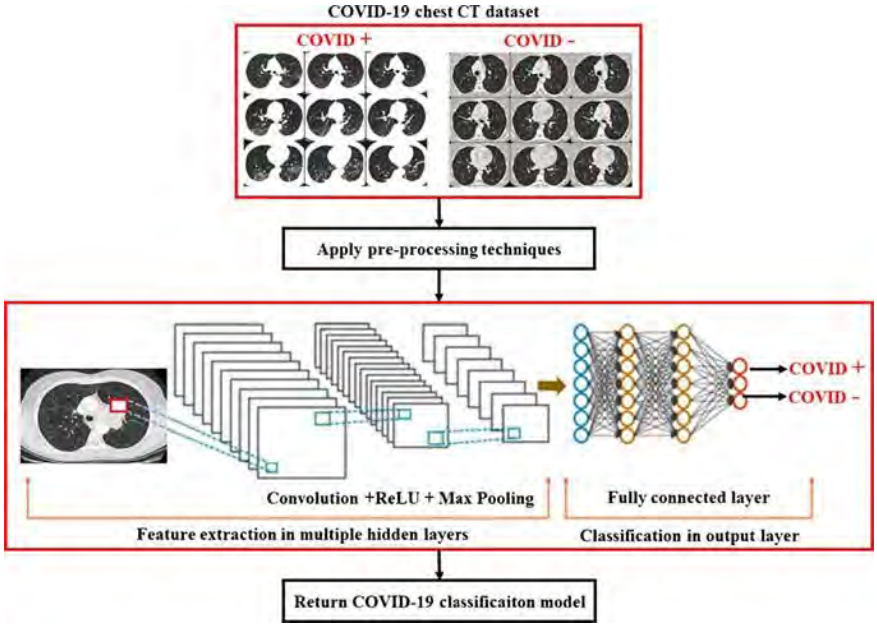


Imagen 2: *Diagrama de bloques del proceso de entrenamiento del modelo de clasificación COVID-19 basado en la CNN [10].*

4.2 ¿Cómo pueden ayudar las RNC a mejorar la velocidad de detección de casos de COVID-19?

- Triaje rápido: Una CXR (Radiografía de Tórax) permite el triaje rápido de pacientes con sospecha de COVID-19 y se puede hacer en paralelo a las pruebas virales (lo que lleva tiempo) para ayudar a aliviar los altos volúmenes de pacientes, especialmente en las áreas más afectadas donde se han quedado sin capacidad (por ejemplo, Nueva York, España e Italia), o incluso de forma independiente cuando las pruebas virales no son una opción (bajo suministro). Además, la CXR puede ser bastante efectiva para el triaje en áreas geográficas donde se les indica a los pacientes que se queden en casa hasta la aparición de síntomas avanzados (p. Ej., La ciudad de Nueva York), ya que a menudo se ven anormalidades en el momento de la presentación cuando llegan pacientes con sospecha de COVID-19 en sitios clínicos [11].
- Disponibilidad y accesibilidad: Las CXR están fácilmente disponibles y accesibles en muchos sitios clínicos y centros de imágenes, ya que se considera un equipo estándar en la mayoría de los sistemas de atención médica [11].
- Portabilidad: la existencia de sistemas CXR portátiles significa que las imágenes se pueden realizar dentro de una sala de aislamiento, lo que reduce significativamente el riesgo de transmisión de COVID-19 durante el transporte a sistemas fijos, como escáneres de TC, así como dentro de las salas que albergan los sistemas de imágenes fijas [11].

- En [9] se puede observar que se logra un alto grado de VPP (Valor de Predicción Positiva) para los casos de COVID-19 (entre 98,4% y 98,9%), lo que indica muy pocas detecciones “falsas positivas de COVID-19”.

4.3 ¿Cómo fue el tratamiento de los orígenes de datos?

En [12] se utilizó en conjunto de datos COVID-CT. Primero se recolectaron 760 preimpresiones sobre COVID-19 de medRxiv1 y bioRxiv2, publicadas del 19 de enero al 25 de marzo. Muchas de estas preimpresiones informan casos de pacientes con COVID-19 y algunas de ellas muestran tomografías computarizadas en los informes. Las tomografías computarizadas están asociadas con subtítulos que describen los hallazgos clínicos en las tomografías computarizadas. Para extraer la información de estructura de bajo nivel de los archivos PDF de preimpresiones se utilizó PyMuPDF3.

Si bien COVID-CT es el dataset más grande que existe de CT sobre COVID-19, aun es pequeño. El entrenamiento de modelos de aprendizaje profundo en un conjunto de datos tan pequeño puede conducir fácilmente a un sobreajuste: el modelo funciona bien en los datos de entrenamiento, pero se generaliza mal en los datos de prueba. Para abordar este problema, se adoptaron dos enfoques: aprendizaje de transferencia y aumento de datos. El aprendizaje de transferencia tiene como objetivo aprovechar una gran colección de datos de un dominio relevante para ayudar con el aprendizaje en el dominio interesado [12].

4.3.1 Dataset Annotation

En [8] aunque se capturaron suficientes datos de las imágenes de TC del tórax COVID-19, las etiquetas anotadas precisas también eran indispensables. Para permitir que ese modelo aprenda sobre anotaciones precisas, crearon un equipo de seis anotadores con antecedentes de radiología profunda y habilidades de anotación competentes para anotar las áreas y límites de las regiones de infección pulmonar y COVID-19. Además, la calidad de las anotaciones finales fue evaluada por un radiólogo senior con experiencia clínica de primera línea de COVID-19.

4.4 ¿Cuáles son las ventajas de las Redes Neuronales Convolucionales sobre otras Redes Neuronales en el manejo de imágenes?

En [6] al concatenar las características de salida de ambas redes, ayudamos a la red a aprender a clasificar la imagen de entrada de ambos vectores de características, y esto ha resultado en una mejor precisión respecto de otro tipo de redes. Se logró una precisión promedio de 99.56% y 80.53% de recuperación para la clase COVID-19, y una precisión general igual a 91.4% entre cinco pliegues.

En [5] el modelo CoroNet logró una precisión general de 89.5%, mientras que la precisión y la medida F para la clase Covid-19 son 96.6% y 98% respectivamente.

4.5 ¿Existen formas de automatizar los resultados?

En [3] se propuso una predicción automática de COVID-19 utilizando una red neuronal de convolución profunda basada en modelos de transferencia pre-entrenados e

imágenes de rayos X de tórax. Para ese propósito, utilizaron los modelos pre-entrenados ResNet50, InceptionV3 e Inception-ResNetV2 para obtener una mayor precisión de predicción para pequeños conjuntos de datos de rayos X. En este estudio se confirma que las imágenes de rayos X del tórax son una buena herramienta para la detección de COVID-19. También se ha demostrado que los modelos pre-entrenados producen resultados muy altos en el pequeño conjunto de datos (50 COVID-19 vs. 50 Normal).

En [10] se plantea la clasificación de imágenes de tomografías computadas usando redes neuronales convolucionales basadas en la evolución diferencial multiobjetivo.

En [8] se presentó una red unificada de alta precisión para la segmentación de la infección por COVID-19 a partir de imágenes de TC de tórax. Esta red consta de dos partes: codificador y decodificador

4.6 ¿Cómo se validan las metodologías?

En [13] para comenzar la fase de capacitación de uno de los siete modelos de aprendizaje profundo seleccionados y / o ajustados, el conjunto de datos preprocesado se divide en 80-20 de acuerdo con el principio de Pareto. Eso significa 20% de los datos de la imagen se utilizarán para la fase de prueba. Nuevamente, al dividir el 80% de los datos se usarán para construir un entrenamiento similar además de los conjuntos de validación.

5 Conclusiones

Consideramos que el principal impacto de este trabajo es que permite conocer, de manera sintética, los avances en el campo de la detección de COVID-19 mediante el análisis de imágenes, identificando las técnicas desarrolladas al momento de la realización de esta primera RSL acerca del tema, esperando sirva como punto de partida para futuras investigaciones.

La velocidad en el diagnóstico es crítica, por tal motivo se evidenció la existencia en simultáneo de trabajos relacionados entre sí, buscando soluciones *rápidas* y *eficaces*. La sugerencia automatizada de un diagnóstico de COVID-19 aplicando RNC para el análisis de tomografías computadas de pecho, aporta una opción con muy buena precisión al profesional de la salud.

Los resultados obtenidos demuestran que el soporte a las unidades médicas a través de la aplicación de redes neuronales convolucionales para la detección temprana de los síntomas agilizan los diagnósticos permitiendo salvar vidas.

El diseño e implementación de arquitecturas de redes neuronales convolucionales en el tratamiento de imágenes para la detección de COVID-19 aportan mejoras en tiempos de entrenamiento de la RNC, precisión en los resultados y mejoras en la utilización de los datasets permitiendo conjuntos de datos pequeños (Menos de 80 imágenes) sin alterar la precisión. Por tal motivo limitamos la investigación al campo de la aplicación de las RNC.

Como trabajo futuro proponemos investigar otras técnicas de análisis de imágenes para la identificación de síntomas de COVID-19.

Referencias

[1] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "COVID-CAPS: A Capsule Network-based Framework for Identification of COVID-19 cases from X-ray Images," Apr. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2004.02696.

[2] "geographical-distribution-2019-ncov-cases." https://www.ecdc.europa.eu/en/geographical- distribution-2019-ncov-cases (accessed May 01, 2020).

[3] A. Narin, C. Kaya, and Z. Pamuk, "Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks," Mar. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2003.10849.

[4] O'Shea, K. & Nash, R. An Introduction to Convolutional Neural Networks. (2015).

[5] A. Iqbal Khan, J. Latief Shah, and M. Bhat, "CoroNet: A Deep Neural Network for Detection and Diagnosis of Covid-19 from Chest X-ray Images."

[6] M. Rahimzadeh and A. Attar, "A New Modified Deep Convolutional Neural Network for Detecting COVID-19 from X-ray Images," Apr. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2004.08052.

[7] A. Abbas, M. Abdelsamea, and M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," medRxiv, p. 2020.03.30.20047456, Apr. 2020, doi: 10.1101/2020.03.30.20047456.

[8] Q. Yan et al., "COVID-19 Chest CT Image Segmentation -- A Deep Convolutional Neural Network Solution," Apr. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2004.10987.

[9] N. S. Punn and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks," Apr. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2004.11676

[10] D. Singh, V. Kumar, Vaishali, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks.," European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology, Apr. 2020, doi: 10.1007/s10096-020-03901-z.

[11] L. Wang and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images," Mar. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2003.09871.

[12] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-Dataset: A CT Scan Dataset about COVID-19," Mar. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2003.13865.

[13] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images," Mar. 2020, Accessed: May 01, 2020. [Online]. Available: http://arxiv.org/abs/2003.11055.

Prototyping A Digital Zen Garden

Nicolás Jofré, Graciela Rodríguez, Yoselie Alvarado,
Jacqueline Fernandez, and Roberto Guerrero

Laboratorio de Computación Gráfica (LCG)
Universidad Nacional de San Luis,
Ejército de los Andes 950
Tel: 02664 420823, San Luis, Argentina
{npasinetti,gbrodriguez,ymalvarado,jmfer,rag}@unsl.edu.ar

Abstract. Since the start of the COVID-19 pandemic, the severity and prevalence of symptoms of psychological distress, fatigue, brain fog, and other conditions have increased considerably, including among people who have not been infected with SARS-CoV-2. Many studies summarize the effect of the pandemic on the availability of mental health services and how this has changed during the pandemic. Concerned that potential increases in mental health conditions, had already prompted 90% of countries surveyed to include mental health and psychosocial support in their post COVID-19 response plans, but major gaps and concerns remain. In this paper we developed a de-stress proposal through a digital zen garden by using an augmented reality sandbox. The system provides patients with flexible interaction and easy control of the scenario, while making real time data recording. An objective evaluation method is proposed to review the effectiveness of the therapy. According to the evaluation results of patients' training, the system is a low cost entertainment tool that augments patients' motivation, and helps to increase the effectiveness of therapy.

Keywords: COVID-19, Mental Health, Cognitive Disorders, Virtual Reality, Augmented Reality.

1 Introduction

In terms of pathophysiology, a closely related coronavirus (SARS-CoV) is reported to be neurotoxic and affect mental health. Furthermore, among the survivors of SARS infection, patients were reported to have persistent elevated stress, and over 64% of the survivors are reported to have a combination of stress, anxiety, and depression [1].

Only in the first year of the COVID-19 pandemic, global prevalence of anxiety and depression increased by a massive 25%, according to a scientific brief released by the World Health Organization (WHO). WHO Director-General, said that the information gathered about the impact of COVID-19 on the world's mental health is just the tip of an iceberg. As a consequence, this is a wake-up call to all countries to pay more attention to mental health and do a better job

of supporting their populations' mental health. One major explanation for the increase is the unprecedented stress caused by the social isolation resulting from the pandemic. Linked to this were constraints on people's ability to work, seek support from loved ones and engage in their communities. Loneliness, fear of infection, suffering and death for oneself and for loved ones, grief after bereavement and financial worries have also all been cited as stressors leading to anxiety and depression [2]. Among health workers, exhaustion has been a major trigger for suicidal thinking [3].

Some studies show that the pandemic has affected the mental health of young people and that they are disproportionally at risk of suicidal and self-harming behaviours. It also indicates that women have been more severely impacted than men and that people with pre-existing physical health conditions, such as asthma, cancer and heart disease, were more likely to develop symptoms of mental disorders [4].

Data suggests that people with pre-existing mental disorders do not appear to be disproportionately vulnerable to COVID-19 infection. Yet, when these people do become infected, they are more likely to suffer hospitalization, severe illness and death compared with people without mental disorders. People with more severe mental disorders, such as psychoses, and young people with mental disorders, are particularly at risk [5].

While the pandemic has generated interest in and concern for mental health, it has also revealed historical under-investment in mental health services. Countries must act urgently to ensure that mental health support is available to all. In this sense, there is an urgent need for tools to address diseases such as stress, anxiety, among others.

The aim of this paper is to propose a de-stress therapeutic tool by approximating the ancient method of Japanese Zen Garden through Augmented Reality (AR).

Section 2 gives a brief overview of Japanese Zen Garden. Section 3 describes the proposed Augmented Reality Sandbox Architecture System. Section 4 details the evaluation method to review the effectiveness of the therapy. Section 5 provides a small discussion and future guidelines.

2 Japanese Zen Garden

Japanese gardens create their own styles and one of the most famous are the so-called Zen gardens that seek to go beyond and create a place conducive to meditation and contemplation (See Figure 1).

The japanese zen garden creates a miniature stylized landscape through carefully composed arrangements of rocks, water features, moss, pruned trees and bushes, and uses gravel or sand that is raked to represent ripples in water. Zen gardens are commonly found at temples or monasteries. A zen garden is usually relatively small, surrounded by a wall or buildings, and is usually meant to be seen while seated from a single viewpoint outside the garden, such as the porch of the hojo, the residence of the chief monk of the temple or monastery. Many,

with gravel rather than grass, are only stepped into for maintenance. Classical zen gardens were created at temples of Zen Buddhism in Kyoto during the Muromachi period. They were intended to imitate the essence of nature, not its actual appearance, and to serve as an aid for meditation [6].

It is important to understand that the word zen means meditation. Monks used zen garden as an ideal place for meditation. They are areas that transmit tranquility, inner serenity and reduce stress through their beauty [7].

Sand represents the vastness of the ocean and rocks represent the mountains. One of the many benefits of zen gardens is to de-stress their owners by playing with the rake, creating shapes in the sand. We can give movement to our garden, for example by creating designs with "stacked stones" which signify stability.

Zen gardens bring us serenity and relaxation; they stimulate creativity and the best thing is that we do not need a large space to create one, we can assemble them in any corner of our home or office.

Fig. 1: Japanese Zen Garden.

3 Augmented Reality Sandbox

When referring to an augmented reality environment it talks about any real-world environment with elements augmented or supplemented by computer-generated input [8].

Since their conception in 2012 [9, 10], the AR sandbox system is used to teach geographical, geological, and hydrological concepts such as how to read topographic maps, the meaning of contour lines, watersheds, catchment areas, levees, etc. It is a tool that combines 3-dimensional visualization applications with a hands-on sandbox. Users can create topography models by shaping real sand, which is then augmented in real time by an elevation color map, topographic contour lines, and simulated water flow.

However, in many cases AR sandbox can be seen as a form of non-verbal therapeutic intervention [11–13]. Patients (often, children) can use sand to portray their experiences that they cannot express verbally. Moreover, many psychologists leverage this tool to treat psychic disorders due to its proven efficiency.

Given this background it would be interesting to evaluate how an AR sandbox can contribute to reducing stress resulting from everyday life problems and the well-known Covid-19 pandemic.

3.1 System Description

As mentioned in section 2, sand represents the vastness of the ocean and rocks represent the mountains. One of the many benefits of zen gardens is to de-stress their owners by playing with the rake, creating shapes in the sand. In relation to this, the sand in our sandbox is initially flattened and unrelieved, and the user has a set of tools such as trowel, rake and ruler. Using these tools, the user can perform different actions such as:

- to create indentations or mountains by the trowel,
- to create ripples that can represent water by the rake, and clean up and redraw them by a ruler,
- to clean and redraw the relief until the desired relief is achieved by a ruler.

It is expected that the user can visualize an analyze the projected textures in each sandbox section so they can determine and materialize the landscape appearance by means of the tools based on their observation.

3.2 Architecture System

Our AR sandbox prototype system comprises the following hardware components:

- A computer with a high-end graphics card, running Linux.
- A Microsoft Kinect 3D camera.
- A digital video short-throw projector with a digital video interface, such as HDMI or DVI.
- A mirror.
- A sandbox with a way to mount the Kinect camera, the mirror and the projector above the sandbox.
- Sand.

The architecture was designed for a medium sandbox of 120x100cm. These measurements determined that the depth sensor is suspended 2.2m above the sandbox. Having a regular projector, the ideal projection aperture is projecting to a screen located at a distance of more than 3 meters, that is, the projector position has to be higher than the position of the depth sensor. This feature gives a resulting sensor shadow on the sandbox. The sensor-projector height problem

can be solved by using a conveniently placed mirror to simplify the structure that supports all the devices. The mirror size results in 60x60cm.

The developed prototype used for this work is shown below (Figure 2). Figure 3 shows in detail the resulting visualized sand landscape.

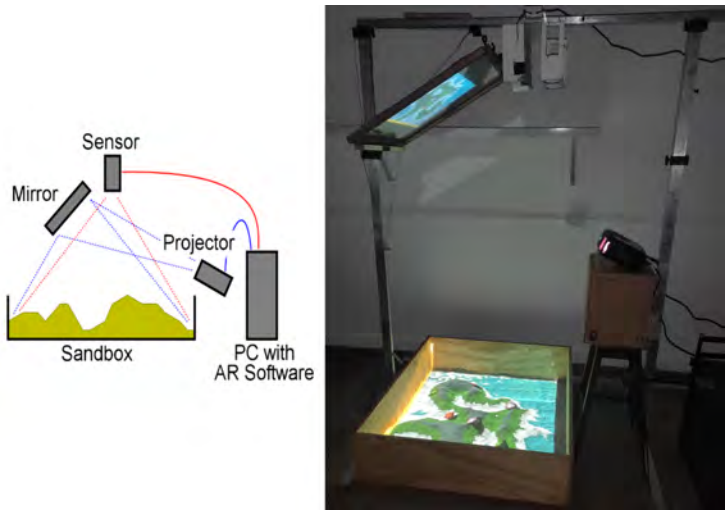


Fig. 2: Sandbox architecture system.

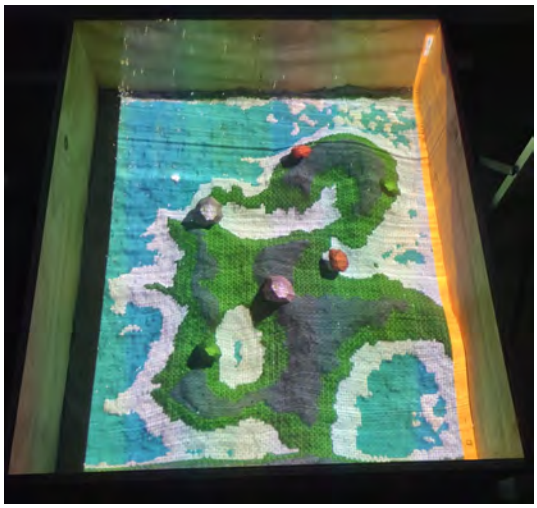


Fig. 3: Sandbox projected.

4 Experiences and Results

Representative stress assessment methods include psychological evaluation methods (questionnaire, etc.), biochemical evaluation method (blood test, etc.), and physiological evaluation methods (heart rate variability, etc.) [14, 15].

The present research carried out an experimental methodology using a group of participants. The user experience was developed with a group of 20 participants between 12 and 60 years old, all of them from San Luis, Argentina.

We performed an objective evaluation method to review the effectiveness of the therapy training. The training consisted of measuring the level of participants' stress before and after the experience with our sandbox prototype through 5-20 minutes user free actions. Actions were detailed in section 3.1.

Stress level monitoring enables to check how much stress the user has. For experimental data acquisition, a smart band was used [16]. The smart band measures the user's heart rate and determines the current stress level based on its variability according to Jachymek et al. work [17]. Ranges are suggested by the smart band application as shown in Table 1 [16].

Levels of stress			
Relaxed	Mild	Moderate	High
0-39	40-59	60-79	80-100

Table 1: Stress level description

Table 2 shows the stress level measured at two different time slices: before and after the sandbox experience (second, third, fifth and sixth columns from left to right).

# User	A Priori Level	A Posteriori Level	# User	A Priori Level	A Posteriori Level
1	35	24	11	43	26
2	29	21	12	47	29
3	34	25	13	44	35
4	29	23	14	42	31
5	40	31	15	30	24
6	64	34	16	41	28
7	38	28	17	35	27
8	27	22	18	44	35
9	38	28	19	49	38
10	59	33	20	42	28

Table 2: Stress level of participants before and after to the experience

5 Discussion and Conclusions

In this paper we developed a therapy system by using the augmented reality technique, a Kinect 3D camera and a sandbox. The system mimics a Japanese Zen Garden de-stress process by using a digital strategy through an augmented reality sandbox. The system provides users with flexible interaction and easy sand control, and also presents real time data recording.

From the objective evaluation method, results indicate that the level of stress after the experience decreases by 20 to 30 percent according to the measurements of the smartwatch used. Gathered data show that the system enables to provide an experience that reduces the participants' stress.

More over, the experiment provides a new insight into the relationship between these types of technologies and traditional mindfullness methods.

The experiences here considered gave users free will at the moment of actions' choice. Robust experiments should consider an actions' protocol to be followed by users. Additionally, real time digital sandbox elements interaction is beyond the scope of this work.

Future studies will take into account the development of a virtual reality Japanese Zen Garden including both headsets and gestural sensing devices. The system should immerse the user inside the virtual garden and give him the possibility to affect the scene through his avatar in the same way as it is done in a real zen garden.

References

1. Arehally M. Mahalakshmi, Bipul Ray, Sunanda Tuladhar, Abid Bhat, Shasthara Paneyala, Duraisamy Patteswari, Meena Kishore Sakharkar, Hamdan Hamdan, David M. Ojcius, Srinivasa Rao Bolla, Musthafa Mohamed Essa, Saravana Babu Chidambaram, and M. Walid Qoronfleh. Does covid-19 contribute to development of neurological disease? *Immunity, Inflammation and Disease*, 9(1):48–58, 2021.
2. Elena Dragioti, Han Li, George Tsitsas, Keum Hwa Lee, Jiwoo Choi, Jiwon Kim, Young Jo Choi, Konstantinos Tsamakis, Andrés Estradé, Agorastos Agorastos, et al. A large-scale meta-analytic atlas of mental health problems prevalence during the covid-19 early pandemic. *Journal of Medical Virology*, 94(5):1935–1949, 2022.
3. G Johns, V Samuel, L Freemantle, J Lewis, and L Waddington. The global prevalence of depression and anxiety among doctors during the covid-19 pandemic: Systematic review and meta-analysis. *Journal of affective disorders*, 298:431–441, 2022.
4. Pınar IRMAK VURAL, Nazife BAKIR, Cuma Demir, and Pınar IRMAK VURAL. The effect of religious coping on geriatric anxiety in a group of older turkish women during the covid-19 pandemic period. *Turkish Journal of Geriatrics*, 25(2):282–290, 2022.
5. Lamiece Hassan, Niels Peek, Karina Lovell, Andre F Carvalho, Marco Solmi, Brendon Stubbs, and Joseph Firth. Disparities in covid-19 infection, hospitalisation and death in people with schizophrenia, bipolar disorder, and major depressive disorder: a cohort study of the uk biobank. *Molecular psychiatry*, 27(2):1248–1255, 2022.

6. M. Locher and T. Fujimori. *Zen Garden Design: Mindful Spaces by Shunmyo Masuno Japan's Leading Garden Designer*. Tuttle Publishing, 2020.
7. M. Locher and U. Shigeru. *Zen Gardens: The Complete Works of Shunmyo Masuno, Japan's Leading Garden Designer*. Tuttle Publishing, 2012.
8. Ronald T. Azuma. A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, 08 1997.
9. https://eos.org/science-updates/augmented-reality-turns-a-sandbox-into-a-geoscience-lesson.
10. O. Kreylos, L. H. Kellogg, S. Reed, S. Hsi, M. B. Yikilmaz, G. Schladow, H. Segale, and L. Chan. The AR Sandbox: Augmented Reality in Geoscience Education. In *AGU Fall Meeting Abstracts*, volume 2016, pages ED51H–0843, December 2016.
11. Mareike Gabele, Simon Schröer, Steffi Husslein, and Christian Hansen. An ar sandbox as a collaborative multiplayer rehabilitation tool for children with adhd. In *Mensch und Computer 2019 - Workshopband*, Bonn, 2019. Gesellschaft für Informatik e.V.
12. Philip Lindner, William Hamilton, Alexander Miloff, and Per Carlbring. How to treat depression with low-intensity virtual reality interventions: Perspectives on translating cognitive behavioral techniques into the virtual reality modality and how to make anti-depressive use of virtual reality–unique experiences. *Frontiers in Psychiatry*, 10, 2019.
13. https://ar-sandbox.com/augmented-reality-sandbox-for-therapy/.
14. Anjana Bali and Amteshwar Jaggi. Clinical experimental stress studies: Methods and assessment. *Reviews in the neurosciences*, 0, 05 2015.
15. Rateb Katmah, Fares Al-Shargie, Usman Tariq, Fabio Babiloni, Fadwa Al-Mughairbi, and Hasan Al-Nashash. A review on mental stress assessment methods using eeg signals. *Sensors*, 21(15), 2021.
16. Xiaomi Inc. Zepp life. Google Play Store, 2022.
17. Magdalena Jachymek, Michał T. Jachymek, Radosław M. Kiedrowicz, Jarosław Kaźmierczak, Edyta Płońska-Gościniak, and Małgorzata Peregud-Pogorzelska. Wristbands in home-based rehabilitation - validation of heart rate measurement. *Sensors*, 22(1), 2022.

XIX Workshop Ingeniería de Software

(WIS)

Coordinadores

Patricia Pesado (UNLP)

Elsa Estevez (UNS)

Alejandra Cechich (UNCOMA)

TRACEM - Towards a Standard Metamodel for Execution Traces in Model-Driven Reverse Engineering

Claudia Pereira[1], Liliana Martinez[1], Liliana Favre[1,2]

[1] Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina
[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina
{cpereira, lmartine, lfavre}@exa.unicen.edu.ar

Abstract. Reverse engineering is a crucial stage in the software modernization process. The current techniques available in existing CASE tools provide forward engineering and limited facilities for reverse engineering, dynamic analysis in particular. The Architecture-Driven Modernization initiative has defined standards to support the modernization process in the model-driven engineering (MDE) context. Standardization increases interoperability between different tools enabling a new generation of solutions to benefit the whole industry and encourage collaboration among complementary vendors. In this paper, we present TRACEM, a metamodel to represent trace information under a standard representation. This metamodel complements a MDE framework for software modernization that aims to integrate static and dynamic analysis techniques during the reverse engineering process. This paper includes a case study that exemplifies how dynamic information combined with static information allows improving the whole reverse engineering process.

Keywords: Architecture-Driven Modernization, Metamodeling, Transformation, Static analysis, Dynamic analysis, Legacy System, Reverse Engineering

1 Introduction

Reverse engineering techniques allow supporting an integral part of software modernization, specifically, the process of analyzing available software artifacts in order to extract information and provide high-level views on the underlying system. Nowadays, many companies are facing the problem of having to modernize or replace their legacy software systems which have involved the investment of money, time and other resources through the ages. Many of them are still business-critical and there is a high risk in replacing them. The growing demand for modernization of software is due to the great advance in mobile technologies and the emergence of the paradigms of Cloud Computing, Pervasive Computing and the Internet of Things. Regarding the systematic modernization process, novel technical frameworks for information integration, tool interoperability and reuse have emerged. Specifically, Model-Driven Engineering (MDE) is a software engineering discipline which emphasizes the use of models and model transformations to raise the abstraction level and the automation degree in software development. Productivity and some aspects of software quality such as maintainability or interoperability are goals of MDE [1].

In the MDE context, the most recent OMG contributions to modernization are in line with the Architecture-Driven Modernization (ADM) proposal. It is defined as "the process of understanding and evolving existing software assets for the purpose of software improvement, modifications, interoperability, refactoring, restructuring, reuse, porting, migration, translation, integration, service-oriented architecture deployment" [2]. The OMG ADM Task Force is developing a set of standards (metamodels) to facilitate interoperability between modernization tools, such as KDM (Knowledge Discovery Metamodel) [3] and ASTM (Abstract Syntax Tree Metamodel) [4]. ADM has emerged complementing OMG standards such as MDA [5], which manages the software evolution from abstract models to implementations. The essence of MDA is the Meta Object Facility Metamodel (MOF) [6] which allows different kinds of artifacts from multiple technologies to be used together in an interoperable way. Metamodeling is an essential technique in MDA and its benefits are well known. The precise standard language definition that is processable by machines may be used to check if models are valid instances. On the other hand, a metamodel defined with the core of UML [7] class diagrams is an accessible language, easy to understand and maintain, therefore it contributes to an easy adaptation allowing language evolution. Based on the level of the meta-metamodel, tools that allow exchanging formats may be developed to manipulate models, regardless of the modeling language used [1].

OMG standards related to ADM allow obtaining models from code that represent static information. Despite the increasing attention to dynamic analysis techniques in reverse engineering, there is no standard for representing information at runtime. A standard for this purpose could be used by tools for visualization and analysis of execution traces, which would facilitate interoperability and data exchange. In previous works, we have shown how to reverse engineering models from code through static analysis, including class, use cases, behavioral and state diagrams [8][9][10]. In this paper, we present TRACEM, a trace metamodel that is the foundation for dynamic analysis in the ADM context. This metamodel allows representing the trace information under a standard representation that supports extensibility, interoperability, abstraction and expressiveness. Moreover, the proposed metamodel along with the specific ASTM aim at automatic instrumentation of the source code. An execution trace model is obtained each time the program runs. Then, by running the program with a significant set of test cases, we obtain a set of trace models that will be analyzed to obtain relevant dynamic information. The ultimate goal is to integrate dynamic and static analysis techniques combining the strengths of both approaches in the reverse engineering process within an MDE framework.

This paper is organized as follows. Section 2 describes a framework for reverse engineering in the MDE context. Section 3 details the TRACEM metamodel. In Section 4, we analyze the impact of dynamic analysis through an example. Section 5 discusses related work. Finally, Section 6 presents conclusions and future work.

2 Reverse Engineering into the MDE Framework

The combination of static and dynamic analysis can enrich the reverse engineering process. Ernst [11] provides a comparison of static and dynamic analysis from the point

of view of their synergy and duality. He argues that static analysis is conservative and sound. Conservatism means reporting weak properties that are guaranteed to be true, preserving soundness, but not strong enough to be useful. Soundness guarantees that static analysis provides an accurate description of behavior, no matter on what input or in what execution environment the program is run. Dynamic analysis is precise given that it examines the actual runtime behavior of the program, however the results of executions may not generalize to other executions. Also, Ernst argues that whereas the main challenge of static analysis is choosing a good abstract interpretation, the main challenge of performing good dynamic analysis is selecting a representative set of test cases. Static or dynamic analyses can enhance one another by providing information that would otherwise be unavailable.

We propose a framework to reverse engineering models that blends the strengths of static and dynamic analysis (Fig. 1). This framework is based on the MDE principles: all artifacts involved can be viewed as models and the process can be viewed as a sequence of model-to-model transformations where the extracted information is represented in a standard way. Each model can be reused, refactored, modified or extended for reverse engineering purposes or for other purposes. Metamodels are defined via MOF and the transformations are specified between source and target metamodels. Then, MOF metamodels "control" the consistency of these transformations.

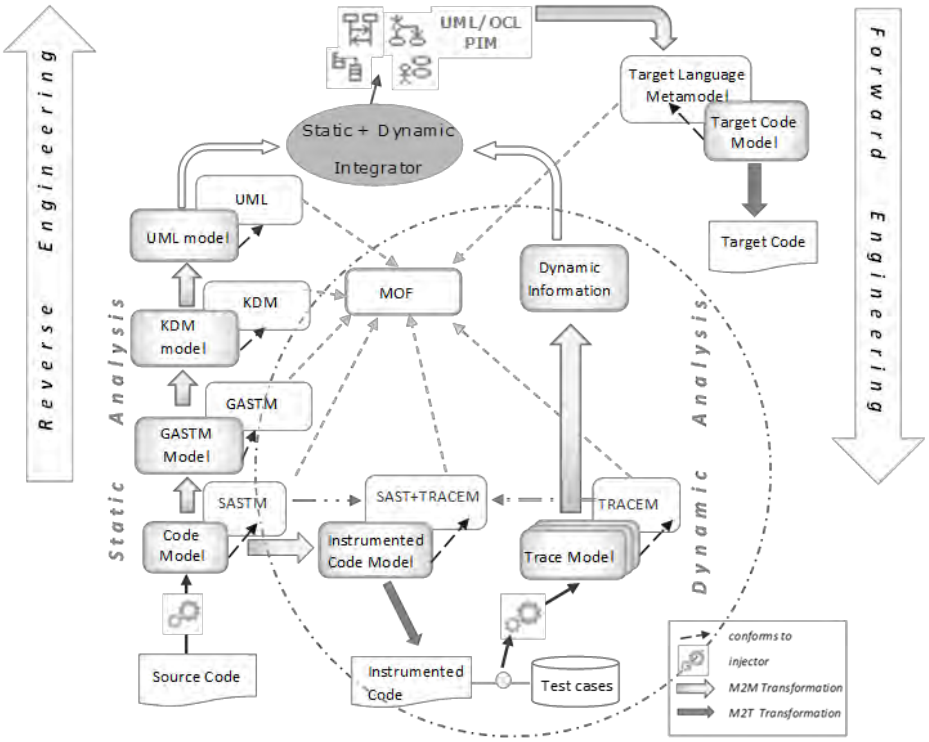


Fig. 1. MDE Modernization Process

In previous works, we present a process to reverse engineering models from code through static analysis, including class diagrams, use cases diagrams, behavioral diagrams and state diagrams [8][9][10]. In the framework, as shown in Figure 1, the first step of the static analysis is to obtain the code model, an abstract syntax tree model instance of the SASTM (Specific ASTM) by using a model injector. Next, an instance of the GASTM (Generic ASTM) is generated from the previous model by a model-to-model transformation. Finally, high-level UML models are obtained by means of a chain of model-to-model transformations, using a KDM model as an intermediate representation of the software system. In the first step of the process, an injector and transformations to obtain the GASTM model must be implemented for each programming language, whereas the sequence of transformations involved in the following steps is independent of the legacy code language .

In this paper we present TRACEM, a trace metamodel that is the foundation for dynamic analysis (see dotted circle in Fig.1). This metamodel allows us to obtain and record trace information. Dynamic analysis provides information about the runtime behavior of software systems, thus, it is a valuable tool for reverse engineering. However, dynamic analysis requires the availability of a full, executable system, which is run with some predefined input data and, on the other hand, it requires the code instrumentation to detect and record relevant events during runtime for later off-line analysis. To reverse engineering models from code, the first stage is to record trace data such as a set of objects, a set of attributes for each object, a location and type for each object, a set of messages, and time stamp for each event. This dynamic information is obtained by instrumenting the source code, a process that inserts additional code fragments into the source code under analysis. An execution trace model, instance of TRACEM, is obtained each time the program runs. Then, by running the program with a significant set of test cases, we obtain a set of trace models. These models will be analyzed to obtain relevant dynamic information that combined with static information allows improving the reverse engineering process. Then, the resulting models are the starting point for the forward engineering process.

3 TRACEM Metamodel

TRACEM allows specifying the trace information under a standard representation supporting extensibility and interoperability. TRACEM was implemented in the Eclipse Modeling Framework [12] that is the core technology in Eclipse for MDE. Figures 2 and 3 partially show this metamodel. The abstract syntax of TRACEM is described by UML class diagram (Fig. 2) augmented with OCL restrictions [13] (Fig. 3). Although the figures show a part of the metamodel, specifically the part focused on the representation of interactions between objects in terms of method calls, it can be extended from the abstract metaclass *Trace* to represent other types of relationships such as inter-process and system-level relationships. The main metaclasses are:

- *ExecutionTrace,* subclass of Trace metaclass, represents a particular execution of a program on a specific test case. Each instance has a name, start and end time, and owns objects and a sequence of statements discovered during the program execution.

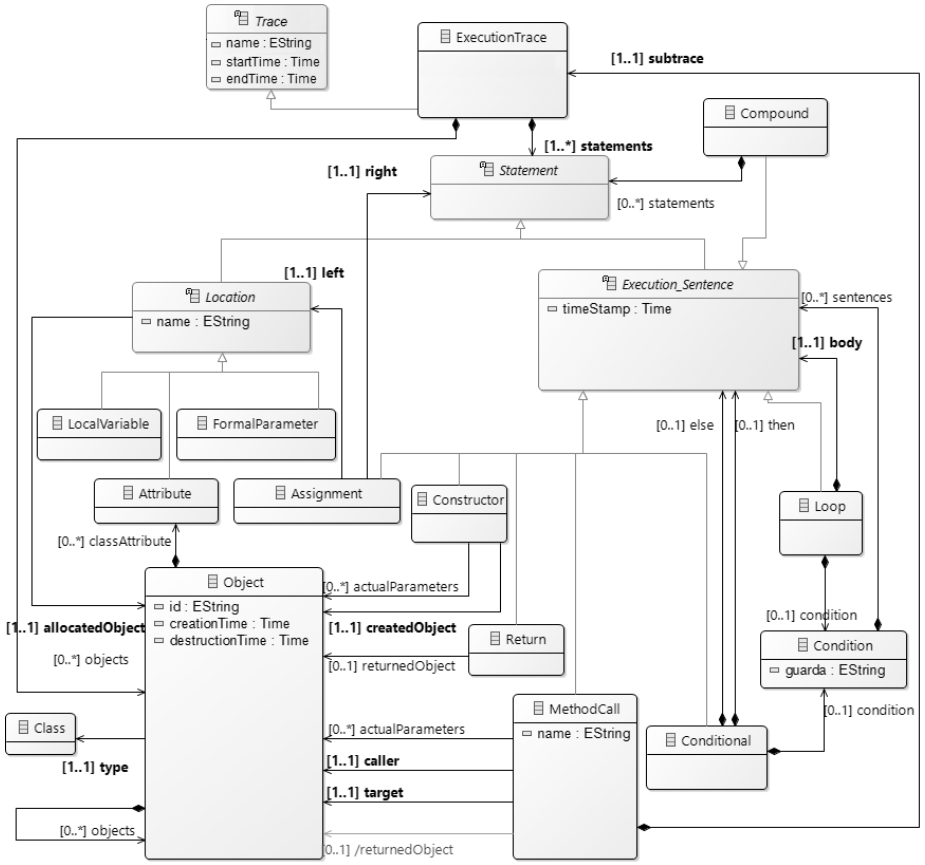


Fig. 2. TRACEM metamodel: Abstract syntax

-- the returned object of a method call corresponds to the object returned by its return sentence
context MethodCall::returnedObject:Object
derived: returnedObject = subtrace.statements->collect (s| s.oclIsTypeOf(return).returnedObject)

context Assignment -- *restrictions on the right and left parts of an assignment*
inv: right.OclIsKindOf(location) **or** right.OclIsKindOf(MethodCall) **or**
right.OclIsKindOf(Constructor) **and** left.allocatedObject =
if right.oclIsTypeOf(Location) **then** right.allocatedObject
else if right.oclIsTypeOf(MethodCall) **then** right.returnedObject
else if right.oclIsTypeOf(Constructor) **then** right.createdObject **endif endif endif**

context Compound -- *compound only has local variables as locations*
inv: statements->select(s | s.oclIsKindOf(Location))-> forAll (l| l.oclIsTypeOf (localVariable))

context MethodCall -- *relationship between formal and actual parameter*
inv: actualParameters->forAll (ap| self.subtrace.statements->
collect(oclIsTypeOf(FormalParameter)) ->exists(fp| fp.allocatedObjet = ap)

Fig. 3. TRACEM metamodel: OCL restrictions

- *Location* is an abstract metaclass that represents a storage that holds an object. Program locations are either local variables, class attributes or method parameters. A *Location* instance has a name and an allocated object which may be changed during program execution.
- *ExecutionSentence* is an abstraction which specifies instructions carried out during the program execution such as *MethodCall*, *Assignment* and *Constructor*.
- *Object* represents objects created during the program execution. An instance has an identifier, a creation and destruction time and owns attributes and objects. It can be stored in different locations throughout the program execution.

4 Recovering Execution Traces from Code: an Example

Dynamic analysis is exemplified in terms of the same case study used in Tonella and Potrich [14], the Java program *eLib* that supports the main library functions (Fig. 4). It contains an archive of documents of different kinds, books, journals, and technical reports. Each of them has specific functionality. Each document can be uniquely identified and library users can request documents for loan. To borrow a document, both user and document must be verified by the Library. As regards the loan management, users can borrow documents up to a maximum number; while books are available for loan to any user, journals can be borrowed only by internal users, and technical reports can be consulted but not borrowed.

```
class Library {
  Map documents = new HashMap();
  Map users = new HashMap();
  Collection loans = new LinkedList();
  ...
  private boolean verifyData (User u, Document d)
  {  if (u == null || d == null) return false;
     if (u.numberOfLoans() <
        MAX_NUMBER_OF_LOANS &&
        d.isAvailable() &&  d.authorizedLoan(u))
                  return true;
     return false;
  }
  public boolean borrowDocument
              (User user, Document doc) {
  if (verifyData (user, doc) {
        Loan loan = new Loan(user, doc);
        addLoan(loan);
        return true;        }
  return false;
  }
  public int numberOfLoans() {
    return loans.size();      }

  private void addLoan(Loan loan) {
    if (loan == null) return;
    User user = loan.getUser();
    Document doc = loan.getDocument();
    loans.add(loan);
    user.addLoan(loan);
    doc.addLoan(loan);   }
...} // end class Library

class Document {...
  public boolean isAvailable() {return loan==null;}
  public boolean authorizedLoan(User user) {
      return true;        }
} // end class Document

class Book extends Document {...}
class InternalReport extends Document {...}
class User {... }
class InternalUser extends User {}
class Loan {
   User user;  Document document;
   public Loan(User usr, Document doc) {
       user = usr;     document = doc;
} // end class Loan
```

Fig. 4. Source code of the *eLib* Program

Tonella and Potrich describe a reverse engineering approach at model level of object-oriented code based on classical compiler techniques and abstract interpretation to obtain UML diagrams from Java code, particularly class, object, interaction, state and

package diagrams. This case study was used in previous works to show the extensions proposed with respect to the approach of Tonella and Potrich [8][9][10]. To highlight the contributions of dynamic analysis, we use the same example.

Dynamic analysis produces a set of execution trace instances, one for each test case. Fig. 5 partially shows an instance of trace metamodel obtained from the execution of the method borrowDocument resulting in a successful loan of the book1 (instance of Book) to the internalUser1 (instance of InternalUser). Each time an object is created, it is identified by the class name concatenated with a numeric value.

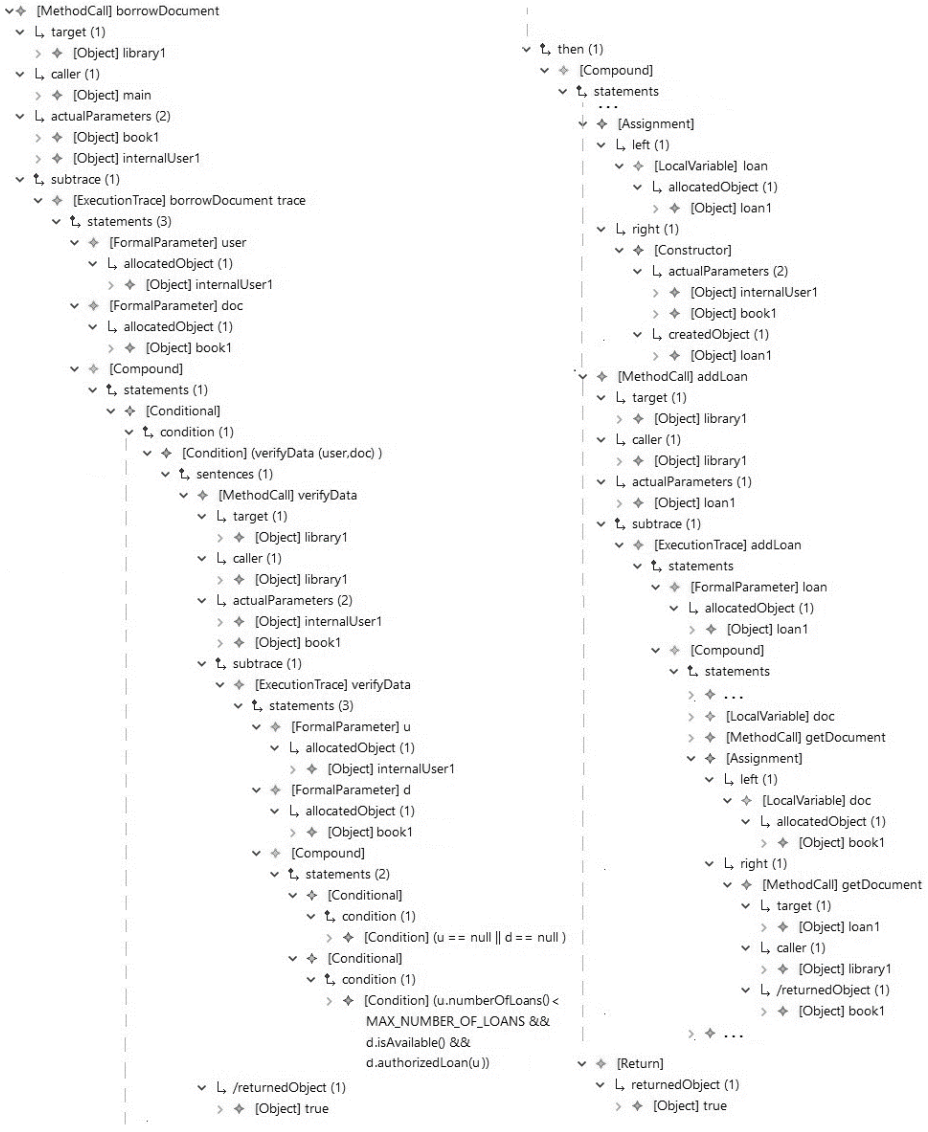


Fig. 5. Trace model: A successful loan

4.1. Dynamic Information Impact

The execution traces provide information that allows complementing the models obtained through static analysis in the aforementioned previous works.

As regards the UML behavioral diagram, it is possible to identify:

- the current object that invokes the method (*caller*) and the one that receives the message (*target*).
- the current parameter linked to each formal parameter, that is, which object is actually stored in each formal parameter for a particular trace. As an example, the objects allocated in the formal parameters *user* and *doc* of the borrowDocument method are the objects *internalUser1* and *book1* respectively (Fig. 5).
- the object flows, that is to say, how an object is passed from one location to another, starting from where it is created. As an example, it is possible to realize that the object *book1* allocated in the formal parameter "*doc*" of the borrowDocument method, is the same object as the one allocated in the formal parameter "*d"* of the verifyData method, the actual parameter "*doc"* of the constructor that creates a new Loan object called loan1, the class attribute "*doc"* of the loan1 object, the local variable "*doc"* in the addLoan method that receives the message addLoan (Fig. 5).
- the kind of dependence relationship between use cases, *include* or *extend*. The common traces reflect primary flow and allow detecting possible *include* relationships between use cases whereas other traces may correspond to *extend* relationships.

As regards the UML structural diagram, it is possible to identify:

- the current objects stored in the generic collections. Containers with weak types (parameterized in abstract types or interfaces) complicate the reverse engineering process. Relationships between classes, such as associations and dependencies, are determined from the declared type for attributes, local variables, and parameters. When containers are involved, the relations to retrieve must connect the given class to the classes of the contained objects. If an attribute type is a generic container, the relationship connects the given class to the class of the contained object, however this information is not directly available in the source code, as a result, the relationship is not depicted in a UML class diagram. Identifying the type of objects that a collection actually stores allows obtaining more complete and accurate class diagrams. As an example, from the trace models, the generic collection *loans* will only contain objects of Loan type, thus, an association between Library and Loan will be inferred.
- composition relationships by analyzing the lifetime of the referenced objects since the metamodel allows recording the creation and destruction time of each object. Within composition, the lifetime of the part is managed by the whole, in other words, when the whole is destroyed, the part is destroyed along with it. As an example, by analyzing the creation and destruction times of the library1 object and the objects of type Loan added to the loans collection of Library, it is possible to infer that the association between Library and Loan is indeed a composition.

Moreover, the execution traces provide information that allows detecting functionality that may never be executed.

5 Related work

Many works have contributed to reverse engineering object-oriented code, dynamic analysis techniques in particular. [14] and [15] perform dynamic analysis to complement the static analysis from java code. Trace information obtained from the program execution is represented with UML models. In the MDE context, [16] presents the first steps towards extending MoDisco with capabilities for dynamic program analysis. MoDisco injects the program structure into a model [17], the authors propose to add execution trace information to the model during program execution. Unlike these works, we propose to represent traces as a new domain in software engineering, independent of any language and providing more expressiveness than those approaches that use UML models to represent the dynamic analysis results.

Following, some works that propose the creation of a standard to represent execution traces are presented. [18] presents a metamodel for representing trace information of routine calls with the aim to develop a standard format for exchanging traces among trace analysis tools. [19] and [20] describe model driven approaches in specific domains that involve dynamic analysis. The former focuses on reverse engineering of AUTOSAR-compliant models using dynamic analysis from trace recordings of a real-time system in the automotive domain. [20] proposes a common metamodel for representing High Performance Computing system traces. Unlike these related works, we propose a MOF-compliant trace metamodel to represent execution traces. This metamodel is the foundation for dynamic analysis within a framework in the MDE context, based on ADM standards in particular.

6 Conclusions

This paper describes the basis for dynamic analysis in the reverse engineering process integrating static analysis, dynamic analysis, and metamodeling in the ADM context. The main contribution is the TRACEM metamodel, which describes concepts and relationships existing in the information obtained from program execution. It allows specifying the execution trace information under a standard representation. Thus, the traces are considered first-class entities, which provide relevant dynamic information that combined with static information allows improving the whole reverse engineering process.

TRACEM together with the metamodels of the different programming languages will allow the automatic instrumentation of code and from this, the injection of trace models that act as decoupling from source technologies. However, there are no available injectors or metamodels for different programming languages and it is necessary to implement them.

We foresee experimenting with different programming languages to implement injector prototypes. Furthermore, we will investigate analysis techniques of execution traces to understand and manipulate the models obtained from program executions.

References

1. Brambilla, M., Cabot, J., & Wimmer, M.: Model-Driven Software Engineering in Practice. Morgan & Claypool Publishers, Second edition (2017)
2. ADM Architecture-Driven Modernization. http://www.omg.org/adm
3. KDM ADM: Knowledge Discovery Meta-Model Version 1.4 OMG Document Number: formal/2016-09-01. http://www.omg.org/spec/KDM/1.4 (2016)
4. ASTM Abstract Syntax Tree Metamodel Version 1.0 OMG Document Number: formal/2011-01-05. Standard document URL: http://www.omg.org/spec/ASTM (2011)
5. The Model-Driven Architecture (MDA). http://www.omg.org/mda/ UML OMG Unified Modeling Language. Version 2.5.1, OMG Document Number: formal/2017-12-05. http://www.omg.org/spec/UML/2.5.1/ (2017)
6. MOF OMG Meta Object Facility (MOF) Core Specification. Version 2.5.1, OMG Document Number: formal/2019-10-01. https://www.omg.org/spec/MOF/2.5.1 (2019)
7. UML OMG Unified Modeling Language. Version 2.5.1, OMG Document Number: formal/2017-12-05. http://www.omg.org/spec/UML/2.5.1/ (2017)
8. Favre, L., Martinez, L. & Pereira, C.: Reverse Engineering of Object-Oriented Code: An ADM Approach. In: Handbook of Research on Innovations in Systems and Software Engineering, pp. 386-410. IGI Global (2015)
9. Martinez, L., Pereira, C. & Favre, L.: Recovering Sequence Diagrams from Object-oriented Code - An ADM Approach. Proc. of the 9th International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE 2014, pp. 188-195 (2014)
10. Pereira, C., Martinez, L., & Favre, L.: Recovering Use Case Diagrams from Object-Oriented Code: an MDA-based Approach. International Journal of Software Engineering (IJSE), vol. 5 (2) (2012)
11. Ernst, M.: Static and Dynamic Analysis: Synergy and duality. Proceedings of ICSE Workshop on Dynamic Analysis. (WODA 2003), pp. 24-27 (2003)
12. EMF EMF. Eclipse Modeling Framework. http://www.eclipse.org/modeling/emf/
13. OCL Object Constraint Language Version 2.4, OMG Document Number: formal/2014-02-03, Standard document URL: http://www.omg.org/spec/OCL/2.4 (2014)
14. Tonella, P., & Potrich, A.: Reverse Engineering of Object-Oriented Code. Monographs in Computer Science. Heidelberg: Springer-Verlag (2005)
15. Systa, T.: Static and Dynamic Reverse Engineering Techniques for Java Software Systems. Ph.D Thesis, University of Tampere, Report A-2000-4 (2000)
16. Béziers la Fosse, T., Tisi, M., & Mottu, JM.: Injecting Execution Traces into a Model-Driven Framework for Program Analysis. In: Software Technologies: Applications and Foundations. STAF 2017. LNCS, vol 10748. Springer, Cham (2018)
17. MoDisco Eclipse MoDisco project.https://www.eclipse.org/MoDisco/
18. Hamou-Lhadj, A., & Lethbridge, T.C.: A metamodel for the compact but lossless exchange of execution traces. Softw Syst Model 11, pp. 77-98 (2012)
19. Sailer, A.: Reverse Engineering of Real-Time System Models from Event Trace Recordings. University of Bamberg Press (2019)
20. Alawneh L., Hamou-Lhadj A. & Hassine J.: "Towards a common metamodel for traces of high performance computing systems to enable software analysis tasks," IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER), 2015, pp. 111-120 (2015)

A flexible and expressive formalism to specify Metamorphic Properties for BIG DATA systems validation

Fernando Asteasuain[1,2]

[1] Universidad Nacional de Avellaneda, Argentina
fasteasuain@undav.edu.ar
[2] Universidad Abierta Interamericana - Centro de Altos Estudios CAETI, Argentina

Abstract. BIG DATA systems represent a huge challenge for software engineering validations tasks since they have been classified as "non testable". Metamorphic Relationships (MR) have been proposed as a technique to overcome this problem. These relationships establish interactions between data that can be used to validate the expected behavior of the system. However, the process of exploring and defining MRs is a very arduous one, and an expressive and flexible specification language is needed to denote them. In this work we show how the Feather Weight Visual Scenarios (FVS) framework can be seen as an appealing tool to specify MRs. We exploit FVS features to model complex MR interactions and analysis, allowing the possibility to perform non trivial operations between MRs such as refinement and consistency checking. FVS is shown in action by introducing a proof of concept example focused on a machine learning system over biology cell images.

Keywords: Formal Verification, BIG DATA, Metamorphic Testing

1 Introduction

The term Software Engineering was coined in 1968 during the so-called "Software Crisis". It was born as a response to crucial aspects that were threatening the computing community such as the repetitive failures and delays of software projects. In a few words, Software Engineering constitutes a toolbox of methods, techniques and processes to build and develop quality software. Since its creation, this Software Engineering's toolbox evolved to cope with different paradigms and challenges that arose in the computing field.

Undoubtedly, one of the most relevant domain nowadays relates to BIG DATA, machine learning and data science systems. These kinds of systems feature distinctive characteristics that urge Software Engineering to evolve in order to guarantee the quality of the developed systems [5, 13, 7, 18, 12, 20, 14]. Some of these characteristics are new software architectures, new protocols of communications, new software interactions, stronger performance and availability concerns and volatile, unstructured, diverse and heterogeneous data, just to name

a few. In this Software Engineering evolution the phase of formal verification and validation is probably the one that needs more attention and contributions. According to [13, 9], only two of nearly one hundred analyzed approaches addressing new software engineering methods for big data were related to formal validation. For example, work in [5, 7] introduces a parallel and distributed tool to perform model checking in big data systems. In this sense, one of the most pinpointed items to be addressed for formal validation in BIG DATA is that these kinds of systems have been defined as "non testeable' software [18, 9] because the lack of a proper testing oracle to check their behavior. The problem can be stated as: How the new version of the system, which now includes the analyzed information, can be tested? How can the new system be checked against its expected behavior? How can the expected behavior be specified? How can the software engineer verify if the system is producing the expected outputs?

One solution to tackle this particular item is known as metamorphic testing [8, 19]. This technique is based on building relationships, called metamorphic relationships (MRs), between the data in the original system and the data in the newer version of the system. For example, for a system focusing on sentiment analysis one can build for every word two MRs: one for its synonyms and another for its antonyms. Under this MRs, the new system can be validated as follows: for every word in the original system the newer version must give a similar response for words in the synonyms list and the opposite response for words in the antonyms list. However, how to specify and define the MR's for every system is an extremely arduous and error-prone task. Some approaches say that defining a lot of MRs solves the problem. Nevertheless, others conclude that defining too many MRs could have a negative impact in the validation phase, since they increase the efforts needed to accomplish this task [9].

Employing expressive and flexible specification languages might be useful to properly explore, understand and define MRs. In particular, in this work we explore the FVS (Feather Weight Visual Scenarios) specification framework [2, 4, 3] as a mechanism to specify MRs for BIG DATA systems. FVS is a very simple yet powerful and expressive graphical language to denote the expected behavior of a system. The behavior can be specified using branching or linear properties, and refinement between specifications is also available. FVS specification can be synthesized providing a controller for the system under analysis. When synthesizing behavior, the controller is automatically built upon the expected behavior of the system and the environment it interacts with. Usually, the controller takes the form of an automaton which decides which actions to take based on the received information (mostly provided by external sensors). The controller is built using game theory concepts, obtaining a winning strategy that takes the system to an accepting state no matter which actions the environment chooses [10, 6]. Finally, FVS can be combined with parallel model checkers to cope with BIG DATA systems performance requirements. In summary, in this work we propose the following contributions for FVS as a formal language to specify MRs:

- MRs can be graphically denoted
- FVS expressive power is strong enough to build all the necessary MRs.

– The possibility to synthesize behavior and obtain a controller can be used to check consistency between all the MRs.
– In FVS complex relationships between MRs can be stated. Refinement between two FVS specifications is well established, as well as analyzing when a MR imposes more restrictions over the system than other candidate MR.

As a proof of concept example, we have analyzed a BIG DATA system introduced in [9]. This system proposes a machine learning service to categorize images of biology cells. Complex MRs are defined between images, taking into account features as size, volume and orientation. All of the MRs were defined using FVS, and we exploited FVS characteristics to understand the interactions between all the MRs. The rest of this work is structured as follows. Section 2 briefly presents FVS and explains how a controller can be obtained. Section 3 develops the proof of concept example and presents the obtained results. Section 4 analyzes some related and future work whereas Section5 enumerates the conclusions of this research.

2 Feather weight Visual Scenarios

In this section we will informally describe the standing features of FVS. The reader is referred to [2] for a formal characterization of the language. FVS is a graphical language based on scenarios. Scenarios are partial order of events, consisting of points, which are labeled with a logic formula expressing the possible events occurring at that point, and arrows connecting them. An arrow between two points indicates precedence. For instance, in Figure 1-(a) A-event precedes B-event. In Figure 1-b the scenario captures the very next B-event following an A-event, and not any other B-event. Events labeling an arrow are interpreted as forbidden events between both points. In Figure 1-c A-event precedes B-event such that C-event does not occur between them. Finally, FVS features aliasing between points. Scenario in 1-d indicates that a point labeled with A is also labeled with $A \wedge B$. It is worth noticing that A-event is repeated on the labeling of the second point just because of FVS formal syntaxis.

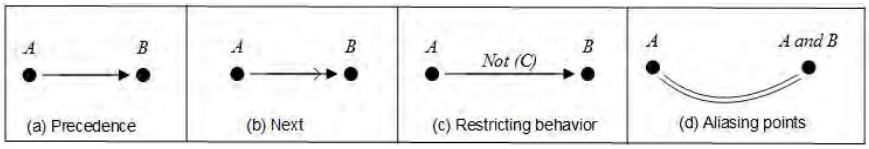


Fig. 1. Basic Elements in FVS

We now introduce the concept of FVS rules, a core concept in the language. The intuition is that whenever a trace "matches" a given antecedent scenario, then it must also match at least one of the consequent ones. In other words,

rules take the form of an implication: an antecedent scenario and one or more consequent scenarios. Graphically, the antecedent is shown in black, and consequent ones in grey. Since a rule can feature more than one consequent, elements which do not belong to the antecedent scenario are numbered to identify the consequent they belong to. An example is shown in Figure 2. The rule describes a requirement for a cooler system. If the cooler is off and the temperature exceeds a certain threshold then two things should happen afterwards: the cooler must be turned on and the observers must be notified.

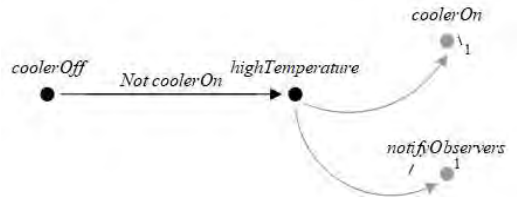


Fig. 2. An FVS rule example

2.1 Behavioral Synthesis in FVS

FVS specifications can be used to automatically obtain a controller employing a classical behavioral synthesis procedure. We now briefly explain how this is achieved while the complete description is available in [4]. Using the tableau algorithm detailed in [2] FVS scenarios are translated into Büchi automata. Then, if the obtained automata is deterministic, then we obtain a controller using a technique [15] based on the specification patterns [11] and the GR(1) subset of LTL. If the automaton is non deterministic, we can obtain a controller anyway. Employing an advanced tool for manipulating diverse kinds of automata named GOAL [21] we translate these automata into Deterministic Rabin automata. Since synthesis algorithms are also incorporated into the GOAL tool using Rabin automata as input, a controller can be obtained.

3 Proof of Concept Example: CMA System

The system under analysis is based on the BIG DATA service implementation described in [9]. This service, called Cell Morphology Assay (CMA) was designed for modeling and analyzing 3D cell morphology and mining morphology patterns extracted from diffraction images of biology cells. Study of 3D morphology can provide rich information about cells that is essential for cell analysis and classification [9]. Relying on different machine learning algorithms and big data tools images are analyzed and explored, and useful scientific information is obtained.

Validation of the system is carried out by defining a set of crucial metamorphic relationships (MR) establishing a proper bond between the original image and the one obtained for validation purposes. These bounds will define whether the system is actually doing a good enough job classifying the biology cells images. In this case, MRs were defined by domain experts users. In what follows, we describe the MR defined in [9] as FVS rules. Finally, a controller for the system is found.

3.1 MRs for the CMA system specified in FVS

The first MR says that if an artificial mitochondrion is added to a stack of original confocal image sections of a cell, then the newly added mitochondrion should be recognized in the new version. Similarly, if an artificial mitochondrion is removed from a stack of the original confocal image sections of a cell, then it must be excluded in the new version. For the FVS specification of this MR we define a set of events such as *image*(standing for the original image), *newImage*(standing for the new version of the image), *artMitoAdded* (standing for the addition of a new artificial mitochondrion), *artMitoRemoved* (a new artificial mitochondrion was removed), *artMitoRecognized* (standing for the recognition of the new artificial mitochondrion) and *artMitoExcluded* (standing for the exclusion of the new artificial mitochondrion). Figure 3 reflects this Inclusion/Exclusion MR.

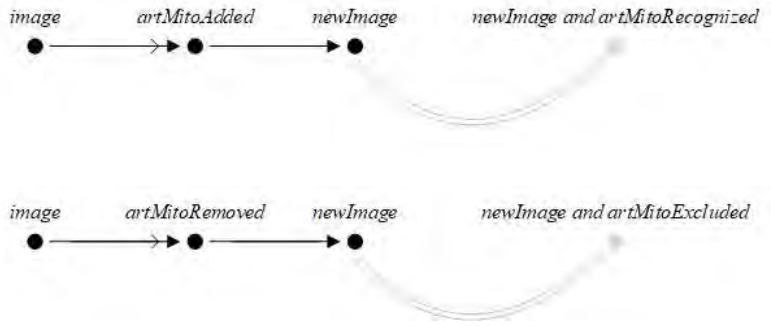


Fig. 3. Inclusion/Exclusion MR as FVS rules

The second MR defines a relationship between size and volume of the images. In a few words, if the size of a mitochondrion in the original image sections is increased then the volume of mitochondria is expected to increase in the new version. The FVS rule for this MR is shown in Figure 4.

The third MR relates lengths between images section. As explained in [9], there is a gap between image sections. This implies that a small mitochondrion may only appear in one section whereas a large one can appear in multiple sections. To specify this MR in FVS we assume that only two sections exist for each image. However, this approach can be easily extended to consider more

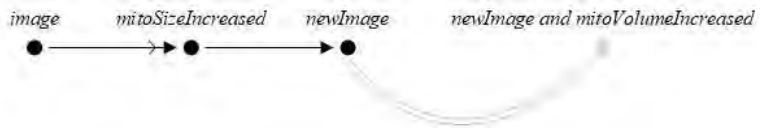


Fig. 4. FVS rule of the Size and Volume MR

sections if necessary. The Lenght-MR is shown in Figure 5. Note that the rule at the bottom of Figure 5 features three consequents: the mitochondrion could be placed in both sections (consequent 1), only in section 1 (consequent 2) or only in section 2 (consequent 3).

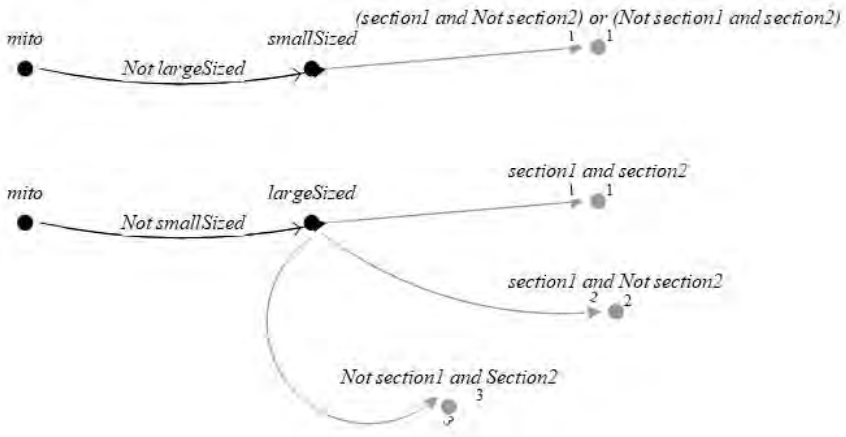


Fig. 5. FVS rule for the Length MR

The fourth MR establishes immutability of the generated shapes. Roughly speaking, the 3D structure of the original one should not be changed in the new version of the image. Considering only two possible shapes, Figure 6 sketches this important MR. As in the previous MR, this specification can be easily extended to consider more than two shapes.

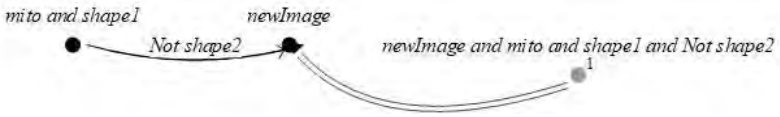


Fig. 6. FVS rule for the immutability MR

The fifth MR focuses on location aspects. Mitochondria that are close to the nuclear should remain in a section near to the nuclear and similarly, mitochondria close to the cell boundary should also appear close to the cell boundary in the new image. Two FVS rules are added to specify this behavior. These are shown in Figure 7.

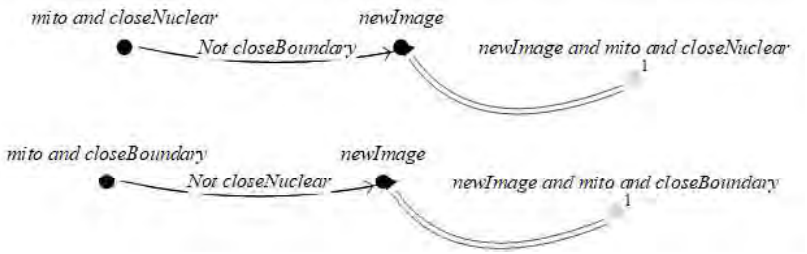


Fig. 7. FVS rules addressing Location MR

MR number six is a simple one. It says that when the size of the sphere in an image becomes larger, the brightness of the texture lines become slimmer. This is reflected in Figure 8.

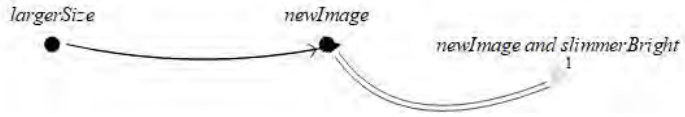


Fig. 8. Size and brightness MR in FVS

Finally, MR number seven establishes that for a sphere shape scatterer the textual pattern should be the same at all orientations. We considered two kinds of scatterers: sphere and irregular ones. The FVS rules tackling this MR are shown in Figure 9.

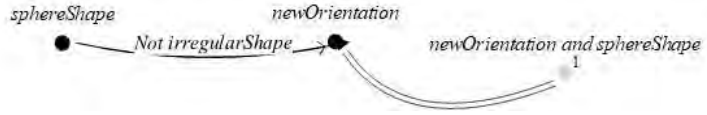


Fig. 9. Sphere shapes MR in FVS

Once all the MR were defined in FVS we were able to obtain a controller as explained in Section 2.1. Since a controller was found we can establish that

there were no inconsistencies in the MRs specifications. A partial view of the controller automaton is shown in Figure 10.

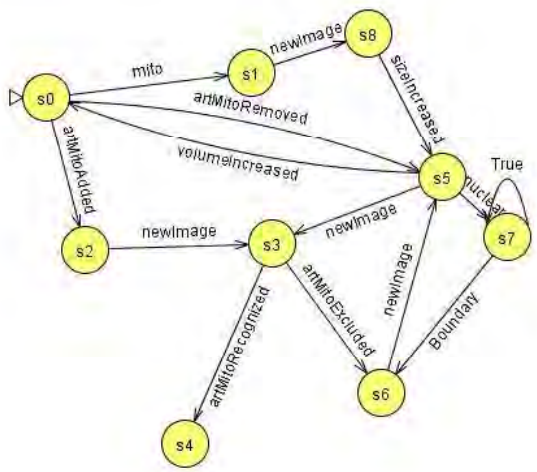


Fig. 10. A controller for the CMA System

3.2 Case Study Analysis, Remarks and Observations

It is relevant to point out that FVS was expressive enough to denote all the MRs defined in [9]. This reflects the richness of FVS's expressive power since this case proof of concept example establishes weighty and meaningful interactions between the images which can be hard to express. In addition, interest analysis of the set of MR can be gathered by analyzing the FVS specification scheme. First of all, visual information relating two or more MR (such as logical subsumption) can be simply noted by visually inspecting the scenarios. For example, both rules in Figure 5 denote equivalent antecedent scenarios but the rule at the bottom holds a "stronger" consequent, since it features more constraints. In consequence, this latter rule can be seen as a specialization of the former one. Secondly, consistency and completeness of the MR set can be stated by the implicit result of whether a controller for the MR specification is found or not. The presence of a controller for the system implies that there were no inconsistencies in all the defined MRs. And finally, other interesting interactions between MRs such as the concept of refinement can also be achieved in the given FVS specification for the system.

A rule featuring multiple consequents as the one shown in the bottom rule of Figure 5 allows the application of the refinement operation between different MR, since a new version with fewer consequents represents a refinement of the original rule [4].

4 Related and Future Work

Previous work in [3] can be seen as a foundation stone for the FVS framework, exhibiting its capabilities to synthesize behavior and to reason about linear an branching behavior together with the formal proofs of soundness and correctness of the approach. In this work we specifically apply FVS to model and verify metamorphic properties in an attempt to formally verify BIG DATA systems.

Work in [9] proposes a very interesting iterative technique to define MRs. Once MRs are defined, more of them can be generated using the notion of refinement between them. The approach was validated against the CMA machine learning system, which is able to categorize a huge amount of biology cells images. We believe FVS graphical flavor could be added to this iterative process to gain power and control to properly define the necessary minimum set of MR for each system. Contrary to FVS, the possibility to check consistency between MRs is not available in this approach.

Other approaches focused on metamorphic properties are [1, 8, 16]. These options are mostly focused on implementation details such as the actual framework tool to deploy the tests. Our approach is addressing a previous phase which is the exploration and specification of MRs.

Regarding future work, we would like to explore the interaction between FVS and other tools. For example, [1] extracts MR in runtime checking the different paths in the execution tree. In this context, FVS scenarios could be used as a monitor in the sense given by the model checking techniques. We also would like to investigate the possibility to automatically generate tests given the set of FVS rules defining the MR. This line of research involves the combination of FVS with automatic code and test generators like [17].

5 Conclusions

In this work we studied FVS as a specification language to denote metamorphic properties. FVS's flexible and expressive notation was able to denote complex MRs behavior. In addition, its formal semantics enables the possibility to perform comparison and formal operations between different MRs such as refinement, or simply comparing which MR imposes more constraints in the expected behavior of the system. We believe these are crucial activities in order to properly define the expected behavior of a BIG DATA system. The results obtained in the proof of concept example are promising enough to consolidate our tool in the formal validation field for BIG DATA systems.

References

1. M. Asrafi, H. Liu, and F.-C. Kuo. On testing effectiveness of metamorphic relations: A case study. In *2011 fifth international conference on secure software integration and reliability improvement*, pages 147–156. IEEE, 2011.
2. F. Asteasuain and V. Braberman. Declaratively building behavior by means of scenario clauses. *Requirements Engineering*, 22(2):239–274, 2017.
3. F. Asteasuain and L. R. Caldeira. A sound and correct formalism to specify, verify and synthesize behavior in big data systems. In *Argentine Congress of Computer Science*, pages 109–123. Springer, 2022.
4. F. Asteasuain, F. Calonge, M. Dubinsky, and P. Gamboa. Open and branching behavioral synthesis with scenario clauses. *CLEI E-JOURNAL*, 24(3), 2021.
5. C. Bellettini, M. Camilli, L. Capra, and M. Monga. Distributed ctl model checking using mapreduce: theory and practice. *CCPE*, 28(11):3025–3041, 2016.
6. R. Bloem, B. Jobstmann, N. Piterman, A. Pnueli, and Y. Sa'Ar. Synthesis of reactive (1) designs. 2011.
7. M. Camilli. Formal verification problems in a big data world: towards a mighty synergy. In *ICSE*, pages 638–641, 2014.
8. T. Y. Chen, S. C. Cheung, and S. M. Yiu. Metamorphic testing: a new approach for generating next test cases. *arXiv preprint arXiv:2002.12543*, 2020.
9. J. Ding, D. Zhang, and X.-H. Hu. A framework for ensuring the quality of a big data service. In *2016 SCC*, pages 82–89. IEEE, 2016.
10. N. DIppolito, V. Braberman, N. Piterman, and S. Uchitel. Synthesising non-anomalous event-based controllers for liveness goals. *ACM Tran*, 22(9), 2013.
11. M. Dwyer, M. Avrunin, and M. Corbett. Patterns in property specifications for finite-state verification. In *ICSE*, pages 411–420, 1999.
12. O. Hummel, H. Eichelberger, A. Giloj, D. Werle, and K. Schmid. A collection of software engineering challenges for big data system development. In *SEAA*, pages 362–369. IEEE, 2018.
13. V. D. Kumar and P. Alencar. Software engineering for big data projects: Domains, methodologies and gaps. In *IEEEBIGDATA*, pages 2886–2895. IEEE, 2016.
14. R. Laigner, M. Kalinowski, S. Lifschitz, R. S. Monteiro, and D. de Oliveira. A systematic mapping of software engineering approaches to develop big data systems. In *SEAA*, pages 446–453. IEEE, 2018.
15. S. Maoz and J. O. Ringert. Synthesizing a lego forklift controller in gr (1): A case study. *arXiv preprint arXiv:1602.01172*, 2016.
16. J. Mayer and R. Guderlei. An empirical study on the selection of good metamorphic relations. In *30th Annual International Computer Software and Applications Conference (COMPSAC'06)*, volume 1, pages 475–484. IEEE, 2006.
17. I. A. Niaz and J. Tanaka. Code generation from uml statecharts. In *Proc. 7 th IASTED International Conf. on Software Engineering and Application (SEA 2003), Marina Del Rey*, pages 315–321, 2003.
18. C. E. Otero and A. Peter. Research directions for engineering big data analytics software. *IEEE Intelligent Systems*, 30(1):13–19, 2014.
19. S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés. A survey on metamorphic testing. *IEEE Transactions on software engineering*, 42(9):805–824, 2016.
20. P. A. Sri and M. Anusha. Big data-survey. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 4(1):74–80, 2016.
21. Y.-K. Tsay, Y.-F. Chen, M.-H. Tsai, K.-N. Wu, and W.-C. Chan. Goal: A graphical tool for manipulating büchi automata and temporal formulae. In *TACAS*, pages 466–471. Springer, 2007.

Derivación de Escenarios por Proximidad

Gladys Kaplan[1] y Jorge Doorn[2]

[1]Departamento de Ingeniería e Investigaciones Tecnológicas, Universidad Nacional de La Matanza. San Justo, Buenos Aires
[2] Escuela de Informática, Universidad Nacional del Oeste. Merlo, Bs As, Argentina.
gkaplan@unlam.edu.ar y jdoorn@uno.edu.ar

Abstract. Los escenarios describen la realidad observable o, dicho de otra manera, el proceso del negocio donde se planifica instalar el nuevo sistema de software. Para obtener la primera versión de estos escenarios se utiliza una heurística de derivación que toma información desde el LEL, la cual ha mostrado en la práctica algunos problemas de gran importancia. El más relevante es que no tiene en cuenta que la estructura del LEL es declarativa mientras que la de los escenarios es procedural, generando escenarios dispersos que no representan la realidad y altamente incompletos. También se realizó un estudio empírico sobre casos existentes donde se detectaron otros problemas no menos graves. Entre ellos la creación de una lista inicial que obstaculiza la identificación de nuevos símbolos y la omisión de características relevantes del contexto. En el presente artículo se analiza el origen de estos problemas y se propone una nueva heurística de derivación desde el LEL con una mirada procedural e incorporando el tratamiento de estados, jerarquías conceptuales y puntos de vista del contexto. Este mecanismo iterativo e incremental analiza cada escenario buscando nuevas situaciones por proximidad, generando un primer conjunto de escenarios reales y sustancialmente más completos.

Keywords: Ingeniería de Requisitos, LEL, escenarios, Derivación.

1 Introducción

Construir escenarios [1] [2] [3] que describan la realidad observable donde se planifica poner en servicio un nuevo sistema de software es una actividad importante en el Proceso de Requisitos [4] ya que los mismos contribuyen de tres maneras diferentes a mejorar la calidad del proceso. Por un lado, la propia construcción de los escenarios hace que los participantes en la actividad precisen su comprensión de esa realidad. Además, estos escenarios son un reservorio confiable de información que permiten solucionar dudas posteriores. Finalmente, estos escenarios, usualmente denominados *escenarios actuales* (EA) [5] son la materia prima básica con la que se pueden construir los *escenarios futuros* (EF) [6] que describen lo que se planifica que ocurrirá cuando el sistema de software este en producción.

El Proceso de Requisitos en el cual se han elaborado los resultados reportados en el presente artículo, construye la primera versión de los EA tomando como guía un modelo anterior denominado Léxico Extendido del Lenguaje (LEL) [7], el que contiene

una descripción del vocabulario específico utilizado en la realidad observable que se intenta modelar. Este pasaje de información se denomina *Derivación de EA desde el LEL* y el resultado obtenido es un conjunto de Escenarios Candidatos Derivados (ECD), los cuales serán luego completados con información del contexto [8].

Que los EA jueguen un rol importante en el proceso trae como consecuencia que toda la calidad del proceso esté fuertemente influenciada por la calidad de los mismos. Obviamente, la calidad de un modelo es una propiedad multifacética donde juegan roles importantes la coherencia, la corrección y la completitud entre varias otras cualidades. Todas estas cualidades han sido estudiadas en profundidad [9] [10], habiéndose encontrado resultados satisfactorios excepto en la completitud [11]. En el estudio empírico realizado se ha detectado que el origen de este problema es, en gran medida, la heurística de derivación existente. Si bien el concepto de completitud puede ser abordado desde diferentes lugares, el primer interrogante que aparece es sin dudas cómo asegurar un grado de completitud compatible con la calidad deseada, o sea procurar que se registre toda la información relevante, minimizando la incorporación de información innecesaria.

En el presente artículo se analizan los problemas e inconvenientes que presenta la heurística de derivación existente y se presentan las mejoras introducidas.

En la sección 2, *Análisis de la heurística de derivación existente*, se detallan los problemas detectados en la heurística existente que han determinado la necesidad de reemplazarla, luego, en la sección 3, *Nueva Heurística de Derivación de Escenarios por Proximidad,* se describe la nueva heurística de derivación. En la sección 4, *Aplicación de la heurística*, se presenta un ejemplo de uso de la nueva heurística y finalmente, en la sección 5 las *Conclusiones y Trabajos Futuros*.

2 Análisis de la heurística de derivación existente

Comparando la lista de ECD con la lista de EA finales que se obtiene luego de haber realizado la organización y descripción de los mismos, se han encontrado varios fenómenos notoriamente reiterados. Particularmente, estos fenómenos no eran fácilmente previsibles en el momento de proponer la heurística y probar la misma en unos pocos casos iniciales. Es más, la percepción de las regularidades que se describen en esta sección sólo fue posible con la realización de una cantidad importante de casos reales.

En orden cronológico, se ha ido observando que: i) el orden de la lista de ECD perturba la descripción de los mismos, ii) es frecuente que varios ECD se consoliden en un sólo escenario final, iii) algunos de los EA se construyen a partir de uno o varios ECD, iv) información relevante registrada en el LEL en los símbolos de tipo estado no es reflejada en los EA, v) información relevante registrada en el LEL relacionada con los puntos de vista del *es* y *deber ser*, o con el *será*, no es registrada apropiadamente en los EA y vi) información relevante relacionada con jerarquías taxonómicas y de composición registrada en el LEL no es registrada en los EA.

2.1 Orden en la lista de ECD

La lista de ECD carece de todo orden o a lo sumo está ordenada siguiendo el orden alfabético de los símbolos del LEL que sugirieron su inclusión. Esta no es una cuestión menor, por el contrario, es esencial. Al evaluar su pertinencia y al intentar agregar la información que corresponde, el ingeniero/a de requisitos carece completamente de contexto ya que los escenarios próximos en la lista tienen escasa relación con el que está considerando. Naturalmente, que esta situación fuerza una reorganización lo que no es trivial en absoluto, ya que en ese momento se conoce poco acerca de la realidad que está siendo modelada[1]. Claramente eso es una fuente de inconvenientes que pueden conllevar errores.

2.2 Consolidación de ECD

Si en un EA interviene más de un actor, cada uno de ellos desempeñará algún rol en el mismo. En el LEL, parte de esta información estará registrada en uno o más impactos de cada uno de los sujetos que actúan en ese escenario. Si la derivación crea un ECD por cada impacto de cada sujeto, este EA se manifestará como varios posibles escenarios. Además, estos posibles escenarios estarán dispersos en la lista de ECD. Claramente la consolidación de estos posibles escenarios tampoco es una tarea trivial y también puede ser causa de errores.

En la Tabla 1 se puede observar que de los impactos del LEL se deriva un conjunto de ECD, los que finalmente, confluyen en un único EA "Planificar la Producción".

Tabla 1. Consolidación de ECD.

Símbolos del LEL que dan origen a los ECD	ECD	Escenario Final
Jefe de Producción (S)	Comunicarse con la Papelera del Sudeste	Planificar la Producción
	Emitir Programa de fabricación extraordinaria	
	Estudiar las fallas ocurridas durante la fabricación	
	Crear secuencias	
Elaborar una OP (V)	Elaborar una OP	
Oficial planificador (S)	Generar los programas de fabricación	
Priorizar OP (V)	Priorizar OP	
Revisar Programa de Fabricación (V)	Revisar programa de fabricación	

El principal problema de la derivación existente radica en la diferencia en los objetivos y en el ordenamiento de la información entre el LEL y los escenarios. Esto sucede porque el LEL describe términos, mientras que los escenarios describen

[1] "There is no sense in being precise about something when you don't know what you are talking about." [4]

situaciones del proceso del negocio. De esta manera, la información de una situación puede estar dispersa en varios símbolos, generando por derivación escenarios candidatos ficticios.

El mecanismo que propone la heurística existente de pasar información desde el LEL a los escenarios de manera textual, se debe descartar. Esto no significa eliminar la derivación, sino aprovechar la información disponible mientras sea beneficiosa. Otra desventaja que se detectó en la heurística existente es que no consulta toda la información disponible, desaprovechando información valiosa. Aun cuando no se disponga de documentación del macrosistema, para construir el LEL fue necesario realizar entrevistas y sus transcripciones son un documento de mucho significado para la IR en general y, para la derivación de escenarios en particular. Utilizar la documentación existente ayuda a reestructurar el orden intrínseco del glosario, desde una perspectiva declarativa a una procedural.

2.3 Aparición de nuevos escenarios

Casi todos los EA son el resultado de haber descripto uno o varios ECD. En otras palabras, la lista de ECD no sólo es una guía acerca de cuáles son los escenarios a considerar, sino que además entorpece seriamente la búsqueda de aspectos de la realidad observable no identificados y, por tanto, no descriptos en los EA. Inicialmente, la existencia de una correlación alta entre la lista de ECD y los EA fue considerada una virtud que ponía en evidencia la calidad de la primera, sin embargo, con la evaluación de sucesivos casos se fue haciendo cada vez más evidente que la completitud del LEL condiciona fuertemente la cantidad total de EA que podrían obtenerse como consecuencia de la inhibición de la detección de nuevos escenarios. Esto reduce notoriamente la habilidad autocorrectiva del Proceso de Requisitos. En la Tabla 2 se describen 5 casos tomados al azar donde se puede observar la incidencia de la derivación en el producto final, donde la aparición de nuevos escenarios es prácticamente nula.

Tabla 2. Aparición de nuevos escenarios.

Caso	Cantidad de escenarios		
	ECD	EA finales	Nuevos
Ofertas de materias a distancia	13	13	0
Control de Stock de la Farmacia de la Clínica	19	9	0
Superintendencia de Riesgos del Trabajo	15	14	2
Gestión de consultas médicas de una clínica	19	18	0
Marketing de Páginas Amarillas	16	9	0

Esta dispersión de las situaciones observables del contexto que no son reparadas oportunamente durante la construcción de los EA llegarán a los EF, generando, en el

mejor de los casos, un importante trabajo adicional para reorganizar las situaciones descriptas.

2.4 Estados

La revisión sistemática de la información registrada en los estados del LEL y su eventual influencia en el contenido de los escenarios donde aparecen los símbolos del LEL afectados por estos estados, ha mostrado que es frecuente la omisión de precondiciones o restricciones en los EA. Obviamente, los EF y los requisitos resultantes también carecerán de esos detalles, lo que puede provocar una pérdida de información no deseable.

2.5 Puntos de vista

Cuando una información está condicionada por un punto de vista, el mismo debe poder percibirse claramente en los EA. Por ejemplo, si una cierta actividad realizada por un sujeto del LEL está condicionada por el punto de vista del *deber ser*, en el EA debe figurar tanto la conducta esperada como la real. En este tipo de casos es muy importante determinar con precisión si el sistema deberá forzar la conducta esperable o si el sistema deberá permitir ambas conductas. Obviamente, los EF serán muy diferentes, según sea el caso y lo mismo ocurrirá con el software que se produzca.

2.6 Jerarquías

Si no se consideran adecuadamente las jerarquías, especialmente las taxonómicas, se corre el fuerte riesgo que se asigne una responsabilidad en los EA a un actor o se asuma que un recurso posee alguna propiedad que en realidad pertenece a una especialización de aquel o del objeto [12]. Si alguna de las responsabilidades del actor o alguno de los usos del recurso es asumido por el sistema de software en los EF, se producirá un software que no se corresponde con la realidad.

3 Nueva Heurística de Derivación de Escenarios por Proximidad

Tomando en cuenta lo descripto en la sección anterior, se decidió construir una nueva heurística de derivación la cual elimina la lista inicial y detecta situaciones del contexto por proximidad. En este nuevo mecanismo se utiliza tanto el LEL como los documentos del dominio (manuales de procedimientos, protocolos, transcripciones de las entrevistas, etc.). Una característica singular del nuevo mecanismo es que se superpone la *Derivación* con la actividad *Describir*, no siendo tan claro cuando termina una y

comienza la otra. La nueva *Derivación* toma los verbos[2] del LEL, los cuales actúan como desencadenantes para orientar la búsqueda de información en la documentación. Esta búsqueda abandona la mirada léxica para abocarse a una estricta mirada procedural. Durante la elaboración de la presente estrategia se tomó en cuenta la dispersión de la información y la necesidad de concentración del ingeniero/a de requisitos en unos pocos aspectos en cada momento. Tomando en cuenta estos dos aspectos surge que se podría trabajar de dos maneras diferentes:

- Describir un escenario por vez recorriendo todas las fuentes de información disponibles.
- Analizar cada fuente de información hasta agotarla, o sea completar todos los escenarios posibles con cada documento.

Las pruebas preliminares, han mostrado que abocarse a un documento favorece la concentración del ingeniero/a de requisitos. De esta manera, se fortalece la construcción de los EA, eligiendo una construcción por proximidad, iterativa e incremental, donde a partir de unos pocos escenarios iniciales se obtiene el resto. Los nuevos escenarios aparecen durante la descripción de los escenarios previamente detectados, lo que implica que debe haber un paso inicial que dé origen a un conjunto reducido de escenarios que se constituirán en las semillas de toda la actividad de derivación. Tanto los escenarios semilla como los que se vayan detectando por proximidad se irán completando a medida que se avanza en la actividad de derivación. Cuando un escenario aparece muy tarde en el proceso, es muy poco probable que sea necesario volver a revisar los documentos ya analizados. Esto ocurre porque una aparición tardía está casi sin excepción asociada a un documento que aborda aspectos del proceso del negocio no detallados en los documentos ya utilizados. La idea es ir construyendo los escenarios de a pequeños grupos fuertemente vinculados. Al finalizar la derivación de los escenarios, se debe consultar el LEL para verificar el cubrimiento del mismo con el objetivo de detectar posibles omisiones.

La nueva estrategia de derivación consiste, como se puede observar en la Fig.1, en tomar información desde el LEL (ver flecha identificada con el "1"), pero también es utilizada como guía para saber "qué buscar" en el "Doc.1". Este primer documento es la transcripción de la entrevista inicial del proyecto. Se construyen los EA a partir de los verbos y los impactos de los sujetos del LEL que contienen verbos que no son símbolos del LEL. Luego, en la flecha "2" se busca en el documento los símbolos seleccionados, analizando la información existente desde una perspectiva de procesos y hasta agotar todo el documento. En la flecha "3" se completan los escenarios hasta donde sea posible con la información encontrada en el paso 2. En la flecha "4" se busca en los documentos restantes los nombres de los escenarios construidos. Se completan los escenarios con la información existente en estos documentos. Se debe tener en cuenta que se analiza un documento por vez hasta agotarlo, recién entonces se pasa al

[2] Los símbolos del LEL se clasifican en Sujetos (personas, organizaciones, sistemas informáticos), Objetos (tangibles e intangibles), Verbos (acciones) y Estados.

siguiente. Finalmente, en la flecha "5", se verifica el cubrimiento del LEL para detectar omisiones.

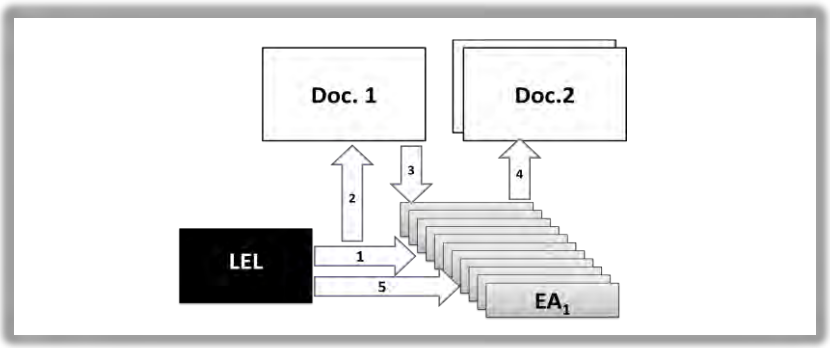


Fig. 1. Estrategia de Derivación de Escenarios por proximidad

A continuación, se describen los pasos de la heurística:

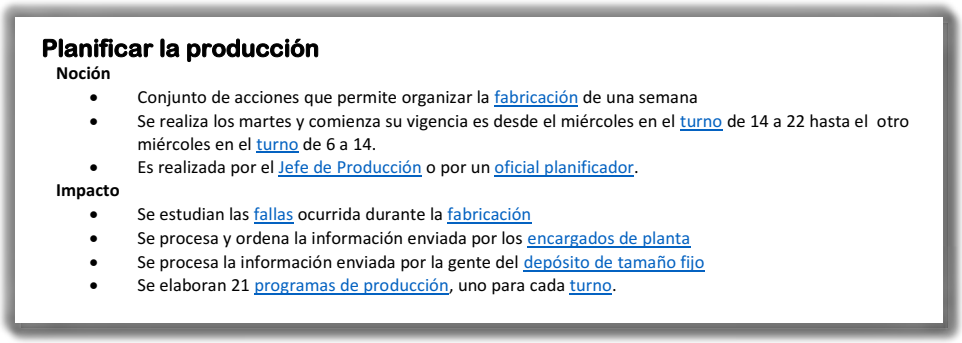
Paso 1:	Seleccionar los símbolos Verbos del LEL más representativos. Incorporar estos posibles escenarios en la *Lista de escenarios pendientes de describir.*
Paso 2:	Seleccionar el documento del macrosistema inicial.
Paso 3:	Se deben crear escenarios, uno por cada símbolo Verbo.
Paso 4:	Crear una cadena de búsqueda por cada símbolo (ej. "Produc*").
Paso 5:	Recorrer el documento ejecutando la cadena de búsqueda.
Paso 6:	Cada vez que se encuentre información en el documento, se debe analizar su pertinencia con la situación que se está describiendo. Cuando la información corresponda a dicha situación incorporarla al escenario.
Paso 7:	Cuando no se encuentra más información para el escenario, buscar en el escenario la presencia de nuevas situaciones próximas que no fueron vistas aún. Incorporar los nuevos posibles escenarios a la *Lista de escenarios pendientes de describir.*
Paso 8:	Para agotar el documento en el cual se está buscando, analizar nuevamente el texto, principalmente aquellas partes no analizadas aún, buscando si se sugieren otras situaciones relacionadas.
Paso 9:	Una vez analizados todos los verbos se deben analizar los impactos de los sujetos que no son verbos. Crear un escenario por cada impacto y repetir desde el Paso 4 hasta completar la *Lista de escenarios pendientes de describir*.
Paso 10:	En este momento se deben analizar los documentos restantes, o sea todos aquellos que no fueron utilizados en la primera parte.
Paso 11:	Cada vez que se encuentre nueva información para un escenario existente, se debe analizar la presencia de puntos de vista ("deber ser/es"), o sea de conflictos entre lo expresado en el documento actual y alguno previamente analizado.
Paso 12:	Buscar en el escenario nuevas situaciones. En este momento es menos probable que aparezcan.
Paso 13:	Recorrer el documento buscando si se sugieren otras situaciones relevantes del contexto.
Paso 14:	Finalizar verificando si se ha cubierto todo el LEL.

En síntesis, para cada documento se deben realizar todas las búsquedas incluyendo aquellas ya realizadas en los documentos anteriores. En este momento, con todos los

documentos analizados, se considera que todos los escenarios ya han sido creados, por lo tanto, se debe concentrar en recorrer cada documento hasta agotarlo buscando nueva información. Es importante destacar que esta actividad se ve muy facilitada si se marca en el documento toda información ya utilizada. También, se debe analizar cada escenario para detectar nuevas situaciones por proximidad.

4 Aplicación de la heurística

En esta sección se presenta un ejemplo de una fábrica de cajas de cartón corrugado. La primera tarea fue definir el Objetivo General del Sistema: "Reducir los errores en las órdenes de producción". Luego, se seleccionó del LEL un símbolo semilla relacionado con dicho objetivo. En este caso se trabajó con un único documento con la descripción del proceso del negocio del cliente.



Planificar la producción

Noción

- Conjunto de acciones que permite organizar la fabricación de una semana
- Se realiza los martes y comienza su vigencia es desde el miércoles en el turno de 14 a 22 hasta el otro miércoles en el turno de 6 a 14.
- Es realizada por el Jefe de Producción o por un oficial planificador.

Impacto

- Se estudian las fallas ocurrida durante la fabricación
- Se procesa y ordena la información enviada por los encargados de planta
- Se procesa la información enviada por la gente del depósito de tamaño fijo
- Se elaboran 21 programas de producción, uno para cada turno.

Fig. 2. Ejemplo Símbolo semilla que inicia la Derivación por Proximidad

Con el símbolo verbo detallado en la Fig. 2 se construyó el ECD vacío. Luego, se creó una cadena de buscada con "Planif*" + "planif*" y se examinó todo el documento detectando dos párrafos donde aparecía dicha cadena de búsqueda. La información pertinente fue incorporada al escenario descripto en la Fig. 3.

Cuando el escenario estuvo completo se lo analizó en busca de situaciones por proximidad. En el escenario de la Fig. 3 se detectó que en el episodio 7 se hablaba de *priorizar* lo que sugirió ser relevante en este contexto. Se decidió ampliar dicha información y se generó la cadena de búsqueda con "Prioriz*" + "prioriz*". Se buscó en todo el documento y se pudo observar que existían dos situaciones relacionadas:

1) se prioriza cuando hay una urgencia de fabricación
2) se prioriza por el desperdicio de cartón corrugado durante la fabricación.

En función de esta información, se crearon ambos escenarios y se modificó el escenario de la Fig. 3 insertaron dos sub-escenarios que reemplazaron al episodio 7 (ver Fig. 4).

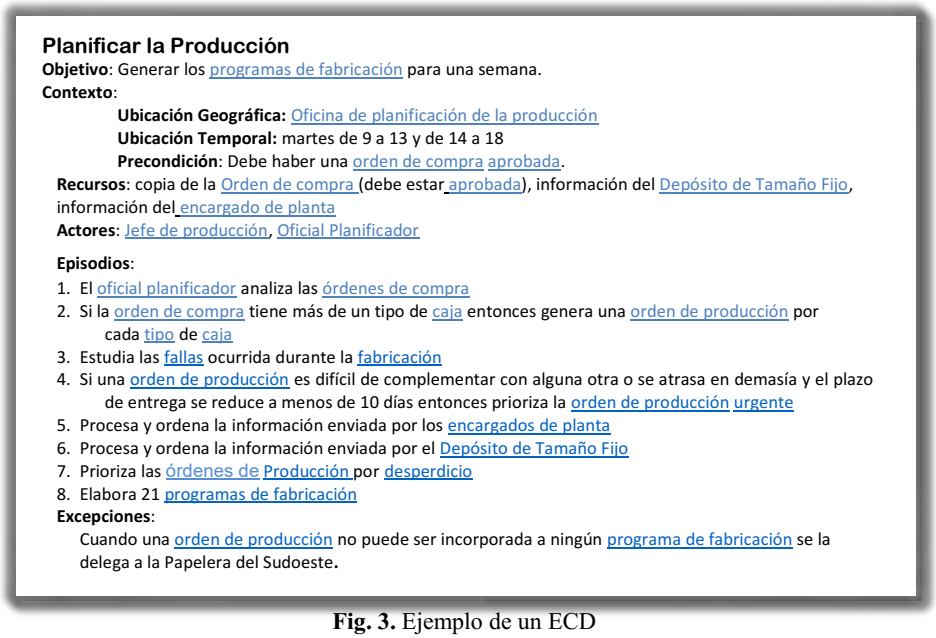


Fig. 3. Ejemplo de un ECD

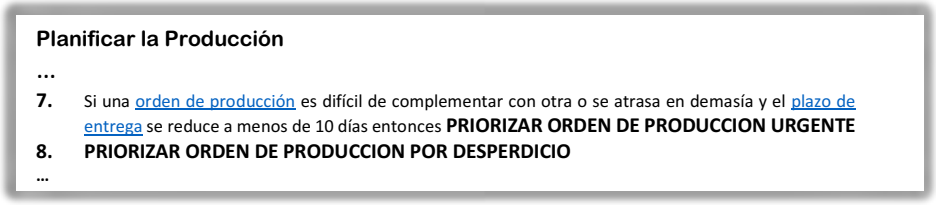


Fig. 4. Ejemplo de situaciones detectadas por Proximidad

Este procedimiento se repitió para todas las acciones relevantes del escenario. De esta manera, de forma iterativa e incremental, se fueron identificando todas las situaciones del contexto. Cabe destacar que se agotó el documento antes de pasar a otro y que los lexemas utilizados para buscar en el primer documento fueron preservados para reutilizarlos en otros documentos de ser necesario.

5 Conclusiones y Trabajos Futuros

Se ha propuesto una nueva heurística de derivación de EA a partir del LEL. Esta heurística sugiere trabajar principalmente con documentos organizacionales, siempre que existan. Además, se propone la utilización de las transcripciones de todas las entrevistas realizadas ya que permiten aprovechar mejor la información que contienen y mejoran la rastreabilidad. Puede observarse que se ha eliminado la lista inicial de ECD. Esta lista se construye muy tempranamente en el proceso, cuando aún no existe suficiente conocimiento del dominio y esta es una desventaja importante ya que no

permite alertar al ingeniero/a de requisitos cuando el camino no es el correcto. La elección de una fuente de información que no sea la "ideal", cuando es la primera fuente consultada, puede resultar más perjudicial. Esto se debe a que la lista, una vez generada, es poco mejorada durante el proceso.

Es probable que el mismo proceso de construcción de los modelos posteriores corrija alguno de estos desvíos, pero existe el importante riesgo de propagar involuntariamente errores a lo largo de todo el Proceso de Requisitos.

Finalmente, como trabajo futuro, se espera probar la heurística en más casos reales y compararlos con los resultados de la heurística anterior. De esta manera se podrá medir con mayor exactitud la mejora en la calidad de los escenarios derivados.

Referencias

[1] Carroll, J., "Introduction: The Scenario Perspective on System Development", en el libro Scenario-Based Design: Envisioning Work and Technology in System Development, editor J. Carroll, John Wiley & Sons, Nueva York, 1995.

[2] Karen L. McGraw, Karan Harbison, "User-centered Requirements: The Scenario-based Engineering Process", 1st Edition, CRC Press, 2020.

[3] Jackson, M., "Software Requirements & Specifications. A lexicon of practice, principles and prejudices", Addison-Wesley, Reading, MA/ACM Press, Nueva York, 1995. Pag. 65

[4] Leite, J.C.S.P., Doorn, J.H., Kaplan, G.N., Hadad, G.D.S., Ridao, M.N., "Defining System Context using Scenarios", en el libro "Perspectives on Software Requirements", Kluwer Academic Publishers, EEUU, ISBN: 1-4020-7625-8, Capítulo 8, pp.169-199, 2004.

[5] Leite, J.C.S.P., Hadad, G.D.S., Doorn, J.H., Kaplan, G.N., "Scenario Construction Process", Requirements Engineering Journal, Springer-Verlag London Ltd., Vol.5, Nº1, pp. 38-61, 2000.

[6] Doorn, J.H., Hadad, G.D.S., Kaplan, G.N., "Comprendiendo el Universo de Discurso Futuro", WER'02 - Workshop en Ingeniería de Requisitos, España, pp.117-131, 2002.

[7] Leite, J.C.S.P., Franco, A.P.M.: O Uso de Hipertexto na Elicitaçao de Linguagens da Aplicaçao. Anais de IV Simpósio Brasilero de Engenharia de Software, SBC, 134-149, 1990

[8] Hadad, G., Kaplan, G., Oliveros, A., Leite, J.C.S.P., "Construcción de Escenarios a partir del Léxico Extendido del Lenguaje", XXVI JAIIO - SoST'97 Simposio en Tecnología de Software, Buenos Aires, pp.65-77, 1997.

[9] Doorn, J., Kaplan, G., Hadad, G., Leite, J.C.S.P., "Inspección de Uso de escenarios en el Desarrollo de Software Referencias 436 escenarios", WER'98 - Workshop de Engenharia de Requisitos, Maringá, Paraná, Brasil, 1998, pp.57-69

[10] Kaplan, G.N., Hadad, G.D.S., Doorn, J.H., Leite, J.C.S.P., "Inspección del Léxico Extendido del Lenguaje", WER'00 – III Workshop de Engenharia de Requisitos, Río de Janeiro, Brasil, pp.70- 91, Julio 2000.

[11] Ridao, M., Doorn, J.H. "Estimación de Completitud en Modelos de Requisitos Basados en Lenguaje Natural". En: IX Workshop on Requirements Engineering (WER'06), Brasil, pp. 151—158, 2006.

[12] Kaplan Gladys y Doorn Jorge, "Jerarquías Naturales en el Contexto del Proceso de Requisitos", IXX Workshop de Engenharia de Requisitos (WER'17), UCA, Bs.As., 2017.

Identificación de Anomalías de APIs web en Mashup

Cinthia Lima, Graciela Vidal y Sandra Casas

GISP - Instituto de Tecnología Aplicada
Universidad Nacional de la Patagonia Austral
Campus universitario Piloto Lero Rivera s/nro. Río Gallegos Santa Cruz
cinty.calderon.15@gmail.com , gvidal@uarg.unpa.edu.ar y sicasas@uarg.unpa.edu.ar

Resumen Las aplicaciones web mashup son el resultado de extraer y combinar información o datos de diversas fuentes externas. Las APIs web son fundamentales en el proceso de desarrollo de una web mashup ya que permiten el acceso a información de distinto tipo (texto, imágenes, videos). Luego, cuando la aplicación mashup está disponible, es posible que se produzcan anomalías en las respuestas de las APIs web, esto ocasiona fallas o pérdidas de información. Por este motivo, es necesario que los desarrolladores identifiquen estas anomalías para manejar estas situaciones. Este trabajo presenta un mashup que integra información de las redes sociales Facebook, Twitter y YouTube, e incorpora un mecanismo para identificar y registrar anomalías conocidas como respuestas vacías y respuestas mal formadas. A partir de un archivo JSON personalizado se registran las anomalías para su análisis. Los resultados se presentan mediante ejemplos sobre la aplicación web y el reporte de anomalías encontradas.

1. Introducción

Un mashup es una aplicación web que combina el contenido de más de una fuente externa y lo integra para que se pueda acceder desde un único lugar [1]. Estas aplicaciones son consideradas híbridas, ya que extraen información o datos de diversas fuentes, con lo cual se obtiene como resultado un nuevo sitio web [2]. Según [3], una web mashup es una aplicación web que integra datos, una aplicación lógica y/o interfaces de usuario de origen web, típicamente un mashup integra y orquesta dos o más elementos. Un componente típico suelen ser las APIs web.

Las APIs son interfaces de programación de aplicaciones que brindan métodos y facilidades para acceder al contenido de alguna aplicación [4]. Las APIs de servicios web, ofrecen un enfoque sistemático y extensible, a diferencia de las APIs vinculadas estáticamente o locales [5][6] para acceder a recursos, servicios y datos a través de la red. En particular, las APIs web facilitan el intercambio de datos inter e intra organizacionales disponibles a través de puntos finales específicos [7].

Las redes sociales se han convertido en los mayores proveedores en este contexto, ofrecen sus APIs y las comparten con el fin de que los desarrolladores integren la información y funcionalidades características de cada una de ellas en nuevas aplicaciones.

Durante la ejecución de una web mashup pueden suceder eventos no esperados, principalmente en el acceso a la respuesta de las APIs, algunos de ellos pueden ser respuestas mal formadas o respuestas vacías [8]. La aplicación mashup se ejecutará erróneamente o presentará información incompleta. Por lo tanto, es necesaria una estrategia para la identificación de estas anomalías con el fin de analizar su alcance sobre la aplicación. De este modo, el desarrollador puede manejar estas situaciones,

ya sea cambiando de API web, modificando la implementación del mashup, o realizar otra acción.

Este trabajo presenta el diseño e implementación de un esquema de resolución para el problema planteado. Se trata de un mashup web que integra y compone las novedades de cuentas en redes sociales (Facebook, Twitter y YouTube) de diferentes áreas de la Universidad Nacional de la Patagonia Austral y además identifica y registra anomalías producidas en las respuestas de las APIs web consumidas (respuestas vacías y respuestas mal formadas).

Este estudio se organiza de la siguiente forma, en la Sección 2 se presenta una definición general de API web y anomalías relacionadas. En la Sección 3 se presenta el caso de estudio, en la Sección 4 se muestran ejemplos del registro de anomalías, en la Sección 5 se exponen los trabajos relacionados, finalmente se presentan las conclusiones en la Sección 6.

2. APIs web y anomalías

El desarrollo de software moderno es inseparable del uso de APIs [9]. Las APIs web pertenecen a una nueva generación de APIs, denominadas API de servicios web, estas ofrecen un enfoque sistemático y extensible, a diferencia de las API vinculadas estáticamente o locales [5][6]. Las APIs web facilitan el intercambio de datos inter e intra organizacionales disponibles a través de puntos finales específicos [7]. Las APIs web se utilizan como un mecanismo de interconectividad clave para acceder a servicios de software a través de Internet. Las aplicaciones interconectadas pueden proporcionar un mejor servicio a un menor costo para sus usuarios [10]. Los catálogos online como API Harmony, PublicAPI o ProgrammableWeb enumeran miles de APIs web y dan cuenta de la proliferación de este tipo recursos, además las grandes empresas IT como Google, Amazon, Facebook, Twitter y YouTube, ofrecen compartir información, recursos y servicios a partir de las APIs web.

Existen diversos problemas relacionados con las respuestas obtenidas de las consultas y/o llamadas a las APIs web, algunos de ellos son las respuestas mal formadas y las respuestas vacías [8]. Estos problemas afectan el funcionamiento de las aplicaciones mashups, por lo tanto, su identificación y análisis es necesario para llevar a cabo su gestión, debido a que impactan de manera negativa en las mismas.

Las respuestas mal formadas se refieren a la generación de respuestas en un formato inadecuado, que no corresponde con el esperado. Por ejemplo, cuando se codifica una cadena JSON que contiene comillas dobles y estas no se ubican correctamente (es decir, "foo": "b" ar " en contraposición al formato válido " foo "=" b\"ar"). Otro ejemplo, ocurre en documentos XML, si se rompe una etiqueta de la respuesta (en este caso, '<data>' se convierte en '<data').

Por otro lado, una API puede devolver respuestas vacías, esto sucede por diferentes razones, por ejemplo si el servidor web está al límite de su capacidad máxima o si la conexión se interrumpe debido a problemas de comunicación, entre otras. En ambas situaciones, la aplicación web mashup no se ejecutará normalmente, presentando información incompleta o no podrá cumplir con otras operaciones [8].

3. Caso de estudio: Portal de publicaciones de redes sociales UNPA

Se desarrolló un portal web de publicaciones realizadas por diferentes áreas de la UNPA en sus redes sociales. Además, sobre este sitio web es posible identificar y reportar anomalías en las respuestas de las APIs web consumidas. En la Fig. 1 se presenta la propuesta desarrollada.

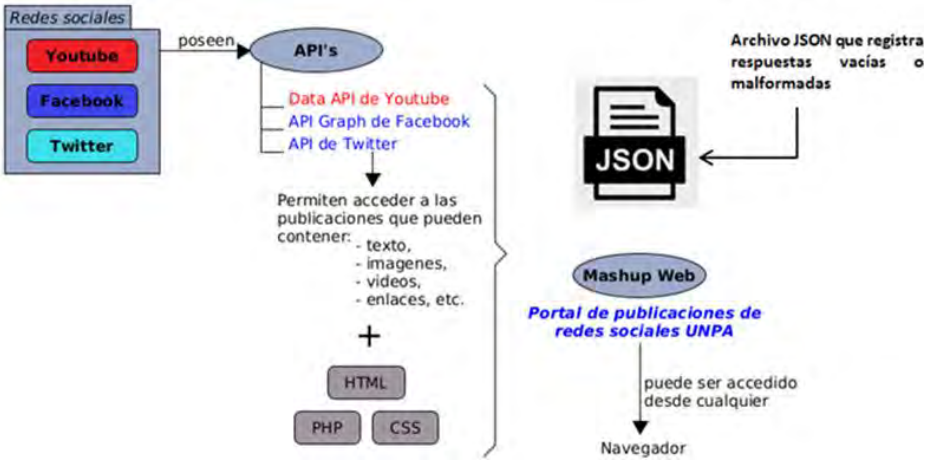


Figura 1: Representación de la propuesta

En este portal se integraron las APIs de las redes sociales, YouTube, Facebook y Twitter. A través de las mismas es posible acceder a diferentes contenidos que luego son presentados en el portal. Además, mientras el portal está en ejecución se registran las anomalías detectadas en las respuestas a las APIs web consumidas.

Para esto se llevaron a cabo tareas considerando los datos incluidos en la respuesta de cada API y los requeridos en el portal. Primero, se analizó la estructura de las respuestas, esto facilitó la identificación de los campos relevantes para el portal.

Aunque el formato de las respuestas de las APIs es el mismo, las estructuras de las respuestas no son idénticas. Por lo tanto, también fue necesario filtrar y ordenar los campos de cada respuesta con el fin de presentar sólo aquellos que el mashup necesitaba mostrar.

Luego, se estudió las anomalías: respuestas vacías y respuestas mal formadas. Se realizaron pruebas sobre cada una de las APIs web con el fin de observar el comportamiento del mashup ante las mismas. Para ello, se incluyó un registro de actividad, el mismo consiste en un archivo JSON, el cual se actualiza cada vez que se detecta pérdida de información en el caso de que se detecten las respuestas vacías o pérdida de la respuesta total, en relación a las respuestas mal formadas.

En la Fig. 2 se muestra la página de inicio del portal de publicaciones de redes sociales UNPA desarrollado. La información se presenta en tres columnas, una para cada red social. Los contenidos incluyen texto, imágenes y videos.

Figura 2: Página de inicio del portal

En la Fig. 3 se describen las consultas y las respuestas obtenidas de la API Graph de Facebook, Twitter y Youtube. En (a) se presenta la Url base (texto negro), seguida del identificador de la página de Facebook, en este caso se trata del área de extensión.unpa.uarg. Luego se especifican los campos que se requieren en la respuesta (texto rojo), en este caso los posts, el texto y la imagen de cada uno. También se indica la cantidad de posts que debería incluir la respuesta (texto naranja), por último se debe incluir el token de acceso del usuario (texto verde). Con el campo access token y los permisos adecuados es posible realizar diversas solicitudes a esta API web (lectura, eliminación, modificación y agregación). En (b) se muestra la respuesta con los 25 últimos posts, con la imagen y texto como se especificó en la consulta.

Luego se replica la consulta y la respuesta de la API de Twitter, en (c) se observa la url base seguida del identificador de la cuenta (texto azul), se indican los campos que se esperan recuperar de los tweets (texto rojo), a continuación se define la cantidad de tweets requeridos (texto violeta). El campo de texto de tweet es retornado por default, sin embargo es posible especificar algunos campos más como la fecha de creación y la imagen asociada si existen (texto color naranja). Por último se debe indicar el tipo de autorización. En (d) se presenta la respuesta al igual que el caso anterior, con los cinco últimos tweets de la cuenta sobre la que se realizó la consulta. Cada tweet posee un id de identificación, la fecha de creación, el texto del tweet y una imagen. La API de Twitter no solicita permisos para realizar lectura de contenido público dentro de la red social. Para las acciones de modificación, eliminación y agregación es necesaria la autenticación OAuth 1.0.

Por último, se especifica la consulta y la respuesta de la API de YouTube, en (e) se identifica la Url base utilizada (texto negro). La búsqueda se realiza con el ID del canal (texto azul), la cantidad máxima de resultados deseados (texto fucsia), a continuación se ordenan por fecha (texto verde), luego se indica el tipo de contenido requerido (color violeta), en este caso video. Finalmente se debe proporcionar la clave de api (texto naranja). En (f) se muestra la respuesta resultante en formato JSON, es

extensa por lo tanto se deben filtrar solo los campos que serán presentados en la aplicación final, a diferencia de las otras APIs que devuelven respuestas resumidas.

Figura 3. Consultas y respuestas a las APIs consumidas.

3.1 Registro de anomalías

El registro de anomalías (RA) genera como resultado un archivo JSON, en esta instancia los algoritmos verifican solo respuestas vacías y respuestas mal formadas. El RA se crea desde el inicio de la aplicación, cada vez que se ejecuta y/o invoca una consulta a una API web, un proceso analiza el archivo de respuesta y crea un objeto por cada anomalía detectada. Cada objeto se describe mediante los siguientes campos: tipo, API, contenido, número de Post o Tweet, fecha y hora. En la Fig. 4 se presenta un objeto creado en el archivo que registra la actividad.

Figura 4. Registro de anomalías

Para verificar la existencia de anomalías en las respuestas de las APIs web consumidas por el mashup desarrollado se han diseñado dos algoritmos. En ambos casos, los posteos se copian en arreglos bidimensionales, y se recorren analizando las condiciones que representan las anomalías. El algoritmo 1 verifica si la respuesta contiene respuestas vacías, y el algoritmo 2 examina la respuesta en busca de respuestas mal formadas. A continuación, se expresan estos algoritmos.

```
Algoritmo 1
Entrada: Json con los posts (JP)
Salida : entradas en el registro de actividades (RA)
MP ← Generar un arreglo bidimensional con JP
FOR EACH fila de MP
  FOR EACH columna de fila
      IF MP[fila][columna] = vacia
         registrar [fila][columna] en RA
      endIf
  endForeach
endForeach
```

```
Algoritmo 2
Entrada: Json con los posts (JP)
Salida : entradas en el registro de actividades (RA)
MP ← Generar un arreglo bidimensional con JP
FOR EACH fila de MP
  FOR EACH columna de fila
      IF jsondecode(MP[fila][columna])
         registrar MP[fila], [columna]en RA
      endIf
  endForeach
endForeach
```

La estructura flexible del RA posibilita su utilización en otros entornos. La información que contiene es de utilidad para tomar de decisiones respecto de las APIs. Con este fin se plantean distintas consultas sobre las anomalías que afectan la aplicación mashup. En la Fig.5 se destacan los campos a ser chequeados para responder a algunas de estas consultas.

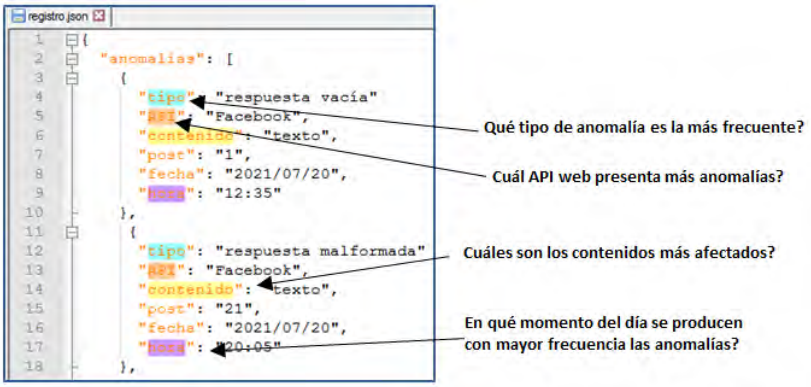


Figura.5. Consultas sobre el registro de anomalías

4. Ejemplos del RA

A continuación, se presentan ejemplos del registro de anomalías detectadas sobre el portal de publicaciones de la UNPA.

En la Fig.6 se expone un ejemplo de respuesta obtenida de la API de YouTube, que retorna cuatro campos con respuestas vacías (a). En el primer post no fue posible recuperar el video, en el segundo la publicación está completa, con texto y video. En el tercer post los campos de texto y video están vacíos por lo tanto se pierde el post completo. En el cuarto posteo el campo de video está vacío al igual que en el primer post. Los detalles de los problemas detectados se presentan en (b).

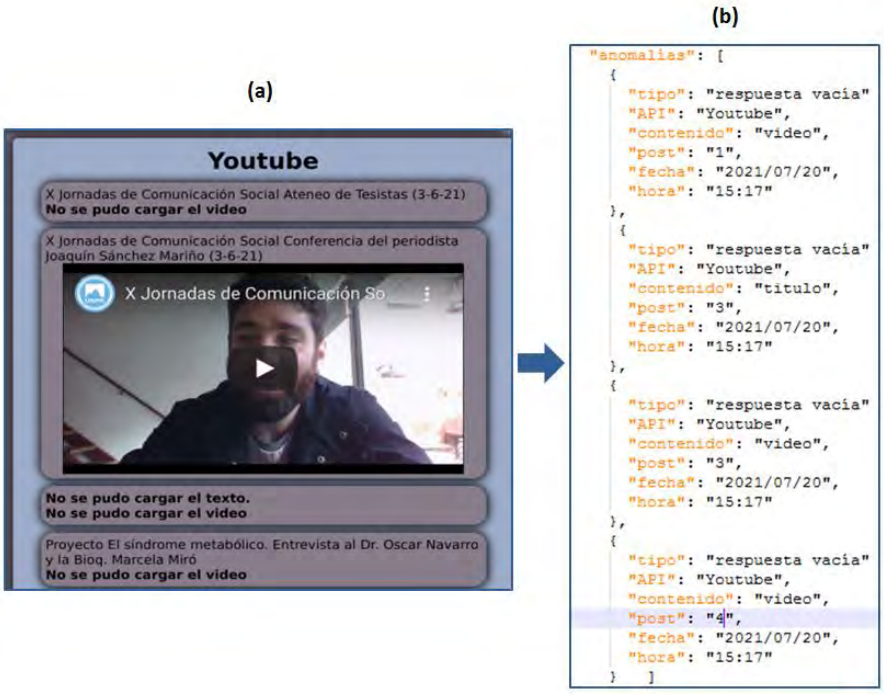


Figura 6. Respuestas vacías en YouTube

En la Fig.7 se muestra el portal de publicaciones y el contenido del RA. En este caso, (a) se obtuvo una respuesta mal formada de la API de Twitter por esta razón la columna correspondiente a esta red social se encuentra vacía y sólo aparece un mensaje al usuario informando que no se pudieron recuperar los tweets. En las otras APIs no se detectaron problemas en el contenido de las respuestas. La anomalía se refleja en (b) con la información sobre la respuesta mal formada identificada. En este caso, además del tipo, del nombre de la API web, la fecha y hora se incorpora un campo con un mensaje.

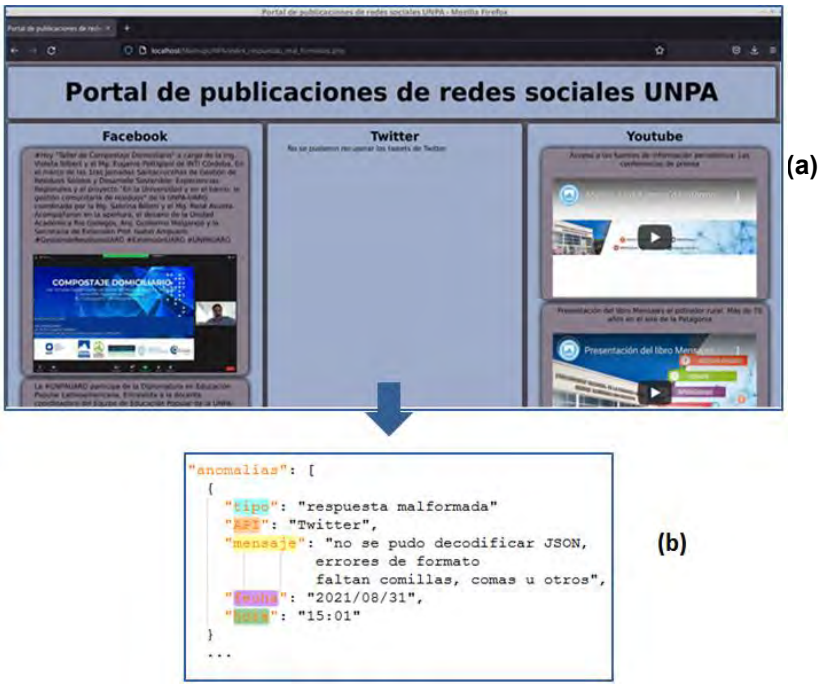


Figura 9. Respuesta mal formada de Twitter.

5. Trabajos relacionados

Existen diversos trabajos que presentan estudios enfocados a problemas originados en el uso de las APIs web. En [11][12] proponen herramientas para identificar errores y usos indebidos de APIs en aplicaciones JavaScripst. Ambas propuestas a partir de las especificaciones de las APIs realizan comprobaciones estáticas, que pueden identificar errores en la codificación de las consultas. Estas herramientas pueden resolver algunos de los errores de ejecución por la evolución de las APIs o los enumerados por [13], sin embargo, no atiende las causas de las anomalías de respuestas vacías o mal formadas

Las APIs web pueden generar diversos problemas de ejecución en la web mashup y/o cualquier aplicación que las consuman, por diversos factores, entre estos, los cambios en las APIs, producto de la evolución de las mismas. Cambios en los puntos finales, en las operaciones, en los parámetros y respuestas de la API web son reportados en diversos estudios, tales como [8][10][14][15]. Algunos de estos cambios podrían generar problemas en la ejecución, pero no parecen estar relacionadas a las anomalías que este trabajo aborda.

En [13] analizan 2,43 millones de respuestas (logs) de error de la API del servicio de pago Adyen. El objetivo del estudio es comprender las fallas que ocurren en la integración de la API web, las cuales potencialmente pueden resultar en problemas de producción para los consumidores de una API. Los resultados muestran que, (a) las fallas en la integración de la API se pueden agrupar en 11 causas generales: entrada de usuario no válida, entrada de usuario faltante, datos de solicitud caducados, datos de

solicitud no válidos, datos de solicitud faltantes, permisos insuficientes, procesamiento doble, configuración, datos de servidor faltantes, internos y de terceros, (b) la mayoría de las fallas pueden atribuirse a solicitudes no válidas o datos faltantes. Observamos que las anomalías de respuestas mal formadas y respuestas vacías, no están explícitamente abordadas, podría ser debido a que no son directamente registradas en el log del servidor de la API o bien porque su frecuencia de ocurrencia es baja o suceden en forma inadvertida para la API web. También podrían ser atribuidas a la tipificación como causa de fallos "internos". Por otro lado, el estudio se basa en el análisis de errores (log) del lado del productor de la API, mientras que nuestra propuesta plantea el análisis de errores del lado del consumidor (mashup).

Otros trabajos [16][17][18] han usado los registros de logs de las APIs web para realizar diversos análisis de estructura, calidad, usabilidad, etc. Nuestro trabajo propone usar la técnica de logs personalizados, pero para uso de los consumidores de APIs web.

6. Conclusiones

Este trabajo abordó la problemática para desarrolladores de web mashup relacionada a anomalías en la ejecución de los APIs web, conocidas como respuestas mal formadas y vacías. En particular, a partir del desarrollo de un web mashup de publicaciones en redes sociales, se implementaron algoritmos para detectar estas situaciones en las respuestas de las APIs de Facebook, Twitter y YouTube y registrar dichos eventos de manera específica.

Este registro de actividades (RA) o log personalizado es una estrategia que permite a los desarrolladores de mashup identificar y manejar estas situaciones. Esta estructura, permite al desarrollador, no solamente identificar los problemas en las respuestas de las APIs que utiliza la aplicación mashup, además es posible utilizar la información detallada en el mismo en otras etapas o por otras aplicaciones.

Una mejora a nuestra propuesta consistirá en proporcionar una solución más reusable y modular, que, a partir de una librería, diversos métodos o funciones, analicen las respuestas de las APIs y en consecuencia actualicen el registro de actividades. Sin embargo, es un gran desafío manejar la heterogeneidad de las estructuras de las respuestas de las APIs web.

El trabajo futuro consiste en continuar la personalización del RA, incorporar más anomalías al análisis, y proponer un enfoque más reusable.

7. Referencias

1. Yee, R. Pro Web 2.0 Mashups: Remixing Data and Web Services. Editorial: Apress. ISBN: 978-1-59059-858-0. (2008)
2. Dávila Macías & A. M.. Tesis. Recuperado a partir de http://repositorio.ug.edu.ec/handle/redug/6767. (2012)
3. Daniel F., Muhammand I., Soi S., De Angeli A., Wikinson C., Casati F., & Marchese M. “Developing Mashup Tools for End.Users: On the Importance of the Application Domain”, International Journal on Next-generation Computing, Vol 2. No 2, (2012).
4. Robbes R., Lungu M. & Janes A.“API fluency” ICSE-NIER '19: Proceedings of the 41st International Conference on Software Engineering: New Ideas and Emerging Results. pp 97–100. https://doi.org/10.1109/ICSE-NIER.2019.00033 (2019)

5. Curbera F., Duftler M., Khalaf R., Nagy W., Mukhi N., & Weerawarana S. “Unraveling the web services web: an introduction to SOAP, WSDL, and UDDI,” Internet Computing, vol. 6, no. 2, pp. 86–93. (2002)
6. Vinoski S. “Restful web services development checklist,” IEEE Internet Computing, vol. 12, no. 6, pp. 96–95. (2008)
7. Dekel U. & Herbsleb J.D. “Reading the documentation of invoked API functions in program comprehension” ICPC’09, pp. 168–177.(2009)
8. Espinha T., Zaidaman A. & Gross HG. Web API Fragility: How Robust Is Your Web API Client?. Software Engineering Research Group - Department of Software Technology. ISSN 1872-5392. (2014)
9. Raemaekers S., van Deursen A., & Visser J. “Measuring software library stability through historical version analysis,” in Proc. Int’l Conf. on Software Maintenance (ICSM). IEEE CS. pp. 378–387. (2012)
10. Sohan S.M, Anslow C. & Maurer F. A Case Study of Web API Evolution. IEEE World Congress on Services. Doi: 10.1109/SERVICES.2015.43.(2015)
11. Writtern E., Ying A., Zheng Y., Dolby J., & Laredo J. Statically checking web API requests in JavaScript. In Proceedings of the 39th International Conference on Software Engineering. IEEE Press.(2017).
12. SungGyeong B., Hyunghun Ch., Inho L. & Sukyoung R. SAFEWAPI: web API misuse detector for web applications. In Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering. DOI:https://doi.org/10.1145/2635868.2635916. (2014)
13. Aué, J., Aniche, M., Lobbezoo, M., & van Deursen, A. An Exploratory Study on Faults in Web API Integration in a Large-Scale Payment Company. In ICSE-SEIP '18: 40th International Conference on Software Engineering: Software Engineering in Practice Track (pp. 13-22). doi.org/10.1145/3183519.3183537.(2018)
14. Li J., Xiong Y., Liu X., & Zhang L. “How does web service API evolution affect clients?” 20th Conf. on Web Services (ICWS). IEEE, pp. 300–307.(2013)
15. Fokaefs M., Mikhaiel R., Tsantalis N., Stroulia E., Lau A.An Empirical Study on Web Service Evolution. IEEE International Conference on Web Services.(2011)
16. Koçi, R., Franch, X., Jovanovic, P., & Abelló, A. Improving Web API Usage Logging. arXiv preprint arXiv:2103.10811.(2021)
17. Suter P. & Wittern E. Inferring web API descriptions from usage data. In HotWeb,IEEE, DOI: 10.1109/HotWeb36442.2015 (2015).
18. Macvean A., Daughtry J., Church J.& Citro C. “API Usability at Scale." Proceedings of the 26th annual workshop of the Psychology of Programming Interest Group (2016)

Strategy for Improving Source Code Compliance to a Style Guide

Pablo Becker, Luis Olsina, and María Fernanda Papa

GIDIS_Web, Facultad de Ingeniería, UNLPam, General Pico, LP, Argentina
[beckerp, olsinal, pmfer]@ing.unlpam.edu.ar

Abstract. This paper illustrates the evaluation and improvement of a Java source code considering the non-compliance with a selected set of items of the Google Java Style Guide. To do this, a strategy was used to understand and improve the Java source code. The strategy has activities that allow specifying non-functional requirements (characteristics and attributes) and designing and implementing measurement, evaluation, analysis, and change. The case was applied in the context of an advanced undergraduate course in System Engineering as a mandatory exam. The evaluation results of attributes' adherence to the aforementioned coding style guide and the improvement of non-compliances are discussed.

Keywords: Java Source Code, Google Java Style Guide, Compliance, Improvement, Evaluation Strategy.

1 Introduction

As stated by Elish *et al*. [6] "The use of agreed-upon coding practices is believed to enhance program comprehension, which directly affects reuse and maintainability". There are agreed coding conventions and style guides for different programming languages that try to improve the readability and maintainability of software source code. Recently, dos Santos *et al*. [4] conducted a field study with a set of 11 coding practices to find out the impact on code readability. Their findings were that 8 out of 11 coding practices had evidence of affecting readability.

Today, it is common for programmers and teams involved in industrial software projects to work with these coding practices. According to Broekhuis [3] "Teams adopt or adapt coding styles, and in some cases, they are mandatory. This means coding practices are an integral part of software development". In 2021, the author conducted a survey with 102 responses from professionals in the Netherlands, including 95 developers, 5 project managers, and 2 testers, and showed that more than 90% of the participants used coding styles in their software projects. As a synthesis of his study, he summarizes "It is, therefore, reasonable to conclude that they [coding conventions and style guides] have a critical role in industries. This could imply the necessity of teaching these coding styles to students".

The present work discusses the evaluation and improvement of a Java source code considering the non-compliance with a selected set of items of the Google Java Style

Guide [7]. This online document serves as the complete definition of Google's coding standards for source code in the Java programming language. As in any other existing coding style guide, in [7], there is a set of items or guidelines mainly in the categories 'Source file structure', 'Formatting', and 'Naming', among others, that the evaluated code must comply with.

The case that we show here was applied in the context of an advanced undergraduate course in System Engineering as a compulsory integrated exam, which regularly lasts around 35 days. The subject called Software Engineering II is taught in the first semester of the 5th year of the degree. The conceptual content of the subject deals with non-functional requirements, measurement, evaluation, and analysis of the quality of a software product or system. To apply these contents and promote the technical and transversal competencies of the students, each year, considering the problem to be solved, an evaluation strategy is selected, from a family of strategies [12].

In the current year (2022), we selected the strategy with the purpose of understanding and improving the quality of a candidate Java source code, considering the compliance of the code with a subset of items of [7]. Note that the code we provided to students was deliberately and slightly modified to partially comply with this coding guide.

The learning objectives were mainly twofold. First, as in any year, apply the concepts of characteristics, attributes, metrics, and indicators, as well as the concept of analysis of the situation for decision-making. These concepts and practices are embedded in the processes and methods of an evaluation strategy. Second, we consider it relevant that students as close future professionals learn software coding styles through practice, as suggested by Broekhuis. In summary, the main contribution of this work is to illustrate both aspects from a practical point of view. Since the study may be of interest to students of other similar degrees, the complete documentation of the case is linked to an additional resource.

The rest of the article is organized as follows. Section 2 overviews the improving strategy and the Google Java Style Guide. Section 3 describes a little more the context of the case study presented. Section 4 shows in detail the application of the aforementioned concepts and practices. Section 5 discusses related work and, finally, Section 6 summarizes conclusions.

2 Overview of the Evaluation Strategy and Coding Style Guide

In [12], a family of evaluation strategies guided by measurement and evaluation activities is presented. Those strategies allow achieving different purposes such as to understand, improve, monitor, compare and adopt, among others. In this work, the strategy called *Goal-Oriented Context-Aware Measurement, Evaluation and Change* (GOCAMEC) is used, which allows us to understand and improve the current state of an entity that in the present case is a Java source code. Fig. 1 shows the GOCAMEC process using the UML activity diagram and the SPEM notation.

As depicted in Fig. 1, the process begins by performing the *Define Non-Functional Requirements* (NFRs) activity (A1), which aims to define the quality attributes and characteristics to be evaluated. A1 has as input a quality model (e.g. those prescribed in [8] and [9]) and produces a "NFRs Specification", which includes "NFRs Tree".

In the *Design Measurement and Evaluation* activity (A2), metrics and indicators are defined or selected from a repository. Then, the *Implement Measurement and Evaluation* activity (A3) implies obtaining the measures and indicator values.

In A4.1 the analysis is designed, which includes, among other aspects, establishing the criteria for the analysis of the results. As seen in Fig. 1, A4.1 can be performed in parallel with A3. Next, in the *Analyze Results* activity (A4.2), the measures, the indicator values, and the "Analysis Specification" are used as input, to produce the "Conclusion/Recommendation Report". The objective of this activity is to detect weaknesses in the evaluated entity and recommend changes.

If there are no recommendations for changes, for example, because the level of satisfaction achieved is optimal, the process ends. But, if there are recommendations for change due to detected weaknesses, *Design Changes* (A5) is carried out, generating an "Improvement Plan" in which the specific changes to be made are indicated. Then, the plan serves as input to *Implement Changes* (A6). The result is a new version of the entity under study.

As shown in Fig. 1, once A6 activity is finished, A3 must be executed again in order to carry out the measurement and evaluation of the new entity. Based on these new results, A4.2 analyzes whether the changes have increased the level of satisfaction achieved by the NFRs. If the improvement is not enough to reach the main business goal, new cycles of change, re-evaluation, and analysis can be carried out until the goal established by the organization is reached.

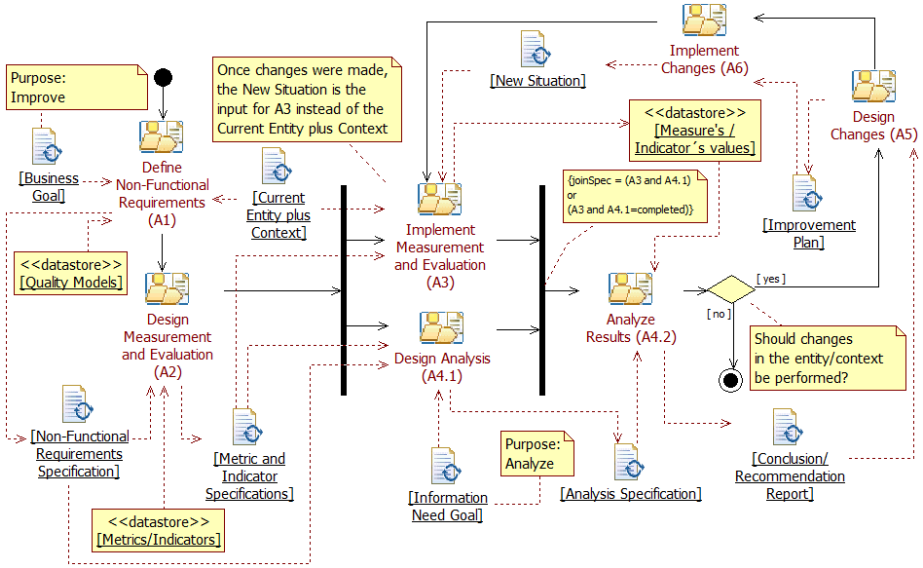


Fig. 1. Generic process specification for the GOCAMEC strategy.

On the other hand, regarding the coding style guide, we use the Google Java Style Guide. This guide includes sections related to: *Source file basics* (that deals with filenames, file encoding, whitespace characters, and special characters), *Source file structure* (that deals with license information, package and import statements, and class member ordering), *Formatting* (that deals with braces, indents, line wrapping,

whitespace, parentheses, enums, arrays, switch statements, annotations, comments, and modifiers), *Naming* (that deals with identifiers such as package, class, method, constant, field, local variable, type variable), *Programming Practices* (that deals with @Override, exceptions, static members and finalizers), and *Javadoc* (pointing out how to format Javadoc and where it is required).

3 More Details of the Context Used for the Practical Case

As commented in the Introduction Section, the case was applied within the framework of the Software Engineering II subject in the first semester of the 5th year of the Systems Engineering degree. The study is illustrated in detail in Section 4 and here we describe a little more about the context.

Firstly, we selected the Java code since the students dedicate about 90 hours to the previous subject called Object Oriented Programming, in the first semester of the 3rd year, using this language to program a video game. The given program for this case named "GUICalculator.java" has 116 lines and is available in Annex IV at http://bit.ly/CACIC_Annexes. We deliberately modified this source code a bit to introduce some violations of the [7] coding conventions. Code lines with at least one evaluated incident are shaded orange in this Annex.

Secondly, the Software Engineering II course assesses students' technical skills on non-functional requirements (using characteristics and attributes to specify non-functional requirements), measurement (using metrics), and evaluation (using indicators) of entities. Note that "establishing software metrics and quality standards" is a specific competency in the new curricular standard in Argentina for Information Systems careers. Thus, the specification of quality requirements and the design and implementation of metrics and indicators are mandatory concepts and practices to pass this course. Consequently, for the presented problem to students, we established that one attribute (as an elementary quality requirement) must be mapped to a single item of the Google Java Style Guide. The maximum number of attributes was 8 mapped to 8 items of the guide. In the present work, we expand the scope of the assignment given to students, by including 11 attributes and by specifying 3 attributes for 1 item of the guide. Specifically, to the "*3.3.3 Ordering and spacing*" item in [7], we evaluate the adherence of the GUICalculator.java code to it by using the following attributes: *1.1.1. Compliance with the ordering of types of imports*; *1.1.2. Spacing compliance between static and non-static import blocks*; and, *1.1.3. Spacing compliance between import sentences*. In total, the present work evaluates 11 attributes mapped to 9 items of the guide. This will be illustrated in detail in the next section.

By allocating this problem to students that represents an integrated exam, we promote group work. In the current year (2022), there were 10 students (one international, by institutional exchange). The sizes of the groups were 1 with three members, 3 with two members, and 1 student who decided to work alone. There were slightly different restrictions regarding the size of the group, such as the number of attributes/characteristics and the size of the monograph as the final report to be examined. For instance, the group of 3 students must have requirements specified by two sub-characteristics of

Compliance [8] and 8 attributes. Additionally, they had to inspect the code beforehand to ensure that the requirements tree included at least 3 attributes that would have to be changed later in the code to fully comply with the guide. The resulting monographs ranged from 33 to 68 pages, including appendices.

Lastly, we give an account of some transversal competencies of the students, encouraging group work, providing all the material in English, and promoting oral and written communication skills in the native language.

4 Application of the Evaluation Strategy to Improve Code Compliance to Google Java Style Guide Items

This section illustrates the application of the GOCAMEC strategy to improve the compliance of the "GUICalculator.java" source code to the Google Java Style Guide. To achieve this goal, the strategy used allows: i) understand the degree of compliance of the source code to the style guide; ii) based on non-compliances apply changes to the current version of the source code (v. 1.0) to improve compliance with the style guide; and iii) understand the degree of source code compliance after the changes (that is, to the version 1.1 of the code). The activities carried out are illustrated below.

Table 1. Google Java Style Guide items mapped to characteristics and attributes related to "*Maintainability*" and "*Compliance*" of a Java source code and its evaluation results (in [%]). Note: The symbol ● means "Satisfactory"; ◆ "Marginal", and ■ "Unsatisfactory". Additionally, op stands for "operator", EI for "Elementary Indicator" and DI for "Derived Indicator".

Google Java Style Guide Items	Characteristics and Attributes (*in italic*)	op	v1.0 EI/DI	v1.1 EI/DI
	1 Maintainability		73.74 ◆	100 ●
	1.1 Compliance	C-+	73.74 ◆	100 ●
3. Source file structure	1.1.1 Source file structure compliance	A	53.33 ■	100 ●
3.3.3 Ordering and spacing	*1.1.1.1 Compliance with the ordering of types of imports*		100 ●	100 ●
	1.1.1.2 Spacing compliance between static and non-static import blocks		100 ●	100 ●
	1.1.1.3 Spacing compliance between import sentences		0 ■	100 ●
3.4.1 Exactly one top-level class declaration	*1.1.1.4 Compliance with the number of top-level class declarations per source file*		33.33 ■	100 ●
4. Formatting	1.1.2 Formatting compliance	*A*	72.10 ◆	100 ●
4.1.1 Use of optional braces	*1.1.2.1 Compliance with the use of optional braces*		9.09 ■	100 ●
4.8.2.1 One variable per declaration	*1.1.2.2 Compliance with the number of variables per declaration*		82.61 ◆	100 ●
4.3 One statement per line	*1.1.2.3 Compliance with the number of statements per line*		88.24 ◆	100 ●
4.4 Column limit: 100	*1.1.2.4 Compliance with the maximum line size*		95.15 ◆	100 ●
5. Naming	1.1.3 Naming compliance	*C--*	82.96 ◆	100 ●
5.2.2 Class names	*1.1.3.1 Class naming compliance*		60.00 ■	100 ●
5.2.3 Method names	*1.1.3.2 Method naming compliance*		100 ●	100 ●
5.2.6 Parameter names	*1.1.3.3 Parameter naming compliance*		100 ●	100 ●

(A1) Define Non-Functional Requirements: For this activity, we consider the quality models prescribed in [8] and [9]. Since adherence to a coding style guide favors the source code maintainability, we use the model for external and internal quality proposed in [8]. This quality model includes the "*Maintainability*" characteristic, which in turn explicitly includes the "*Compliance*" sub-characteristic that was removed from [9]. The "*Compliance*" characteristic is defined as "*The degree to which the software product* (e.g. the source code) *adheres to standards or conventions relating to maintainability*". Then, from the Google Java Style Guide we select a set of items to evaluate (see Table 1, 1st column) and for each item, we define one or more attributes. E.g., for item *3.3.3 Ordering and spacing* we define 3 attributes while for item *3.4.1 Exactly one top-level class declaration* we define a single attribute. The 2nd column of Table 1 shows the identified attributes and their mapping to the guide items. All the characteristic and attribute definitions can be seen in Annex I at http://bit.ly/CACIC_Annexes.

(A2) Design Measurement and Evaluation: In this activity, a set of metrics were defined to quantify all the attributes. E.g., to quantify the attribute "*Compliance with the number of top-level class declarations per source file"* (coded 1.1.1.4 in Table 1) the indirect metric "*Percentage of top-level class declarations per source file*" (%TLC) was defined. Table 2 shows the specification of this indirect metric. Additionally, for each indirect metric, one or more direct metrics were specified. E.g., for the indirect metric %TLC, the direct metric "*Availability of valid top-level class*" (AVTLC) was defined. The measurement procedure for this metric was defined as follows: "*AVTLC = 0; if (class declaration defines a top-level class) and (class name is equal to the source file name) then AVTLC = 1;*"

It is important to say that to clarify the measurement procedures, sometimes some notes were included. E.g., for the previous measurement procedure we include the following notes: "*1. A top-level class is any class that is not a nested class. A nested class is any class whose declaration occurs within the body of another class or interface. 2. The keyword "class" is the tag for any class declaration in Java. 3. Class name and source file name are case-sensitive*".

The rest of the indirect and direct metrics defined for this work can be found in Annex II of the document available at http://bit.ly/CACIC_Annexes.

Table 2. Indirect metric specification to quantify the "*Compliance with the number of top-level class declarations per source file*" attribute coded 1.1.1.4 in Table 1.

Metric Name: Percentage of top-level class declarations per source file (**%TLC**)
Objective: Determine the percentage of valid top-level classes with respect to the total of top-level classes in the source code to be measured. **Author:** Pablo Becker and Luis Olsina **Version:** 1.0
Calculation Procedure Formula: $\%TLC = \left(\frac{\sum_{i=1}^{\#JF}\sum_{j=1}^{\#TLC} AVTLC_{ij}}{\sum_{i=1}^{\#JF} \#TLC_i}\right) * 100$
Scale: Numeric Scale Type name: Ratio Value Type: Real Representation: Continuous
Unit: Name: Percentage Acronym: %
Related Direct Metrics: AVTLC: Availability of valid top-level class; #TLC: Number of top-level classes; #JF: Number of Java files

Since the measured values do not represent the level of satisfaction of an elementary requirement (attribute), a transformation must be performed that converts the measured

value into a new value that can be interpreted. Therefore, for each attribute, an elementary indicator was specified. For this work, the elementary indicator for the attribute 1.1.1.4 is specified in Table 3. The rest of the elementary indicators can be found in Annex III at http://bit.ly/CACIC_Annexes.

Derived indicators were also defined to interpret the requirements with a higher level of abstraction, that is, the characteristics and sub-characteristics documented in Table 1. For all these indicators, an aggregation function named Logic Scoring of Preference (LSP) [5] was used whose function is:

$$DI(r) = (w_1 * I_1^r + w_2 * I_2^r + ... + w_m * I_m^r)^{1/r}$$

where DI represents the derived indicator to be calculated and I_i are the values of the indicators of the immediate lower level, or grouping in the tree, in a range $0<=I_i<=100$; w_i represents the weights that establish the relative importance of the elements within a grouping and must comply with $w_1 + w_2 + ... + w_m = 1$, and $w_i > 0$ for $i = 1 ... m$; and r is a coefficient for LSP operators. These operators model different relationships among the inputs to produce an output. There are operators (op) of simultaneity or conjunction (operators C), replaceability or disjunction (operators D), and independence (operator A). LSP operators for this work are shown in the 3rd column of Table 1.

As shown in Table 3, the elementary and derived indicators have the same three acceptability levels. We decided to use the traffic light metaphor to facilitate the visualization of the levels of satisfaction achieved: ■ red / Unsatisfactory (values less than or equal to 60%), ◆ yellow / Marginal (values greater than 60% and less than 100%) and ● green / Satisfactory (values equals to 100%).

Table 3. Elementary indicator specification for the attribute "*Compliance with the number of top-level class declarations per source file*" (coded 1.1.1.4 in Table 1). Note: %TLC stands for the "*Percentage of top-level class declarations per source file*" metric.

Name: Performance Level of the Compliance with the number of top-level class declarations per source file (PL_TLC)	
Author: Santos L.	**Version:** 1.1
Elementary model: **Specification:** the mapping is PL_TLC = %TLC **Decision criterion (3 acceptability levels):** **Name 1:** ■ Unsatisfactory; **Range:** [0 ; 60] **Description:** Indicates that corrective actions must be performed with high priority. **Name 2:** ◆ Marginal; **Range:** (60 ; 100] **Description:** Indicates that corrective actions should be performed. **Name 3:** ● Satisfactory; **Range:** [100 ; 100] **Description:** Indicates that corrective actions are not necessary since the attribute meets the required quality satisfaction level.	
Numerical Scale: **Value Type:** Real **Scale Type:** Ratio **Unit:** **Name:** Percentage **Acronym:** %	

(A3) Implement Measurement and Evaluation: This activity produces the measures and indicators' values. E.g., for attribute 1.1.1.4 the value was 33.33%. This derived measure is produced by applying the calculation procedure specified in Table 2. All the base measures (used to calculate this and other derived measures) can be seen in Annex V of the document at http://bit.ly/CACIC_Annexes.

Then, the derived measures were used to calculate the values of elementary indicators and the latter to calculate the derived indicators during the evaluation. Indicators' values both for elementary and derived indicators are shown in Table 1, 4th column.

(A4.1) Design Analysis: Concurrently to A3, the A4.1 activity was carried out. In our case, it was decided to classify the attributes following the decision criteria defined for the indicators (see Table 3). Those attributes that fall into the Unsatisfactory range (■) would be the first to receive attention, and then those that fall into the Marginal range (◆). It is important to say that a guide item reaches the Satisfactory level only if all the mapped attributes fall into the Satisfactory (●) level.

(A4.2) Analyze Results: Following the guidelines of the "Analysis Specification", the values of the 4th column of Table 1 were analyzed and improvements were recommended for the attributes with a low level of performance (values marked with ■ and ◆). E.g., under the "*1.1.1 Source file structure compliance*" characteristic there are 2 attributes with a low level of performance. So, for the "*Spacing compliance between import sentences attribute"* (coded 1.1.1.3) which reached 0%■ the recommendation was "*Blank lines between import sentences must be eliminated*", and for the attribute coded 1.1.1.4 which reached 33.33%■, the recommendation was "*Each top-level class must be defined in a file named as the class considering that names are case-sensitive*".

Considering the "*1.1.3 Naming compliance*" characteristic, the recommendation for the "*Class naming compliance*" (1.1.3.1) –which reached 60%■- was: "*All class names must be in upper camel case*". Similarly, recommendations for each attribute under the "*1.1.2 Formatting compliance*" characteristic were made.

(A5) Design Changes: Using the "Recommendation Report" generated in A4.2, the changes to be made were designed. E.g., to improve the level of satisfaction achieved by the attribute coded 1.1.3.1, it was proposed that the classes named "calculatorFrame" and "calculatorpanel" were renamed as "CalculatorFrame" and "CalculatorPanel", respectively. Additionally, to improve the attribute coded 1.1.1.4, it was proposed that the classes named "CalculatorFrame" and "CalculatorPanel" (which are top-level classes) be defined in other source files, which should be called "CalculatorFrame.java" and "CalculatorPanel.java", respectively. All proposed changes were recorded in the "Improvement Plan" document.

(A6) Implement Changes: In this activity, the changes proposed in the "Improvement Plan" were made. The reader can find the new version (v1.1) of the source code in Annex VIII at http://bit.ly/CACIC_Annexes.

As prescribed by the GOCAMEC process (recall Fig. 1), once A5 and A6 activities were completed, a re-evaluation must be performed. Therefore, A3 and A4.2 activities were enacted again to determine the level of satisfaction achieved by the new version of the source code after the changes.

(A3) Implement Measurement and Evaluation: In this second execution of A3, the same metrics and indicators were used on the new source code (v 1.1). The results obtained are shown in the 5th column of Table 1.

(A4.2) Analyze Results: As can be seen in column 5th of Table 1, all the attributes reached 100% (●), that is, the new version of the source code satisfies all the Google Java Style Guide items considered for this work. Since new cycles of change, re-evaluation, and analysis are not required because the goal was successfully achieved, the process is finished.

5 Related Work and Discussion

Coding conventions for programmers to follow have been proposed since the mid-1970s. One of the most cited examples is Kernighan *et al*. [10], which gave many hints on how to write readable code in C language using real software cases.

Particularly for the Java language, the best-known coding style guidelines and conventions emerged from the work of Sun Microsystems in 1997 [15], and of Reddy in 2000 [14], who was also a member of this company. After these proposals, the Google Java Style Guide [7] appeared. We selected this guide for the current case, as it has the main categories and items to make code readable, as well as easy formatting and online access. Another recent reference for Java coding conventions and practices is Bogdanovych *et al*. [2], to name just a few.

Many studies and experiments of different coding styles that affect code readability have emerged, such as those by Lee *et al*. [11], dos Santos *et al*. [4], to mention just a couple of those works carried out so far. As commented in the Introduction Section, according to the survey conducted by Broekhuis [3], out of 102 responses from industry professionals, only 3% did not use a coding style in their software projects. This can emphasize the role that the learning process in academia should continue to play in these beneficial concepts and practices. For the case shown, this was one of the learning objectives established in Software Engineering II.

To the best of our knowledge, what is not present in related works is the mapping of non-functional requirements in the form of characteristics and attributes with categories and items from the coding style guides, as illustrated in Section 4. This mapping enables systematic understanding and improvement of source code compliance by using metrics, indicators, and refactoring as methods for performing the measurement, evaluation, and change activities. In turn, these methods and activities are well established and specified in GOCAMEC. The employment of these concepts and practices was another of the learning objectives established in Software Engineering II. For this learning objective, the use of tools and analyzers was not promoted as is done in other works, but the design of metrics and indicators, and the elaboration of code changes manually from the data recorded from the implementation. However, the implementation of all these methods and activities for code evaluation and refactoring could be automated.

6 Final Remarks

In this paper, we have discussed the quality evaluation and improvement of a typical Java source code by considering the compliance with internal quality attributes properly mapped to a set of Google Java Style Guide elements. To carry out this study, the GOCAMEC strategy was used, which allows us not only to understand the current state of the entity but also to design and implement changes that positively affect the quality of the new version of the entity.

For the problem posed to the students in the context of an undergraduate course, the resulting Java source code was improved by justifying the different steps and results.

Additionally, the learning objectives of the subject and the skills and capabilities expected after passing were commented as well.

To conclude, we would like to highlight that what is not present in the literature regarding related works is the correspondence of quality (compliance) requirements in the form of characteristics and attributes with categories and items of the coding style guides and conventions, as illustrated in the previous sections. In future work, we are planning the automation of the presented approach, which can be an assignment for a student thesis in System Engineering.

Acknowledgment. This line of research is supported partially by the Engineering School at UNLPam, Argentina, in the project coded 09/F079.

References

1. Becker, P., Tebes, G., Peppino, D., Olsina, L.: Applying an Improving Strategy that embeds Functional and Non-Functional Requirements Concepts, Journal of Computer Science and Technology, 19:(2), pp. 153–175, doi: 10.24215/16666038.19.e15, (2019).
2. Bogdanovych, A., Trescak, T.: Coding Style and Decomposition. In: Learning Java Programming in Clara's World. Springer Nature Switzerland, Chap. 4, pp. 83-100, https://doi.org/10.1007/978-3-030-75542-3_4, (2021).
3. Broekhuis, S.: The Importance of Coding Styles within Industries, 35th Twente Student Conference on IT (TScIT 35), pp. 1-8, (2021).
4. dos Santos, R. M., Gerosa, M. A.: Impacts of coding practices on readability. In: International Conference on Software Engineering, pp. 277-285, (2018).
5. Dujmovic, J.: Continuous Preference Logic for System Evaluation, IEEE Transactions on Fuzzy Systems, (15): 6, pp. 1082-1099, (2007).
6. Elish, M., Offutt J.: The adherence of open source java programmers to standard coding practices. In: 6th IASTED International Conference on Software Engineering and Applications, pp. 1-6, (2002).
7. Google Java Style Guide. Available at https://google.github.io/styleguide/javaguide.html, and Last Accessed June (2022).
8. ISO/IEC 9126-1: Software Engineering – Software Product Quality – Part 1: Quality Model, International Organization for Standardization, Geneva, (2001).
9. ISO/IEC 25010: Systems and Software Engineering – Systems and software product Quality Requirements and Evaluation (SQuaRE) – System and software quality models, (2011).
10. Kernighan, B. W., Plauger, P. J.: The elements of programming style. McGraw-Hill, New York, 1st Ed., (1974).
11. Lee, T., Lee, J. B., In, H. P.: A study of different coding styles affecting code readability. Int'l Journal of Software Engineering and Its Applications, 7:(5), pp. 413-422, (2013).
12. Olsina, L., Becker, P.: Family of Strategies for Different Evaluation Purposes. In XX CIbSE'17, Published by Curran Associates, pp. 221–234, (2017).
13. Oman, P. W., Cook, C. R.: A paradigm for programming style research. ACM SIGPLAN Notices 23:(12), pp 69-78, https://doi.org/10.1145/57669.57675, (1998).
14. Reddy, A.: Java™ coding style guide. Sun MicroSystems, (2000).
15. Sun Microsystems: Java code conventions, Available at https://www.oracle.com/technetwork/java/codeconventions-150003.pdf, (1997).

UN MÉTODO PARA DEFINIR REQUISITOS DE CALIDAD DE DATOS EN CONTEXTO DEL DESARROLLO ÁGIL CON SCRUM

Carrizo Claudio[*], Javier Saldarini[*], Angélica Caro[#], Carlos Salgado[+], Alberto Sánchez[+], Mario Peralta[+]

[*]Facultad Regional San Francisco – Universidad Tecnológica Nacional
Av. de la Universidad 501 - San Francisco - Córdoba - Tel. 03564-421147
{cjcarrizo77, saldarinijavier}@gmail.com
[#]Departamento de Ciencias de la Computación y Tecnologías de la Información - Facultad de Ciencias Empresariales, Universidad del Bio Bio
Casilla 447, 3780000, Chillán, Chile
{mcaro}@ubiobio.cl
[+] Departamento de Informática - Facultad de Ciencias Físico-Matemáticas y Naturales Universidad Nacional de San Luis
Ejército de los Andes 950 – C.P. 5700 – San Luis – Argentina
{csalgado, alfanego, mperalta}@unsl.edu.ar

Abstract. Los Sistemas de Información son los encargados de proveer información a los usuarios dentro de las organizaciones, a fin de que estos puedan llevar adelante los procesos comerciales y la toma de decisiones. En muchas ocasiones, resulta un desafío contar con información adecuada, debido a problemas en la calidad de los datos, sobre todo en el momento en que se produce el dato. En el ámbito del desarrollo ágil, existen escasas propuestas que tengan un enfoque hacia el resguardo de la calidad de los datos. En este sentido, se observa que no es muy común la incorporación de requisitos de calidad de datos en etapas tempranas del desarrollo. El presente trabajo propone un método que permita guiar la especificación temprana e implementación de Requisitos de Calidad de Datos, definidos desde la perspectiva de la producción de datos. Este método está basado en un conjunto de Normas de la Serie de Estándares ISO/IEC 25000, y está enfocado en el contexto del desarrollo ágil, guiado por Scrum. También se proporciona una herramienta que permite facilitar y/o agilizar la aplicación del Método. Esta propuesta pretende ser un aporte de valor en pos de resguardar la calidad de datos de los productos de software.

Keywords: Requisitos de Calidad de Datos – SQuaRE – Desarrollo de Software – Metodologías Ágiles – Scrum

1. Introducción

Los Sistemas de Información (de aquí en adelante, SI) están presentes en todo tipo de organizaciones y son un componente fundamental para el éxito de los negocios, ya que su implementación permite obtener una ventaja competitiva [1]. En [2] se menciona que en un modelo de sistema, los sistemas informáticos son parte de un SI, y están compuestos por los siguientes elementos: hardware, sistema operativo, software de aplicación y datos. El software de aplicación, tiene entre otras funciones, la de procesar datos de entrada con el propósito de poder brindar información útil a

los usuarios en las organizaciones, para llevar adelante los procesos comerciales y la toma de decisiones [3]. Por otra parte, en [4] se representa la pirámide del conocimiento, en donde se denota que la información es un conjunto de datos que, al ser procesados, producen un significado sobre algún fenómeno en particular; por lo tanto, los datos son la materia prima de su sucesor. De lo anterior, se deduce la importancia de contar con datos que tengan un alto nivel de calidad, a fin de que la información resultante sea útil para los usuarios que la consuman. En [5] se menciona que la mala calidad de datos genera un alto costo e impacto en las organizaciones. En [6] se define la calidad de datos como "grado en que las características de los datos satisfacen necesidades implícitas y/o establecidas cuando son usados en condiciones específicas". Estas características, también conocidas como "dimensiones", son atributos que llevan a la calidad, y son medibles. En el ámbito de la calidad de datos, existe la Familia de Estándares de Calidad ISO/IEC 25000 [7], la cual propone un marco para la definición de requisitos de la calidad de software/datos y evaluación de la calidad del software/datos, apoyados por un proceso de medición de la calidad de software/datos. Dentro de este marco, existen las Normas ISO/IEC 25012:2008 [6] e ISO/IEC 25024:2015 [8], las cuales permiten respectivamente definir un modelo general de características de calidad de datos, y un conjunto de medidas de calidad asociadas a dichas características, para datos conservados en formato estructurado dentro de un sistema informático.

En lo que concierne a los procesos de desarrollo de software, en [9] se menciona que muchos de los problemas de calidad de datos provienen de los procesos de generación o producción de datos. En [10] también se destaca la importancia de incorporar aspectos de calidad de datos en el desarrollo de los requisitos, debido a que esta fase es crucial para el éxito o fracaso del SI. Actualmente, existe la Norma ISO/IEC 25030 [2] la cual proporciona un proceso que permite analizar y definir Requisitos de Calidad de Datos (RCD) en un contexto específico. Si bien existen trabajos que abordan la calidad de datos desde los requisitos, desde diferentes perspectivas como por ejemplo, portales web [11], aplicaciones web [12], modelos de procesos de negocio [13], en el ámbito de las metodologías ágiles se han encontrado pocas evidencias al respecto. En este sentido, en [14] se propone el desarrollo de un modelo de calidad de datos utilizando la metodología de sistemas blandos (SSM); en [15] se propone una adaptación del marco de trabajo de Scrum para incorporar calidad de producto; en [16] se menciona la posibilidad de enriquecer las historias de usuario, desde la característica de calidad "usabilidad".

Lo antes expuesto motivó a realizar una propuesta en el ámbito de la especificación e implementación de RCD, desde la perspectiva de la producción de datos, en el contexto del desarrollo ágil, usando Scrum. Para ello, se definió un método, que permite arribar a una especificación e implementación de RCD, los cuales están enfocados en producción de datos, y son provistos desde SQuaRE. Para facilitar el uso y/o aplicación del método, se desarrolló una herramienta para tal fin. Con esta propuesta, se desea realizar un aporte en el sentido de incorporar aspectos de calidad de datos, en el contexto del desarrollo, mediante el uso de metodologías ágiles.

El resto del artículo se organiza de la siguiente manera. En la Sección 2 se presentan las metodologías ágiles. En la Sección 3 se presenta la Serie ISO/IEC 25000. En la Sección 4 se presenta el método propuesto. En la Sección 5 se realiza

una discusión luego de la instanciación del Método en Scrum. Finalmente, en la Sección 6 se presentan las conclusiones y trabajos futuros.

2. Metodologías Ágiles

En la década de los noventa surgieron metodologías de desarrollo de software ligeras, más adelante nombradas como "metodologías ágiles". Estas se caracterizan por el desarrollo iterativo e incremental, las entregas frecuentes, la priorización de los requisitos, la constante interacción con el cliente, la adaptación al cambio, el trabajo colaborativo en equipo, etc. Estan basadas en el Manifiesto Ágil [17], que establece los siguientes valores: Individuos e interacciones, software funcionando, colaboración con el cliente y respuesta al cambio. Autores como Sommerville [18] y Pressman [19] coinciden en que las más reconocidas en la industria del software son: Scrum [20], Programación Extrema [21] y Crystal [22], entre otras. Por otra parte, según un estudio realizado por la Scrum Alliance [20], Scrum es la metodología ágil más utilizada actualmente, debido a que alrededor del 95% de los encuestados aseguran que utilizan prácticas de Scrum en la gestión de proyectos de software ágiles.

3. ISO/IEC 25000

La familia ISO/IEC 25000 [7], también conocida como "SQuaRE" (del inglés, System and Software Quality Requirements and Evaluation), proporciona una guía que permite la definición de requisitos y evaluación de la calidad de sistemas y del software. En el marco de este trabajo, se presenta en detalle las normas ISO/IEC 25012, ISO/IEC 25024 e ISO/IEC 25030.

3.1. ISO/IEC 25012: Modelo de Calidad de Datos

Esta Norma define un modelo general de calidad de datos, que persisten en formato estructurado. El modelo clasifica los atributos de calidad en 15 características desde dos puntos de vista: inherentes y dependientes del sistema (Tabla 1).

Tabla 1. Clasificación de características de calidad de datos según puntos de vista

Característica	Inherente	Dependiente del Sistema
Exactitud	X	
Completitud	X	
Consistencia	X	
Credibilidad	X	
Actualidad	X	
Accesibilidad	X	X
Conformidad	X	X
Confidencialidad	X	X
Eficiencia	X	X
Precisión	X	X
Trazabilidad	X	X
Comprensibilidad	X	X
Disponibilidad		X
Portabilidad		X
Recuperabilidad		X

3.2. ISO/IEC 25024: Medición de Calidad de Datos

Esta Norma define medidas de calidad para medir cuantitativamente la calidad de datos en términos de características definidas en la Norma ISO/IEC 25012. Dichas medidas incluyen métodos de medición y elementos de medida de calidad.

3.3. ISO/IEC 25030: Requerimientos de Calidad

Esta Norma permite definir requisitos de calidad, a través de un proceso de análisis y definición, que involucra características (ISO/IEC 25012) y medidas de calidad (ISO/IEC 25024). Dentro de los tipos de requisitos de calidad que existen, se encuentran los RCD, los cuales permiten especificar los niveles de calidad requeridos para datos que están asociados con el producto.

4. Método Propuesto

En esta sección se presenta el método, el cual permite guiar el trabajo de especificación temprana e implementación de RCD, los cuales están enfocados en procesos de producción de datos, y se obtienen a través de la Serie SQuaRe. El método está compuesto por 3 etapas y 5 actividades (Tabla 2).

Tabla 2. Etapas y Actividades del método

Etapa 1: Definición del Gestor de Calidad de Datos (DGCD)
DGCD.A1. Designación del Gestor de Calidad de Datos
Etapa 2: Especificación de Requisitos de Calidad de Datos (ERCD)
ERCD.A1. Incorporación de Historias de Usuario con Producción de Datos
ERCD.A2. Definición de Requisitos de Calidad de Datos (RCD)
ERCD.A3. Definición de Criterios de Aceptación para RCD
Etapa 3: Implementación de Requisitos de Calidad de Datos (IRCD)
IRCD.A1. Verificación de Cumplimiento de Criterios de Aceptación de RCD

Este Método se aplicó en el contexto de desarrollo guiado por la metodología ágil Scrum, debido a que es la más adoptada a nivel mundial. En la Figura 1 se pueden observar los momentos en que se llevan a cabo las 3 etapas dentro del flujo de Scrum.

Figura 1. Etapas del método aplicadas dentro del flujo de trabajo de Scrum

A continuación, se brindará mas detalle respecto de las etapas aplicadas en el contexto de Scrum. La Etapa 1 (E1) se aplica al comienzo del proyecto, antes del relevamiento de necesidades de las Partes Interesadas; el responsable de aplicar esta etapa es el Equipo Scrum. La Etapa 2 (E2) se aplica una vez que las historias de usuario están especificadas con un cierto grado de detalle, sobre todo, que incluyan los datos correspondientes; el responsable de aplicar esta etapa es el Gestor de Calidad de Datos (GCD). La Etapa 3 (E3) se aplica despues del desarrollo, más precisamente, cuando se realiza el aseguramiento de la calidad; el responsable de aplicar esta etapa es el Asegurador de la Calidad. Cabe acotar que el GCD también participa en esta etapa, accediendo al resultado de la implementación de los RCD.

El método propuesto cuenta con una herramienta denominada "Metodi", la cual fue desarrollada de manera ad-hoc por los autores de este trabajo, que permite dar soporte al despliegue de las actividades propuestas en el método.

También se llevó adelante la aplicación del método en 3 casos de estudio reales, a fin de poder poder validar no sólo el método, sino también la herramienta.

Por cuestiones de espacio, en este trabajo se presentará 1 caso de estudio, con la incorporación de 1 historia de usuario con producción de datos, 2 datos asociados y 3 requisitos de calidad por cada dato. El caso en cuestión consistió en el desarrollo de un sistema web de gestión de actividades agropecuarias.

Etapa 1: Definición del Gestor de Calidad de Datos (DGCD)

DGCD.A1. Designación del Gestor de Calidad de Datos

De acuerdo a lo sugerido en el método, en una reunión que tuvo una duración aproximada de 10 minutos, el Equipo Scrum designó al "Analista Funcional" como "Gestor de Calidad de Datos (GCD)", debido a que este miembro contaba con experiencia tanto en desarrollo de requisitos, como en aseguramiento de la calidad.

Etapa 2: Especificación de Requisitos de Calidad de Datos (ERCD)

En esta etapa se obtuvo como resultado un conjunto de documentos de especificación de RCD, asociados a historias de usuario con producción de datos.

ERCD.A1. Incorporación de Historias de Usuario con Producción de Datos

En primera instancia, el GCD identificó las HU con producción de datos en la Pila del Producto; luego las incorporó al método, haciendo uso de la herramienta, ingresando su identificador y nombre. Por último, ingresó el nombre de cada dato asociado a las HU, los cuales fueron identificados y consensuados con las partes interesadas, de acuerdo a su relevancia o criticidad para el negocio. En este trabajo se presentará la incorporación de la HU "US01. Registro de Usuarios", para los datos "Cuit" y "TelefonoCelular".

ERCD.A2. Definición de Requisitos de Calidad de Datos (RCD)

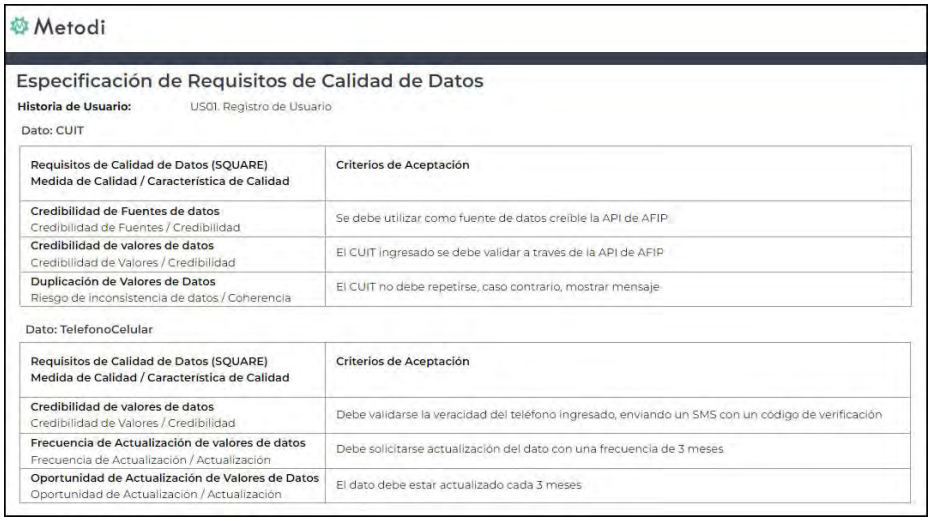
Por cada dato asociado a cada HU incorporada, el GCD se encargó de definir los RCD (obtenidos a través de ISO/IEC 25030), los cuales se expresan de manera amigable, para que puedan interpretarse facilmente. Además, se registró también la regla de negocio asociada al RCD, la cual fue relevada en instancia de entrevistas con las Partes Interesadas. En la Tabla 3 se puede observar el resultado de esta actividad.

Tabla 3. Resultado de la definición de RCD para "US01. Registro de Usuario"

US01. Registro de Usuarios	
Cuit	
Requisitos de Calidad de Datos	**Regla de Negocio**
El valor del dato debe provenir de fuentes de datos creíbles	Utilizar API de AFIP
El valor del dato debe ser verdadero	Obtener desde AFIP
El valor del dato no debe repetirse o duplicarse	
TelefonoCelular	
Requisitos de Calidad de Datos	**Regla de Negocio**
El valor del dato debe ser verdadero	Enviar SMS con código de verificación
El valor del dato debe actualizarse con frecuencia	Solicitar actualización cada 3 meses
El valor del dato debe estar oportunamente actualizado	Cada 3 meses

ERCD.A3. Definición de Criterios de Aceptación para RCD

El GCD definió el Criterio de Aceptación (CA), en base al RCD en sí, y de acuerdo a su regla de negocio asociada, en caso de corresponder. Cómo resultado de esta acción, se obtuvo una especificación de RCD, la cual puede visualizarse en la Figura 2. Esta especificación contiene: el identificador y nombre de la HU; por cada dato de la HU, se exhibe una tabla compuesta por los RCD (junto con su medida y característica de calidad), y los CA definidos para cada RCD. Cabe acotar que, para cada HU con producción de datos incorporada al método, se obtiene como resultado un documento de especificación de RCD.


Metodi

Especificación de Requisitos de Calidad de Datos

Historia de Usuario: US01. Registro de Usuario

Dato: CUIT

Requisitos de Calidad de Datos (SQUARE) **Medida de Calidad / Característica de Calidad**	**Criterios de Aceptación**
Credibilidad de Fuentes de datos Credibilidad de Fuentes / Credibilidad	Se debe utilizar como fuente de datos creíble la API de AFIP
Credibilidad de valores de datos Credibilidad de Valores / Credibilidad	El CUIT ingresado se debe validar a través de la API de AFIP
Duplicación de Valores de Datos Riesgo de inconsistencia de datos / Coherencia	El CUIT no debe repetirse, caso contrario, mostrar mensaje

Dato: TelefonoCelular

Requisitos de Calidad de Datos (SQUARE) **Medida de Calidad / Característica de Calidad**	**Criterios de Aceptación**
Credibilidad de valores de datos Credibilidad de Valores / Credibilidad	Debe validarse la veracidad del teléfono ingresado, enviando un SMS con un código de verificación
Frecuencia de Actualización de valores de datos Frecuencia de Actualización / Actualización	Debe solicitarse actualización del dato con una frecuencia de 3 meses
Oportunidad de Actualización de Valores de Datos Oportunidad de Actualización / Actualización	El dato debe estar actualizado cada 3 meses

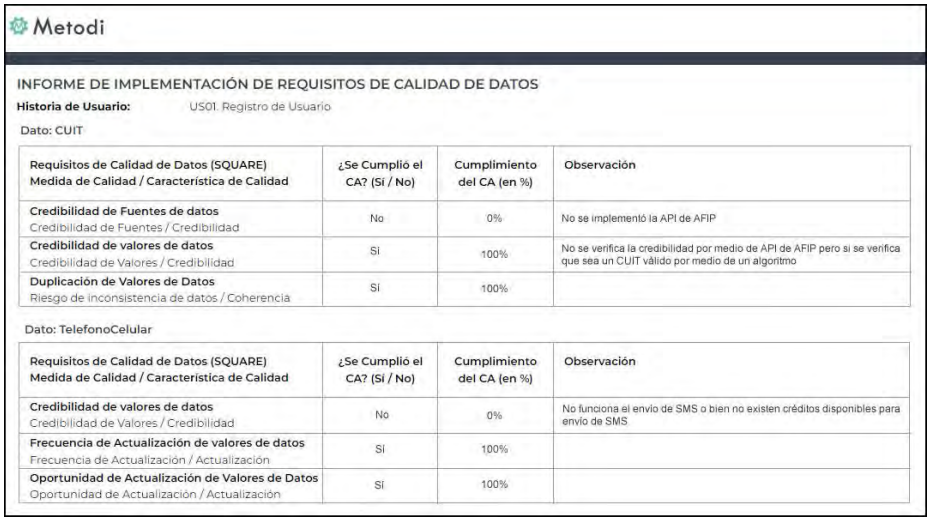
Figura 2. Extracto de Documento de Especificación de RCD (extraído de Metodi)

Finalmente, el GCD accedió a la herramienta “Gitea”, a fin de vincular dicho documento de especificación, con su historia de usuario correspondiente.

Etapa 3: Implementación de Requisitos de Calidad de Datos (IRCD)
En esta etapa se obtiene como resultado un conjunto de documentos de implementación de RCD, asociados a historias de usuario con producción de datos.

IRCD. A1. Verificación de Cumplimiento de Criterios de Aceptación

Esta actividad la lleva a cabo el Asegurador de Calidad (AC) y se realiza una vez finalizado el trabajo de diseño y ejecución de los casos de prueba. Con el resultado de esta última tarea, y a través del uso de Metodi, el AC indicó si los criterios de aceptación se cumplieron o no, registrando además el porcentaje de cumplimiento (0% para CA no cumplidos y 100% para CA cumplidos). Cabe acotar que no existen porcentajes con valores intermedios (entre 0 y 100), ya que se se considera que un CA se cumple o no se cumple. Por último, se puede ingresar una observación, ya sirva de apoyo al cumplimiento o no cumplimiento del CA. En la Figura 3 se puede observar el resultado de esta actividad.



Metodi

INFORME DE IMPLEMENTACIÓN DE REQUISITOS DE CALIDAD DE DATOS

Historia de Usuario: US01. Registro de Usuario

Dato: CUIT

Requisitos de Calidad de Datos (SQUARE) Medida de Calidad / Característica de Calidad	¿Se Cumplió el CA? (Si / No)	Cumplimiento del CA (en %)	Observación
Credibilidad de Fuentes de datos Credibilidad de Fuentes / Credibilidad	No	0%	No se implementó la API de AFIP
Credibilidad de valores de datos Credibilidad de Valores / Credibilidad	Si	100%	No se verifica la credibilidad por medio de API de AFIP pero si se verifica que sea un CUIT válido por medio de un algoritmo
Duplicación de Valores de Datos Riesgo de inconsistencia de datos / Coherencia	Si	100%	

Dato: TelefonoCelular

Requisitos de Calidad de Datos (SQUARE) Medida de Calidad / Característica de Calidad	¿Se Cumplió el CA? (Si / No)	Cumplimiento del CA (en %)	Observación
Credibilidad de valores de datos Credibilidad de Valores / Credibilidad	No	0%	No funciona el envío de SMS o bien no existen créditos disponibles para envío de SMS
Frecuencia de Actualización de valores de datos Frecuencia de Actualización / Actualización	Si	100%	
Oportunidad de Actualización de Valores de Datos Oportunidad de Actualización / Actualización	Si	100%	

Figura 3. Extracto de Documento de Implementación de RCD (extraído de Metodi)

El GCD accedió a los documentos de implementación de RCD, a fin de visualizar los resultados obtenidos. Luego procedió a vincular dichos documentos con su HU correspondiente, a través de la herramienta “Gitea”. Posteriormente participó en la “Definición de Hecho”, liberando los criterios de aceptación relacionados a calidad de los datos.

Con el resultado del compromiso descripto anteriormente, el GCD participó de la “Revisión del Sprint”, informando a las Partes Interesadas acerca del trabajo de incorporación de aspectos de calidad de datos realizado en el incremento.

Finalmente, el GCD estuvo presente en la "Retrospectiva del Sprint", a fin de receptar aspectos de mejora por parte del Equipo, para ser introducidos en la planificación del próximo sprint.

5. Resultados

En la Tabla 4 se pueden observar algunos resultados obtenidos, luego de la aplicación del método en 3 casos de estudio reales, haciendo uso de la herramienta "Metodi".

Tabla 4. Resultados de la aplicación del método en 3 casos de estudios

	Caso de Estudio 1	**Caso de Estudio 2**	**Caso de Estudio 3**
HU en Total	15	28	40
HU con producción de datos incorporadas	7	8	10
Cantidad de Datos asociados a HU	42	76	95
Cantidad de RCD Definidos (en promedio)	19	18	20
Cantidad de CA Definidos (en promedio)	90	105	120
Documentos de Especificaciones de RCD	7	8	10
Porcentaje de Implementación de RCD	100%	98%	95%

Como puede observarse en la tabla anterior, existe un alto porcentaje de implementación de RCD para los 3 casos de estudios llevados a cabo, lo que permite garantizar la incorporación de aspectos de calidad de datos para el incremento. Cabe destacar que, el promedio de la cantidad de RCD y CA definidos, se realizó en función de la HU "US01. Registro de Usuario".

También se destacaron algunos aspectos que pueden considerarse como un aporte de valor, luego de la aplicación del Método en contexto del desarrollo ágil con Scrum, los cuales se detallan a continuación:

- Posibilidad de contar con un rol responsable de cuidar la calidad de los datos durante todo el desarrollo del proyecto.
- Posibilidad de poder contrastar la calidad de los datos frente a un estándar reconocido a nivel internacional, como lo es SQUARE.
- Posibilidad de poder definir RCD en forma temprana que, en proyectos anteriores, no se definían o especificaban.
- Existencia de mayor cobertura de casos de prueba diseñados y ejecutados en relación a garantizar la calidad de los datos.
- Implementación de RCD en fase de desarrollo, en base a una especificación temprana.
- Mayor confianza del cliente respecto de la calidad de los datos de su producto, cuando se despliegue en producción.
- Detección temprana de errores relacionados a datos.

Los miembros del equipo que participaron en la aplicación del Método coincidieron que les llevó tiempo y esfuerzo extra, aunque estas dos variables se fueron reduciendo a medida que se llevaron adealnte los sprint del proyecto. También

mencionaron que, de no haberse utilizado el método, se hubieran definido muchos menos requisitos de calidad y criterios de aceptación, en relación a los datos; también hicieron hincapié en la amplia cobertura de RCD definidos e implementados.

6. Conclusiones y Trabajos Futuros

La calidad de la información es crucial para llevar adelante los procesos de negocio y la toma de decisiones en las organizaciones, es por esto que resulta de gran importancia poder incorporar de manera temprana en los procesos de desarrollo de software, aspectos relacionados a garantizar la calidad de los datos.

En este artículo se ha presentado un Método, basado en SQuaRe, e instanciado en el contexto del desarrollo ágil, mediante el uso de Scrum, el cual permite guiar la especificación temprana e implementación de RCD, a fin de obtener productos de software con un alto nivel de calidad de datos.

También se ha desarrollado una herramienta web que permite informatizar las etapas y actividades del Método, a fin de poder facilitar y/o agilizar su aplicación, en el contexto de un proyecto real, desarrollado a través de Scrum. Para quien esté interesado en el uso del Método, en la herramienta se pone a disposición un conjunto de tutoriales que explican en detalle la aplicación de cada una sus actividades.

En cuanto a los resultados obtenidos, luego de la aplicación del Método en 3 casos de estudio reales, se observa que el uso del Método, permitió incrementar la cantidad de requisitos de calidad definidos en relación a los datos, debido a que el Método se basa en un estándar de calidad de referencia, que proporciona este tipo de requisitos. Otra de las ventajas es la cuantificación de la implementación de los RCD, lo que permite determinar un nivel de aseguramiento de la calidad de los datos. Según los resultados obtenidos en los 3 casos de estudio, el procentaje de implementación de RCD supera el 95%, por lo que este valor refleja un alto nivel de calidad de los datos.

En cuanto a líneas de trabajo a futuro, se pueden mencionar las siguientes:

- Llevar adelante más casos de estudio en el contexto del desarrollo ágil mediante Scrum, a fin de optimizar el uso y/o aplicación del Método, y la herramienta.
- Instanciar el Método en otras metodologías ágiles que utilicen el formato de historias de usuario, por ejemplo, Extreme Programming (XP).
- Posiblidad de intercambiar las características y/o medidas de calidad de datos, utilizando otros modelos y estándares.

Bibliografía

[1] F. Soto, “Análisis de la problemática asociada con la baja calidad de datos en los sistemas de información,” 2014.

[2] “ISO/IEC 25030:2007, Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Quality requirements., ISO, 2007.” 2007.

[3] E. A. L. D Cohen Karen, *Tecnologías de información en los negocios*. 2009.

[4] J. H. Bernstein, “Antithesis JH Bernstein 2009 Bernstein , J . H . (2009). The data-information-knowledge- wisdom hierarchy and its antithesis . How does access to this work benefit you ? Let us know !,” 2009.
[5] M. Fernández and J. Vilalta, “La calidad de los datos y las decisiones empresariales,” *Libr. Empres.*, vol. 5, pp. 9–10, 2008.
[6] “ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model.” 2012.
[7] “ISO/IEC 25000:2014, Systems and software engineering-Systems and software Quality Requirements and Evaluation (SQuaRE), ISO-Guide to SQuaRE.” 2014.
[8] “ISO/IEC 25024:2015 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality.” 2015.
[9] R. Wand, Y and Wang, “Anchoring Data Quality Dimensions in Ontological Foundations,” *Commun. ACM*, vol. 39, no. 11, pp. 86–95, 1996.
[10] Yulma Fernanda Torres Alonso, "Especificación de requerimientos de software con un enfoque de calidad de datos", Universidad Autónoma de San Luis de Potosí, 2020.
[11] C. A. Guerra García, I. Caballero, M. Cardenas Juarez, and J. R. Juárez Ramírez, “A proposal to consider aspects of quality in the software development,” *J. Adv. Theor. Appl. Informatics*, vol. 2, no. 2, p. 12, 2016, doi: 10.26729/jadi.v2i2.2103.
[12] A. Rodríguez, A. Caro, C. Cappiello, and I. Caballero, “A BPMN extension for including data quality requirements in business process modeling,” *Lect. Notes Bus. Inf. Process.*, vol. 125 LNBIP, pp. 116–125, 2012, doi: 10.1007/978-3-642-33155-8_10.
[13] A. Caro, A. Fuentes, and M. A. Soto, “Desarrollando sistemas de información centrados en la calidad de datos,” *Ingeniare. Rev. Chil. Ing.*, vol. 21, no. 1, pp. 54–69, 2013, doi: 10.4067/s0718-33052013000100006.
[14] M. B. N William, WK Ivins, “Data quality & agile methods: A BT perspective,” in *11th International Conference on Information Quality (ICIQ-2006)*, 2006, pp. 10–12.
[15] C. Tona, R. Juarez-Ramirez, S. Jimenez, A. Quezada, C. Guerra-Garcia, and R. G. Pacheco Lopez, “Scrumlity: An Agile Framework Based on Quality Assurance,” *Proc. - 2021 9th Int. Conf. Softw. Eng. Res. Innov. CONISOFT 2021*, pp. 88–96, 2021, doi: 10.1109/CONISOFT52520.2021.00023.
[16] A. M. Moreno and A. Yagüe, “Agile user stories enriched with usability,” *Lect. Notes Bus. Inf. Process.*, vol. 111 LNBIP, pp. 168–176, 2012, doi: 10.1007/978-3-642-30350-0_12.
[17] “Manifesto for Agile Software Development.” http://agilemanifesto.org/.
[18] I. Sommerville, *Software engineering [9ª ed.]*. Boston, MA, USA, 2010.
[19] R. Pressman, *Ingeniería de Software*, 6 ed. 2005.
[20] “Scrum Alliance”. https://resources.scrumalliance.org/Article/quick-guide-things-scrum.
[21] “Extreme Programming”. https://www.agilealliance.org/glossary/xp.
[22] “Crystal, Agile project management”. http://crystalmethodologies.org/.

UN MODELO DE CALIDAD DE SOFTWARE CON LA SOSTENIBILIDAD COMO CARACTERÍSTICA TRANSVERSAL

Rosana Leo (1); Carlos Salgado(2); Alberto Sánchez(2); Mario Peralta(2)
(1) Universidad Nacional de La Rioja. La Rioja, La Rioja
(2) Universidad Nacional de San Luis. San Luis, San Luis
*E-mail del autor de contacto: leorosana@gmail.com

RESUMEN: El software es la herramienta necesaria para las más diversas y variadas gestiones en la actualidad. La elección del mismo se hace en función a requerimientos específicos y respetando ciertos criterios de calidad que se evalúan mediante modelos o estándares. La sostenibilidad, habitualmente se relaciona con el medio ambiente, pero si lo entendemos desde sus dimensiones: ambiental, técnica, económica y social, y como el concepto que trasciende múltiples disciplinas, es posible relacionarlo con la calidad del software como aquella característica transversal al modelo o estándar que la caracteriza. En el presente trabajo se define un modelo de calidad en base a la ISO/IEC 25010 y a la sostenibilidad como característica transversal. Se definen las métricas e indicadores para analizar con criterios de sostenibilidad: en qué medida un software es sostenible.

Palabras clave: Modelo de Calidad de Software. ISO 25010. Sostenibilidad. Métricas e Indicadores. Recursos.

1. INTRODUCCIÓN

En la actualidad, en todos los ámbitos y en la mayoría de las actividades, los sistemas se informatizaron, esto conlleva a la necesidad de que el software reúna ciertos criterios de calidad que permitan satisfacer las necesidades de los usuarios.
Pressman, en [1], define a la calidad del software como: *"Proceso eficaz de software que se aplica de manera que crea un producto útil que proporciona valor medible a quienes lo producen y a quienes lo utilizan."*
En ISO/IEC 25000 [2] *es la capacidad del producto de software para satisfacer necesidades declaradas e implícitas cuando se utiliza en determinadas condiciones.*
En general, toda la literatura apunta a que es el cumplimiento y/o el grado de satisfacción de los requisitos tanto explícitos como implícitos.
La especificación y evaluación de la calidad del software es factor clave para garantizar el valor a las partes interesadas. Esto se puede lograr mediante la definición de las características de calidad necesarias y deseadas asociadas con las metas y objetivos del sistema. Es importante que dichas características se especifiquen, midan y evalúen siempre que sea posible utilizando medidas y métodos de medición validados o ampliamente aceptados [3].

Para garantizar la calidad de software, es importante implementar algún modelo o estándar que permita la gestión de atributos en el proceso de construcción de software, teniendo en cuenta que la concordancia de los requisitos y su construcción son la base de las medidas de calidad establecidas [4].

La ventaja de estos estándares es que la calidad se convierte en algo concreto, que se puede definir, medir y, sobre todo, planificar. Ayudan también a comprender las relaciones que existen entre las diferentes características de un producto de software.

Si bien la norma ISO 14001:2015 [5] proporciona a las organizaciones un marco con el que proteger el medio ambiente y responder a las condiciones ambientales cambiantes, manteniendo el equilibrio con las necesidades socioeconómicas, este estándar no incluye requisitos específicos para otros sistemas como lo son los de gestión de la calidad.

La Norma Internacional ISO/IEC 25010:2011 [3], una de las divisiones de la serie SQUARE, describe el modelo de calidad para el producto software, presentando características y subcaracterísticas de calidad, criterios a tener en cuenta al momento de la evaluación.

Por otro lado, si se habla de satisfacer necesidades, es importante tener en cuenta un término que se escucha cada vez con más frecuencia, como lo es **Sostenibilidad**. Surge en el año 1987 en el Informe de Brundtland [6], titulado "Nuestro futuro común", y por la necesidad de estudiar y delimitar el impacto de las actividades humanas sobre el medio ambiente.

En el 2015 un manifiesto expone principios y compromisos vinculados con el diseño sostenible [7]. Infiriendo que la sostenibilidad tiene múltiples dimensiones (social, medioambiental, económica, técnica y humana) y que todas deben analizarse. La dimensión humana abarca la libertad y el albedrío individuales (capacidad de actuar en un entorno), la dignidad humana y la realización. La dimensión social abarca las relaciones entre individuos o grupos. La dimensión económica abarca aspectos financieros, crecimiento de capital y la liquidez, inversiones y operaciones financieras. La dimensión técnica abarca la capacidad de mantenimiento y evolución de los sistemas a lo largo del tiempo. Y la dimensión ambiental abarca el uso y la administración de los recursos naturales, incluye temas de residuos, consumo de energía, equilibrio de los ecosistemas, cambio climático, etc.

Al hablar de tecnologías sostenibles, nos imaginamos un menor consumo de energía, empleo de menor cantidad de recursos, la no contaminación, el reciclado o reutilización, siempre enfocados en satisfacer las necesidades de la sociedad. De allí se define un **producto sostenible** como aquel que aporta beneficios ambientales, sociales y económicos resguardando la salud pública, el bienestar y el medio ambiente en todo su ciclo de vida.

En este contexto, el software también puede ser sostenible, cuando su desarrollo se basa en el uso adecuado de recursos y cuando su impacto negativo en la economía, la sociedad y el medio ambiente es mínimo o, en el mejor de los casos, resulta positivo respecto del desarrollo sostenible [8].

En 2017, se aplica un modelo de calidad del software al Módulo de Talento Humano del Sistema Informático Integrado Universitario de la Universidad Técnica del Norte de Ecuador. Mediante las normas ISO/IEC 25000, se aplicó el modelo que permitió medir y evaluar el módulo, definiendo una propuesta de mejora [9].

En 2018, se propone el "Catálogo de Diseño de Sostenibilidad del Software" [10] como la herramienta que permite integrar la sostenibilidad en el diseño. Es un conjunto de criterios derivados de los nueve principios del manifiesto de Karlskrona, basados en el análisis cruzado de diferentes sistemas. Para cada criterio se derivan indicadores relacionados con las dimensiones de la sostenibilidad y su orden de impactos.
Un aporte interesante surge del trabajo de Condori Fernández y Lago [11], quienes determinaron qué atributos de calidad del modelo ISO/IEC 25010 son más relevantes para las dimensiones de la sostenibilidad. Asignan niveles de contribución del atributo a la dimensión, que pueden ser: altamente contributiva, contributiva, ligeramente contributiva o no contributiva. Se comparan los resultados obtenidos en dos instancias diferentes obteniendo tres valores posibles (0, 1 o 2) que indican qué característica es más importante para la dimensión en estudio y cuál es el atributo más relevante.
Un trabajo más reciente, "Método para la evaluación de la sostenibilidad del software para el proceso de Compra y Contratación del sector público", propone un método para que las entidades del sector público evalúen al adquirir software, la sostenibilidad del mismo, asegurando la calidad y favoreciendo la inversión óptima de los recursos del estado [12].
En base a lo anterior, se propone un modelo de calidad de software con la sostenibilidad como característica transversal junto con un conjunto de métricas e indicadores. Este instrumento sirve como apoyo a la toma de decisiones por parte de los encargados de evaluar el impacto del software en el medio ambiente. El modelo sirve también como una guía de las características deseables o esperables en un producto software que aporte a la sostenibilidad y al cuidado del medioambiente.
La estructura de este trabajo se organizó en las siguientes secciones: en la sección 2, un breve resumen describe la estructura de la norma de referencia, ISO/IEC 25000 (SQuaRE). En la sección 3, se desarrolla la propuesta de este trabajo. En la sección 4, se presenta el software, objeto de estudio. Y en la sección 5 se resumen conclusiones y trabajos futuros.

2. NORMAS ISO/IEC 25000

SQuaRE (Software Quality Requirements and Evaluation) es una familia de normas que tiene por objetivo crear un marco de trabajo para evaluar la calidad del producto software. Se basa en las anteriores ISO 9126 [13]. (que describe las particularidades de un modelo de calidad del producto software) y en ISO 14598 [14] (que trataba la evaluación del software). Uno de los principales objetivos de la serie SQuaRE es la coordinación y armonización del contenido de ISO 9126 y de ISO 15939:2002 [15] (Measurement Information Model). ISO 15939 tiene un modelo de información que ayuda a determinar que se debe especificar durante la planificación, performance y evaluación de la medición.
SQuaRE está formada por las divisiones siguientes:

- **ISO/IEC 2500n.** División de gestión de calidad. Los estándares que forman esta división definen todos los modelos, términos y definiciones comunes que se referencian en los demás apartados de SQuaRE.

- **ISO/IEC 2501n.** División del modelo de calidad. El estándar que conforma esta división presenta un modelo de calidad detallado, incluyendo características para la calidad interna, externa y en uso del producto software.
- **ISO/IEC 2502n.** División de medición de calidad. Los estándares pertenecientes a esta división incluyen un modelo de referencia de medición de la calidad del producto software, definiciones matemáticas de las métricas de calidad y guías prácticas para su aplicación.
- **ISO/IEC 2503n.** División de requisitos de calidad. Los estándares que forman parte de esta división ayudan a especificar los requisitos de calidad. Estos requisitos pueden ser usados en el proceso de especificación de requisitos de calidad para un producto software que va a ser desarrollado o como entrada para un proceso de evaluación.
- **ISO/IEC 2504n.** División de evaluación de la calidad. Estos estándares proporcionan requisitos, recomendaciones y guías para la evaluación de un producto software, tanto si la llevan a cabo evaluadores, como clientes o desarrolladores.
- **ISO/IEC 25050–25099.** Estándares de extensión SQuaRE. Incluyen requisitos para la calidad de productos de software "Off-The-Self" y para el formato común de la industria (CIF) para informes de usabilidad.

3. MODELO DE CALIDAD PROPUESTO

Tomando como base el modelo de calidad propuesto por SQuaRE [3] y teniendo en cuenta el concepto de sostenibilidad y los principios del Manifiesto de Karlskrona, se definió un modelo (Figura 1) que incluye *transversalmente* a la sostenibilidad, esto significa que no será una característica aislada, sino que estará inmersa en todo el modelo mediante métricas que permitan evaluar el producto, teniendo en cuenta sus particularidades y objetivos.
La figura 1 ilustra el modelo de calidad con la nueva característica de "Sostenibilidad" que se incorpora y atraviesa toda la estructura complementándose con cada una de las características originales al incluir criterios sostenibles en la medición y evaluación de atributos.
En principio se analizó en qué subcaracterísticas del modelo se pueden definir métricas con criterios de sostenibilidad, teniendo en cuenta la relevancia de cada atributo para las distintas dimensiones de la sostenibilidad, como lo definieron en [9]. A los fines de este trabajo se considera que los atributos a evaluar para inferir la sostenibilidad son los detallados en la Tabla 1. Allí también se los relaciona con las dimensiones ambiental, económica, técnica y social.

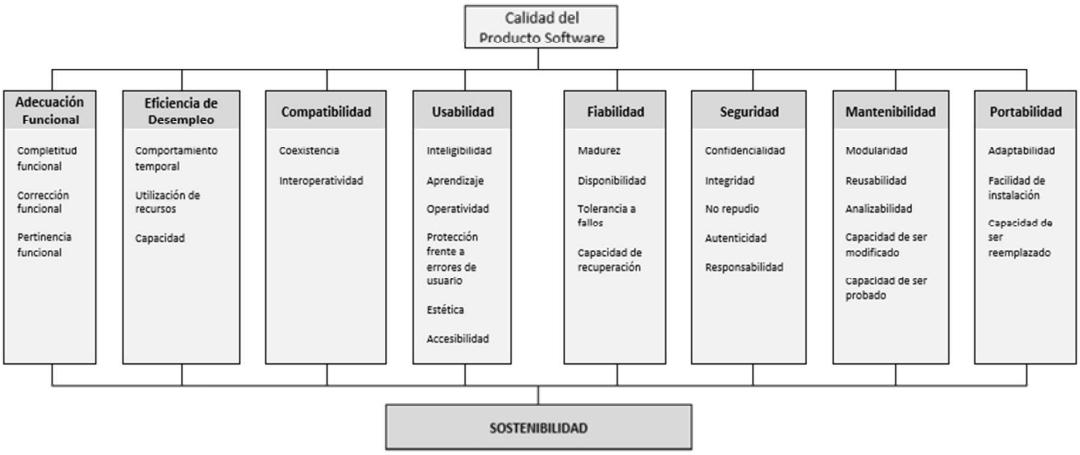


Figura 1. Modelo de Calidad del Producto Software propuesto

Tabla 1. Características y Atributos de Calidad del Software relacionados con las dimensiones de la sostenibilidad

CARACTERÍSTICA	ATRIBUTO	DIMENSIÓN
Adecuación Funcional	Corrección funcional	Técnica
Eficiencia de Desempeño	Capacidad	
	Utilización de recursos	Ambiental
Compatibilidad	Interoperabilidad	Técnica
Usabilidad	Protección frente a errores de usuarios	Social
	Accesibilidad	
Fiabilidad	Disponibilidad	Económica
	Capacidad de recuperación	
Seguridad	Confidencialidad	Social
	Autenticidad	
Mantenibilidad	Modularidad	Técnica
	Reusabilidad	Ambiental
	Capacidad de ser modificado	Técnica
Portabilidad	Adaptabilidad	
	Capacidad de ser reemplazado	

A continuación, en las figuras 2 y 3, se presentan ejemplos de métricas basadas en las desarrolladas en la Norma ISO/IEC 25023 [16]. En el modelo de calidad propuesto, se definieron y redefinieron métricas para las características “Eficiencia de desempeño” y “Seguridad”

EFICIENCIA DE DESEMPEÑO					
Subcaracterística	**Métrica**	**Propósito**	**Función de medición**	**Elementos de medida de la calidad utilizados**	**Valor deseado**
UTILIZACION DE RECURSOS	Uso de papel	Medir el uso de papel	X = A / B	A=Cantidad de informes en papel B=Cantidad total de informes (B>0)	Lo más cercano a 0 es mejor.
CAPACIDAD	Usuarios conectados simultáneamente	Medir la cantidad de usuarios conectados	X = A / T	A=Número máximo de accesos simultáneos T= Tiempo de operación (T>t)	El más lejano a 0/t es el mejor.

Figura 2. Métricas para la característica Eficiencia de Desempeño

SEGURIDAD					
Subcaracterística	**Métrica**	**Propósito**	**Función de medición**	**Elementos de medida de la calidad utilizados**	**Valor deseado**
CONFIDENCIALIDAD	Control de accesos	Conocer el numero de accesos ilegales	X = A / B	A=Número de ingresos ilegales detectados B=Número total de accesos (B>0)	0<=X<=1 El más cercano a 0 es el mejor
AUTENTICIDAD	Autenticación de la identidad del usuario	Conocer de qué manera se autentica la identidad del usuario	X = A	A=Número de métodos de autenticación	X>=0 X mayor o igual a 2 es el mejor

Figura 3. Métricas para la característica Seguridad

4. CASO DE ESTUDIO

En la provincia de La Rioja, en el poder ejecutivo, se está procurando tener un gobierno digital. En esta idea de un gobierno 4.0, hay varias instancias o etapas de desarrollo de software. Una de las condiciones que se debía cumplir era tener el menor impacto ambiental posible. En este sentido, la Directora de Compras y Contrataciones de la provincia bajó las directivas al equipo de desarrollo sobre qué necesidades se debían satisfacer. De las reuniones entre los diversos interesados surgió la necesidad de definir qué características debía tener el producto software a desarrollar. Esto llevó a tener que pensar qué características de sostenibilidad eran necesarias de considerar. Para ello, se definió un modelo de calidad que permitiera evaluar en qué medida el desarrollo que se llevaba a cabo estaba respetándolo. Junto al modelo de calidad se definió un conjunto de métricas e indicadores para cada atributo de calidad considerado. En el área de desarrollo se está implementando el "Sistema de Contrataciones".

Para cumplir el requerimiento solicitado se realizó un análisis de los procesos de negocio que se venían llevando a cabo. Había que identificar puntualmente qué procesos eran necesarios considerar para digitalizar para disminuir el impacto ambiental. Se trabajó en el módulo de gestión de proveedores, juntamente con las publicaciones de las contrataciones.

El modelo de calidad propuesto, junto con las métricas e indicadores, se instanciaron tanto para los procesos de negocio manuales como digitales. Del análisis llevado a cabo surge que había un gran desperdicio de recursos tipo papel, consumo eléctrico, mala disposición de los recursos humanos, limitaciones operativas, entre otras. Por ejemplo, luego del estudio, y ya con el sistema implementado siguiendo las recomendaciones de sostenibilidad que el modelo de calidad propuesto satisfacía, se procedió a hacer una nueva evaluación de las ventajas obtenidas.

En particular, para validar el modelo de calidad propuesto junto con las métricas e indicadores se lo aplicó al Sistema de Contrataciones de la provincia de La Rioja, específicamente al módulo de Gestión de Proveedores y Contratistas del Estado.

Una de las características en la que se centró el interés de la gerencia era lo referido al medioambiente. Luego del análisis de la situación actual versus el estado que se deseaba alcanzar, que era el menor impacto en el medioambiente, se procedió a instanciar el modelo de calidad.

Al haber implementado el legajo electrónico de proveedores se eliminó tanto el comprobante de registro como toda la documentación respaldatoria de la persona (física o jurídica) que en principio se presentaba o emitía en papel, actualmente son archivos digitales. El promedio de documentación por proveedor para su registro es de diez archivos, los cuales varían según el tipo y características de cada persona. Actualmente, si bien existe la posibilidad de impresión de la documentación, se ha logrado despapelizar por completo el proceso de registro.

También existía el inconveniente de la cantidad de gente en paralelo que se podía registrar en determinados períodos de licitaciones, compras, etc. Se debían apersonar a la Dirección de Compras y Contrataciones para realizar el trámite con toda la documentación respaldatoria impresa. En el mejor de los casos, se atendían seis personas simultáneamente por hora, En la actualidad se podría establecer como límite de accesos en paralelo a cien usuarios en el mismo periodo de tiempo.

A modo de ejemplo a continuación se muestra que, como parte del modelo de calidad, se tiene la característica Eficiencia de Desempeño que cuenta con los atributos: Uso de Recursos y Capacidad. Y la característica Seguridad con las subcaracterísticas: Confidencialidad y Autenticidad. Al instanciar el modelo de calidad y aplicar las métricas e indicadores se obtuvieron los resultados que se detallan en las figuras 4 y 5 que ponen de manifiesto las necesidades que ahora se estaban satisfaciendo, quedando detalles para mejorar, por ejemplo, respecto de la Autenticidad, debería agregarse un segundo factor de autenticación para algunas operaciones del sistema:

EFICIENCIA DE DESEMPEÑO						
Subcaracterística	**Métrica**	**Función de medición**	**Elementos de medida de la calidad utilizados**	**EJERCICIO**	**Valores**	**Valor deseado**
UTILIZACION DE RECURSOS	Uso de papel	X = A / B	A=Cantidad de informes, documentación o comprobantes de proveedor impresos B=Cantidad total de informes, documentación o comprobantes de proveedores (B>0)	2019	X = 9 /10 X = 0,9	Lo más cercano a 0 es mejor.
				2021	X = 0 /10 X = 0	
CAPACIDAD	Usuarios conectados simultáneamente	X = A / T	A=Número máximo de accesos simultáneos para determinada gestión T= Tiempo de operación medido en minutos(T>0)	2019	X = 6 /60 X = 0,1	El más lejano a 0/t es el mejor.
				2021	X = 100 /60 X = 1,66	

Figura 4. Valores de métricas de Eficiencia de Desempeño

SEGURIDAD						
Subcaracterística	**Métrica**	**Función de medición**	**Elementos de medida de la calidad utilizados**	**EJERCICIO**	**Valores**	**Valor deseado**
CONFIDENCIALIDAD	Control de accesos	X = A / B	A=Número de ingresos ilegales detectados B=Número total de accesos (B>0)	2019	X = 6 / 2000 X = 0,003	0<=X<=1 El más cercano a 0 es el mejor
				2021	X = 3 / 3500 X = 0,0008	
AUTENTICIDAD	Autenticación de la identidad del usuario	X = A	A=Número de métodos de autenticación	2019	X = 1	X>=0 X mayor o igual a 2 es el mejor
				2021	X = 1	

Figura 5. Valores de métricas de Seguridad

Solamente con los valores obtenidos de las métricas, no es posible estimar un concepto calculable, para ello son necesarios puntos de referencia que permitan comparar y determinar niveles de cumplimiento, además facilitan la evaluación a partir de criterios de decisión asociados. Así, se propusieron los Indicadores de figura 6.
Observando los resultados obtenidos en las métricas respecto, por ej. del uso del papel, se analiza, se compara de acuerdo a la valoración de la figura 6, que evoluciona de "Malo" a "Muy Bueno", se concluye en las ventajas de implementar un software que favorezca la despapelización mediante la gestión digital de los proveedores del Estado, sobre todo porque en la provincia todavía es un tema a implementar en muchas gestiones y trámites.

PUNTUACIÓN (valor óptimo: el más cercano a 1)	VALORACIÓN	PUNTUACIÓN (valor óptimo: el más cercano a 0)
1	Muy bueno	0
0,8	Bueno	0,2
0,5	Regular	0,5
0,2	Malo	0,8
0	Muy malo	1

Figura 6. Valoración de resultados de métricas

Un análisis similar fue realizado para otros recursos que pueden impactar en el medio ambiente como el consumo energético, la emisión de residuos, la capacidad de las personas, y se está trabajando en la detección de otros elementos que puedan ser analizados. Los mismos no son detallados en el presente trabajo por motivos de espacio.

5. CONCLUSIONES

El modelo de calidad propuesto permitió obtener mejoras en lo que respecta al impacto del producto software sobre el medioambiente. No se encontraron en la bibliografía consultada trabajos científicamente probados sobre la evaluación de la sostenibilidad del software. Algunos autores coinciden en que se puede caracterizar a través del estándar ISO/IEC 25000, evaluando sus atributos más relevantes respecto de las dimensiones ambiental, económica, técnica y social. Siguiendo esa línea de estudio, se definió el modelo de calidad que se propone en este trabajo donde la sostenibilidad sea una característica transversal al modelo, estando inmersa en la evaluación a través de criterios sostenibles dependiendo de las particularidades y objetivos del software.
La utilización de la Norma ISO/IEC 25000 como base para el modelo de calidad propuesto fue de gran aceptación por parte del equipo de desarrollo de software por ser un estándar ampliamente conocido y utilizado por la empresa.
Al implementarse el sistema, se observaron ventajas en distintos ámbitos. Desde el Estado, eficiencia transaccional, mayor competencia, eficiencia en el ciclo de contratación y satisfacción de los usuarios. Desde el punto de vista de los proveedores, facilidad de acceso al mercado público, mayor participación y facilidad de registro. Y desde el punto de vista social en general, transparencia en la gestión pública.
Aplicar el modelo propuesto, evaluando métricas, como las del ejemplo presentado anteriormente, en dos momentos diferentes del Sistema de Compras y Contrataciones de la provincia, permitió inferir que actualmente el software es sostenible respecto de las dimensiones técnica y ambiental (en este caso), debido a promover mediante la gestión digital de los proveedores del Estado, el uso responsable de recursos.
Al momento de esta presentación, nos encontramos trabajando en el módulo de Licitaciones Pública. Para lo cual se está utilizando el modelo de calidad definido, tanto como guía de desarrollo de la experiencia adquirida como instrumento de evaluación.

6. BIBLIOGRAFIA

1. R. Pressman; "Ingeniería de Software. Un enfoque práctico" 9ª Ed. McGraw-Hill Interamericana, 2021
2. ISO/IEC 25000 Systems and software engineering-Systems and Software Quality Requirements and Evaluation (SQUARE)
3. ISO/IEC 25010 Systems and software engineering-Systems and Software Quality Requirements and Evaluation (SQUARE) System and software quality models
4. Callejas-Cuervo, Mauro; Alarcon-Aldana, Andrea Catherine; Álvarez-Carreño, Ana María. Modelos de calidad del software, un estado del arte. En: Entramado. Enero - Junio, 2017. vol. 13, no. 1, p. 236-250
5. ISO/IEC 14001 Environmental management systems - Requirements with guidance for use
6. CMMAD. Comisión Mundial de Medio Ambiente y Desarrollo de la Organización de las Naciones Unidas (ONU) (1987). Informe Brundtland: "Nuestro futuro común".
7. Christoph Becker. Manifiesto Karlskrona. Sustainability design and software. 2015
8. Naumann - Dick - Kern – Johan. El modelo GREENSOFT: un modelo de referencia para el software verde y sostenible y su ingeniería.
9. Vaca Sierra, Tulia Nohemi. Modelo de Calidad de Software aplicado al módulo de Talento Humano del Sistema Informático Integrado Universitario. Universidad Técnica del Norte. Ecuador. 2017.
10. Oyedeji, Shola – Seffah Ahmed – Pensestadler Birgit. Catálogo de Diseño de Sostenibilidad del Software (Suiza). 2018.
11. Fernández, N.C. y Lago P. "Characterizing the contribution of quality requirements to software sustainability". Revista de Sistemas y Software, Vol. 137 pag 289-305. Marzo 2018.
12. Carrascal Vergara, Carlos David. Método para la evaluación de la Sostenibilidad del Software para el proceso de Compra y Contratación del Sector Publico. Universidad de Antioquia. 2021.
13. ISO/IEC 9126 Software engineering – Product quality.
14. ISO/IEC 14598 Information technology – Software product evaluation.
15. ISO/IEC 15939 Software engineering – Software measurement processs.
16. ISO/IEC 25023 Systems and software engineering - Systems and Software Quality Requirements and Evaluation (SQUARE) - Measurement of system and software product quality.

Sistemas de gestión de calidad y Blockchain en la era de la industria 4.0: Revisión de literatura

Kristian Petkoff Bankoff, Rocío Muñoz, Ariel Pasini, and Patricia Pesado

III-LIDI, Facultad de Informática
Universidad Nacional de La Plata, Argentina
{kpb,rmunoz,apasini,ppesado}@lidi.info.unlp.edu.ar

Resumen Los sistemas de gestión de la calidad surgieron durante el siglo XX catalizados por una corriente de concientización sobre la calidad en las organizaciones para que las mismas adopten cambios culturales que promuevan la mejora de la calidad no ya del producto o servicio prestado sino de todos los procesos e incluso de las personas mismas. La evolución misma de la computación a la par de los modelos de calidad ayudó a que se utilicen herramientas informáticas para asistir a la implementación de los sistemas de gestión de la calidad; sin embargo, la inercia de cambio de la informática es superior a la de los modelos de mejora utilizados en la gestión de la calidad y propicia que se reflexione sobre la necesidad de adoptar nuevas tecnologías en estos sistemas como parte integral de los mismos y no solo por el contexto del negocio. En este sentido, se propone analizar publicaciones que permitan relacionar tecnologías disruptivas como Blockchain con sistemas o modelos ya bien establecidos como los definidos en ISO 9001 de cara a la cuarta revolución industrial, en la que la calidad de los productos, de los servicios y la optimización de los procesos es central.

Keywords: Quality Management, 4.0 Industry, Blockchain, Smart Contracts

1. Introducción

Los sistemas de gestión de la calidad surgieron durante el siglo XX catalizados por una corriente de concientización sobre la calidad en las organizaciones para que las mismas adopten cambios culturales que promuevan la mejora de la calidad no solo del producto o servicio prestado sino de todos los procesos e incluso de las personas mismas. La evolución misma de la computación, a la par de los modelos de calidad, llevó a que se utilicen herramientas informáticas para asistir a la implementación de los sistemas de gestión de la calidad; sin embargo, la inercia de cambio de la informática es superior a la de los modelos de mejora utilizados en la gestión de la calidad[1] y propicia que se reflexione sobre la necesidad de adoptar nuevas tecnologías en estos sistemas como parte integral de los mismos y no solo por el contexto del negocio. En este sentido, se

propone analizar publicaciones que permitan relacionar tecnologías disruptivas como Blockchain con sistemas o modelos ya bien establecidos como los definidos en ISO 9001 de cara a la cuarta revolución industrial, en la que la calidad de los productos, de los servicios y la optimización de los procesos es central.

2. Contexto

Esta revisión de literatura surge a partir de hallazgos en el marco del trabajo de tesina de licenciatura titulado "Asistencia a la auditoría de sistemas de gestión basados en normas ISO, un enfoque orientado a la industria 4.0 utilizando contratos inteligentes". Durante la fase de análisis del estado del arte no se hallaron publicaciones que vincularan de forma directa a los tres temas fundamentales que se abordan; sin embargo, en [1] se trata la adaptabilidad de los modelos de calidad frente a los cambios que se dan en la cuarta revolución industrial y se concluye que se requieren actualizaciones profundas en éstos para que la evolución constante sea posible.

2.1. Sistemas de gestión de la calidad

Existen diferentes versiones de calidad que giran en torno a conceptos como la conformidad con los requerimientos, el cumplimiento de expectativas y la adecuación al uso. En la búsqueda de garantizar la calidad de sus productos y servicios, han surgido diversas corrientes de gestión de la calidad a lo largo del siglo XX, motorizadas por los así llamados gurúes de la calidad[2]. Los aspectos clave de la gestión de la calidad que se plasman en el modelo de la norma ISO 9001 implican la definición de una política de calidad, la planificación, el control, el aseguramiento y la mejora. Un sistema de gestión de la calidad basado en ISO 9001 consiste en un conjunto de definiciones estratégicas, de productos, de servicios, de clientes y de procesos necesarios para la producción de los bienes y servicios con los que se espera satisfacer al cliente. Tales procesos (además de los productos) son medidos, controlados y mejorados utilizando herramientas como el ciclo PDCA (del inglés *plan, do, check, act*: planificar, hacer, verificar, actuar). Este modelo se basa en la premisa de que aplicar procesos de calidad permiten generar productos de calidad, y que las organizaciones deben buscar permanentemente la mejora para optimizar sus procesos productivos en pos de lograr la calidad. Una de las más grandes dificultades que representa la implementación de un sistema de gestión de la calidad es el esfuerzo de generar y mantener un conjunto de documentos, procedimientos y registros exhaustivos sobre sus procesos y su historia de ejecución. Sin embargo, una organización solo puede mejorar y garantizar calidad si se conoce primero a sí misma, y esto solo se logra a través de la disponibilidad de datos e información.

2.2. Industria 4.0

El concepto de Industria 4.0, o cuarta revolución industrial, se menciona por primera vez en Alemania en el año 2011[3] y representa cómo la incorporación

de recursos de hardware, de software y de redes modifica la forma en que las industrias llevan adelante sus operaciones, produciendo fábricas inteligentes y no solo procesos asistidos con computadoras. Las principales características de los modos de producción y de las organizaciones enmarcadas en esta revolución son la cooperación global, el intercambio permanente de gran cantidad de datos, la generación de resúmenes e información de forma automatizada, el uso de sensores y otros recursos digitales para registrar más y mejor todas las actividades desarrolladas. Como resultado de la integración de las herramientas digitales en todo el proceso de producción, control y decisión, se obtienen beneficios que son rectores en las industrias 4.0: reducción de costos operativos, optimización del desperdicio, optimización de la comunicación entre todos los stakeholders, involucramiento de los clientes y la obtención de productos de mayor funcionalidad y calidad.

2.3. Blockchain

Blockchain es una tecnología de base de datos descentralizada que provee un mecanismo de garantía de inmutabilidad y de ausencia de una entidad centralizada para la toma de decisiones en las operaciones sobre los datos. Esto se logra mediante los algoritmos de consenso, que establecen protocolos para que los nodos participantes decidan de manera autárquica cuándo un bloque de datos es válido y cuándo no[4]. La descentralización y la inmutabilidad (garantizada mediante criptografía asimétrica) han dado paso a algunas implementaciones exitosas, sobre todo en el negocio de la banca y las finanzas[5]. Con el tiempo se ha dado paso a más negocios en los que las herramientas basadas en esta tecnología hacen un aporte beneficioso, llegando incluso a plantearse que en cierta medida es propulsor de una revolución en la informática[6]. Una de las herramientas más relevantes basada en Blockchain son los contratos inteligentes. Estos contratos son aplicaciones de software programadas para ejecutar acciones predefinidas al alcanzarse condiciones acordadas por dos entidades o nodos, sin necesidad de tener confianza entre sí, registrando todo en una base de datos que garantiza la inmutabilidad. Es interesante mencionar que el registro es tanto de la evidencia que permite establecer que las condiciones están dadas, como el propio contrato en sí mismo. El contrato, por lo tanto, no podrá ser alterado por ninguna de las partes ni por terceros ya que violaría uno de los principios de la red. Es en este sentido que esta tecnología, que se puede nutrir de datos que surgen de otras piezas de hardware, de software o de la red, está alineada con principios de optimización de procesos, de negocio y de integración de herramientas digitales en el núcleo de los más diversos negocios.

3. Revisión Bibliográfica

En esta sección se presentan los objetivos y los criterios bajo los cuales se realizó la búsqueda para el relevamiento. La misma está orientada a establecer nexos entre pares de temas para, finalmente, vincular los tres tópicos entre sí, teniendo en cuenta que no existen publicaciones que aborden esta vinculación.

3.1. Objetivos

El principal objetivo de este trabajo es verificar si existe, dentro de la literatura de cada uno de los tres temas principales, publicaciones a partir de las cuales se pueda establecer una red que vincule a los sistemas de gestión de calidad (u otros similares compatibles como, por ejemplo, sistemas de gestión ambiental) con la industria 4.0 y con tecnología Blockchain, con la mira puesta en la adaptación de los primeros a estas corrientes industriales e informáticas que surgen como novedad. Para organizar la investigación se dividió a este objetivo en los siguientes tres sub-objetivos:

- Encontrar relaciones entre publicaciones para definir si los sistemas de gestión de la calidad son compatibles o adaptables a los procesos de las industrias 4.0.
- Analizar la situación de los sistemas de gestión de la calidad al contexto de la cuarta revolución industrial.
- Determinar si existen otras líneas de investigación relacionadas con la implementación de sistemas de gestión de la calidad con tecnología Blockchain.

3.2. Búsqueda

Para recopilar publicaciones relacionadas con la gestión de calidad, industrias 4.0 y Blockchain se realizó una búsqueda en los motores Scopus y ResearchGate utilizando las expresiones "quality management" AND blockchain, "quality management" AND 4.0 AND (industries OR industry), blockchain AND 4.0 AND (industries OR industry). Los resultados fueron sometidos a un primer análisis en base a su abstract para filtrar aquellos mejor ajustados al alcance de la investigación.

3.3. Alcance

Se acotó el alcance de la revisión a publicaciones relacionadas a sistemas de gestión de la calidad con modelos compatibles o basados en las normas ISO 9001, ISO 14001, ISO/IEC 17025, ISO 45001; también a experiencias de aplicación de sistemas de gestión en industrias, a la incidencia de la digitalización y a la temática de industrias 4.0 en los procesos de organizaciones, tengan o no implementados sistemas de gestión. Por otra parte, se hizo una selección de aquellas publicaciones sobre Blockchain o Smart Contracts que tienen relación con aspectos clave de un sistema de gestión, tales como el manejo de evidencia, versionamiento y control de procesos. La recopilación de material fue realizada entre diciembre de 2021 y febrero de 2022, seleccionando los artículos que tuvieran una mejor interrelación entre pares de los tres tópicos principales, obteniendo así 26 publicaciones a revisar. A su vez, de la base de literatura encontrada, se tomó un criterio de selección de los artículos más recientes por tratarse de una línea de trabajo sobre temas de actualidad. La carencia de publicaciones que vinculen a la totalidad de la temática abordada se observa como resultado de las búsquedas aún antes de tomar la decisión de acotar el alcance de la revisión.

3.4. Publicaciones identificadas

La tabla 1 presenta las 26 publicaciones recopiladas con su categorización según los temas que se abordan dentro del interés de esta revisión.

Tabla 1: Publicaciones relevadas

Título	**B**	**I4**	**C**
Data quality certification using ISO/IEC 25012: Industrial experiences[7]			X
The Computerized Maintenance Management System an Essential Tool for World Class Maintenance[8]			X
A process approach to ISO/IEC 17025 in the implementation of a quality management system in testing laboratories[9]			X
A proposal of model for a quality management system in research testing laboratories[10]			X
ISO 9001:2015 Adoption: A Multi-Country Empirical Research[11]			X
Proposal for a maintenance management system in industrial environments based on ISO 9001 and ISO 14001 standards[12]			X
Quality of the ISO 9000 series of standards-perceptions of quality management experts[13]			X
Recertification of a Quality Management System based on ISO 9001 - Is it a must for a modern manufacturing company?[14]			X
System proposal for implementation of risk management in the context of ISO/IEC 17025[15]			X
System quality and security certification in seven weeks: A multicase study in Spanish SMEs[16]			X
Blockchain Enabled Quality Management in Short Food Supply Chains[17]	X		X
Evidence Management System Using Blockchain and Distributed File System (IPFS)[18]	X		X
Are QM models aligned with Industry 4.0? A perspective on current practices[1]		X	X
Procedure for Defining the System of Objectives in the Initial Phase of an Industry 4.0 Project Focusing on Intelligent Quality Control Systems[19]		X	X
Quality Culture of Manufacturing Enterprises: A Possible Way to Adaptation to Industry 4.0[20]		X	X
Quality management in the 21st century enterprises: Research pathway towards Industry 4.0[21]		X	X
Re-Engineering of Logistics Business Processes Influenced by the Digitalization[22]		X	
Blockchain Enterprise: Use Cases on Multiple Industries[23]	X	X	
Blockchain Technology: A Fundamental Overview[24]	X	X	

Blockchain Technology Applications for Next Generation[25]	X	X	
EPS-ledger: Blockchain hyperledger sawtooth-enabled distributed power systems chain of operation and control node privacy and security[26]	X	X	
Significance of Blockchain Technologies in Industry[27]	X	X	
Blockchain. La revolución industrial de internet[6]	X	X	
Construction of Blockchain Technology Audit System[28]	X		
Decentralized collaborative business process execution using blockchain[29]	X		
A blockchain-based integrated document management framework for construction applications[30]	X		

4. Resultados

4.1. Representación gráfica

En la figura 1 se muestra la relación de los artículos relevados, permitiendo visualizar fácilmente la ubicación de cada publicación en el contexto de las dimensiones planteadas y el nivel de cohesión entre los términos.

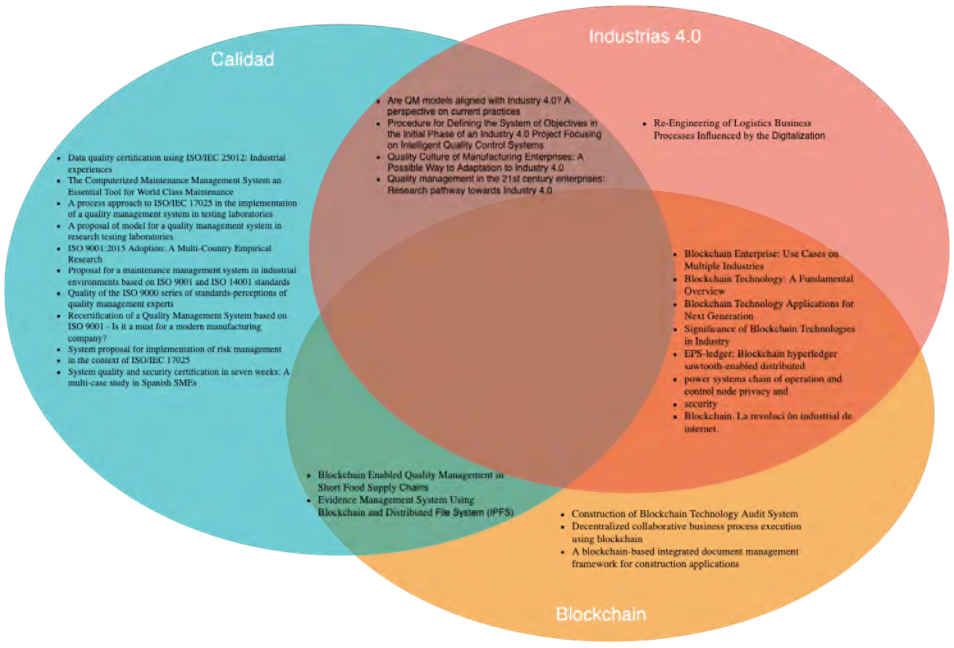


Figura 1. Diagrama de las tres dimensiones y los artículos relevados

4.2. Análisis

Al analizar los artículos relevados se encontró que, del total de publicaciones seleccionadas, un 15 % respondía de forma directa a los tópicos de calidad y de industria 4.0 como temas principales, un 24 % coincidía en las claves Blockchain e industria 4.0 y un 8 % relacionaba de forma directa Blockchain con gestión de calidad. En la tabla 2 se listan los temas y la proporción de publicaciones que los abordaban, de forma aislada del resto de manera que una publicación relacionada con dos temas es sumada en los dos tópicos correspondientes de la tabla. De la literatura seleccionada que aborda temas individualmente se encuentran relaciones en el texto que permiten establecer vínculos entre Blockchain y gestión de calidad (8 %) y entre calidad e industria 4.0 (8 %).

Tabla 2: Resumen por temas

Tema	**Cantidad**	**%**
Blockchain	11	42,3 %
Industria 4.0	11	42,3 %
Gestión de calidad	16	61,5 %
Total	**26**	**100 %**

También se observa que las búsquedas realizadas no produjeron resultados que vinculen de forma directa a los tres tópicos. Sin embargo, al realizar un análisis en función de una lectura del cuerpo de las publicaciones y en particular de las conclusiones y discusiones que se plantean en los mismos, sí puede establecerse una red de conceptos que conectan a los temas buscados. Para comenzar, Asif en [1] concluye que la gestión de la calidad como fue concebida en el siglo XX naturalmente no contiene los fundamentos de la industria 4.0 y que una adaptación sustancial incorporando nuevas tecnologías es necesaria para llevarla adelante de cara a las industrias digitales del siglo XXI, ya que estas producen cambios y procesos inteligentes que son más rápidos que los métodos tradicionales de gestión de calidad de hoy en día. Algunos aspectos a considerar al integrar la gestión de la calidad con un modelo de negocio alineado con la industria 4.0 es el control de la calidad en la tercerización y los procesos de cadena de suministro que integran a distintas organizaciones, de manera que se garantice la optimización de los costos, del uso de los recursos y de la calidad misma. En [21] se proponen preguntas de investigación futura como qué cambios son necesarios en la gestión de la calidad para comprometer el desarrollo humano con tecnologías emergentes como Blockchain, IoT, Big Data, cadenas de suministro inteligentes; cómo gestionar la tercerización e integración de procesos multi organizacionales en la era de la industria 4.0; cómo desarrollar intercambios de calidad con el soporte de métodos sofisticados para las cadenas de suministro inteligentes. Es de interés para este análisis remarcar que la adquisición de recursos es uno de los aspectos clave que se deben atender al implementar un sistema de gestión de la calidad.

La norma ISO 9001 para la implementación de sistemas de gestión de calidad puede ser aplicado a cualquier tipo de organización, sin importar el tipo de

industria en el que opera. Actualmente existen más de un millón de compañías e instituciones certificadas bajo la norma ISO 9001 (año 2021). Muchas organizaciones han comenzado a tratar a los sistemas de gestión de la calidad no solo como un mecanismo para alcanzar la performance organizacional sino también como un punto de partida para la construcción de sistemas integrados más complejos de producción. La implementación de sistemas de gestión de la calidad mejora la predictibilidad del comportamiento de los procesos y, por lo tanto, mejora la eficiencia de las organizaciones y optimiza el uso de recursos[14], lo que coincide con preceptos fundamentales de las industrias 4.0. Entre los aspectos más resaltados durante un relevamiento presentado en [14] se menciona que la reducción de costos y la eliminación de registros que afectan la eficiencia de trabajo sin aportar información relevante a los procesos en pos de optimizarlos.

Otros sistemas de gestión son compatibles con ISO 9001 sin ser específicamente de calidad, ya que en los últimos años se ha realizado una adaptación de muchos de los modelos para que tengan un cuerpo documental estandarizado. Uno de estos casos son los sistemas de gestión ambiental definidos en la familia ISO 14000. El control de los materiales utilizados es un aspecto clave para la certificación del sistema de gestión ambiental; Castillo-Martinez y col. en [12] proponen una arquitectura de sistema en la que se optimice el intercambio de información entre distintas áreas mediante la adopción de herramientas y tecnología que automatice estos procesos, haciendo hincapié en el registro de cada uso de los distintos materiales y el seguimiento del ciclo de vida de los mismos. Si bien los sistemas de gestión ambiental buscan como objetivo principal mejorar las prácticas ambientales, la optimización de costos sobre todo en la gestión de residuos está dentro de sus incumbencias y es consecuente con las industrias 4.0.

La relación del tópico de Blockchain surge al analizar los requerimientos de los sistemas de gestión de la calidad, que basan su método en el registro de evidencia sobre las acciones para propiciar una toma de decisiones y un seguimiento de las acciones de mejora o de corrección que se propongan durante su mantenimiento. Las auditorías suponen también el control de la correcta ejecución de los procedimientos predefinidos y la integridad del registro correspondiente, que son tareas en general bien cubiertas por herramientas implementadas sobre Blockchain[18][30]. Por otra parte, la tecnología Blockchain está ganando cada vez más lugar en diversos casos de uso y por lo tanto al garantizar la calidad de una organización cuyos procesos se soporten en esta tecnología permitiría catalizar una migración de los sistemas de gestión de la calidad hacia tecnologías Blockchain como por ejemplo los contratos inteligentes[25][26].

5. Conclusiones

Se identificó y relevó una cantidad de artículos relacionados con los términos Blockchain, Industria 4.0 y Gestión de calidad, realizando búsquedas con distintas combinaciones de pares de temas. Una vez organizados según la temática en un orden que permite pasar desde la calidad hacia Blockchain a través de las coincidencias en la dimensión de industrias 4.0, se establecieron puntos de

contacto entre las tres temáticas que no eran evidentes a simple vista ni a través de las palabras clave. La calidad es resultado y a la vez es requerida por las industrias digitales y al mismo tiempo la tecnología Blockchain (por ejemplo a través de los contratos inteligentes) es una herramienta cada vez más presente en las industrias 4.0, por lo que en cuanto a los objetivos planteados podemos concluir que:

Encontrar relaciones entre publicaciones para definir si los sistemas de gestión de la calidad son compatibles o adaptables a los procesos de las industrias 4.0. Los sistemas de gestión de la calidad son adaptables a la industria 4.0 pero deberían realizar una transformación que permita una mejor resiliencia ante los cambios motorizados por la evolución digital, y en este sentido podrían realizar su propia revolución de la calidad en el marco de la revolución de la industria.

Analizar la situación de los sistemas de gestión de la calidad al contexto de la cuarta revolución industrial. Los sistemas de gestión de la calidad se concibieron como herramientas que permitían garantizar al cliente que el producto que se le suministraba cumpliría con sus expectativas aún antes de evaluarlo. En cierta medida funcionaron como un herramienta comercial; en el contexto de la cuarta revolución industrial, los sistemas de gestión de la calidad corren riesgo de quedar relegados por los cambios que promueve la propia digitalización de las industrias, lo que les permitiría introducir mejoras en otros aspectos clave como la optimización de costos. Sin embargo los sistemas de gestión de la calidad son compatibles con la industria 4.0 porque no introducen restricciones de ningún tipo al objeto ni a los objetivos del sistema.

Determinar si existen otras líneas de investigación relacionadas con la implementación de sistemas de gestión de la calidad con tecnología Blockchain. Si bien existen investigaciones o implementaciones de algunos aspectos claves de la gestión de la calidad (como el control documental o el registro de evidencia controlada), no se han hallado otros autores o instituciones que traten específicamente este tema, pero la relación entre las publicaciones relevadas permite catalogarla como investigación futura.

Referencias

[1] Muhammad Asif. *Are QM models aligned with Industry 4.0? A perspective on current practices*. Vol. 258. 2020, pág. 120820. DOI: https://doi.org/10.1016/j.jclepro.2020.120820. URL: https://www.sciencedirect.com/science/article/pii/S0959652620308672.

[2] Mario G. Piattini Velthuis, Félix O. García Rubio e Ismael Caballero Muñoz-Reja. *Calidad de Sistemas Informáticos*. Alfaomega Grupo Editor, 2007. Cap. 1. ISBN: 84-7897-734-1.

[3] Klaus Schwab. *The Fourth Industrial Revolution*. World Economic Forum, 2016. Cap. 1. ISBN: 978-84-9992-699-5.

[4] Anamika Chauhan y col. *A Deep Dive into Blockchain Consensus Protocols*. Ed. por Yu-Dong Zhang y col. Singapore: Springer Singapore, 2022, págs. 571-581. ISBN: 978-981-16-4016-2.

[5] Nirmal K. Gupta y col. *State of the Art and Challenges in Blockchain Applications*. Ed. por Arun K. Somani y col. Singapore: Springer Singapore, 2022, págs. 311-320.

[6] Alexander Preukschat y col. *Blockchain. La revolución industrial de internet*. Barcelona, Spain: Centro Libros PAPF, S. L. U., 2017. ISBN: 978-84-9875-448-3.

[7] Fernando Gualo y col. *Data quality certification using ISO/IEC 25012: Industrial experiences*. Vol. 176. 2021, pág. 110938. DOI: https://doi.org/10.1016/j.jss.2021.110938. URL: https://www.sciencedirect.com/science/article/pii/S0164121221000352.

[8] Michael Wienker, Ken Henderson y Jacques Volkerts. *The Computerized Maintenance Management System an Essential Tool for World Class Maintenance*. Vol. 138. SYMPHOS 2015 - 3rd International Symposium on Innovation and Technology in the Phosphate Industry. 2016, págs. 413-420. DOI: https://doi.org/10.1016/j.proeng.2016.02.100. URL: https://www.sciencedirect.com/science/article/pii/S1877705816004641.

[9] Inês Hexsel Grochau y Carla Schwengber ten Caten. *A process approach to ISO/IEC 17025 in the implementation of a quality management system in testing laboratories*. 2012.

[10] S Martínez-Perales, Ortiz-Marcos I. y Ruiz J.J. *A proposal of model for a quality management system in research testing laboratories*. Vol. 26. Dic. de 2021, págs. 237-248.

[11] Luis Miguel Ciravegna Martins da Fonseca y col. *ISO 9001:2015 Adoption: A Multi-Country Empirical Research*. 2019.

[12] A. Castillo-Martinez y col. *Proposal for a maintenance management system in industrial environments based on ISO 9001 and ISO 14001 standards*. Vol. 73. 103453. 2021.

[13] Piotr Rogala y Sławomir Wawak. *Quality of the ISO 9000 series of standards-perceptions of quality management experts*. 2021, págs. 509-525.

[14] Katarzyna Midor y Grzegorz Wilkowski. *Recertification of a Quality Management System based on ISO 9001 - Is it a must for a modern manufacturing company?* Vol. 27. Sep. de 2021, págs. 217-222. DOI: 10.30657/pea.2021.27.29.

[15] Fabiane Rodrigues da Silva, Inês Hexsel Grochau y Hugo Marcelo Veit. *System proposal for implementation of risk management in the context of ISO/IEC 17025*. 2021.

[16] Domingo Gaitero, Marcela Genero y Mario Piattini. *System quality and security certification in seven weeks: A multi-case study in Spanish SMEs*. Vol. 178. 110960. Ago. de 2021.

[17] Patrick Burgess, Funlade Sunmola y Sigrid Wertheim-Heck. *Blockchain Enabled Quality Management in Short Food Supply Chains*. Vol. 200. Ene. de 2022, págs. 904-913. DOI: 10.1016/j.procs.2022.01.288.

[18] Shritesh Jamulkar y col. *Evidence Management System Using Blockchain and Distributed File System (IPFS)*. 2021, págs. 337-359.

[19] Albert Albers y col. *Procedure for Defining the System of Objectives in the Initial Phase of an Industry 4.0 Project Focusing on Intelligent Quality Control Systems*. Vol. 52. The Sixth International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV2016). 2016, págs. 262-267. DOI: https://doi.org/10.1016/j.procir.2016.07.067. URL: https://www.sciencedirect.com/science/article/pii/S2212827116308666.

[20] Pavol Durana y col. *Quality Culture of Manufacturing Enterprises: A Possible Way to Adaptation to Industry 4.0*. Vol. 8. 4. 2019. DOI: 10.3390/socsci8040124. URL: https://www.mdpi.com/2076-0760/8/4/124.

[21] Angappa Gunasekaran, Nachiappan Subramanian y Wai Ting Eric Ngai. *Quality management in the 21st century enterprises: Research pathway towards Industry 4.0*. Vol. 207. 2019, págs. 125-129. DOI: https://doi.org/10.1016/j.ijpe.2018.09.005. URL: https://www.sciencedirect.com/science/article/pii/S092552731830375X.

[22] V. Dubolazov y col. *Re-Engineering of Logistics Business Processes Influenced by the Digitalization*. Vol. 246. 2021, págs. 539-547. ISBN: 978-303081618-6.

[23] T. Narayanaswamy, P. Karthika y Kandappan Balasubramanian. *Blockchain Enterprise: Use Cases on Multiple Industries*. Cham: Springer International Publishing, 2022, págs. 125-137. ISBN: 978-3-030-76216-2. DOI: 10.1007/978-3-030-76216-2_8. URL: https://doi.org/10.1007/978-3-030-76216-2_8.

[24] Ashraf Jaradat, Omar Ali y Ahmad AlAhmad. *Blockchain Technology: A Fundamental Overview*. Singapore: Springer Singapore, 2022, págs. 1-24. ISBN: 978-981-16-6301-7. DOI: 10.1007/978-981-16-6301-7_1. URL: https://doi.org/10.1007/978-981-16-6301-7_1.

[25] N. Puri, V. Garg y R Agrawal. *Blockchain Technology Applications for Next Generation*. Springer Science y Business Media Deutschland GmbH, 2022. Cap. Blockchain Technology Applications for Next Generation, págs. 53-73.

[26] A.A. Khan y col. *EPS-ledger: Blockchain hyperledger sawtooth-enabled distributed power systems chain of operation and control node privacy and security*. 2395. Oct. de 2021.

[27] R. S. M. Lakshmi Patibandla y Lakshman Narayana Vejendla. *Significance of Blockchain Technologies in Industry*. Springer Science y Business Media Deutschland GmbH, 2022. Cap. Signifcance of Blockchain Technologies in Industry, págs. 19-31.

[28] Q. Liu. *Construction of Blockchain Technology Audit System*. Vol. 97. Springer Science y Business Media Deutschland GmbH, 2022. Cap. Construction of Blockchain Technology Audit System.

[29] Faiza Loukil y col. *Decentralized collaborative business process execution using blockchain*. 2021, págs. 1645-1663.

[30] Moumita Das y col. *A blockchain-based integrated document management framework for construction applications*. Vol. 133. 2022, pág. 104001. DOI: https://doi.org/10.1016/j.autcon.2021.104001. URL: https://www.sciencedirect.com/science/article/pii/S0926580521004520.

Strategies for agile software development based on technical and environmental complexity factors

Fernando Pinciroli[1]

[1]Instituto de Investigaciones, Universidad Nacional de San Juan,
San Juan, Argentina
fernando.pinciroli@gmail.com

Abstract. Project management strategies require a perfect understanding of the problem to be faced. For this, the Cynefin framework is a good starting point, offering a way to classify reality according to its complexity. However, there are other aspects that impact and increase the complexity of projects. One of them is the number of triple constraint elements that are fixed; the other, corresponds to project characteristics such as size, criticality, time constraint, etc. In this work we offer a strategy to classify the best tools, techniques and approaches for project management depending on technical and environmental complexity factors.

Keywords: project management, agile methodologies, Cynefin framework.

1 Introduction

The management of software development projects is complex. It requires dealing with numerous unforeseen events that constantly arise along the way and that go against the expectations that had been established at the beginning. A good project leader is not so much the one who carries out what is planned, but rather the one who is able to deal with the inconveniences that arise and, in the end, achieve a decent outcome [1].

On the other hand, we are used to cling to the tools that gave us the best results, although many times we continue to do so even when the context has changed.

Also, software development is complex, both because software is inherently so and because people, who are an essential and intensive part of that development, and our relationships are even more so.

Finally, for some decades we have extrapolated many proven techniques, tools and approaches for project management from other fields to computer science [2]. Surely many of them were of value, but others did not give the expected results.

The appearance of the Manifesto for Agile Software Development [3], in 2001, radically changed the way of seeing project management and was like a breath of fresh air, but there are still several difficulties that sometimes arises:

- Agile tools and techniques applied in projects, but without an agile approach.

- Reality is not analyzed to determine the best tools to use.
- Agility implies adaptation, but there are not too many cases in which tools and techniques are adapted or combined depending on the problem to be faced.
- Many people are trained in a single agile method and use it systematically.

2 Mapping tools, techniques, and approaches

Snowden created the Cynefin framework [4], which describes a way to categorize reality based on its complexity. The model consists of five contexts that are defined by the nature of the cause-effect relationship present in them. The contexts are called "clear", "complicated", "complex", "chaotic", and "aporetic". The first four can be identified by the characteristics that are observed, but the fifth corresponds to the situation in which it was not possible to characterize the context for whatever reason.

It is also widely known the triple restriction on projects: scope, time and cost [2] [5]. In the projects, some of these three restrictions are fixed, forcing the rest to be variables, in order to be able to manage the project. This characterizes the type of management that can be carried out, whether predictive or empirical.

Based on Cynefin and the characterization of projects regarding their fixed and variable restrictions, we proceeded to locate different project management techniques, tools, and approaches in the four contexts of the former. The objective of this mapping is to guide the selection of project management strategies.

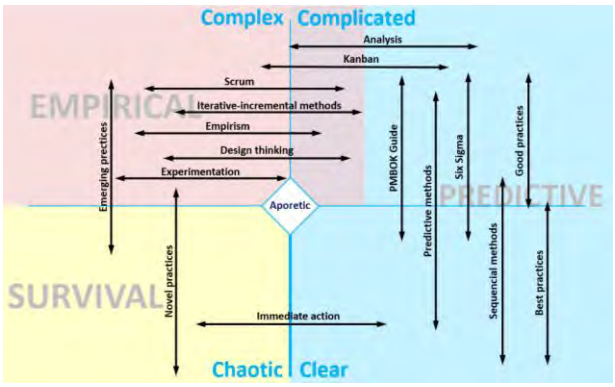


Fig. 1. Tools, techniques, and approaches for project management in Cynefin framework.

In this line of thought, Kenneth Rubin makes a first approach to locate the techniques in each Cynefin's context [6]. It places the waterfall style, in the clear context; an analytical method, like Six Sigma, in the complicated context; and an iterative and incremental approach, like Scrum, in the complex one. Paulus et al. [7] also classify sequential development, based on a plan, as an appropriate approach for a clear context. Pelrine [8] understands that a project is not complex or complicated in

itself, but rather its constituent parts that can be classified independently in the four Cynefin's contexts.

3 Technical and environmental complexity factors

There are other factors that influence the complexity of the projects and that would move them from context to one of higher complexity: time constraints, knowledge of the problem domain, aspects of interest to stakeholders, size of the project (in people, requirements, duration, etc.), criticality, global software development, culture and maturity of the organization, etc. [9].

Alistair Cockburn establishes a principle that the size of the project management methodology to be used will be determined by the presence of these factors [10] [11]. Along with Jim Highsmith, Alistair states that agile development is more difficult with larger teams [12], referring to projects involving hundreds of people, indicating that extra precautions must be taken due to the increasing complexity of the project.

Thus, an agile development complemented with the necessary methodological elements to manage the complexity added by the size of the problem, could be an adequate solution, since a waterfall model would not be adequate in these cases [13]. Serrador and Pinto also found evidence that larger projects require a combination of agile management with a higher degree of advance planning [14].

Alistair [10] mentions that the criticality of the problem demands a greater methodology and McConnell confirms it through another way [15], by presenting the productivity that can be obtained according to the criticality of the problem. Other authors recognize the contribution made by approaches based on a plan to projects that require important levels of security, reliability and protection [7] [14].

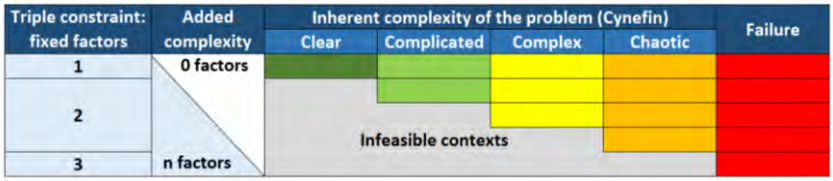


Fig. 2. Possible scenarios depending on the project's complexity.

We have designed a problem classification scheme in which possible scenarios are presented in the presence these complexity factors (Fig. 2). We must add to the Cynefin's contexts the complexity added by the number of fixed elements of the triple restriction and the number of factors that impact a project complexity. Thus, we present the contexts that are possible and those that are not, as the technical and environmental complexities of a project grow.

5 Conclusions and future work

The nature of a project should be characterized in order to select the best strategies to manage it. This characterization could be done based on the triple restriction and the added complexity factors present in the challenge to be faced. We have proposed a strategy to locate every project in the Cynefin framework depending on its complexity and then choosing the most appropriate techniques, tools, and management approaches. This selection must continue throughout the life of the project because reality changes and the project can move from one Cynefin context to another, with the consequent need to employ new techniques, tools, and management approaches to keep the project under control.

As future work, it remains to obtain more evidence from the literature and describe with more detail the impact of each factor of technical and environmental complexity.

References

1. McConnell, S. Software project survival guide. Redmond, Microsoft Press, 1998.
2. Project Management Institute. A guide to the project management body of knowledge. 6th ed. Project Management Institute, 2017.
3. "Manifesto for Agile Software Development". https://agilemanifesto.org/
4. Snowden, D.J. and Boone, M.E. "A Leader's Framework for Decision Making". Harvard Business Review, p. 10, 2007.
5. Atkinson, R. "Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria". International Journal of Project Management, vol. 17, no. 6, pp. 337–342, 1999.
6. Rubin, K.S. Essential Scrum: a practical guide to the most popular agile process. Upper Saddle River, Addison-Wesley, 2012.
7. Paulus, S., Mohammadi, N.G. and Weyer, T. "Trustworthy Software Development". In: Communications and Multimedia Security, vol. 8099, De Decker, B. et al. eds. Berlin, Springer, pp. 233–247, 2013.
8. Pelrine, J. "On Understanding Software Agility – A Social Complexity Point Of View", vol. 13, p. 13, 2011.
9. Špundak, M. "Mixed Agile/Traditional Project Management Methodology – Reality or Illusion?". Procedia - Social and Behavioral Sciences, vol. 119, pp. 939–948, 2014.
10. Cockburn, A. "Selecting a project's methodology" IEEE Software, vol. 17, no. 4, pp. 64–71, 2000.
11. Cockburn, A. "People and Methodologies in Software Development," PhD Thesis, University of Oslo, 2003.
12. Cockburn A. and Highsmith, J. "Agile software development, the people factor". Computer, vol. 34, no. 11, pp. 131–133, 2001.
13. Petersen, K., Wohlin, C. and Baca, D. "The Waterfall Model in Large-Scale Development" In: Product-Focused Software Process Improvement, vol. 32, Bomarius, F. et al. eds. Berlin, Springer, pp. 386–400, 2009.
14. Serrador P. and Pinto, J.K. "Does Agile work? – A quantitative analysis of agile project success". International Journal of Project Management, vol. 33, no. 5, pp. 1040–1051, 2015.
15. McConnell, S. Software estimation: demystifying the black art. Redmond, Microsoft Press, 2006.

Hacia la Recomendación Automática de Patrones de Diseño Ontológico

Tomás Quiñonez[1], Christian Gimenez[1], Laura Cecchi[1], and Pablo Fillottrani[2,3]
tomas.quinonez@est.fi.uncoma.edu.ar
{christian.gimenez,lcecchi}@fi.uncoma.edu.ar
prf@cs.uns.edu.ar

[1] Grupo de Investigación en Lenguajes e Inteligencia Artificial
Facultad de Informática
UNIVERSIDAD NACIONAL DEL COMAHUE
Neuquén, Argentina
[2] Laboratorio de I&D en Ingeniería de Software y Sistemas de Información - Departamento de Ciencias e Ingeniería de la Computación
UNIVERSIDAD NACIONAL DEL SUR, Bahía Blanca, Argentina
[3] Comisión de Investigaciones Científicas de la provincia de Buenos Aires (CIC), Buenos Aires, Argentina

Resumen En la Ingeniería Ontológica, los modeladores que desean construir una ontología por medio del reuso de patrones poseen poca asistencia en las herramientas de desarrollo. Por ello, se propone una metodología que permita recomendar a los usuarios Patrones de Diseño Ontológico (ODP) para una ontología en etapa de diseño.
La metodología propuesta posee dos etapas: el análisis del patrón para extraer información relevante para detectarlos parcial o totalmente en una ontología, y el análisis de la ontología del usuario para buscar y sugerir patrones.
De esta manera, se presenta una metodología implementable para sugerir patrones a medida que es diseña una ontología. Así, los modeladores sin conocimiento previo de los ODP, podrán seleccionar patrones ampliamente aceptados mejorando el proceso de desarrollo y la calidad.

Palabras Clave: Ingeniería de Software basada en Conocimiento, Patrones de Diseño Ontológico, Ontologías, Lógicas Descriptivas.

1. Introducción

La Ingeniería Ontológica estudia los métodos y metodologías que guían a los modeladores en el diseño, desarrollo, implementación, mantenimiento, uso y publicación de ontologías [7].Una posible forma de asistir al modelador es a través del uso de patrones de diseño ontológicos (ODP) [4,5]. Estos patrones son considerados pequeñas ontologías bien definidas y aceptadas por la comunidad, y son usados como modelos o *templates* para ser incorporados a la ontología en desarrollo.

Sin embargo, los enfoques para la Ingeniería Ontológica basados en patrones requieren, por un lado, de la existencia de un conjunto de patrones adecuados y aceptados

en la comunidad para ser reusados. Y por el otro, de metodologías apropiadas que soporten la elicitación de estos patrones y su aplicación en la construcción de nuevos modelos.

Por otra parte, un modelador que desea construir, mantener o validar una ontología a través del reuso de patrones, posee poca asistencia en las herramientas de desarrollo, respecto del uso de estructuras lógicas, que generalmente son poco amigables, haciendo así que las ontologías sean difíciles de comprender [1].

En la literatura, existen pocas herramientas desarrolladas que incluyan soporte para el modelado basado en patrones: CoModIDE [9] y ODPReco [10]. Estas herramientas utilizan un lenguaje gráfico *ad-hoc* basado en grafos para el modelado, lo que implica familiarizarse con este nuevo lenguaje o no proveen soporte gráfico. Asimismo, existen herramientas visuales, pero que no proveen metodologías para el uso de patrones.

Por consiguiente, las tareas de Ingeniería Ontológica en conjunción con la integración de metodologías y buenas prácticas en el uso de patrones, que pueda ser seleccionado en forma amigable de un catálogo, en ambientes gráficos de modelado ontológico, es una arista no explorada en profundidad [6].

En este trabajo, se propone una metodología para sugerir ODP, considerando un modelo ontológico en desarrollo. En particular, se trabaja con Patrones de Contenido, considerando que esta clase de patrones resultan muy útiles en el proceso de modelado [1]. Los Patrones de Contenido se obtendrán de un listado ampliamente aceptado por la comunidad de expertos en el tema [8].

En este sentido, se espera ampliar con la implementación de dicha metodología, el soporte a la Ingeniería Ontológica de las herramientas con ambiente gráfico, como por ejemplo crowd [3,2]. De este modo, los modeladores podrán utilizar esta metodología para diseñar sus ontologías, a partir del reuso, extensión e integración de uno o varios patrones del catálogo propuesto.

2. Metodología

La metodología propuesta se diseñó estructurándola como pasos o procedimientos, donde cada uno de ellos recibe y emite productos o información. La Figura 1 presenta el diseño completo, dividido en dos etapas. En dicha imagen, se representa la información de entrada y salida con figuras geométricas ovaladas.

La primer etapa, *Preprocesamiento de Patrones*, tiene como objetivo la extracción de información relevante de los patrones para la detección de los mismos en la ontología del usuario. El *Procesamiento de Nombres* extraerá los nombres de todos los elementos presentes en todos los patrones, y el *Procesamiento de Axiomas* extraerá todos los axiomas de dichos patrones. Esta etapa utilizará una lista de sinónimos para los nombres utilizados en cada uno de los patrones, en caso de que la ontología del usuario utilice algunos de estos sinónimos. Luego, con los nombres y los axiomas extraídos de los patrones, se procede a formar dos grupos de consultas: El primer grupo contendrá todas las consultas de nombres a realizar al razonador; y el segundo grupo contendrá todas las consultas de axiomas presentes en los patrones.

La segunda etapa, *Analizador de Ontologías*, utilizará la información obtenida del submódulo anteriormente mencionado y de la ontología del usuario, para formular con-

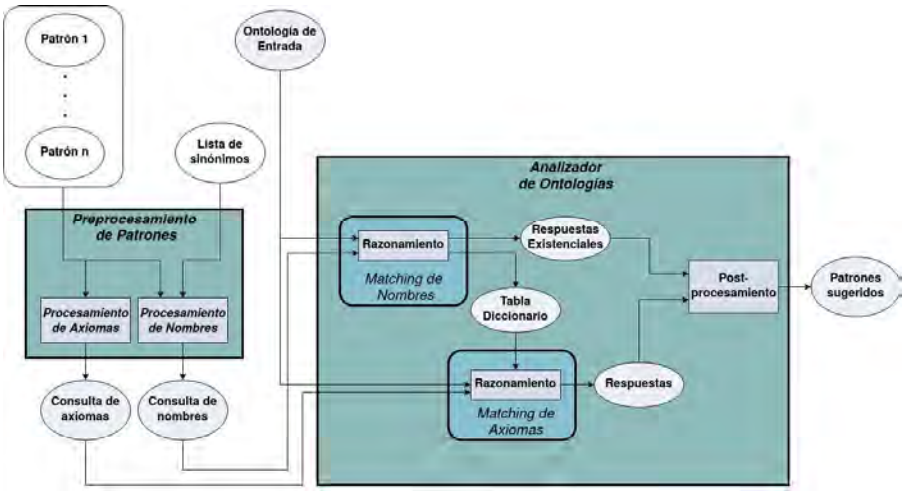


Figura 1. Descripción gráfica de la metodología.

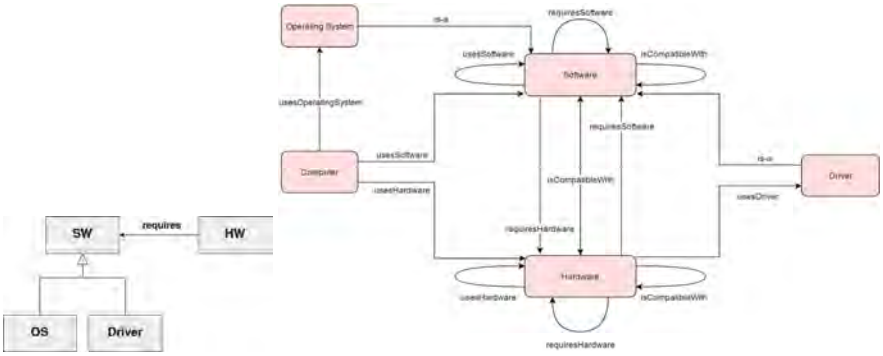
sultas a los razonadores. Inicialmente, se realizará el *Matching de Nombres*, en el que se tomará como entrada las consultas de nombres, junto con la ontología y se le ejecutarán dichas consultas al razonador. Esta tarea realizará una comparación de los nombres (o sinónimos) de los elementos involucrados en la ontología con cada uno de los nombres de los elementos de los patrones para encontrar coincidencias.

A medida que se obtienen los resultados de las consultas, se irán registrando los nombres con sus respectivos sinónimos a la Tabla Diccionario. Esta información será utilizada posteriormente por la tarea *Matching de Axiomas* para verificar que los elementos de un axioma están presentes en la ontología, y para reformular las consultas en base a los sinónimos en caso de ser necesario.

Esta tarea compara los axiomas que existen en la ontología con los axiomas de los patrones. Esto incluye: cardinalidades de las relaciones entre clases, tipos de las relaciones de herencia (disjunta, *covering*), tipos de datos, etc., realizando los cambios necesarios según la Tabla Diccionario poblada durante la tarea *Matching de Nombres*.

Finalmente, luego de este análisis, se hará un postprocesamiento de los resultados obtenidos a fin de determinar los patrones que serán sugeridos al modelador. En este sentido, se realizarán dos tipos de evaluaciones. Primero, se analizará el porcentaje de clases de cada patrón presente en la ontología. Segundo, se analizará la cantidad de nombres de relaciones y de consultas afirmativas de los axiomas del patrón, que están presentes en la ontología. El patrón será sugerido, si los valores obtenidos en ambas evaluaciones supera un umbral particular establecido para cada condición.

Así, la salida de esta etapa será una lista con los patrones sugeridos, que ocurren total o parcialmente en la ontología. Un punto interesante a mencionar es que la complejidad computacional de los algoritmos que implementen la metodología se vislumbra intratable respecto de la cantidad de clases y axiomas.



(a) Parte de una ontología de sistemas informáticos (b) Patrón completo *Computer System*

Figura 2. Vista parcial de una ontología y patrón ontológico

3. Ejemplo de aplicación de la Metodología

En la Figura 2 se presenta una vista parcial de una ontología simple, que representa un sistema informático y el patrón *Computer System*.

Inicialmente, la etapa de *Preprocesamiento de Patrones*, el *Procesamiento de Nombres* recorre el patrón *Computer System* y extrae los siguientes nombres: `Operating System`, `Software`, `Driver`, `Hardware`, `requiresSoftware`, etc. Por otra parte, selecciona de la Lista de Sinónimos los siguiente nombres: SW, HW, Controlador y OS. El *Procesamiento de Axiomas* extrae los siguientes axiomas: `Operating System` es subclase de `Software`, `Driver` es subclase de `Software`, `Hardware` se relaciona con la clase `Software` a través de la relación `requiresSoftware`.

A partir de esto, se formulan las consultas que se realizan a los razonadores sobre la ontología del usuario, para detectar la presencia de los elementos del patrón. Las consultas a elaborar son las siguientes: la clase `Operating System`, ¿es subclase de `Software`?, la clase `Driver`, ¿es subclase de `Software`?, y la clase `Hardware`, ¿se relaciona con la clase `Software` a través de la relación `requiresSoftware`?

Luego del preprocesamiento, el sistema se encuentra listo para analizar ontologías del usuario. En este caso, el usuario ingresa la ontología de la Figura 2 (a) obteniendo una Tabla Diccionario con los valores SW = Software, HW = Hardware, OS = Operative System y Driver = Driver y las Respuestas Existenciales de que dichas clases existen. A continuación, el proceso de *Matching de Axiomas* resuelve las consultas elaboradas previamente. En otras palabras, se confirma que Hardware está en relación `requiresSoftware` con Software y las consultas de herencia entre Software y sus subclases en esta ontología de entrada. Todas las Respuestas y las Respuestas Existenciales serán Post-procesadas.

En el Post-procesamiento, se realizan las evaluaciones, detectando un $80\,\%$ de las clases del patrón en la ontología. Asimismo, el análisis permite concluir que el nombre de la relación `requiresSotftware`, y tres axiomas del patrón se encuentran presentes en la ontología. Suponiendo un umbral del $50\,\%$ para la cantidad de clases en el

patrón detectadas en la ontología y considerando los resultados del análisis, el patrón *Computer System* es sugerido al usuario.

4. Conclusiones y Trabajos Futuros

En este trabajo se presentó el desarrollo de una metodología que recomienda ODP, en la que se consideran Patrones de Contenido obtenidos de un listado ampliamente aceptado. La metodología está dividida en dos etapas. La primer etapa analiza los patrones para obtener una serie de consultas necesarias para detectarlos. En la segunda etapa, recibe la ontología del usuario y sugiere los patrones utilizando las consultas generadas. De esta forma, la metodología permitirá sugerir patrones a medida que el usuario modela su ontología.

Actualmente, se encuentra en desarrollo la implementación de la metodología en una herramienta que puede ser consultada utilizando un *API REST*. Esta decisión fue tomada, teniendo en cuenta que la funcionalidad de la misma no debería depender de una aplicación o software preexistente. Entre nuestros trabajos futuros, se propone una extensión de la herramienta visual Web crowd [3,2], para la Ingeniería Ontológica, permitiendo la utilización de las funcionalidades de dicha API.

Este soporte permitirá a los modeladores sin experiencia en sistemas formales, seleccionar patrones ya aceptados en la comunidad, como base desde donde comenzar sus diseños, acelerando el proceso de desarrollo y mejorando la calidad de sus ontologías al hacerlas más modulares y reusables.

Referencias

1. E. Blomqvist, A. Gangemi, and V. Presutti. Experiments on pattern-based ontology design. In *Proceedings of the fifth international conference on Knowledge capture*, pages 41–48, 2009.
2. Germán Braun, Christian Gimenez, Laura Cecchi, and Pablo Fillottrani. `crowd`: A Visual Tool for Involving Stakeholders into Ontology Engineering Tasks. *KI - Künstliche Intelligenz*, 2020.
3. Germán Braun, Elsa Estevez, and Pablo Fillottrani. A Reference Architecture for Ontology Engineering Web Environments. *Journal of Computer Science and Technology*, 19(01), Apr. 2019.
4. Aldo Gangemi and Valentina Presutti. Ontology design patterns. In Steffen Staab and Rudi Studer, editors, *Handbook on ontologies*, pages 221–243. Springer, 2009.
5. Pascal Hitzler, Aldo Gangemi, and Krzysztof Janowicz. *Ontology engineering with ontology design patterns: foundations and applications*, volume 25. IOS Press, 2016.
6. Pascal Hitzler and Cogan Shimizu. Modular ontologies as a bridge between human conceptualization and data. In *International Conference on Conceptual Structures*, pages 3–6. Springer, 2018.
7. C. M. Keet. *An Introduction to Ontology Engineering*. University of Cape Town, 2018.
8. NeON Project. Ontology Design Patterns - category proposed contents OP. http://ontologydesignpatterns.org/ Visitado por última vez el 22 de agosto del 2022.
9. Cogan Shimizu and Karl Hammar. CoModIDE–The Comprehensive Modular Ontology Engineering IDE. In *ISWC 2019 Satellite Tracks*, volume 2456. CEUR-WS, 2019.
10. Maleeha Arif Yasvi and Raghava Mutharaju. ODPReco-A Tool to Recommend Ontology Design Patterns. In *WOP@ ISWC*, pages 71–75, 2019.

Evaluación de la usabilidad de APIs web

Ariel Machini[1] and Sandra Casas[2]

[1] Centro de Investigaciones y Transferencias Santa Cruz (CONICET/UNPA/UTN)
amachini@conicet.gov.ar
[2] Instituto de Tecnología Aplicada (ITA) - Universidad Nacional de la Patagonia Austral (UNPA UARG)

Abstract. En los últimos años las APIs web se han convertido en componentes clave para agilizar el proceso de construcción de aplicaciones. Debido a esto, resulta importante que estas sean fáciles de aprender y utilizar para no reducir la productividad de los programadores ni obstaculizar el proceso de desarrollo. La usabilidad de las APIs es entonces considerada un factor fundamental para su correcta adopción, y si bien existen estudios que proponen soluciones para evaluarla y mejorarla, estos abarcan mayormente APIs locales; están más centrados sobre la documentación; suelen contemplar una limitada cantidad de características de usabilidad y, además, los resultados de investigaciones sobre usabilidad de APIs en general aún son insuficientes. Por estas razones, a través del presente trabajo se propone resolver algunos de los problemas recién mencionados mediante el estudio de la especificación OpenAPI y la construcción de un framework de evaluación de la usabilidad de APIs web.

Keywords: APIs Web · Métricas · Usabilidad · Framework · OpenAPI

1 Motivaciones, problemas y objetivos

En la última década, las APIs web se han convertido en uno de los pilares del desarrollo de aplicaciones modernas [1]. Además de permitir la reutilización de software y el acceso a servicios y recursos *de* y *entre* organizaciones, han producido una nueva y real perspectiva de negocio, la **Economía API** [2]. El mercado de APIs web es muy competitivo, por lo que se considera clave la facilidad de uso de las APIs para los desarrolladores potenciales; lo cual va en línea con [3], que resuelve que el valor de una API se identifica con su usabilidad. Por lo tanto, cualquier buena API debería ser fácil de aprender y usar y traducirse en la productividad de los desarrolladores [3]. Según la ISO 9241-11:1998, se define como "la medida en que un producto puede ser utilizado por usuarios específicos para lograr objetivos específicos con efectividad, eficiencia y satisfacción en un contexto de uso específico". [5] afirman que "los diseñadores de API deben agregar la usabilidad como un criterio explícito de diseño y evaluación para no crear APIs inutilizables sin darse cuenta". Así, existen diversos estudios que proponen métodos y modelos de análisis, mejora y evaluación de la usabilidad de APIs, pero se observa: (i) están enfocados mayormente en APIs locales, (ii) están centrados principalmente en la documentación de las APIs, (iii) en cuanto a los enfoques

métricos, el problema común es que los trabajos están "algo" asociados con un modelo de usabilidad, pero las métricas derivadas no están relacionadas con una característica de usabilidad o están asociadas solo con algunas de estas características de usabilidad [6] y (iv) los resultados de investigaciones en usabilidad de APIs en general, y web en particular aún son insuficientes [6]. En la búsqueda de estrategias que sirvan tanto a proveedores y a consumidores de APIs web, se propone resolver algunos de los problemas planteados. Concretamente, se estudiará la especificación OpenAPI para el análisis y evaluación de la usabilidad de APIs web. ¿Qué aspectos de la usabilidad pueden medirse desde este artefacto?, ¿cómo hacerlo?, y ¿qué métricas de usabilidad se pueden obtener? son algunas preguntas que este trabajo pretende responder. El **objetivo general** de este trabajo es proporcionar soluciones consistentes, integrales y automáticas de análisis y evaluación de usabilidad de APIs web, y los **objetivos específicos** son (i) estudiar, identificar e integrar, métodos y heurísticas de análisis y evaluación de la usabilidad de APIs; (ii) formular teórica y conceptualmente un framework para el análisis y evaluación de la usabilidad de APIs web, aplicando el paradigma GQM (Goals-Questions-Metrics) y (iii) desarrollar una herramienta software que de soporte de automatización al framework propuesto.

2 Antecedentes

APIs como interfaces de librerías de códigos, frameworks o fuentes de datos, para liberarse de las tareas de programación de bajo nivel y/o acelerar el desarrollo [8]. A diferencia de las API locales, una nueva generación de APIs, las denominadas APIs de servicios web, ofrecen un enfoque sistemático y extensible para integrar servicios en aplicaciones [9, 10]. La cantidad de servicios y recursos que proporcionan las APIs web disponibles al público ha aumentado rápidamente [11]. Varios estudios señalan que los desarrolladores han pasado de SOAP o RPC a la implementación de servicios web de transferencia de estado representacional (REST) como medio para que los consumidores utilicen sus servicios [11–14]. Esto puede corroborarse con compañías IT como Google, Facebook o Amazon, las cuales implementaron servicios REST para proporcionar un fácil acceso a sus recursos de datos mientras promueven sus negocios [11]. La arquitectura REST fue introducida en el año 2000 y se basa en los principios que sustentan la WWW [15]. A pesar de esto, REST es simplemente un estilo arquitectónico sin especificaciones estándar. Los desarrolladores de APIs web deben decidir y definir cómo se debe exponer una API, qué características debe poseer o cómo se proporciona la documentación. Este último factor ha sido problemático, ya que sin medios estándares para documentar las APIs, la tendencia fue utilizar texto para describir la API (muchas veces en lenguaje natural), que puede ser diverso en forma, estructura o profundidad y es obviamente un problema para sus consumidores [9]. Para superar este obstáculo, se propusieron varias especificaciones de la interfaz de las APIs web (WADL, API Blueprint, RAML), pero la que se ha consolidado como un estándar fue la especificación OpenAPI. Esta

es abierta, portátil, neutral e independiente del lenguaje, de definir los recursos y las operaciones de una API REST, ya sea en JSON o YAML.

3 Usabilidad y APIs web

Ejecutar tareas con efectividad, eficiencia y satisfacción [16] son factores determinantes en el éxito de las aplicaciones de software, por ello, la usabilidad adquiere una importancia cada vez mayor en el desarrollo de software [17]. Varias disciplinas y métodos proponen un diseño centrado en el usuario para asegurar que se cumplan los principios básicos de la usabilidad en una interfaz: facilidad de aprendizaje, facilidad de uso, flexibilidad y robustez. La usabilidad de un software se evalúa realizando pruebas para recopilar datos y hallar debilidades relacionadas con el uso de la misma. Existen cuatro formas básicas de evaluación: automática, empírica, formal e informal. Este proceso de evaluación implica actividades acordes al método empleado, sin embargo, estos comparten [18]: Captura, análisis y crítica.
La primera investigación en usabilidad de APIs [19] afirma que los programadores son usuarios de las APIs en algunos casos y, por ende, necesitan librerías que sean fáciles de aprender y usar, y proponen una lista de atributos a considerar al analizar la usabilidad de las APIs. Otros autores siguen un enfoque metodológico para la usabilidad de APIs, como [6]. Clarke [20] basó su trabajo en la adaptación del framework de *13 dimensiones cognitivas* de [21]. También propone seguir un enfoque centrado en el usuario para diseñar APIs usables que empleen escenarios, para así garantizar que reflejen las tareas que los usuarios desean realizar en lugar de sus detalles de implementación; y establece perfiles de desarrollador, que afectan la forma de analizar la usabilidad de la API. Sin embargo, el trabajo de Clarke tiende a ser demasiado abstracto para poder ser aplicado por programadores [22] y excluye aspectos importantes de usabilidad. Hay aportes como [3, 23] más enfocados en las necesidades de los programadores, que abordan problemas específicos o elecciones que deben tomarse en el diseño de API y proponen soluciones y cursos de acción. Otros trabajos, como [3], ofrecen una comprensión de qué son las APIs y por qué son utilizadas por desarrolladores. Establecen que el valor de una API se identifica en gran medida por su poder y, más importante, su usabilidad. Para que una API sea usable, debe ser eficiente, satisfactoria, fácil de aprender y memorizar y tener errores mínimos [4]. Otro tipo de contribuciones, refiere al desarrollo de directrices, lineamientos o guías para ayudar a los programadores a desarrollar APIs usables [24–28]. En una línea diferente, el problema de la usabilidad de las APIs se aborda desde el punto de vista de las métricas de software. [29] propuso 12 métricas estrictamente de complejidad. [30] propusieron medir la usabilidad de APIs como una función de su complejidad. [31] propusieron 9 métricas estructurales que se crearon específicamente para medir la usabilidad de la API. [32] definieron el *API Concepts Framework*, un marco extensible para medir la complejidad de la interfaz. Si bien el trabajo ignora algunos aspectos de usabilidad por la dificultad que representa el medirlos automáticamente, los otros aspectos de la usabilidad

están bien medidos. [33] publicaron una revisión del estado del arte de los estudios de usabilidad de API y afirman que "los diseñadores de API deben agregar la usabilidad como un criterio explícito de diseño y evaluación para no crear una API inutilizable sin darse cuenta" y promueven el uso de métodos centrados en el ser humano para mejorar la API.

4 Metodología y actividades

Como marco metodológico general se usará el enfoque de investigación DSR [34]. Siguiendo este enfoque, las actividades principales y técnicas que se aplicaran son: ***Recopilación y análisis bibliográfico***. Se realizará una revisión sistemática para identificar e integrar el corpus relacionado a usabilidad de APIs (en general) y particularmente a APIs web. El resultado de esta actividad es la sistematización integrada y compatibilizada del vocabulario que es heterogéneo, modelos y métodos de análisis y evaluación, enfoques métricos, guías, directrices y pautas; ***Formulación del framework de usabilidad de APIs web***. Para esta actividad se aplicará el paradigma GQM [7]. Como resultado se obtiene un conjunto de atributos/factores de usabilidad relevantes a APIs web y métricas asociadas; ***Validación del framework y ajustes***. El framework obtenido en la actividad anterior será validado mediante una técnica empírica como las entrevistas a desarrolladores (proveedores y consumidores de APIs) y/o expertos en el tema. Los resultados obtenidos se usarán para mejorar el framework; ***Desarrollo de una herramienta*** que de soporte de automatización al framework propuesto; ***Validación***. Se realizará una evaluación de un conjunto de al menos 100 APIs web reales con la herramienta y los resultados se analizarán estadísticamente.

References

1. Raemaekers, S., Van Deursen, A., & Visser, J. (2012). *Measuring software library stability through historical version analysis.* 28th IEEE International Conference on Software Maintenance (pp. 378-387). IEEE.
2. Tan, W., Fan, Y., Ghoneim, A., Hossain, M. A., & Dustdar, S. (2016). *From the service-oriented architecture to the web API economy.* IEEE Internet Computing, 20(4), 64-68.
3. Stylos, J., & Myers, B. (2007). *Mapping the space of API design decisions.* IEEE Symposium on Visual Languages and Human-Centric Computing (pp. 50-60). IEEE.
4. Nielsen, J. (1992). *The usability engineering life cycle.* IEEE Computer, 25(3), 12-22.
5. Myers, B. A., & Stylos, J. (2016). *Improving API usability.* Communications of the ACM, 59(6), 62-69.
6. Mosqueira-Rey, E., Alonso-Ríos, D., Moret-Bonillo, V., Fernández-Varela, I., & Álvarez-Estévez, D. (2018). *A systematic approach to API usability: Taxonomy-derived criteria and a case study.* Information and Software Technology, 97, 46-63.
7. Van Solingen, R., Basili, V., Caldiera, G., & Rombach, H. D. (2002). *Goal Question Metric (GQM) approach.* Encyclopedia of software engineering.

8. Dagenais, B., & Robillard, M. P. (2011). *Recommending adaptive changes for framework evolution*. ACM Transactions on Software Engineering and Methodology, 20(4), 1-35.
9. Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002). *Unraveling the web services Web: an introduction to SOAP, WSDL, and UDDI*. IEEE Internet Computing, 6(2), 86-93.
10. Vinoski, S. (2008). *RESTful web services development checklist*. IEEE Internet Computing, 12(6), 96-95.
11. Maleshkova, M., Pedrinaci, C., & Domingue, J. (2010). *Investigating web APIs on the World Wide Web*. 8th IEEE European Conference on web Services (pp. 107-114). IEEE.
12. Renzel, D., Schlebusch, P., & Klamma, R. (2012). *Today's top "RESTful" services and why they are not RESTful*. International Conference on web Information Systems Engineering (pp. 354-367). Springer, Berlin.
13. Bülthoff, F., & Maleshkova, M. (2014). *RESTful or RESTless–current state of today's top web APIs*. European Semantic web Conference (pp. 64-74). Springer, Cham.
14. Kopecký, J., Fremantle, P., & Boakes, R. (2014). *A history and future of web APIs*. IT-Information Technology, 56(3), 90-97.
15. Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. Universidad de California, Irvine.
16. Harms, P., & Grabowski, J. (2014). *Usage-based automatic detection of usability smells*. International Conference on Human-Centred Software Engineering (pp. 217-234). Springer, Berlin.
17. Grau, X. F. (2000). *Principios Básicos de Usabilidad para Ingenieros Software*. JISBD (pp. 39-46).
18. Ivory, M. Y., & Hearst, M. A. (2001). *The state of the art in automating usability evaluation of user interfaces*. ACM Computing Surveys, 33(4), 470-516.
19. McLellan, S. G., Roesler, A. W., Tempest, J. T., & Spinuzzi, C. I. (1998). *Building more usable APIs*. IEEE Software, 15(3), 78-86.
20. Clarke, S. (2005). *Describing and measuring API usability with the cognitive dimensions*. Cognitive Dimensions of Notations 10th Anniversary Workshop (p. 131).
21. Green, T. R. G., & Petre, M. (1996). *Usability analysis of visual programming environments: A 'cognitive dimensions' framework*. Journal of Visual Languages & Computing, 7(2), 131-174.
22. Bore, C., & Bore, S. (2005). *Profiling software API usability for consumer electronics*. 2005 Digest of Technical Papers. International Conference on Consumer Electronics, 2005. (pp. 155-156). IEEE.
23. Ellis, B., Stylos, J., & Myers, B. (2007). *The factory pattern in API design: A usability evaluation*. 29th International Conference on Software Engineering (ICSE'07) (pp. 302-312). IEEE.
24. Jacques, M. (2004). *API Usability: Guidelines to improve your code ease of use*. Recuperado en Julio del 2021, de http://www.codeproject.com/Articles/8707/APIUsability-Guidelines-to-improve-your-code-ease
25. Henning, M. (2007). *API: Design Matters: Why changing APIs might become a criminal offense*. Queue, 5(4), 24-36.
26. Zibran, M. (2008). *What makes APIs difficult to use*. International Journal of Computer Science and Network Security (IJCSNS), 8(4), 255-261.

27. Zibran, M. F., Eishita, F. Z., & Roy, C. K. (2011). *Useful, but usable? Factors affecting the usability of APIs*. 18th Working Conference on Reverse Engineering (pp. 151-155). IEEE.
28. Grill, T., Polacek, O., & Tscheligi, M. (2012). *Methods towards API usability: A structural analysis of usability problem categories*. International conference on human-centred software engineering (pp. 164-180). Springer, Berlin.
29. Doucette, A. (2008). *On API usability: An analysis and an evaluation tool*. CMPT816-Software Engineering, Canada. Universidad de Saskatchewan.
30. de Souza, C. R., & Bentolila, D. L. (2009). *Automatic evaluation of API usability using complexity metrics and visualizations*. 31st International Conference on Software Engineering-Companion Volume (pp. 299-302). IEEE.
31. Rama, G. M., & Kak, A. (2015). *Some structural measures of API usability*. Software: Practice and Experience, 45(1), 75-110.
32. Scheller, T., & Kühn, E. (2015). *Automated measurement of API usability: The API concepts framework*. Information and Software Technology, 61, 145-162.
33. Daughtry, J. M., Farooq, U., Myers, B. A., & Stylos, J. (2009). *API usability: Report on special interest group at CHI*. ACM SIGSOFT Software Engineering Notes, 34(4), 27-29.
34. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). *A Design Science Research methodology for information systems research*. Journal of Management Information Systems, 24(3), 45-77.

Requisitos de Calidad de Software en Organizaciones Agiles

María Fernanda Burdino+, Carlos Salgado+, Mario Peralta+, Alberto Sánchez+

+Departamento de Informática Facultad de Ciencias Físico-Matemáticas y Naturales Universidad Nacional de San Luis
Ejército de los Andes 950 – C.P. 5700 – San Luis – Argentina
fburdino@gmail.com {csalgado, mperalta, alfanego }@unsl.edu.ar

Resumen: Desde la pandemia en el 2020, las organizaciones de desarrollo de software ahora son conscientes de los nuevos desafíos y oportunidades que se presentan, para las cuales deben estar preparadas para la era digital, adoptando metodologías ágiles de desarrollo. Adoptar metodologías ágiles significa mejorar el seguimiento del desarrollo del producto, cumpliendo con las funcionalidades y las fechas acordadas. La calidad del producto está íntimamente ligada a la calidad del proceso que se usa para desarrollarlo, es por ello que las organizaciones también deben pensar en contar con una certificación de calidad, la cual permita acreditar la calidad del proceso utilizado para el desarrollo del producto ofrecido. El enfoque a procesos establecido en la ISO 9001 bajo el ciclo Planificar-Hacer-Verificar-Actuar, permite a las organizaciones gestionar sus procesos de manera sistémica, permitiendo conocer las interrelaciones y dependencias entre los diferentes procesos organizacionales. Ni una implementación ágil cumplirá con los requisitos de la norma ISO 9001 por casualidad, ni la norma ISO puede abarcar, adoptar o embeber las metodologías ágiles por accidente. La norma ISO 9001 debe tener una adaptación a las metodologías ágiles y poder aun así cumplir con sus requisitos para obtener la certificación, permitiendo a las organizaciones que utilizan la ISO 90003 como guía para implementarla, contar con las prácticas habituales de las metodologías ágiles más habituales. El presente trabajo tiene como objetivo desarrollar una matriz de trazabilidad entre los requisitos para la implementación de la Norma ISO 9001:2015 en base a las prácticas comúnmente implementadas en organizaciones de desarrollo de software que utilizan las metodologías agiles

Palabras Claves: Calidad Proceso, Scrum, Requisitos xxxx, Certificación, Metodologías Ágiles, Iso 90003, iso9001

1. Introducción

La industria del software es uno de los sectores de mayor generación de empleo en la Argentina y según un informe de CESSI publicado en junio de 2022, se prevé generar exportaciones por $10.000 millones de dólares anuales incrementales, es por esta razón que

para cumplir con esta demanda no solo se necesitan recursos, sino la aplicación de metodologías que ayuden a organizar esos recursos y permitan entregar el producto a tiempo y con la calidad esperada. Para el desarrollo de estos productos las empresas deben utilizar un marco de procesos y actividades que sirven como base para comprender y comunicar sobre el sistema de software, según el estándar ISO/IEC/IEEE 12207:2017 [1], se lo denomina ciclo de vida. Este estándar proporciona requisitos relacionados con un marco de proceso común para describir el ciclo de vida de los sistemas de software.

El enfoque Agile para el desarrollo de software promueve un entorno de mejora continua y confianza dentro de una organización, lo que permite que las organizaciones respondan a los requisitos cambiantes. Ayuda a crear software en un entorno flexible a través de la colaboración de equipos multifuncionales autoorganizados.

La adopción del desarrollo de software ágil fue creciendo a lo largo de la última década, ya que se entendió que es una forma de mejorar la colaboración, capitalizar las fortalezas personales y la responsabilidad personal. Los enfoques tradicionales fueron perdiendo confianza en base a problemas conocidos como falta de precisión en las estimaciones, demoras en las entregas, dificultades en la adaptación al cambio de requisitos, requisitos del producto mal relevados, entre muchos otros. Según el State Agile Report 2020 [2], la adopción de metodologías ágiles ha crecido del 37% al 86% en el 2021 en todo el mundo y el objetivo de su implementación en las organizaciones es lograr la gestión del cambio y la entrega del software lo antes posible. La norma ISO 90003 proporciona las pautas para la aplicación de la norma ISO 9001:2015 [3] en las organizaciones de desarrollo de software. Los lineamientos brindados por esta norma aplican para la compra, abastecimiento, desarrollo, operación y mantenimiento del software. Esta norma es independiente de la tecnología, modelos del ciclo de vida, procesos de desarrollo, secuencia de actividades y estructura, usados por una organización. Esta norma no es certificable, pero proporciona los lineamientos necesarios para que una empresa de desarrollo de software pueda aplicarla, y de esta forma conseguir la certificación de ISO 9001:2015, logrando de esta manera una mejora en la calidad de los productos o servicios que ofrece.

El presente trabajo tiene como objetivo desarrollar una matriz de trazabilidad entre los requisitos para la implementación de la Norma ISO 9001:2015 en base a las prácticas comúnmente implementadas en organizaciones de desarrollo de software que utilizan la metodología Scrum, como así también se pretende obtener una lista de chequeo que facilite a los auditores controlar la adherencia de las organizaciones ágiles a los requisitos de la norma. En base a la investigación de la implementación combinada de metodologías ágiles y modelos de calidad, se identificaron alternativas basadas en Scrum a los requisitos de implementación planteados en la guía ISO 90003.

2. Estrategia Propuesta

Para realizar la investigación entre otros recursos/herramientas se recurrió a una encuesta

que permitió identificar prácticas y la adopción de las mismas en las organizaciones que implementan metodologías ágiles. Nos facilitó conocer en la industria de desarrollo de software en la Argentina qué metodologías ágiles son las más utilizadas, cuáles prácticas aportan mayor valor agregado a la gestión, las razones que llevó a estas empresas a implementar metodologías ágiles y cuáles fueron los resultados obtenidos al implementarlas. La encuesta constaba de tres secciones, en la primera sección se solicita información general de la organización como actividad principal, cantidad de empleados, exportación, certificaciones de calidad obtenidas, conocimiento y/o utilización de la norma ISO 90003 (ISO ORG), nivel de dificultad para implementar procesos agiles cumpliendo los requisitos de ISO 9001, y el tipo de metodología de desarrollo utilizada. En la segunda sección, se consultaba por el uso de metodologías ágiles. En la última sección se pregunta sobre las distintas prácticas de planificación, control y seguimiento, requerimientos, diseño y desarrollo, verificación y validación, correspondientes a la metodología SCRUM si es que era utilizada en las organizaciones.

El resultado de la encuesta fue utilizado para identificar qué requisitos de la norma ISO 90003 podrían estar cubiertos con los principios ágiles y las prácticas de SCRUM. Para ello se preparó una matriz de trazabilidad entre los requisitos de la ISO 90003 y las prácticas de SCRUM, además, se tuvieron que agregar prácticas de ingeniería de software para poder cumplir con la totalidad de los requisitos en ciertos puntos de la norma ISO.

3. Casos de Estudio

Para el presente trabajo, se han entrevistado a 29 organizaciones de desarrollo de software argentinas, intentando conocer la adopción de las metodologías ágiles en las mismas, ventajas y desventajas obtenidas, principales prácticas utilizadas y si las mismas estuvieron o están certificadas bajo la ISO 9001:2015 como así también si utilizaron la ISO 90003 para implementar la misma en la empresa y de esta manera obtener la certificación. A continuación, se presenta un resumen de los resultados obtenidos: El 62% de las empresas encuestadas exporta software por lo cual contar con una certificación de calidad como la ISO 9001:2015 les permite acceder a los mercados extranjeros con mayor facilidad, ya que cuenta con una garantía de calidad de su producto como es la implementación de la norma ISO. En consonancia con lo expresado anteriormente el 69% de las empresas, cuenta con certificación de calidad, de las cuales el 50% tiene certificación ISO 9001:2015. La norma ISO 90003 proporciona las pautas para la aplicación de la norma ISO 9001:2015 en las organizaciones de desarrollo de software, pero solo el 24% de las organizaciones encuestadas la utilizó como guía a la hora de certificar, y tan solo el 17% la conocía, pero no la utilizó. La definición de un proceso de desarrollo de software basado en la ISO 9001:2015, es algo que puede llevar tiempo y mucho trabajo a una organización, pero no es algo imposible, es por ello que las empresas en una escala que iba del 1 al 5, siendo uno el nivel de complejidad menor y 5 el de mayor complejidad, manifestaron que obtener la

certificación de calidad tiene un nivel medio de complejidad. Ahora, surge el planteamiento de que metodología o modelo de desarrollo es conveniente implementar, o si lograr la certificación de calidad bajo la norma ISO 9001 requiere de la utilización de los modelos tradicionales de desarrollo de software, la respuesta es no, es posible obtener la certificación, implementando metodologías ágiles, así lo demuestran, ya que el 86% de las empresas, utilizan algún tipo de metodología ágil, y tan sólo el 14% continúan desarrollando software bajo los modelos tradicionales, como son cascada, o en V. La metodología ágil mayormente implementada es SCRUM (45%), le sigue Kanban (34%), Lean (13%) y XP (8%). Esta misma elección se puede ver reflejada en el State Agile Report 2020, en dónde Scrum es la metodología más popular con el 66% de adeptos, siguiendo la combinación de Scrum con otras metodologías como son Scrum y Kanban (Scrumban) y Scrum y XP.

Otros indicadores utilizados por las empresas son historias de usuario planeadas vs actuales; satisfacción del cliente/usuario. En relación a la problemática tan recurrente respecto a las especificaciones de requerimientos ambiguas, incompletas, cambiantes, el 87% de las empresas utilizan para especificar sus requerimientos las historias de usuario, definen un diseño de alto nivel durante el sprint planning, el cual es refinado mientras transcurre el sprint. Respecto al involucramiento del Product Owner con el proyecto, las empresas mencionaron que el 54% de las mismas cuentas con la participación activa del Product Owner, por otra parte, el 42% cuenta a veces con este rol en su equipo y el 4% restante nunca. El Product Owner debería participar del reléase y del sprint planning, es quién debe asegurar que el Product Backlog está actualizado. El proyecto no debería comenzar sin un Product Owner involucrado. Las historias de usuario deben contar el criterio Done definido, para asegurar que todo el equipo conoce el criterio de aceptación. El 75% de las empresas encuestadas, lo define, y el resto lo define a veces.

Pero las metodologías ágiles no son explícitas en el uso de prácticas de ingeniería, ya que dejan en la decisión del equipo, como van a trabajar para desarrollar el producto. Las prácticas de ingeniería de software son tan importantes como el sprint planning, la sprint review, etc. Por ello se consultó a las empresas por las prácticas de ingeniería utilizadas y el resultado fue que en mayor medida aplican testing unitario, integración continua y testing automatizado. Pero no se quedan atrás prácticas como, la prueba del sistema, utilización de estándares de codificación, entrega continua, programación de pares y refactoring. También mencionaron Test Driven Development y testing de aceptación automatizado.

3.1. Trazabilidad ISO 90003-SCRUM

En la siguiente tabla, de manera resumida podremos observar en base a los requisitos de la ISO 90003, qué prácticas de SCRUM podrían ser utilizadas para cumplir con los mismos. Además, veremos que otras prácticas de ingeniería de software deberían ser incorporadas en el proceso definido para poder obtener la certificación ISO 9001:2015. Cabe aclarar que solo se presenta un ítem de los trabajados como ejemplo.

ISO 90003	SCRUM	Observaciones
8.4 Control de los procesos, productos y servicios suministrados externamente	No existen prácticas de Scrum para cubrir específicamente el requisito.	La organización debe definir cómo serán controlados los productos y servicios que son suministrados externamente.
8.5 Producción y provisión del servicio	Planificación del Sprint Planificación del Release Ejecución del Sprint Revisión del Sprint Entrega Continua	Este punto se deberá complementar aplicando Integración continua y haciendo Gestión de Configuración. Las herramientas de Gestión de Configuración permiten controlar la trazabilidad del producto.
8.6 Liberación de los productos y servicios	Sprint Review Backlog del Producto	
8.7 Control de las salidas no conformes	Control del cumplimiento del Criterio "Done" para dar por finalizado el sprint. Criterio de aceptación de las historias de usuario definido.	Se deben implementar prácticas de ingeniería de software como prueba de Sistema y prueba de aceptación.

4. 4. Conclusiones

Es importante tener en cuenta que, tanto Scrum como las otras metodologías ágiles, nos ayudan a organizarnos, a realizar un buen seguimiento al proyecto, a trabajar mejor y de esta forma poder entregar el producto prometido a tiempo, pero para que ese producto tenga la calidad esperada por el cliente no nos debemos olvidar de las prácticas de ingeniería. Scrum es la metodología ágil que más adeptos tiene en Argentina como en el mundo, dado que con sus prácticas ayudan a resolver problemáticas que la industria del software ha tenido durante años. Por esta razón es importante facilitar la combinación entre Scrum y la ISO 9001, para que de esta forma nos aseguremos una gestión ágil entregando productos de calidad. Como se puede observar en la matriz de trazabilidad si bien implementando Scrum se pueden cubrir gran parte de los requisitos para obtener una certificación ISO 9001, es importante destacar que sin las prácticas de ingeniería de software y la definición de ciertos procesos organizaciones sería imposible obtenerla.

Referencias

1. (IEEE), I. o. E. a. E. E. (2017). IEEE/ISO/IEC 12207-2017. 35.080 : Desarrollo de software y documentación de sistemas, IEEE: 462.
2. Sstate Agile Report https://www.scrum.org/resources/blog/nuestro-analisis-de-la-agilidad-en-2021-partir-del-15th-annual-state-agile-report
3. ISO 9001:2015. Sistemas de gestión de calidad. Requisitos. Madrid: AENOR, 2015
4. Ken Schwaber & Jeff Sutherland, 2020. La Guía Definitiva de Scrum: Las Reglas del Juego https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-Spanish-European.pdf.

XIX Workshop Bases de Datos y Minería de Datos **(WBDDM)**

Coordinadores

Rodolfo Bertone (UNLP)

Hugo Alfonso (UNLPam)

Nora Reyes (UNSL)

Quality Flaws Prediction in Wikipedia by Using Deep Learning Approaches

Gianfranco Capodici[1], Gerónimo Bazán Pereyra[1], Rodolfo Bonnin[1], and Edgardo Ferretti[1,2]

[1] Universidad Nacional de San Luis (UNSL), San Luis - Argentina
[2] Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (UNSL)
e-mail: ferretti@unsl.edu.ar

Abstract. Quality flaws prediction in Wikipedia is an ongoing research trend. In particular, in this work we tackle the problem of automatically predicting four out of the ten most frequent quality flaws; namely: *No footnotes*, *Notability*, *Primary Sources* and *Refimprove*. Different deep learning state-of-the-art approaches were evaluated on the test corpus from the 1st *International Competition on Quality Flaw Prediction in Wikipedia*; a well-known uniform evaluation corpus from this research field. Particularly, the results show that *TabNet* reachs or improves the existing benchmarks for the *Notability* and *Refimprove* flaws, and performs in a very competitive way for the other two remaining flaws.

Keywords: Wikipedia, Information Quality, Quality Flaws Prediction, Deep Learning

1 Introduction

The evaluation of the information quality (IQ) on the Web has become a crucial task today, since entities from different areas make decisions on the information available on this source. In turn, the amount of information has increased exponentially, and in part, this is due to the growing popularity of websites that allow ordinary users to generate content very easily. The latter has driven the need to automate the evaluation of the quality of information on the Web. Wikipedia is one of the best examples we have from these sites. It is a free content encyclopedia, generated from the contributions of millions of registered and anonymous users. These users write, correct and edit articles; and they are heterogeneous in aspects such as: their education level, age, culture, writing skills and specialization area. This fact makes this encyclopedia one of the 20 most visited sites in the world, but at the same time, it generates the challenge of finding a way to automatically improve the IQ of its articles; viz. a multi-dimensional concept which combines criteria such as accuracy, reliability and relevance.

A widely accepted interpretation of IQ is the "fitness for use in a practical application", i.e. the assessment of IQ requires the consideration of context and use case. Particularly, in Wikipedia the context is well-defined by the encyclopedic genre, that forms the ground for Wikipedia's IQ ideal, within the so-called

featured article criteria.[3] Having a formal definition of what constitutes a high-quality article, i.e. a featured article (FA), is a key issue; however, as indicated in [1], in 2012 less than 0.1% of the English Wikipedia articles were labeled as featured. At present, this ratio still remains, since there are 6 114 featured articles out of 6 525 174 articles on the English Wikipedia.[4]

In the literature, a variety of approaches have been proposed to automatically assess different quality aspects in Wikipedia, such as: featured articles identification; development of quality measurement metrics; vandalism detection, among others. In particular, in this paper we will concentrate on the quality flaws prediction research trend [2–10], since this approach provides concrete hints for human editors about what has to be fixed in order to improve articles' quality. The detection of quality flaws is based on user-defined cleanup tags, which are commonly used in the Wikipedia community to tag content that has some shortcomings. Thus, the tagged articles serve as human-labeled data that is exploited by a machine learning (ML) approach to predict flaws in untagged articles.

This paper extends [7] by doing a deeper study on the Deep Neural Networks (DNN) and Stacked-LSTM models previouslty evaluated on this work and by also exploring *TabNet* [11], a novel high-performance and interpretable canonical deep tabular data learning architecture, that to the best of our knowledge, has not been previously studied in the Wikipedia domain of quality flaws prediction.

The rest of the article is organized as follows. Section 2 introduces the context of the problem faced in this work. Then, in Sect. 3, we present the formal problem statement and the different prediction approaches evaluated are briefly described. Also, the document model used to represent the articles is discussed. Section 4 reports on the experimental setting carried out and the obtained results. Finally, Sect. 5 offers the conclusions.

2 Related Work

In 2012, the first exhaustive study of quality flaws for the English Wikipedia [1] gave rise to the generation of a well-formed data set (for its use in IQ research by the scientific community related to Wikipedia), in the context of the 1st *International Competition on Quality Flaw Prediction in Wikipedia* [12]. That same year, in the international competition "ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)", AlexNet [13] –a system based on Deep Convolutional Neural Networks (DCNN)– emerged as the broad winner. In this way, since 2012, deep learning approaches are consolidated as the state of the art in the field of visual recognition and then spread their supremacy to other ML fields as well.

In this respect, according to our literature review, we can observe that from 2012 to middle 2021, the state of the art regarding IQ in Wikipedia has been mostly determined by research works that use classical approaches ([3–8, 10]). The differences between these works are mainly found in the applied classification

[3] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`

[4] `https://en.wikipedia.org/wiki/Wikipedia:Featured_articles` (accessed June 2022)

algorithms (semi-supervised or supervised), the underlying document representation model (number of features, their complexity and conceptualization made of each flaw, among others). Having this great diversity, it makes difficult to establish a conceptual comparison on which approach is the state of the art to be improved.

For example, [3–7, 10] have followed working methodologies close to the original one proposed by Anderka et al. [2]. In [3], the quality flaw prediction task was faced as a one-class classificacion problem and in [4], the same document model used in [3] was evaluated on the corpus from the "1st International Competition on Quality Flaw Prediction in Wikipedia", where a modified version of the PU-learning winning approach was proposed. The obtained results showed an improvement of 18.31%, averaged over the ten flaws. From among the ten flaws of the competition, the so-called *Refimprove* flaw –which alerts that the tagged article needs additional citations for verification–, has been particularly studied in [5–7]. It is worth mentioning that this information quality flaw, ranks among the five most frequent flaws and represents 12.4% of the flawed articles in the English Wikipedia [3].

In particular, [6] and [7] use the same document model proposed by Anderka [3] and these works were also evaluated on the corpus from the 1st international competition mentioned above. In [6], three different state-of-the-art binary approaches were used with the aim of handling the existing imbalances between the number of articles' tagged as flawed content, and the remaining untagged documents that exist in Wikipedia. These approaches were under-bagged decision trees, biased-SVM and centroid-based balanced SVM. The results showed that under-bagged decision trees with the *min* rule as aggregation method, perform best achieving an F_1 score of 0.96 for the *Refimprove* flaw. In addition, [7] extends the work performed in [6] by incorporating deep neural methods to the study (DNN and Stacked-LSTM) and tackles other quality flaws as well. Stacked-LSTM performed well and reached the existing benchmark for the *Refimprove* flaw. For the other flaws (*No footnotes*, *Notability*, *Primary Sources* and *Wikify*), under-bagged decision trees with different aggregation rules perform best.

Finally, regarding the aforementioned original methodology proposed by Anderka et al., we found that [10] studies different ML approaches, both traditional and deep learning methods; all using as learning experience manually constructed document models and/or automatically extracted features. From among the 12 studied classifiers, the deep approach Bi GRU (bidirectional gated recurrent unit) is the one that achieved the best classification performance: $F_1 = 0.99$ for *Notability*, *No Footnotes* and *Refimprove* flaws; and $F_1 = 0.98$ for *Primary Sources*. It is important to note that the classification approach addressed in this work, is the so-called optimistic approach by Anderka et al. [2], which uses FAs as negative class, while the approach of the competition is the so-called pessimistic, and therefore more challenging.

To be best of our knowledge, Anderka's document model [3] and the one proposed in [14], are the most comprehensive document models built so far based on a features engineering approach. In particular, Bassani and Viviani document

model [14] is composed of 264 features and in principle it seems to contain the 95 features from Anderka's document model. They evaluated their model with the aim of building a suitable ground truth for a (single-label) multi-class classification task, where each article is assigned exactly to one of the seven classes from the quality grading scheme that Wikipedia employed at the time of that paper writing.[5] They evaluated eight state-of-the-art classifiers and Gradient Boosting performed best achieving and accuracy of 90% in some experiments.

A similar classification problem to that reported in [14] was evaluated by Zhang et al. [15], since that a 6-class classification task was performed –considering the Wikipedia quality grading scheme mentioned above–, but where `AC` was skipped on the grounds that is not a real quality class and it overlaps with `FA` and `GA` classes. The proposed history-based article quality assessment model combines feature engineering with learned features by a Recurrent Neural Network (RNN); and it only contains 16 features. Zhang et al. argue that this can be one of the reasons why the best-achieved accuracy value rounds 69%.

In [16], the same 6-class classification task performed in [15] was tackled but with a different document model that relies on explicitly defined features. Moreover, as classification method it was used XGBoost. Furthermore, a deep learning-based baseline was used for assessing the performance of XGBoost given the same feature set. In this respect, the accuracy achieved by XGBoost was 73% against 67% of the deep learning-based baseline. Additionally, XGBoost was also compared against the RNN-LSTM evaluated by Dang and Ignat in [17], where the classification of Wikipedia articles in English, French, and Russian languages in different quality grading schemes was promising without the need of a feature extraction phase. In particular, for the English dataset, XGBoost outperformed the RNN-LSTM by 5%; i.e. RNN-LSTM achieved an accuracy of 68%.

Finally, in [18], following a feature engineering approach to build articles' document models –composed of 68 features–, Wang and Li present a comparative study of state-of-the-art deep-learning approaches by distinguishing high quality articles from low quality. With this aim, a 6-class classification problem on the Wikipedia quality grading scheme mentioned above, was reduced to a binary classification problem where the high-quality class includes `FA`, `AC` and `GA`; and the low-quality class includes `BC`, `SC` and `SB`. Stacked-LSTM networks achieved the best performance ($F_1 = 0.8$). Also, the influence of different features and feature sets on the proposed models were extensively investigated.

3 Problem Statement and Flaw Prediction Approaches

We start with a formal definition of the problem faced in this paper, namely the algorithmic prediction of quality flaws in Wikipedia (Section 3.1). We then provide the theoretical background of the flaw prediction approaches used in our work (Section 3.2) and finally, we introduce the document model used to represent articles (Section 3.3).

[5] At present, this quality grading scheme has been refined; cf. `https://en.wikipedia.org/wiki/Template:Grading_scheme`

3.1 Problem Statement

Following [3], quality flaw prediction is treated here as a classification problem. Let D be the set of English Wikipedia articles and let f_i be the specific quality flaw that may occur in an article $d \in D$. Let $\mathbf{d}$ be the feature vector representing article d, called document model, and let $\mathbf{D}$ denote the set of document models for D. Hence, for flaw f_i, a specific classifier c_i is learned to decide whether an article d suffers from f_i or not; that is, $c_i : \mathbf{D} \rightarrow \{1, 0\}$. For flaw f_i a set $D_i^+ \subset D$ is available, which contains articles that have been tagged to contain f_i (so-called *labeled* articles). However, no information is available about the remaining articles in $D \setminus D_i^+$—these articles are either flawless or have not yet been evaluated with respect to f_i (so-called *unlabeled* articles).

As originally proposed (see e.g. [2, 3]) c_i is modeled as a one-class classifier, which is trained solely on the set D_i^+ of labeled articles. However, in the Wikipedia setting, the large number of available unlabeled articles may provide additional knowledge that can be used to improve classifiers training. Thus, addressing the problem of exploiting unlabeled articles to improve the performance of c_i lead us to cast the problem as a binary classification task.

3.2 Flaw Prediction Approaches

Despite its theoretical one-class nature, quality flaw prediction has been tackled in prior studies as a binary classification task –which relates to the realm of supervised learning– and the results achieved in practice have been quite competitive [5–7]. Supervised learning deals with the situation where training examples are available for all classes that can occur at prediction time. In *binary classification*, the classification $c_i(\mathbf{d})$ of an article $d \in D$ with respect to a quality flaw f_i is defined as follows: given a sample $P \subseteq D_i^+$ of articles containing f_i and a sample $N \subseteq (D \setminus D_i^+)$ of articles not containing f_i, decide whether d belongs to P or to N. The binary classification approach tries to learn a class-separating decision boundary to discriminate between P and a particular N. In order to obtain a sound flaw predictor, the choice of N is essential. N should be a representative sample of Wikipedia articles that are flawless regarding f_i.

ANN An Artificial Neural Network (ANN) is just a collection of units (mathematical model that it simply "fires" when a linear combination of its inputs exceeds some hard or soft threshold; that is, it implements a linear classifier) connected together; the properties of the network are determined by its topology and the properties of the "neurons". In this work, we will refer as an ANN, a feed-forward network; that is, every unit receives inputs from "upstream" units and delivers output to "downstream" units; there are no loops —like in the case of Recurrent Neural Networks [19]. A feed-forward network represents a non-linear function of its current input; thus, it has no other internal state than the weights themselves.

DNN As stated in [20], the quintessential example of a deep learning model is the feedforward deep network (DNN), or multilayer perceptron; that is an ANN with more than one hidden layer. The input of the model is presented to the so-called "input layer", because it contains the variables that we are able to observe. Then a series of hidden layers extracts increasingly abstract features from the input. These layers are called "hidden" because their values are not given in the data; instead the model must determine which concepts are useful for explaining the relationships in the observed data.

Stacked-LSTM Long short-term memory (LSTM) [21] are a modification of the original Recurrent Neural Networks, which includes three types of gates: the forget gate, the input gate, and the output gate. The original LSTM model is comprised of a single hidden LSTM layer followed by a standard feedforward output layer. Stacked-LSTM model [22] extends the reach of this type of network, to the realm of deep neural architecture, in that it has multiple hidden LSTM layers where each layer contains multiple memory cells. Every LSTM in the stack obtains all the information from the preceding layer only.

TabNet It is a recently proposed canonical DNN architecture for tabular data [11]. It inputs raw tabular data without any preprocessing and is trained using gradient descent-based optimization, enabling flexible integration into end-to-end learning. TabNet uses sequential attention to choose which features to reason from at each decision step, enabling interpretability and better learning as the learning capacity is used for the most salient features. That is, feature selection is instance-wise, given that it can be different for each input.

3.3 Document Model

To model the articles, we used the document model proposed in [3], one of the most comprehensive document model proposed so far for quality flaw prediction in Wikipedia—it comprises 95 article features. Formally, given a set $D = \{d_1, d_2, \ldots, d_n\}$ of n articles, each article is represented by 95 features $F = \{f_1, f_2, \ldots, f_{95}\}$. A vector representation for each article d_i in D is defined as $d_i = (v_1, v_2, \ldots, v_{95})$, where v_j is the value of feature f_j. A feature generally describes some quality indicator associated with an article.

In [3] four such subsets were identified by organizing the features along the dimensions *content*, *structure*, *network* and *edit history*. Content features are computed based on the plain text representation of an article and mainly address aspects like writing style and readability. Structure features rely on an article's wiki markup and are intended to quantify the usage of structural elements like sections, templates, tables, among others. Network features quantify an article's connectivity by means of internal and external links. Edit history features rely on an article's revision history and model article evolution based on the frequency and the timing of edits as well as on the community of editors. In [3], a detailed description for each feature is provided including implementation details. Due to space constraints, these features are not explicitly described in this paper.

Table 1. Four out of the top ten quality flaws of English Wikipedia articles that are comprised in the PAN-WQF-12 corpus.

Flaw name	Flaw description	Training corpus		Test corpus	
		tagged articles	untagged articles	tagged articles	untagged articles
No footnotes	The article's sources are unclear because of its inline citations.	6 068	–	1 000	1 000
Notability	The article does not meet the general notability guideline.	3 150	–	1 000	1 000
Primary sources	The article relies on references to primary sources.	3 682	–	1 000	1 000
Refimprove	The article needs additional citations for verification.	23 144	–	999	999
Additional random (untagged) articles		–	50 000	–	–

4 Experiments and Results

To perform our experiments, we have used the corpus available in the above-mentioned Competition on Quality Flaw Prediction in Wikipedia [12], which has been released as a part of PAN-WQF-12,[6] a more comprehensive corpus related to the ten most important article flaws in the English Wikipedia, as pointed out in [1]. The training corpus of the competition contains 154 116 tagged articles (not equally distributed) for the ten quality flaws, plus additional 50 000 untagged articles. The test corpus (19 010 articles) contains a balanced number of tagged articles and untagged articles for each of the ten quality flaws, and it is ensured that 10% of the untagged articles are FAs. Table 1 introduces a brief description for each flaw evaluated in our work. Moreover, for each flaw, the numbers of tagged and untagged articles in the training and test corpus of the 2012-competition is specified. The training corpus does not contain untagged articles for the individual flaws, but it comprises 50 000 additional randomly selected untagged articles.

4.1 Experimental Setting

As mentioned in Sect. 1, in this paper we extend [7], where an initial study on deep learning approaches applied to quality flaws prediction in the Wikipedia domain was carried out. In that work, also classical (non-neural) approaches were evaluated and were in fact, the ones which reported in general the best

[6] The corpus is available at `https://webis.de/data/pan-wqf-12.html`

performing measures, except for Stacked-LSTM that reached the existing benchmark for the *Refimprove* flaw of $F_1 = 0.96$ from [6]. In [7], only *random search* (RS) over the different variables that influence each model was evaluated. In our current work, we have also tried two other search strategies, viz. *HyperBand* (HB) and Bayesian optimization (BO) for the DNN and Stacked-LSTM models, maintaining the same number of epochs (10) for training. However (cf. Table 2), only for *Notability* flaw BO or HB performed better than RS for all the models. We conjecture that this may be due to the low number of epochs –given the computational cost of each trial–, this number is not large enough to allow these more sophisticated methods to show some more advantageous parametric configurations.

Moreover, we also extend [7] by evaluating TabNet, a deep-learning achitecture specially suited for tabular data, as it is our case. We also evaluated an ANN as a baseline. Due to resource and time-execution constraints, in the validation stage we used a split of 80%-20% of the dataset. All the networks (ANN, DNN and Stacked-LSTM) consist of an input layer of 95 units and a sigmoid layer output. All the neurons in the hidden layers use ReLU activation functions. For the case of the ANN, different hidden layer widths were tried (from 512 to 2018, in 512 units steps) and values 0.001 and 0.005 were evaluated as Adam's learning rate. For the DNN, a variable number of hidden layers (up to three) were evaluated with optional dropout layers. Similarly, for the Stacked-LSTM, a variable number of LSTM layers (up to five) was tried. The width of each hidden / LSTM layer was set from 128 units up to 2048, in 128 units steps and the learning rate was varied from 0.0001 to 0.005. Finally, the *Wikify* flaw is not addressed in our study as it was in [7], given that on August 2021 its associated cleanup tag was revised in the template index and replaced for more detailed tags indicating more specifically which layout aspects must be corrected.

4.2 Results

The state-of-the-art F_1 score of 0.96 for the *Refimprove* flaw on the test set of the 1st *International Competition on Quality Flaw Prediction in Wikipedia* was achieved in [6] by using under-bagged decision trees with the min rule as aggregation method. Besides, it was also achieved by a Stacked-LSTM deep approach in [7]. As we can see in Table 2, only two models have surpassed this value; a new configuration of a Stacked-LSTM ($F_1 = 0.97$) and TabNet ($F_1 = 0.98$). It may seem small improving the state-of-the-art result by 1% and 2.1%, respectively; but it is worth considering than the benchmark is high and increasing by 2.1% the current F_1 score, reduces notably the gap to the optimum score. Moreover, our results are directly comparable to the values found in [4, 6, 7], since we have used the same data set and document model for representing the articles. In this respect, we also reached the benchmark of $F_1 = 0.99$ for the *Notability* flaw and remain 0.01 below from the benchmark of $F_1 = 0.99$ for the *No footnotes* (from [7]) and *Primary Sources* (from [4] and [7]) flaws.

As expected, the values reported on the test set correspond to the configurations which achieved the best values on the validations sets. Due to space con-

straints, we only report next the configurations of combinations which achieved the best values –highlighted in bold in Table 2– viz. ANN-HB (learning rate 0.005, 2048 neurons in the hidden layer), DNN-BO (learning rate 0.001, [1536, 2048, 2048][7]) and DNN-HB (learning rate 0.001, [1024, 512]) for *Notability* flaw, and Stacked-LSTM-HB (learning rate 0.001, [384, 512, 512, 256, 128]) for *Primary Sources* flaw. We evaluated TabNet with its default parameters.[8]

Table 2. F_1 values on the test set of the 1st International Competition on Quality Flaw Prediction in Wikipedia for all the evaluated models.

Flaws / Models	**ANN**			**DNN**			**Stacked LSTM**			**TabNet**
	RS	**BO**	**HB**	**RS**	**BO**	**HB**	**RS**	**BO**	**HB**	
No Footnotes	0.79	0.72	0.67	0.75	0.68	0.79	0.97	0.95	0.93	**0.98**
Notability	0.98	0.93	**0.99**	0.98	**0.99**	**0.99**	0.95	0.93	0.98	**0.99**
Primary Sources	0.86	0.76	0.86	0.92	0.89	0.87	0.76	0.93	**0.97**	**0.97**
Refimprove	0.64	0.64	0.63	0.63	0.63	0.64	0.97	0.93	0.81	**0.98**

5 Conclusions

In this work, we carried out a comparative study of three deep state-of-the-art approaches to automatically assess information quality; in particular, to identify four out of the ten quality flaws most frequent in Wikipedia, and the task was carried out by binary classification. The results obtained showed that the new benchmark of $F_1 = 0.98$ for the *Refimprove* flaw prediction was achieved by using the default configuration of TabNet architecture. Moreover, for the remaining flaws, very competitive results were obtained.

Acknowledgments

This work has been partially founded by PROICO 03-0620, UNSL, Argentina.

References

1. Anderka, M., Stein, B.: A breakdown of quality flaws in Wikipedia. In: 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12), ACM (2012) 11–18
2. Anderka, M., Stein, B., Lipka, N.: Detection of text quality flaws as a one-class classification problem. In: Proceedings of the CIKM'11, ACM (2011) 2313–2316
3. Anderka, M.: Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. PhD thesis, Bauhaus-Universität Weimar (June 2013)

[7] This list notation should be understood as: 3 layers, 1536 neurons in the first layer, and 2048 in second and third layers.

[8] `https://pypi.org/project/pytorch-tabnet/`

4. Ferretti, E., Errecalde, M., Anderka, M., Stein, B.: On the use of reliable-negatives selection strategies in the pu learning approach for quality flaws prediction in wikipedia. In: 11th Intl. Workshop on Text-based Information Retrieval. (2014)
5. Ferretti, E., Cagnina, L., Paiz, V., Donne, S.D., Zacagnini, R., Errecalde, M.: Quality flaw prediction in spanish wikipedia: A case of study with verifiability flaws. Information Processing & Management **54**(6) (2018) 1169–1181
6. Bazán-Pereyra, G., Cuello, C., Capodici, G., Jofré, V., Ferretti, E., Errecalde, M.: Automatically assessing the need of additional citations for information quality verification in Wikipedia articles. In: Actas del XXV Congreso Argentino de Ciencias de la Computación (CACIC). (2019) 42–51 ISBN: 978-987-688-377-1.
7. Bazán Pereyra, G., Cuello, C., Capodici, G., Jofré, V., Ferretti, E., Bonnin, R., Errecalde, M.: Predicting information quality flaws in wikipedia by using classical and deep learning approaches. In Pesado, P., Arroyo, M., eds.: Computer Science – CACIC 2019, Cham, Springer International Publishing (2020) 3–18
8. Herrera, J., Funes, A., Ferretti, E., Cagnina, L.: Selección de Características para Clasificación de Clase Única de Fallas de calidad de Información en Wikipedia. In: Actas del XIII CoNaIISI. (2020)
9. Guda, B.P.R., Seelaboyina, S.B., Sarkar, S., Mukherjee, A.: NwQM: A neural quality assessment framework for Wikipedia. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, ACL (2020) 8396–8406
10. Wang, P., Li, M., Li, X., Zhou, H., Hou, J.: A hybrid approach to classifying wikipedia article quality flaws with feature fusion framework. Expert Systems with Applications **181** (2021)
11. Arik, S.Ö., Pfister, T.: TabNet: Attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 35. (2021)
12. Anderka, M., Stein, B.: Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. In Forner, P., Karlgren, J., Womser-Hacker, C., eds.: Working Notes Papers of the CLEF 2012 Evaluation Labs. (2012)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems 25. (2012)
14. Bassani, E., Viviani, M.: Quality of Wikipedia articles: Analyzing features and building a ground truth for supervised classification. In: 11th International Joint Conference, IC3K. (2019) 338–346
15. Zhang, S., Hu, Z., Zhang, C., Yu, K.: History-based article quality assessment on Wikipedia. In: IEEE 5th Intl. Conference BigComp. (2018) 1–8
16. Schmidt, M., Zangerle, E.: Article quality classification on Wikipedia: introducing document embeddings and content features. In: 15th Intl. OpenSym. (2019)
17. Dang, Q.V., Ignat, C.L.: An end-to-end learning solution for assessing the quality of wikipedia articles. In: 13th Intl. Symposium on Open Collaboration. (2017) 1–10
18. Wang, P., Li, X.: Assessing the quality of information on Wikipedia: A deep-learning approach. Journal of the Association for Information Science and Technology **71**(1) (2020) 16–28
19. Rumelhart, D., Hinton, G., Williams, R.: Learning representations by back-propagating errors. Nature **323** (1986)
20. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8) (1997)
22. Graves, A., Mohamed, A., Hinton, G.E.: Speech recognition with deep recurrent neural networks. CoRR **abs/1303.5778** (2013)

On the Importance of Data Representation for the Success of Text Classification

Carolina Y. Cuello [1,2], Vanessa Jofre Caradonna [2],
Ma. José Garciarena Ucelay [1,2], and Leticia C. Cagnina [1,2,3]

[1] LIDIC, Universidad Nacional de San Luis, San Luis, Argentina,
[2] Universidad Nacional de San Luis, San Luis, Argentina,
[3] Consejo Nacional de Investigaciones Científicas y Técnica (CONICET), Argentina
{carolina.yamile.cuello,vane.jofre.caradonna,
mjgarciarenaucelay,lcagnina}@gmail.com

Abstract. Text mining approaches use natural language processing to automatically extract patterns from texts. Tasks as topic labeling, news classification, question answering, named entity recognition and sentiment analysis, usually require elaborate and effective document representations. In this context, word representation models in general, and vector-based word representations in particular, have gained increasing interest to alleviate some of the limitations that Bag of Words exhibits. In this article, we analyze the use of several vector-based word representations besides the classical ones, in a polarity analysis task on movie reviews. Experimental results show the effectiveness of more elaborate representations in comparison to Bag of Words. In particular, Concise Semantic Analysis representation seems to be very robust and effective because independently the classifier used with, the results are really good. Dimension and time of getting the representations are also showed, concluding in the efficiency of the classifiers when Concise Semantic Analysis is considered.

Keywords: text mining, text representations, text classification, movie reviews, sentiment analysis, polarity analysis

1 Introduction

Since the creation of the World Wide Web in 1989, the history of the communication has being evolved year by year. In 1990, with few computers connected around the world, the first web browser was presented. By the early year 2000, only some countries had access to the information through Internet. But the use of this technology grew rapidly and, in 2016, more than 80% of people in USA, UK, Australia, Spain, Israel and Germany between others, had Internet access[1]. This massive use of Internet can be translated in communication, re-

[1] Pew Research Center. Last access: July 4th, 2022. Available on https://www.pewresearch.org/global/2016/02/22/internet-access-growing-worldwide-but-remains-higher-in-advanced-economies/

sources, information and *data*. Thus, we started to live the exploitation of digital technologies.

Nowadays, 2.5 quintillion bytes of data are uploaded at day through Internet[2], and machine learning (ML) methods allow to use them for different purposes. Particularly, if the online resource is text, tasks as topic labeling, news classification, question answering, named entity recognition and sentiment analysis [1] are frequently solved with ML.

There are some interesting papers reviewing text classification algorithms [2–5]. They range from traditional algorithms such as Naïve Bayes, Decision Trees and Support Vector Machines to Deep Learning methods such as Recurrent and Convolutional Neural Networks, and Transformers. Besides the classifiers and the proposed systems, we believe that it is equally important to discuss the way in which data is represented for the understandability and efficacy of ML methods. For that reason, we analyze different representations for texts in a particular case of study, aiming to show the adequacy of each one with standard ML classifiers.

Contributions. We review traditional *document-level representations* such as those that rely on term frequencies (Bag of Words (BoW) with different weighting schemes [6, 7]), and more elaborate and compact ones (Concise Semantic Analysis (CSA) [8] and Latent Semantic Analysis (LSA) [9]). Bags of tokens[3] condense the document into a single count vector. Each position of that vector has associated a particular token and its value represents the frequency in the document. CSA extracts concepts from the texts and relates each document with the words in the corresponding concept space. LSA discovers latent topics of the texts and represents those as a mixture of topics which are probability distributions over words. Finally, we move to *featurized word-level representations* which capture the semantic and syntactic correlation between words. Common used features are aspects, relations and words, and are represented by dense vectors of fixed size named *embeddings*. Some examples of words embeddings are Word2vec [10] and GloVe [11]. Thus, we analyze several dimensions of each representation like the performance obtained with state-of-the-art classifiers on a particular problem to solve: polarity analysis of IMDB reviews related to movies. We also discuss the size of the obtained representations and the time required to get them. We conclude the work with some highlights about the adequacy of the text representations for this particular problem.

Organization. In Section 2 we describe the polarity analysis task and the corresponding dataset used in this paper. In Section 3 we briefly define the document-level and featurized word-level representations for the texts analyzed. Then, the experimental study carried out is in Section 4. Finally, Section 5 summarizes the main conclusions obtained and future work.

[2] Earthweb web site. Last access: July 4th, 2022. Available on https://earthweb.com/how-much-data-is-created-every-day/

[3] We assume that tokens would be words, n-grams, concepts, classes, topics, etc.

2 Case Study: Polarity Analysis on IMDB Movie Reviews

Sentiment Analysis is one of the most popular applications of ML in Natural Language Processing since it is relevant in many domains such as marketing, politics, research, affective computing and so on [12–15]. However, it is a challenging field because when people express their feelings, emotions and opinions, language can be used to persuade, encourage and communicate a positive or negative form of thinking. Especially when analyzing written opinions, due to the use of non-verbal language and the missing context, it is even more difficult to detect ambiguity and sarcasm.

In general, Sentiment Analysis encompasses several aspects like polarity, subjectivity, intensity, intentions and aspect-based analysis, as well as specific emotions identification. Depending on the task selected, regressors or classifiers are trained [16–18]. Particularly, *Polarity Analysis* task can consider continuous or discrete values for documents labels. When the labels are continuous numbers, usually they vary between -1 (negative) and 1 (positive). Although when the labels are categorical they are mostly classified as positive, negative or neutral.

In this work, we used IMDB Movie Review Dataset [19] which was built in view of the latter approach and taking into account only positive and negative classes. Therefore, it is a task review-level polarity classification, where a classifier must predict if a review is *positive* or *negative* given its text. IMDB Movie Review Dataset consists of 50,000 reviews about movies collected from Internet Movie Database (IMDb[4]). This website contains information related to films, television series, videogames and streaming content, information about cast, production crew, plot summaries, ratings, also fan and critical reviews.

When the authors built this dataset they extracted only ratings and reviews. Hence, they considered only highly polarized reviews, that is, a negative review has a score less than or equal to 4, and a positive review has a score greater than or equal to 7 (considering a scale from 1 to 10). Neutral reviews were not included in the collection. Additionally, there are not more than 30 reviews assigned per movie, this is to avoid correlation of common terms (for example, characters, actors, events, etc.). The dataset is divided into two sets of 25,000 reviews. One set is used to train and the other one to test the different proposed models. Also each collection is balanced in their amount of positive and negative labeled reviews and there are no movies in common between the two sets. We choose this well-known corpus because it is the most utilized in polarity analysis task.

3 Evaluated Models

A *document* is a unit of textual data that belongs to a collection and constitutes one of the main lexical resources of text mining. To extract relevant information from a document, ML approaches should receive the data in an adequate form, that is, a proper representation of the text.

[4] www.imdb.com

As mentioned in Section 1, the objective of this work focuses on studying different types of representations, such as the simple *BoW*, more elaborate and compact ones such as *CSA* and *LSA*, and the featurized word-level vectors as *Word2vec* and *GloVe*. These text representations were combined with standard classifiers to obtain the models. We considered *Support Vector Machines* with Linear and RBF kernels, the *Multinomial* version[5] of the probabilistic model *Naïve Bayes* and the decision trees classifier named *Random Forest*. Next, we describe briefly these representation techniques considered in our experimental study.

The *BoW* strategy builds a set with the words of the documents, that is, the *vocabulary* of the dataset. Formally, a document d is represented by a vector of weights $d_{BoW} = (w_1, w_2, ..., w_n)$ where w_i depends on the selected weighting scheme (boolean, term frequency, term frequency-inverse document frequency, etc.) and n is the size of the vocabulary of the collection. This representation is one of the most used in text categorization tasks because it is simple to implement, the classifiers perform relatively well and it is possible to consider different weighting schemes for the terms. However, *BoW* have known limitations such as the order of the words in the document is lost and, context and grammar are not considered, causing the loss of semantic and conceptual information.

The *CSA* technique proposes a vector representation of low dimensionality with a high level of representativeness. A space of concepts is built according to the categories used for the classification task. Then, for each word of the vocabulary, a value that indicates the relationship between the word and each concept is calculated and stored in a vector. Finally, the representation of the document is obtained by adding the vectors corresponding to all the words included in the document, weighted by the relative frequency of each word in the document.

LSA is a method that uses the contexts in which a word appears and associates it with a meaning. In order to find a small subset of concepts, *LSA* analyzes the relationship between the meanings of the documents and the words in those (the latent space). The matrix word-document is created, where the rows correspond to the words and the columns to the documents. The components f_{ij} of that matrix indicate how many times the word i is in the document j. By applying a weighting scheme, the model is improved to give relevance to the word that appears in the context of the documents. This representation faces polysemy and synonymy problems, which are alleviated by applying the Singular Value Decomposition technique. The new matrix corresponds to a concept-document one which is the representation of each document.

Word2vec is a neural model that computes feature vectors, usually named word embeddings, by processing a large collection of documents. Formally, given a set of n words and a context window of size r (the r words before and after the target word), the model tries to predict which is the target word based on their neighbors. This is the Continuous Bag-of-Words version. On the other hand, Skip-Gram approach consists in predict the neighboring words from the target word. The resultant embeddings are vectors of dimensions $1 \times W$ with

[5] The multinomial variant of Naïve Bayes is usually employed with textual data.

W denoting the number of neurons in the hidden layer of the neural network. Finally, the document representation can be obtained averaging or summing the embeddings of each word in the text. *Word2vec* efficiently models related words, which appear in similar contexts, and the representation captures the semantic relationship between them. An advantage of considering *Word2vec* embeddings is the possibility to use pre-trained vectors to represent the data for a particular task. In this way, the hard process to obtain the embeddings is avoided. However, this model has some problems representing out-of-vocabulary words, even there are no shared representations at sub-word levels.

GloVe, stands for Global Vectors, is an unsupervised learning algorithm for obtaining vector representations of words. The technique generates a continuous vector space using matrix factorization and local context windows. The training is performed on global statistics of word-word co-occurrence of texts. The computed representations are based on the proportion of the co-occurrence probabilities that are obtained from the semantic relationship between words. When that relationship exists, the probability is large, whereas that in the case of words that are related to only few words, the value obtained is small. After computing the word embeddings, the representation of the document is obtained averaging or summing the word vectors of the text. *GloVe* captures syntactic and semantic information with good results, but it provides limitations related to the detection of word analogies. The model is straightforward to obtain and reproduces correctly sub-linear relationships in the vector space. There are evidence that *GloVe* performs better than *Word2vec* in word analogy tasks, although the problem of words out-of-vocabulary is still present.

4 Experimental Study

The execution of the experiments was carried out with *Google Colaboratory*[6] platform. We employed *Jupyter notebooks* with Python language (version 3.7.13) to code the routines to generate, train, test and evaluate the models. In particular, we utilized the following main libraries for Natural Language Processing[7]: *sklearn* (version 1.0.2), *gensim* (version 3.6.0), *nltk* (version 3.7), *spacy* (version 3.3.1), *numpy* (version 1.21.6) and *pandas* (version 1.3.5).

The corpus used to generate and evaluate the models was described in Section 2. It consists of 50,000 movie reviews extracted from IMDd site. We consider the original training and testing sets, that is, 25,000 samples to train and 25,000 more to evaluate. In each subset there is the same amount of reviews with positive and negative polarities. The reviews were pre-processed by converting all the characters to lowercase and removing stop words, numbers, punctuation marks and any special character. The hyperparameters for each representation are described below.

In order to obtain the vocabulary of the BoW representation, we selected several subsets of different amount of terms. We tested with the 100, 500, 1,000,

[6] https://colab.research.google.com/
[7] This information is important for reproducibility isuues.

2,500, 5,000, 7,500 and 10,000 most frequent terms. Then, we evaluated *unigrams* and *bigrams* of words, *unigrams + bigrams* of words and *trigrams* of characters. We considered those proposals with *Boolean*, *Term Frequency* and *Term Frequency–Inverse Document Frequency* (TF-IDF) weighing schemes.

Since reviews are classified into two classes, i.e. positive and negative, the resulting vectors of CSA representation are two-dimensional. Meanwhile, in LSA, we considered 100 and 300 concepts. In addition, this last representation was calculated considering TF-IDF weights.

Regarding word embeddings representations, for Word2vec and GloVe approaches, we made use of pre-trained vectors provided by Google[8] and Stanford University[9], respectively. Word2vec pre-trained vectors were generated from *Google English News* corpus, with Skip-Gram architecture, a context windows size of 5 words and 300 dimensions. On the other hand, for GloVe pre-trained embeddings, we considered the ones learned from *Twitter* with vector's sizes of 25, 50, 100 and 200. We also used the GloVe embeddings trained with *Common Crawl* data which have 300 components. For both methods, the document embeddings were obtained by averaging the word vectors present in the corresponding text.

The different representations were evaluated using *Support Vector Machine* classifiers with linear (SVM-Linear) and RBF (SVM-RBF) kernels, *Multinomial Naïve Bayes* (MNB) and *Random Forest* (RF). It should be noted that we employed the default parameters of each classifier. Finally, we used *precision*, *recall* and *F-Measure* to evaluate the performance of the classifiers in a single run.

In Table 1 we present a summary of the best F-Measure values obtained for each model (representation and classifier). The highest values are highlighted in bold.

Table 1. Summary of the best F-Measure values obtained with different models and the dimension of each document representations.

Document Representation	Size	SVM-Linear	SVM-RBF	MNB	RF
BoW$_{(Unigram+Bigram)}$	10,000	0.86	0.88	0.84	0.84
CSA	2	**0.93**	**0.92**	**0.92**	**0.88**
LSA	300	0.86	0.87	0.82	0.80
Word2vec	300	0.85	0.86	0.74	0.81
GloVe	300	0.85	0.83	0.83	0.83

Based on all the experiments executed for BoW representation, the combination of *unigrams + bigrams* of words with TF-IDF weighting scheme and

[8] Word2vec pre-trained embeddings downloaded from https://code.google.com/archive/p/word2vec/

[9] GloVe pre-trained embeddings available at https://nlp.stanford.edu/projects/glove/

considering the 10,000 most frequent words, was the best. The models using BoW obtained good performance (between 0.86 and 0.88) with SVM classifiers.

Slightly lower performance was achieved with models that include embeddings as text representations. The best metrics were reached with vectors of 300 dimensions and, in the case of GloVe, using the pre-trained embeddings from *Common Crawl*.

The highest result obtained with models considering LSA utilized 300 concepts when building the representation. The performance of SVM classifier with LSA is similar to that of BoW, but with the other classifiers are barely lower.

The best F-Measure values in our experimental study were achieved with CSA representation which has only 2 dimensions. All those values are equal or higher than 0.92 with the exception of RF classifier which is 0.88.

Table 2 presents a summary of the time taken to generate the different document representations (see *Generation* column) and the execution time of the classifiers, expressed in seconds. The columns corresponding to the classifiers indicate the sum of the training and the testing stages of the best models stated in Table 1. This allows us to take into account the temporal complexity of each model. A value of 0 (second) in Table 2 means a period of time less than 1 second.

Table 2. Time consumed by the different representations and classifiers expressed in seconds.

Representation	**Generation**	**SVM-Linear**	**SVM-RBF**	**MNB**	**RF**
BoW	25	0	1,397	3	63
CSA	1,257	0	18	0	3
LSA	29	60	372	0	60
Word2vec	4,551	8	211	0	200
GloVe	62	124	1,271	0	221

The representation that took the longest time was Word2vec (4,551 seconds). The reason, possibly, is that it includes the time considered to load the full set of pre-trained vectors into memory. Meanwhile, despite of CSA is the second slowest in computing the representation, it is only a quarter of the time Word2vec took. On the other hand, if we consider the time spent for the models using CSA, all classifiers are very fast.

Analyzing the execution time involved in generating the embeddings of documents, GloVe has minimal execution time compared to Word2vec, and their performance (see Table 1) is similar.

Finally, it is worth mentioning that the times to obtain BoW, LSA and GloVe representations are really good (62 seconds or less). Regarding the classifiers, MNB always was the fastest, followed by SVM-Linear. Whereas SVM-RBF generally took the longest time to train and test the models.

5 Conclusions and Future Work

In this article, we analyzed different text representations for Polarity Analysis task. From the experimental study, we can conclude that models including CSA representation obtained the best F-Measure values. The CSA vectors representing the documents are really small (just 2 dimensions) and CSA-based models are the fastest to execute. CSA preserves only the necessary information for the classifiers extracting few concepts from categories and then interprets those concisely on the words of the documents. Thus, this representation can be considered robust and efficient demonstrating to be adequate for this task.

Regarding BoW with unigrams+bigrams of words, TF-IDF weighting scheme and vocabulary of 10,000, we found that it is fast to obtain and the models generally run rapidly (except when SVM classifier is used with RBF kernel). Nevertheless, the size of the representation is large, more precisely 10,000 dimensions.

Another text representation that can be obtained quickly is LSA. With just 300 dimensions, the analyzed models including LSA as representation are fast to execute beyond the good performance they showed.

Finally, the models including embeddings-based representations perform similarly well although Word2vec is slower to obtain compared to GloVe.

In relation to the time complexity of the classifiers, MNB and SVM demonstrated to be very efficient in comparison to RF when evaluating the training and testing stages. Also, if we focus in their performance, we observed that they were really good.

As future work we plan to continue the analysis of these text representations combining them with neural networks-based classifiers. We are also interested in deeply study the models which use embeddings to fine-tune their pre-trained word vectors for this particular tasks. Finally, we want to include more advanced models such as BERT in our study.

Acknowledgments. Authors thank project PROICO 03-0620 of the Universidad Nacional de San Luis and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina for the continuous support to the research.

References

1. Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A.: A Survey on Text Classification Algorithms: From Text to Predictions. Information. 13(2), 83 (2022)
2. Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., He, L.: A Survey on Text Classification: From Traditional to Deep Learning. ACM Transactions on Intelligent Systems and Technology (TIST). 13(2), 1–41 (2022)
3. Li, R., Liu, M., Xu, D., Gao, J., Wu, F., Zhu, L.: A Review of Machine Learning Algorithms for Text Classification. In: Lu, W., Zhang, Y., Wen, W., Yan, H., Li, C. (eds) Cyber Security. CNCERT 2021. Communications in Computer and Information Science, vol 1506, pp. 226-–234. Springer, Singapore (2022)

4. Riduan, G. M., Soesanti, I., Adji, T. B.: A Systematic Literature Review of Text Classification: Datasets and Methods. In: 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 71–77. IEEE, Purwokerto, Indonesia (2021)
5. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text Classification Algorithms: A Survey. Information. 10(4), 150 (2019)
6. Salton, G., McGill, M. J.: Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York (1983)
7. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, (2007)
8. Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K.: Fast Text Categorization Using Concise Semantic Analysis. Pattern Recognition Letters. 32(3), 441–448 (2011)
9. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. 41(6), 391–407 (1990)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: 1st International Conference on Learning Representations (ICLR) 2013. Workshop Track Proceedings, Scottsdale, Arizona, USA (2013)
11. Pennington, J., Socher, R., Manning, C. D.: GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha (2014)
12. Rambocas, M., Pacheco, B. Online Sentiment Analysis in Marketing Research: A Review. Journal of Research in Interactive Marketing. 12(2), 146–163 (2018)
13. Bose, R., Dey, R. K., Roy, S., Sarddar, D.: Analyzing Political Sentiment Using Twitter Data. In: Satapathy, S., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies, vol 107, pp. 427–436. Springer, Singapore (2019)
14. Yousif, A., Niu, Z., Tarus, J. K., Ahmad, A.: A Survey on Sentiment Analysis of Scientific Citations. Artificial Intelligence Review. 52(3), 1805—1838 (2019)
15. Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A.: Affective Computing and Sentiment Analysis. In: Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (eds) A Practical Guide to Sentiment Analysis, vol. 5, pp. 1–10. Springer, Cham (2017)
16. Yadav, A., Vishwakarma, D. K.: Sentiment Analysis Using Deep Learning Architectures: A Review. Artificial Intelligence Review. 53(6), 4335-4385 (2020)
17. Birjali, M., Kasri, M., Beni-Hssane, A.: A Comprehensive Survey on Sentiment Analysis: Approaches, Challenges and Trends. Knowledge-Based Systems. 226, 107–134 (2021)
18. Wankhade, M., Rao, A. C. S., Kulkarni, C.: A Survey on Sentiment Analysis Methods, Applications, and Challenges. Artificial Intelligence Review. 55, 1–10 (2022).
19. Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., Potts, C.: Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150. Association for Computational Linguistics, Portland, Oregon, USA (2011)

Democratizing Argentine Marine Science Data Through Linked Open Data

Marcos Zárate[1,2,⋆][0000−0001−8851−8602] Carlos Buckle[2][0000−0003−0722−0949], Mirtha Lewis[1,3][0000−0001−6262−6226], Claudio Delrieux[4][0000−0002−2727−8374], Dario Ceballos[1], and Gustavo Nuñez[2]

[1] Centre for the Study of Marine Systems, Patagonian National Research Centre (CESIMAR-CENPAT-CONICET), Puerto Madryn, Argentina.
zarate@cenpat-conicet.gob.ar

[2] Laboratorio de Investigación en Informática (LINVI) - Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB), Puerto Madryn, Argentina.
cbuckle@unpata.edu.ar

[3] Centro de Investigaciones y Transferencia Golfo San Jorge, (CIT-GSJ-CONICET), Comodoro Rivadavia, Argentina.
mirtha@cenpat-conicet.gob.ar

[4] Computer Science and Engineering Department, Universidad Nacional del Sur (DIEC-UNS), Bahia Blanca, Argentina
cad@uns.edu.ar

Abstract In this paper we expose experiences carried out during the last five years in the domain of Argentine marine sciences. Specifically data generated by Pampa Azul Argentine initiative to improve the publication of data using the advantages provided by Linked Open Data (LOD), Knowledge Graph (KG) and FAIR principles. The focus is on: a) to provide a conceptual analysis of traditional data publication in marine science, b) to describe projects based on LOD that involve information from Argentina, we mainly focus on the OceanGraph KG project, c) generate recommendations for data management for its best use in marine science.

Keywords: Linked Open Data · FAIR · Pampa Azul · Knowledge Graph · Marine Science.

1 Introduction and Motivation

In July 2020, Pampa Azul initiative was relaunched [1] aimed at promoting scientific knowledge, technological development and productive innovation in the South Atlantic Ocean, in order to develop a culture of the sea in Argentine society, promote the sustainable use of marine natural assets and strengthen the growth of the associated national industry.

⋆ Corresponding author.

However, data management of data generated during first launch of Pampa Azul (2014) made it clear that data management and modeling of online data needed to be planned, safeguarded and shared, so products generated can be used by the scientific community for an adequate understanding of the functioning of our marine spaces. Furthermore, integration of these data with global federated databases was not taken into account, which makes their comprehensive scientific use very difficult.

Prior to Pampa Azul, the Ministry of Science, Technology and Productive Innovation developed the National Biological Data System [2] (SNDB by its acronym in Spanish) on a platform called Integrated Publishing Toolkit (IPT) [3] developed by Global Biodiversity Information Facility (GBIF) for biological data based on the Darwin Core standard [4] which is widely adopted by the international community. For marine data, the National Sea Data System (SNDM by its acronym in Spanish) was created on the platform developed by the International Oceanographic Data and Information Exchange (IODE), in order to visualize the information of the national satellite-type oceanographic data producing centers of Argentina.

From political and scientific contexts, a demand for data management is identified within the context of Pampa Azul, but the platform used by the SNDM provides information from the different resources in a fragmented manner and there are not sufficiently detailed resources for data entry. In addition, it does not allow integration with other scientific data repositories, this generated a lack of interest in contributing data sets. In 2018 only three new data sets were recorded and in 2019 there were no new data. These difficulties in data management are because they involve conceptual frameworks from different disciplines, such as Oceanography (physical, chemical and biological), Geosciences, and Meteorology, among which there is a great diversity in the types and formats of data to be managed. Nowadays, this data portal is down (see: https://www.argentina.gob.ar/ciencia/sistemasnacionales/datos-del-mar), among other things because it does not allow interactive viewing of data or other types of information, making it an unattractive tool for researchers or the general public interested in oceanography. Figure 1 shows the timeline with the most important milestones in marine science data management at local level.

Several causes can be mentioned that led to the failure of the SNDM, but we consider that the most important is due to the fact that marine sciences generate large volumes of data, due to advances in remote acquisition technology and the permanent emergence of new oceanographic campaigns [5]. Thus, it is necessary to develop systems capable of managing their integration and communication, both for comprehensive and secondary use by the participating groups and institutions, as well as for external users who require information.

One of the promising approaches to address the problems associated with heterogeneous data management is to store the information in a structured way and to represent data sets as graphs [6] which has been used in research and business, generally in close association with Semantic Web technologies [7], Linked Open Data (LOD) [8], large-scale data analytics, and cloud computing.

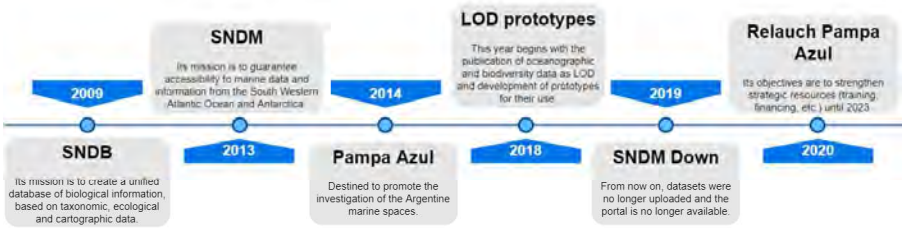


Fig. 1. Relevant milestones on marine data management in Argentina.

The main contribution of this paper is to focus on how LOD contributed the opening and democratization of marine science information, which is financed with public funds and show the lessons learned, so that future developments take into consideration from the beginning the opening of the data, complying with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) [9] for its best use.

The remainder of this paper is structured as follows: Section 2, presents details of the main developments using LOD in the field of Argentine marine sciences. Section 3, enumerates preliminary results obtained which can be reused in future developments. In Section 4, we discuss the principles of LOD that tie everything together. Finally, in Section 5, we present some lessons learned in these years to serve as experience in future research related to marine science.

2 LOD prototypes

Taking into account the limitations of the conventional systems described in Section 1, we have developed different proofs of concept in the oceanographic domain and marine biodiversity, which were reported in different scientific journals and conferences. We summarize these efforts below.

2.1 Linked Data in oceanography

Regarding the management and modeling of Argentine marine science data such as KG and LOD we can enumerate:

- 2018: The first development was based on the publication of metadata from oceanographic campaigns related to Pampa Azul [10].
- 2019: an oceanographic KG prototype called OceanGraph was defined in [11]. It is currently under development integrating new data sources. A simplified view of the proposed KG is shown in Figure 2.
- 2020: in [12] the potential uses of OceanGraph were demonstrated with a concrete example by specialists.
- 2022: a LOD dataset of observational data and hydrographic profiles of the South Atlantic Ocean was published as LOD in [13]. This data set was integrated into the structure of OceanGraph KG.

Based on the experience gained in these works, a series of recommendations related to the interoperability and integration of information from the Global Ocean Observing System (IOOS) were published in [14], this work being carried out in collaboration with the U.S. Integrated Ocean Observing System, the Norwegian Institute of Marine Research (IMR) and the National Centers for Environmental Information (NCEI).

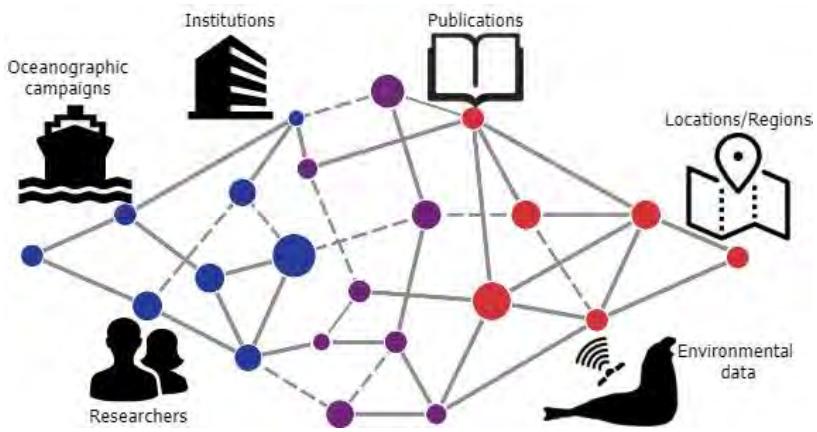


Fig. 2. Overview of *OceanGraph*: integrates information from oceanographic campaigns, scientific publications, researchers, institutions, marine locations and regions, and environmental variables from sensors placed on animals.

2.2 Linked Data in marine biodiversity

In addition to previous developments, we have published a marine biodiversity LOD dataset [15] and developed a linked data dashboard to visualize and complement information on certain species with other linked datasets [16]. In [15], we collaborate with Research Group on Languages and Artificial Intelligence (GILIA-UNCOMA) of Universidad Nacional del Comahue and Department of Computer Science and Engineering (DCIC) of Universidad Nacional del Sur, while [16], was a collaboration with the Argentine node of the Global Biodiversity Information Facility (GBIF)[5] and VertNet[6].

Finally, an ontology-based system called BiGe-Onto [17] was developed for the integrated management of marine Biodiversity and Biogeography information and a LOD dataset publicly available through DOI 10.5281/zenodo.3235548.

[5] https://www.gbif.org/es/country/AR/summary
[6] http://vertnet.org/

3 Preliminary results

As we detailed in the previous section, since 2018 we have developed different proposals using ontologies, LOD and publication of oceanographic data following the FAIR principles, in this section we discuss the preliminary results obtained and how these developments can be reused by other initiatives whose data pertain to marine sciences.

From a theoretical point of view, a survey and selection of standard ontologies was carried out for the extension of BiGe-Onto, which will later be the central model of OceanGraph. The survey was carried out from public repositories of ontologies and the subsequent analysis of the conceptual models underlying said ontologies. From a practical point of view, a web application is being implemented to present metadata results visually on data from oceanographic surveys, types of sampling carried out, people and institutions involved, and recorded environmental variables. This implementation requires the design of a software architecture and the subsequent selection of current web development and semantic web technologies. In this last group, we work on the selection of knowledge graph storage systems that also provide efficient search engines. A logical reasoner is being developed for data validation with a temporal dimension that allows the modeling of data in temporal logic from relational databases. Also, work is being done on the validation and scalability of this approach with case studies from different domains. As a summary, we can list the following results:

- a network of ontologies (semantic model), modeling marine science domain, with focus on sensor data and biodiversity.
- a LOD dataset.
- a proof-of-concept application to explore visually the KG.
- a set of running examples that potential consumers can use as training material. They consist of natural language Competency Questions (CQs) and their corresponding SPARQL queries [18].
- a SPARQL endpoint[7] to explore the resource, run tests, etc.

Regarding the semantic model we consider that the main contribution to highlight it is the main component of OceanGraph is a KG, intended as the union of the ontology network defined in Web Ontology Language (OWL) [19] and LOD data. Nevertheless, the KG is released as part of a package including accompanying material (documentation and online services) that support its consumption, understanding and reuse. OceanGraph bases its main structure on the relationships established between the selected datasets. The main classes that we define and reuse are: *campaigns, occurrences, papers, researchers, environmental variables and positions*. If a researcher consults OceanGraph, the expected results could recover one or more oceanographic campaigns in which she/he was involved from SNDM, datasets they collected from GBIF and Ocean

[7] https://linkeddata.cenpat-conicet.gob.ar/snorql/

Biogeographic Information System (OBIS)[8], and papers written by themself (from Springer Nature SciGraph)[9].

In the same way, the user could query data related to the occurrence of a species and the KG must retrieve in which campaigns it was observed, the information of the person who collected it, the exact place and date and associated variables that may be of importance (*e.g.*, weather or other environmental conditions during the collection).

We reuse only the elements from these ontologies that are necessary for modeling our data, adopting a *soft reuse* strategy [20] instead of importing the whole ontologies. OceanGraph ontology network consists of several ontologies modules connected by `owl:imports` axioms (See Figure 3). A list of prefixes and their corresponding URIs are listed in Table 1.

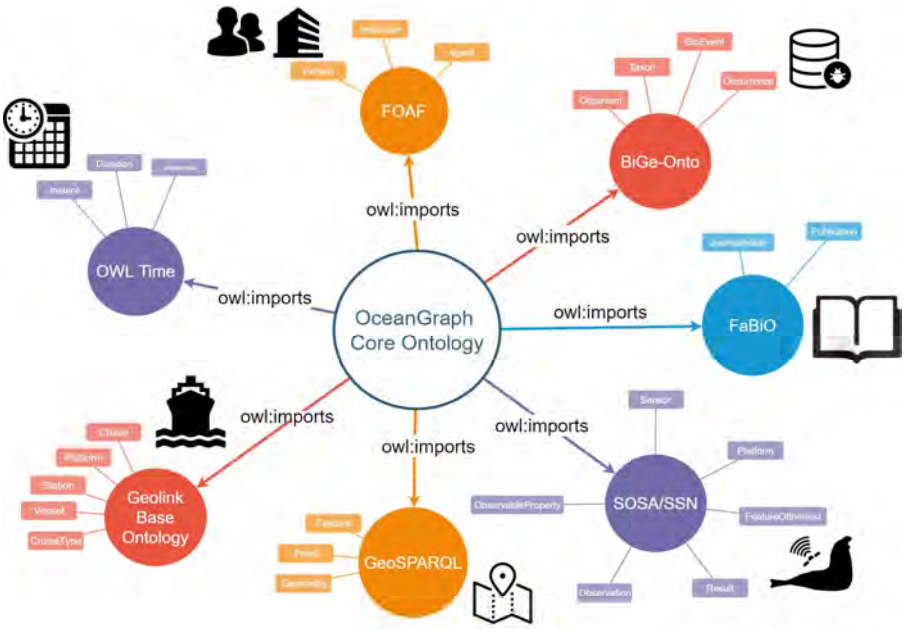


Fig. 3. OceanGraph ontology network intended to be adapted to different domains and reused by different marine science projects.

To synthesize the results obtained, we can highlight that the semantic model is generic enough to incorporate new data sources and be reused in other projects. Compliance with the FAIR principles allows the information from Argentine marine sciences to be visible and reusable by third-party applications that are interested in its exploitation.

[8] http://www.iobis.org/
[9] https://www.springernature.com/gp/researchers/scigraph

Table 1: Reused vocabularies and ontologies.

Ontology/Vocabulary name	Prefix
BiGe-Onto ontology	bigeonto
Semantic Sensor Network Ontology	ssn
Sensor, Observation, Sample, and Actuator Ontology	sosa
Darwin Core (literal values)	dwc
Darwin Core (IRI values)	dwciri
GeoSPARQL ontology	geosparql
W3C Time Ontology	time
FRBR-aligned Bibliographic Ontology	fabio
NERC vocabulary server (measured phenomena)	P01
NERC vocabulary server (biological entity sex)	S10
Quantities, Units, Dimensions and Types Ontology (v1.1) vocabulary	qudt
Quantities, Units, Dimensions and Types Ontology (version 1.1) schema	qudts
GoodRelations (v1.0)	gr
Simple Knowledge Organization System	skos

4 Discussion

In this section we discuss the aspects related to fulfilment of LOD principles of and aspects related to the semantic model used in OceanGraph.

LOD [21] is an idea from the Semantic Web [7] aimed at ensuring that data published on the Web is reusable, discoverable, and more importantly, that data published by different entities can work together. LOD principles are summarized in:

- Use URIs as names for things.
- Use HTTP URIs so people can look up these things.
- When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
- Include links to other URIs so they can discover more things.

We have followed these guidelines when creating all datasets described in Section 2. Below we discuss each of these points separately.

Usage of URIs as resource identifiers Each instance is uniquely identifiable by an HTTP URI. For example, we define the result of a measurement of the average depth of the water column measured by an instrument as: http://linkeddata.cenpat-conicet.gob.ar/data/result/id-233/avg_depth. All instance identifiers follow this scheme.

Usage of HTTP URIs and dereferencing : According with linked data principles, we use dereferenceable HTTP URIs for our resources. For example for the average depth URI above, we generate a human-readable version of the dereferenced version using the URL: http://linkeddata.cenpat-conicet.gob.ar/page/result/id-233/avg_depth to dereference the URI.

Linking to other resources All resources in OceanGraph form a graph (there are no disconnected parts). In addition, resources are linked to external databases via properties like `owl:sameAs`, `skos:broader` and `skos:exactMatch`. These identifiers can be: ORCIDs, Wikidata entities, DBPedia resources, NERC vocabulary server ID´s, etc. We have created links between people and their ORCID records, publications and their OpenCitations records, as well as the environmental variables were related to the identifier in NERC. For example, average water temperature corresponds to the NERC identifier `SDN:P02::TEMP` through `skos:broader` property. See: `http://linkeddata.cenpat-conicet.gob.ar/resource/observableProperty/id-233/avg_temp` to understand this implementation.

Availability, sustainability, and licensing One of the most important design decisions when developing a KG is the platform that supports it. After several performance comparisons, we decided to use GraphDB[10] since it allows a quick integration of new sources of information, analyzes structured data in CSV, XLS, JSON, XML or other formats, it allows to generate data in RDF and store it in a local or remote SPARQL endpoint, and last but not least, it allows to clean the input data with a generic script language. GraphDB allows users to explore the hierarchy of RDF classes and its instances (Class hierarchy menu). In the same way, we can check the relationships between the KG classes and visually explore how many links were created between different class instances (Class relationship). To access the OceanGraph dataset, the user must authenticate themselves on `http://web.cenpat-conicet.gob.ar:7200/login`, using the following credentials (user: **oceangraph** password: **ocean.user**). *OceanGraph KG* is also available for download in DOI: 10.17632/9t5xkt9wwk.1 under CC BY 4.0 license.

5 Conclusions

Tim Berner-Lee suggested LOD principles [7] for judging data quality by its accessibility (open data access), by its format and structures, and by its interoperability with other data sources. The FAIR data principles have been introduced for similar reasons with a greater emphasis on achieving reuse. LOD gives a clear mandate to the opening of the data, while FAIR requires an established license for access and therefore includes the concept of reuse under consideration in the license agreement. In addition, FAIR makes a strong reference to contextual information required to improve data reuse. In accordance with LOD principles, such metadata would be considered interoperable data as well, however the requirement to augment the data with metadata indicates that FAIR is an extension of the LOD [22]. Our recommendation based on mistakes mentioned in Section 1 for data management in marine sciences is: it is not enough to develop useful applications for specific users, conception of these applications

[10] `http://graphdb.ontotext.com/`

must contemplate compliance with the FAIR principles so that they are truly useful. In particular the use of LOD from the beginning, this facilitates reuse by scientists and non-expert users, on the other hand it facilitates interoperability with other systems allowing more complex analyses.

As explained in Section 4, publication of the LOD version of OceanGraph allows compliance with a large part of the FAIR principles. There is a description of the data online, the data is available as RDF, and there are many links to structured vocabularies, and metadata about the collection is made available.

We envision OceanGraph as an integral part of the existing semantic network of marine science knowledge in Argentina, based on HTTP identifiers and controlled vocabularies. By enhancing and semantically linking OceanGraph knowledge to existing machine-readable data, we increase the quality of marine science data and increase the potential for reuse.

Acknowledgments: This research received funding from project *Linked Open Data Platform for Management and Visualization of Primary Data in Marine Science*. Project No. PI-1562. Financed by Secretariat of Science and Technology of the National University of Patagonia San Juan Bosco (UNPSJB).

References

1. Se relanzo la iniciativa pampa azul. https://www.argentina.gob.ar/noticias/se-relanzo-la-iniciativa-pampa-azul, 2020. [Online; accessed 23-Feb-2022].
2. Portal de datos de biodiversidad argentina. https://datos.sndb.mincyt.gob.ar/, 2014. [Online; accessed 23-Mar-2022].
3. Tim Robertson, Markus Döring, Robert Guralnick, David Bloom, John Wieczorek, Kyle Braak, Javier Otegui, Laura Russell, and Peter Desmet. The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLoS ONE*, 2014.
4. John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core: An evolving community-developed biodiversity data standard. *PLoS ONE*, 2012.
5. Tanu Malik and Ian Foster. Addressing data access needs of the long-tail distribution of geoscientists. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 5348–5351. IEEE, 2012.
6. Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2, 2016.
7. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.
8. Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.
9. Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

10. Marcos Zárate, Pablo Rosales, Pablo Fillottrani, Claudio Delrieux, and Mirtha Lewis. Oceanographic data management: Towards the publishing of pampa azul oceanographic campaigns as linked data. In *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2018)*, 2018.
11. Marcos Zárate, Pablo Rosales, Germán Braun, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Oceangraph: Some initial steps toward a oceanographic knowledge graph. In Boris Villazón-Terrazas and Yusniel Hidalgo-Delgado, editors, *Knowledge Graphs and Semantic Web*, pages 33–40, Cham, 2019. Springer International Publishing.
12. Marcos Zárate, Carlos Buckle, Renato Mazzanti, Mirtha Lewis, Pablo Fillottrani, and Claudio Delrieux. Harmonizing big data with a knowledge graph: Oceangraph kg uses case. In Enzo Rucci, Marcelo Naiouf, Franco Chichizola, and Laura De Giusti, editors, *Cloud Computing, Big Data & Emerging Topics*, pages 81–92, Cham, 2020. Springer International Publishing.
13. Marcos Zárate, Germán Braun, Mirtha Lewis, and Pablo Fillottrani. Observational/hydrographic data of the south atlantic ocean published as lod. *Semantic Web*, 13(2):133–145, 2022.
14. Derrick Snowden, Vardis M Tsontos, Nils Olav Handegard, Marcos Zarate, Kevin O'Brien, Kenneth S Casey, Neville Smith, Helge Sagen, Kathleen Bailey, Mirtha N Lewis, et al. Data interoperability between elements of the global ocean observing system. *Frontiers in Marine Science*, 6:442, 2019.
15. M. Zárate, G. Braun, and P. Fillottrani. Adding biodiversity datasets from argentinian patagonia to the web of data. *CEUR Workshop Proceedings*, 1963, 2017.
16. Marcos Zárate, Paula F Zermoglio, John Wieczorek, Anabela Plos, and Renato Mazzanti. Linked open biodiversity data (lobd): A semantic application for integrating biodiversity information. *Biodiversity Information Science and Standards*, 4:e58975, 2020.
17. Marcos Zárate, Germán Braun, Pablo Fillottrani, Claudio Delrieux, and Mirtha Lewis. Bige-onto: an ontology-based system for managing biodiversity and biogeography data. *Applied Ontology*, 15(4):411–437, 2020.
18. Sparql 1.1 overview. w3c recommendation 21 march 2013. https://www.w3.org/TR/sparql11-overview/, 2013. [Online; accessed 16-Jul-2022].
19. W3C Owl Working Group et al. Owl 2 web ontology language document overview. *http://www. w3. org/TR/owl2-overview/*, 2009.
20. Mariano Fernández-López, María Poveda-Villalón, Mari Carmen Suárez-Figueroa, and Asunción Gómez-Pérez. Why are ontologies not reused across the same domain? *J. Web Semant.*, 57, 2019.
21. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI global, 2011.
22. Ali Hasnain and Dietrich Rebholz-Schuhmann. Assessing fair data principles against the 5-star open data principles. In *European Semantic Web Conference*, pages 469–477. Springer, 2018.

Un Estudio de Procesos de Diseño de Bases de Datos NoSQL

Luciano Marrero[1], Verena Olsowy [1], Fernando Tesone[1], Pablo Thomas[1], Leandro Corbalán[1], Juan Fernández Sosa[1], Patricia Pesado[1]

[1] Instituto de Investigación en Informática LIDI
Facultad de Informática - Universidad Nacional de La Plata – Argentina
Centro Asociado Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

{lmarrero, volsowy, ftesone, pthomas, corbalan, jfernandez
ppesado}@lidi.info.unlp.edu.ar

Resumen: Los Sistemas de Gestión de Bases de Datos (SGBDs) NoSQL no poseen un proceso tradicional para el modelado de datos. Para los SGBDs relacionales, generalmente, existe un proceso de diseño que se inicia con un modelo conceptual, luego es derivado a un modelo lógico relacional y finaliza con un modelo físico de la Base de Datos. La ausencia de documentación en un proceso de diseño de Bases de Datos puede llevar a una interpretación errónea sobre la semántica de los datos que se encuentran almacenados. En este trabajo se plantea una revisión bibliográfica sobre los procesos de diseño para obtener un esquema de los datos en Bases de Datos NoSQL. El principal objetivo es el de realizar un análisis comparativo a partir de las propuestas existentes.

Keywords: Proceso de Diseño de Bases de Datos, Modelado de Datos en Bases de Datos NoSQL, Almacenamiento no estructurado de datos.

1. Introducción

Las metodologías tradicionales de diseño y construcción de Bases de Datos relacionales han sido ampliamente estudiadas, aplicadas y refinadas por décadas. Inicialmente, en base a una especificación de requerimientos, se presenta una descripción conceptual de alto nivel. Posteriormente, se realiza un mapeo del modelo conceptual a algún tipo de modelo lógico aplicando las reglas adecuadas para un determinado tipo de Base de Datos, siendo el más popular, el modelo lógico relacional [26]. Finalmente, se obtiene una representación del esquema físico de la Base de Datos. Los principios o reglas que se aplican a un modelo de datos relacional no resultan apropiados para una Base de Datos NoSQL. Esto se debe a que NoSQL no representa un tipo de Base de Datos, sino que son un conjunto de tipos de Bases de Datos y cada una de ellas posee su propia forma de estructurar y almacenar sus datos. En general, NoSQL, se basa en la redundancia, la desnormalización y las consultas que se impactarán sobre el sistema. Las Bases de Datos relacionales y NoSQL

proponen modelos diferentes, por lo tanto, su proceso de diseño también tendrá que ser diferente [27, 28].

En este trabajo se propone realizar una revisión bibliográfica sobre procesos y/o metodologías de diseño para Base de Datos NoSQL. Además, se realiza un análisis comparativo con el objetivo de brindar una guía rápida que muestre las distintas características de cada enfoque propuesto.

A partir de la sección 2, el trabajo se organiza del siguiente modo: se detallan las características de NoSQL, en la sección 3 se presentan los trabajos relacionados, en la sección 4 se realiza el estudio propuesto, en la sección 5 se presenta un análisis comparativo, y finalmente, en la sección 6 y 7 se expresan las conclusiones generales y trabajos futuros.

2. Características de las Bases de Datos NoSQL

Las Bases de Datos NoSQL representan un conjunto de tipos de Bases de Datos que poseen sus propias implementaciones para el almacenamiento de datos. NoSQL se distingue de los tradicionales sistemas de gestión de Bases de Datos relacionales en diversos aspectos; no poseen un lenguaje de consulta estructurado (SQL) como lenguaje principal, no requieren de una estructura fija y tabular, no soportan operaciones JOIN, no garantizan por completo las propiedades de ACID (atomicidad, consistencia, aislamiento y durabilidad), y en general, su estructura se basa en la escalabilidad horizontal [27, 29, 30].

NoSQL propone un sistema llamado "BASE (Básicamente Disponible, Estado Suave, Consistencia Eventual)". A través de estas propiedades se logra disponibilidad básica (Base Availability), esto significa que el sistema se encontrará disponible la mayoría del tiempo. Con el estado débil (Soft State) el sistema se vuelve más flexible en cuanto a consistencia y con la consistencia eventual (Eventual Consistency) se garantiza que el sistema eventualmente se volverá consistente [28, 29, 30].

Dependiendo de la forma en que se almacene la información, existen cuatro categorías principales de almacenamiento para Bases de Datos NoSQL.

Almacenamiento Clave/Valor: simples en cuanto a su implementación, almacenan datos como un conjunto de pares "clave/valor" (key-value). La clave representa un identificador único que puede retornar un objeto complejo y arbitrario de información, denominado valor (value). Por ejemplo, Redis y Amazon DynamoDB, entre otros, implementan este tipo de almacenamiento [38, 39].

Almacenamiento Documental: el concepto central de este tipo de almacenamiento es el documento. Una Base de Datos NoSQL Documental, almacena, recupera y gestiona datos de documentos. Estos documentos encapsulan y codifican datos o información bajo algún formato estándar (XML, YAML, JSON, BSON). Por ejemplo, MongoDB y Apache CouchDB, entre otros, son implementaciones de Bases de Datos Documentales [34, 35].

Almacenamiento de Familia de Columnas: en este tipo de almacenamiento los datos se encuentran organizados por columnas, en lugar de filas. Las Bases de Datos que utilizan esta forma de almacenamiento tienden a ser un híbrido entre las Bases de

Datos Relacionales y la tecnología orientada a columna. Por ejemplo, Cassandra y Apache HBase, entre otros, utilizan este tipo de almacenamiento [40, 41].

Almacenamiento de Grafos: en este tipo de almacenamiento se representa la Base de Datos bajo el concepto de un grafo, permitiendo almacenar la información como nodos y sus respectivas relaciones, con otros nodos, mediante aristas. Se aplica la teoría de grafos para recorrer la estructura. Son útiles para almacenar información en modelos que poseen numerosas relaciones entre sus datos. Neo4j y OrientDB, entre otros, implementan este tipo de almacenamiento [36, 37].

3. Trabajos relacionados

Se han recopilado tres trabajos relacionados que realizan revisiones bibliográficas sobre el diseño de Bases de Datos NoSQL.

En [24] se realiza una revisión sistemática de la literatura para el diseño de Bases de Datos no Relacional. El análisis realizado en este trabajo es un aporte importante para futuras investigaciones y la evolución de los métodos de diseño.

En [25] se realiza una revisión de los enfoques existentes para el diseño de Bases de Datos en general con el objetivo de detectar consideraciones y necesidades en comparación al diseño tradicional de Base de Datos (Relacional).

En [42] se realiza un estudio exhaustivo sobre decisiones de diseño para el almacenamiento no estructurado de información; además se incluyen otras temáticas como modelo de consistencia, partición de datos y teorema de CAP.

En este trabajo, a diferencia de [24], [25] y [42], se presenta una revisión bibliográfica con el aporte de una grilla que permite realizar una comparación rápida y precisa al momento de tener que seleccionar o analizar las distintas técnicas o procesos para el diseño de Bases de Datos NoSQL.

4. Enfoques de Diseño para Bases de Datos NoSQL

En esta sección se presenta un estudio realizado sobre 23 propuestas de diseño de Bases de Datos que utilizan almacenamiento no estructurado de datos, con el objetivo de obtener el panorama actual para el modelado y el diseño de este tipo de Base de Datos.

Cuando se piensa en el diseño de una Base de Datos NoSQL, generalmente, se presenta directamente su estructura a nivel físico teniendo en cuenta el problema a resolver. Realizar un modelo de datos en un nivel conceptual para NoSQL es algo que aún es tema en discusión. Generalmente, un modelo de datos conceptual describe sus componentes en términos de entidades y cómo éstas se relacionan, independientemente de la tecnología a utilizar. Existen aspectos como la redundancia, la desnormalización y las consultas del sistema que en una etapa conceptual aún no son consideradas.

A continuación, se presenta un breve resumen sobre distintas publicaciones consultadas que han abordado la temática de procesos de diseño para Bases de Datos NoSQL desde diferentes perspectivas. Para la búsqueda de los artículos, se han consultado diversas fuentes, por ejemplo, Google Scholar [33], IEEE Xplore [31], Springer [32], entre otros. Además, se tuvieron en cuenta las referencias que expone cada uno con el objetivo de seguir un lineamiento temporal del trabajo.

Los autores de "*ToNoSQLModel Process / NoSQLToUML*" [1, 14], proponen un enfoque automático para obtener un modelo físico (en formato JSON) a partir del motor de Base de Datos No Relacional MongoDB [34]. Este proceso se lleva a cabo a través de una secuencia de transformaciones formales utilizando QVT (Query/View/Transformation) y siguiendo un conjunto de reglas bien definidas.

Los autores de "*ToConceptualModel*" [4], amplían [1] proponiendo un enfoque de Ingeniería Inversa para Bases de Datos NoSQL orientadas a documentos realizando una transformación del modelo físico en un modelo conceptual mediante un diagrama de clases UML (Unified Modeling Language). Para su implementación utilizan Eclipse Modeling Framework (EMF, por sus siglas en Inglés) y QVT (Query/View/Transformation).

Los autores en [2] presentan "*JSON Discoverer*", una herramienta que permite mapear un documento JSON y obtener su esquema implícito a través de un diagrama UML. La herramienta generada se puede ver en la siguiente URL: http://som-research.uoc.edu/tools/jsonDiscoverer/#/. La herramienta tiene como objetivo asistir a los desarrolladores en la realización de tareas que involucren inferir y visualizar el esquema implícito de los datos que poseen un documento JSON.

En [3] se presenta un nuevo enfoque de extracción de esquemas de documentos JSON mediante la introducción de un algoritmo para la extracción de esquemas que operan fuera de un motor de Bases de Datos NoSQL. En lugar de diseñar el esquema de forma adelantada, se realiza una extracción, algo que puede verse como un proceso de Ingeniería Inversa.

En [5] se presenta un enfoque basado en una Arquitectura Dirigida por Modelos (Model-Driven Architecture o MDA) que permite lograr una Ingeniería Inversa de las Bases de Datos orientadas a Grafos [36, 37]. Parte de un esquema físico descrito mediante el lenguaje Cypher propuesto por Neo4j [36]. Luego, se deduce un grafo lógico y finalmente se mapea el grafo lógico obtenido a un esquema conceptual a través de un Diagrama de Entidad-Relación Extendido (EERD).

En [6] se propone un método para el diseño no relacional denominado "*NoAM (NoSQL Abstract Model)*". Se plantea un proceso que incluye una fase inicial conceptual (expresada en UML), seguida de una fase de diseño lógico NoAM, propuesto por los autores e independiente del sistema, y una fase final en la que se tiene en cuenta las características de los sistemas individuales. NoAM explota los puntos comunes existentes en distintos sistemas NoSQL e introduce abstracciones para equilibrar diferencias y variaciones.

En [7, 8] se presenta una solución de Ingeniería Inversa para inferir esquemas de Bases de Datos NoSQL teniendo en cuenta el versionado de los datos. El esquema obtenido se expresa en UML y la solución se ha implementado con MDE (Model-Driven Engineering) para conseguir independencia del tipo de Base de Datos.

Los autores en [9] proponen un marco en que, dado un cambio en el modelo conceptual, identifica lo que se debe modificar en un esquema de Base de Datos NoSQL y los datos subyacentes. El trabajo se centra en el estudio de siete tipos de cambios del modelo conceptual y para cada cambio se describe la transformación requerida en el esquema de la Base de Datos para mantener la consistencia entre el esquema y el modelo.

En [10] se plantean nuevas directrices de modelado para Bases de Datos NoSQL Documental. Estas directrices abarcan tanto las etapas lógicas como las físicas de los diseños. Cada una de ellas se desarrolla sobre conocimiento empírico desarrollando un enfoque exploratorio mediante consultas a expertos.

Los autores en [11] proponen "*Schema Design Advisor Model (SDAM)*" y un algoritmo automático para el problema de diseñar un esquema de Base de Datos Documental. Se logra un modelo de asesoramiento para el diseño de esquemas documentales, a partir de un esquema relacional teniendo en cuenta las consultas del sistema.

En [12] se presenta "*Mortadelo*", un proceso de diseño de Base de Datos NoSQL basado en modelos, en donde, a partir de un esquema conceptual se puede generar de forma automática una implementación para una Base de Datos NoSQL. Este proceso se puede personalizar, de modo que algunas compensaciones de diseño se pueden gestionar de manera diferente según las necesidades de cada contexto. Con "*Mortadelo*" se generaron implementaciones para Cassandra [40] y MongoDB [34] a partir de un mismo modelo conceptual.

En [13] se presenta el diseño de un método estructurado para la selección y el diseño del modelo de Base de Datos basado en una variedad de factores (relaciones entre los datos, requisitos funcionales y requisitos no funcionales). El método recomendará qué modelos de Base de Datos son los más apropiados para una aplicación y sugerirá un diseño para los modelos recomendados.

Los autores en [15] proponen una metodología de modelado basada en consulta para el motor de base de datos NOSQL Apache Cassandra [40]. Definen importantes principios de modelado, reglas de mapeo y patrones de mapeo, para lograr un modelo lógico de datos. Además, diseñaron e implementaron una herramienta de modelado de datos basada en la web denominada KDM (www.cs.wayne.edu/andrey/kdm).

En [16] se propone un sistema para recomendar esquemas de Bases de Datos no relacionales a partir del modelo conceptual de datos de la aplicación. Además, se implementó un prototipo basado en este enfoque para el motor de Base de Datos Apache Cassandra [40]. Este prototipo denominado "*NoSQL Schema Evaluator (NoSE)*" produce esquemas eficientes y permite examinar más alternativas de las que serían posibles con un enfoque manual.

Los autores en [17] proponen un modelo conceptual común para diversos tipos de Bases de Datos NoSQL. También han ideado un lenguaje de especificación de datos NoSQL para representar un modelo de datos de nivel lógico equivalente e independiente de cualquier representación de nivel físico. Además, se han formalizado e ilustrado distintas reglas de validación relativas al modelo conceptual propuesto, utilizando un caso de estudio adecuado.

En [18] se propone un mecanismo para transformar un modelo conceptual en un modelo lógico para Bases de Datos NoSQL. El modelo lógico se representa en notación JSON. La representación lógica propuesta es capaz de representar datos estructurados y semiestructurados, y es capaz de una mayor transformación hacia tipos heterogéneos de Bases de Datos NoSQL.

Los autores de [19] abordan el problema de diseño de Bases de Datos NoSQL. Proporcionan métodos que asignan un modelo de datos estándar a un modelo de datos admitido por los servicios de datos NoSQL, utilizando "*AWS DynamosDB*" (Clave-Valor) como Base de Datos específica [39].

En [20] se muestra una estrategia de diseño en esquemas para el motor de Base de Datos "DynamoDB" (Clave-Valor) [39], teniendo en cuenta los patrones de acceso a los datos. Además, se presenta una simulación del enfoque.

En [21] se presenta una metodología de apoyo al diseño para Bases de Datos NoSQL Clave-Valor. Esta metodología propone un modelo lógico que considera los patrones de acceso y/o consulta a una Base de Datos Clave-Valor [38, 39].

Los autores en [22] proponen un método de modelización lógica adecuado para las Bases de Datos NoSQL Documentales, incorporando un nuevo tipo de diagrama denominado Diagrama de Interrelación de Documentos (DID). Este trabajo se basa en la modelización conceptual (DER) y suman patrones de consulta que surgen del análisis de requisitos.

En [23] se presenta un método de mapeo para el modelado de Bases de Datos no relacionales. Mediante la aplicación de reglas, se unifica la forma de mapear esquemas Documentales, Familia de Columnas y Grafos.

5. Diseño de Bases de Datos NoSQL. Análisis Comparativo

Luego de realizar el análisis de cada uno de los trabajos recopilados, se presenta una grilla comparativa para visualizar las distintas características que poseen los diferentes enfoques de procesos y/o métodos de diseño para motores de Bases de Datos NoSQL (clave-valor, documental, familia de columnas y grafos). En este trabajo se han definido los siguientes criterios:

- **Aplicable a**: estrategias tratadas (clave-valor, documental, familia de columnas o grafos) y/o motor de Bases de Datos NoSQL específico.
- **Enfoque**: define formalmente un proceso o es una metodología propuesta sobre el tema.
- **Herramienta de automatización**: Si implementan una herramienta para automatizar alguna parte del proceso.
- **Pautas bien definidas**: Si define todos los pasos necesarios para el proceso o técnica que describe.
- **Diseño completo**: si parte con la realización de modelo inicial o utiliza algún modelo conocido.
- **Diagramas**: diagramas que se utilizan en el proceso descripto.
- **Referencia**: artículo y/o trabajo analizado.

A continuación, en las tablas 1, 2, 3, 4 y 5 se presenta el resultado del análisis realizado.

Tabla 1 - Enfoques aplicables a almacenamiento ***Documental***

Aplicable a	Enfoque	Herramienta de automatización	Pautas definidas	Diseño completo	Diagramas	Referencia
MongoDB	Ingeniería inversa	No	Si	No	Físico definido por los autores y UML	[1, 4, 14]
General	Herramienta web	Si	No	No	UML y Físico JSON	[2]
General	ingeniería inversa	No	Si	No	XML y JSON	[3]
General	Grupo de normas	No	Si	Si	UML	[10]
MongoDB	Relacional a documental	No	Si	No	Físico	[11]
General	Modelado de datos	No	Si	No	DID (Diagrama de interrelación de documentos)	[22]

Tabla 2 - Enfoques aplicables a almacenamiento ***Clave-Valor***

Aplicable a	Enfoque	Herramienta de automatización	Pautas definidas	Diseño completo	Diagramas	Referencia
General	Modelo de datos	No	No	Si	DER y UML	[21]

Tabla 3 - Enfoques aplicables a almacenamiento de ***Grafos***

Aplicable a	Enfoque	Herramienta de automatización	Pautas definidas	Diseño completo	Diagramas	Referencia
Neo4j	Ingeniería inversa	No	Si	No	Esquema físico, Grafo lógico y Modelo conceptual	[5]

Tabla 4 - Enfoques aplicables a almacenamiento de ***Familia de Columnas***

Aplicable a	Enfoque	Herramienta de automatización	Pautas definidas	Diseño completo	Diagramas	Referencia
Cassandra	Actualización	No	Si	No	No aplica	[9]
Cassandra	Modelado de datos	Si	Si	Si	Modelo conceptual y Diagrama Chebotko	[15]
General	Modelado de datos	Solo en Cassandra	Si	Si	Modelo conceptual y físico	[16]

Tabla 5 - Enfoques aplicables a más de un tipo de almacenamiento

Aplicable a	Enfoque	Herramienta de automatización	Pautas definidas	Diseño completo	Diagramas	Referencia
Documental, Clave-Valor y Familia de Columnas	Ingeniería inversa	No	Si	No	JSON y UML	[7, 8]
Documental, Clave-Valor y Familia de Columnas	Proceso de diseño	No	Si	Si	UML, Lógico NoSQL y Físico	[6]
General	Proceso de diseño	No	Si	Si	Generic Data Metamodel (GDM), Metamodelo para Familia de Columnas y Metamodelo para Documental	[12]
General	Asesor de base de datos	No	No	Si	Ninguno	[13]
General	Proceso de diseño	No	Si	Si	Conceptual NoSQL y Lógico JSON	[17, 18]
DynamoDB	Modelado de datos	No	Si	Si	Modelo de datos estándar y Físico DynamoDB	[19]
DynamoDB	Modelado de datos	No	Si	Si	Hipergrafo de modelo conceptual, Esquema jerárquico y Esquema Físico	[20]

Documental, Familia de Columnas y Grafos	Modelado de datos	No	Si	No	TPC-Hbenchmark	[23]

6. Conclusiones

Este trabajo se centra en una revisión de la bibliografía existente para el estudio y análisis sobre distintos procesos y/o metodologías de diseño para Bases de Datos NoSQL. Para ello, se consultaron diversas fuentes, como Google Scholar [33], IEEE Xplore [31], Springer [32] entre otros. Los criterios de exploración y selección se basaron principalmente en la identificación de artículos que traten sobre diseño Bases de Datos NoSQL para alguna de las 4 categorías principales de almacenamiento no estructurado de información (documental, clave-valor, familia de columnas y grafos). En este contexto, se analizaron más de 30 artículos. De este grupo, se han seleccionado unos 23 trabajos que proponen o se aproximan al desarrollo de un proceso o metodología para el diseño de Bases de Datos NoSQL con líneas diferenciadas entre sí, abarcando distintas alternativas y características. Se pretende obtener un panorama tentativo sobre cómo ha evolucionado el diseño de Bases de Datos NoSQL en los últimos años.

A través de este estudio, se puede observar que existe una gran variedad de enfoques y para un mismo tipo de almacenamiento, se tienen distintas variantes según la visión del autor. No contar con la posibilidad de tener un proceso de diseño unificado, como sucede para las Bases de Datos Relacionales, probablemente implique tener que realizar grandes modificaciones al modelo conceptual (si es que ha sido realizado) ante un cambio en el modelo físico. Además, algunos de los enfoques, son específicos para un motor de Base de Datos NoSQL, o el proceso de diseño propuesto, involucra numerosas pautas o reglas que los diseñadores deben aplicar de forma manual, algo que puede ser propenso a cometer errores.

En resumen, si bien desde hace varios años existen diversos estudios sobre el diseño de Bases de Datos NoSQL, aún no se ha encontrado un proceso de diseño adoptado formalmente que contenga una serie de pasos establecidos y que sea genérico a casi todas las formas de almacenamiento no estructurado de datos, similar a lo que sucede en Bases de Datos Relacionales (Modelado Conceptual, Modelado Lógico y Modelado Físico). Además, existen propiedades que las Bases de Datos NoSQL poseen, que son muy poco consideradas, como la disponibilidad, la consistencia eventual y la tolerancia a particiones.

7. Trabajo Futuro

Como trabajo futuro, se prevé extender la revisión bibliográfica realizada con el objetivo de tener un panorama más amplio sobre el diseño de Bases de Datos NoSQL.

Además, se pretende seleccionar aquellos enfoques y/o procesos que abarcan todas las etapas del diseño (Conceptual, Lógico y Físico) e incorporarlos a proyectos reales con el objetivo evaluar su aplicabilidad y aportar mejoras o nuevas características que aún no se tienen en cuenta durante el diseño de la Base de Datos, por ejemplo, la escalabilidad horizontal.

8. Bibliografía

1. Brahim, A.; Ferhat, R. and Zurfluh, G. (2019). Model Driven Extraction of NoSQL Databases Schema: Case of MongoDB.In Proceedings of the 11th International Joint Conference on Knowledge Discovery. V.1: KDIR, ISBN 978-989-758-382-7, pages 145-154.
2. Izquierdo, J. L. C., & Cabot, J. (2016). JSONDiscoverer: Visualizing the schema lurking behind JSON documents. Knowledge-Based Systems, 103, 52-55.
3. M. Klettke, U. St ¨orl, S. Scherzinger, Schema Extraction and Structural Outlier Detection for JSON-based NoSQL Data Stores, in: BTW conf., 2015, pp. 425–444.
4. Fatma ABDELHEDI, Amal AIT BRAHI, Rabah TIGHILT FERHAT, Gilles ZURFLUH. Reverse engineering approach for NoSQL databases. 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020.
5. Isabelle Comyn-Wattiau;Jacky Akoka. Model driven reverse engineering of NoSQL property graph databases: The case of Neo4j.2017 IEEE International Conference on Big Data..
6. Paolo Atzeni, Francesca Bugiotti, Luca Cabibbo, Riccardo Torlone. Data Modeling in the NoSQL World. HAL open science. https://hal.archives-ouvertes.fr/hal-01611628.
7. Severino Feliciano Morales, Jesús García Molina,Diego Sevilla Ruiz.Inferencia del esquema en bases de datos NoSQL a través de un enfoque MDE.
8. Diego Sevilla Ruiz, Severino Feliciano Morales, Jesús García Molina. Inferring Versioned Schemas from NoSQL Databases and Its Applications. https://link.springer.com/chapter/10.1007/978-3-319-25264-3_35 (Springer).
9. Pablo Suárez Otero, Michael J. Mior, Maria José Suárez Cabal, Javier Tuya. Maintaining NoSQL Database Quality During Conceptual Model Evolution.
10. Paolo Atzenia,Francesca Bugiottib,Luca Cabibboa,Riccardo Torlonea.Data Modeling Guidelines for NoSQL Document-Store Databases. International Journal of Advanced Computer Science and Applications.
11. Basant Namdeo,Ugrasen Suman. Schema Design Advisor Model for RDBMS to NoSQL Database Migration.
12. Alfonso de la Vega, DiegoGarcía,Saiz Carlos Blanco,Marta Zorrilla, Pablo Sánchez. Mortadelo: Automatic generation of NoSQL stores from platform-independent data models.Future Generation Computer Systems. Volume 105, April 2020, Pages 455-474.
13. Noa Roy-Hubara.The Quest for a Database Selection and Design Method.CAiSE 2019 (31st International Conference on Advanced Information Systems Engineering).3-7 June 2019, Rome, Italy.
14. Amal AIT BRAHIM; Rabah TIGHILT FERHAT; Gilles ZURFLUH. Extraction process of conceptual model from a document-oriented NoSQL database. 2019 11th International Conference on Knowledge and Systems Engineering (KSE).
15. Artem Chebotko, Andrey Kashlev, Shiyong Lu.A Big Data Modeling Methodology for Apache Cassandra. 2015 IEEE International Congress on Big Data.
16. Michael Joseph Mior, Kenneth Salem;Ashraf Aboulnaga, Rui Liu.NoSE: Schema Design for NoSQL Applications. IEEE Transactions on Knowledge and Data Engineering.
17. Shreya Banerjee, Anirban Sarkar. Modeling NoSQL Databases: From Conceptual to Logical Level Design. 3rd International Conference on Applications and Innovations in Mobile Computing (AIMOC – 2016) At: Kolkata, India.

18. Shreya Banerjee, Anirban Sarkar. Logical level design of NoSQL databases.2016 IEEE Region 10 Conference (TENCON).
19. Sudarshan S. Chawathe. Data Modeling for a NoSQL Database Service. 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference 2019 (New York).
20. W. Y. Mok.A Feasible Schema Design Strategy for Amazon DynamoDB: A Nested Normal Form Approach. Industrial Engineering and Engineering Management (IEEM 2020).
21. Gerardo ROSSEL, Andrea MANNA. A Modeling methodology for NoSQL Key-Value databases.Database Systems Journal, Bucharest, Romania. 12-18, August.
22. Diseño de Bases de Datos basadas en Documentos: Modelo de Interrelación de Documentos. Rossel Gerardo, Manna Andrea. XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). Páginas 662-671. http://sedici.unlp.edu.ar/handle/10915/56746.
23. Afef Gueidi, Hamza Gharsellaoui, Samir Ben Ahmed.Towards Unified Modeling for NoSQL Solution Based on Mapping Approach. 25th KES-2021. http://kes2021.kesinternational.org/
24. Noa Roy-Hubara, Arnon Sturm.Design methods for the new database era: a systematic literature review. Software and Systems Modeling | Issue 2/2020 (https://www.springerprofessional.de/en/software-systems-modeling/5337962).
25. Noa Roy-Hubara, Arnon Sturm.Exploring the Design Needs for the New Database Era. Springer (https://link.springer.com/chapter/10.1007/978-3-319-91704-7_18).
26. Carlo Batini, Stefano Ceri, Shamkant B. Navathe. Diseño Conceptual de Bases de Datos, un enfoque de entidades-interrelaciones. ISBN 0-201-60120-6 (1994).
27. Análisis de performance en Bases de Datos NoSQL y Bases de Datos Relacionales. Pesado P., Thomas P., Delía L., Marrero L., Olsowy V., Tesone F.. XXVI Congreso Argentino de Ciencias de la Computación (CACIC 2020). ISBN 978-987-4417-90-9. http://sedici.unlp.edu.ar/handle/10915/114202.
28. Un estudio comparativo de bases de datos relacionales y bases de datos NoSQL. Pesado P., Thomas P., Delía L., Marrero L., Olsowy V., Tesone F., Fernandez S. J. XXV Congreso Argentino de Ciencias de la Computación (CACIC 2019). ISBN 978-987-688-377-1. http://sedici.unlp.edu.ar/handle/10915/91403.
29. Aspectos de Ingeniería de Software, Bases de Datos Relacionales, y Bases de Datos No Relacionales y Bases de Datos Como Servicios en la Nube para el desarrollo de Software Híbrido. XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021). ISBN 978-987-24611-3-3; 978-987-24611-4-0. http://sedici.unlp.edu.ar/handle/10915/120139.
30. NoSQL: modelos de datos y sistemas de gestión de bases de datos. XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste). http://sedici.unlp.edu.ar/handle/10915/67258.
31. IEEE Xplore.https://ieeexplore.ieee.org/Xplore/home.jsp. Mayo de 2022.
32. Springer. https://link.springer.com/. Mayo de 2022.
33. Google Scholar. https://scholar.google.com/. Mayo de 2022.
34. MongoDB. https://www.mongodb.com/es. Mayo de 2022.
35. CouchDB. https://couchdb.apache.org/. Mayo de 2022.
36. Neo4j. https://neo4j.com/. Mayo de 2022.
37. OrientDB. https://orientdb.org/. Mayo de 2022.
38. Redis. https://redis.io/. Mayo de 2022.
39. Amazon DynamoDB. https://aws.amazon.com/es/dynamodb/. Mayo de 2022.
40. Apache Cassandra. https://cassandra.apache.org/_/index.html. Mayo de 2022.
41. Apache HBase. https://hbase.apache.org/. Mayo de 2022.
42. Davoudian, Ali and Chen, Liu and Liu, Mengchi. A survey on NoSQL stores. ACM Computing Surveys (CSUR), volume=51, number=2, pages=1--43, year=2018, publisher=ACM New York, NY, USA.

Instance retrieval from non-labeled data as a strategy for automatic classification of imbalanced e-mail datasets

Juan Manuel Fernández[1], Marcelo Errecalde[2]

[1] Department of Basic Sciences, National University of Lujan, Argentina
[2] LIDIC, National University of San Luis, Argentina
jmfernandez@unlu.edu.ar, merreca@unsl.edu.ar

Abstract. One of the main challenges in automatic email classification problems occurs when it is necessary to work with a relatively large number of classes and the classes are highly imbalanced. That happens even when non-labeled textual bases are available because manual labeling is costly. In this respect, all automatic text classification strategies –to a greater or lesser extent– are sensitive to the problems of imbalance between classes.
The most widely used approaches for learning from unbalanced databases consists of resampling techniques, either by undersampling or oversampling the datasets. However, existing techniques have some problems to be solved.
In this work we present a new proposal that consists of balancing the classes of the data set by retrieving unlabeled instances (e-mails) that are similar to those of the minority classes. It is shown that, for the data set used, it is a valid, viable and competitive strategy with respect to the resampling strategies currently used to learn from imbalanced email databases.

Keywords: imbalanced data, automatic classification, information retrieval

1 Introduction

Text analysis and processing techniques face very complex problems within the area of computer science, mainly due to the difficulty of language analysis. That is caused by its ambiguity, mainly in the semantic analysis stage, and the relatively scarce training materials and the computational capacity required for the analysis to run certain algorithms very demanding in hardware resources. [29][5]. In particular, emails have specific characteristics concerning other textual elements that present some differences and problems between traditional text mining and *email mining*.

Regarding the problem of automatic classification of emails, it consists of assigning an email to a set of automatically predefined classes using, in general, a machine learning technique. The classification is generally performed on the

basis of relevant words or features extracted from the e-mail text and, since the classes are predefined and training instances are class-labeled, it is usually addressed as a supervised machine learning task [11].

Approaches to email classification include neural networks [1], techniques based on support vector machines, Naive Bayes and TF-IDF classifiers [28], among others. More recently, Deep Learning-based approaches like *Long-Short-Term-Memory* are gaining attention to classify spam emails [10].

Finally, as an evolution in the previous strategy, in 2017, a new neural network architecture, simple and parallelizable, called *Transformer* [30] was proposed. It is exclusively based on attention mechanisms and completely dispenses with recurrence and convolutions. From these ideas arise what is known in the literature like the current state of the art of language representation models, called BERT (Bidirectional Encoder Representations from Transformers) [12]. There are uncountable studies on text classification with this representation model and, in email classifications, it has shown improvements in performance compared to previous strategies [15].

All of those automatic classification strategies - to a greater or lesser extent - are affected by class imbalance problems. Class imbalance is present in many real-world classification datasets and consists of a disproportion in the number of examples of different classes in the problem. This situation hampers the performance of classifiers due to their accuracy-oriented design, which generally results in the minority class being overlooked [14].

In this work, different well-known strategies for learning from imbalanced data are evaluated and compared against a new one, in the specific domain of e-mail classification. This proposal consists in using the set of manually labeled data that constitute the (imbalanced) training set, to select the most representative words of each minority class and then using them to retrieve new instances from a repository of unlabeled data to balance the training dataset.

The rest of the article is organized as follows. Section II presents some related works, the addressed research gap, and our working hypothesis. Section III presents the research methodology with its involved tasks and Section IV describes the experimental study and the analysis of the results. Finally, Section V gives some conclusions, contributions of our work, and possible future work.

2 Background

Most machine learning algorithms work best with balanced datasets but the problem arises when the given datasets are highly imbalanced in nature [26]. Classification of these imbalanced datasets is a complex task for traditional classifiers, as they generally tend to favor the samples of the majority classes over the minority ones. A large number of techniques have been developed [21] [25] to correctly distinguish the minority classes. These techniques can be categorized into four main groups, depending on how they deal with the problem [14]:

– Algorithm level approaches (also called internal): try to adapt existing classifier learning algorithms to bias the learning toward the minority class.

– Data level (or external) approaches: aim at rebalancing the class distribution by resampling the data space.
– Cost-sensitive approaches: allow the definition of costs associated with each of the classes in order to generate a weighting in the classification.
– Ensemble-based methods: usually consist of a combination of an ensemble learning algorithm and one of the above techniques.

One of the most widely used is the data-level approach, which consists of resampling techniques that are used to balance the data by either *undersampling* or *oversampling* the dataset [25].

First, *undersampling* is the process of decreasing the number of instances (or samples) in the majority classes. Some of the most commonly used undersampling methods consist of using the KNN algorithm, clustering or ensemble techniques. In the case of the KNN (k-nearest neighbors) algorithm, it is used to eliminate data where the target class is not equal to the majority of its "nearest neighbor instances" [25]. The use of the *k-means* clustering method aims at balancing the instances of imbalanced classes by reducing the number of majority instances [22]. In turn, in random *undersampling* methods [7], instances of majority classes are generally randomly sampled without label replacement to create a fully balanced training set [23]. Finally, there are assembly methods such as *EasyEnsemble* [31] where the majority class is divided into several subsets where the size of each subset is equal to the size of a minority class.

Secondly, *oversampling* consists of increasing the number of instances or samples of minority classes by producing new instances or repeating pre-existing ones. The most common technique is known as SMOTE (Synthetic Minority Over-sampling Technique) [8], where, to oversample, a sample is taken from the data set and the k nearest neighbors are considered based on the feature space, creating a synthetic data point from the multiplication of one of the feature vectors and a random value, usually between 0 and 1. Another example of oversampling methods is Borderline-SMOTE [17] whose objective is to identify minority samples located at the decision boundary and use them for oversampling, avoiding the potential risks of overgeneralization that occur with SMOTE. RAMOBoost (Ranked Minority Oversampling in Boosting) [9] is a technique that systematically generates synthetic samples using an ordered sampling probability distribution. There are also other synthetic sample generation approaches, such as ADASYN [19] and MWMOTE [3], which have obtained good results based on modifications to the synthetic data generation mechanisms.

Finally, some studies have shown that the combination of oversampling and undersampling methods allows better classifier performance than methods used separately [8]. In any case, the number of approaches proposed to solve these problems allow us to infer the importance of the topic for the evolution of supervised machine learning techniques.

It is important to observe that techniques based on undersampling are not an alternative when minority classes have very few identified instances because it has been shown that to accurately characterize the effectiveness of such systems, they must be evaluated at the operational scale at which they will be used in

practice [20]. On the other hand, most oversampling techniques are based on the generation of new synthetic instances that are not part of the real observations, which clearly seems to be a limitation.

However, fundamentally as a result of the massification of Internet access, millions and millions of data are generated every day, and the amount of data available for training classification algorithms is not a restriction [13]. The limitations here are given by the capacity to label those available data with the traditional (manual) strategy performed by a human. Hence, while expert labels provide the traditional cornerstone for training and evaluating classifier models, limited or expensive access to experts represents a practical bottleneck [20].

In that context, we present a new alternative for learning from unbalanced data that generates new training samples, not artificially, but by identifying unlabeled instances in the original dataset. In this paper, we present a new approach, previously used as a semi-supervised labeling strategy [16], which consists of starting from a manually labeled dataset and, using feature selection strategies, extracting representative terms from each minority class to retrieve new instances from a repository of unlabeled data and thus balancing the dataset with non-synthetic examples.

3 Research Methodology

As discussed above, the general objective of this research is to present a new strategy for learning from imbalanced data and to evaluate its performance for automatic email classification in relation to oversampling and undersampling strategies widely used in the scientific community. Figure 1 shows the schematic diagram of the process developed.

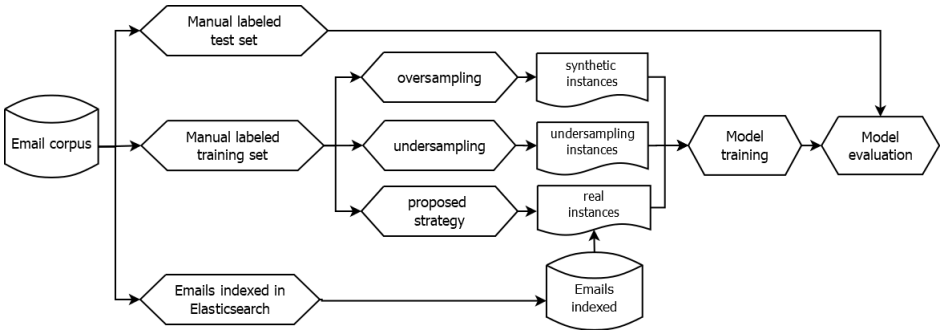


Fig. 1. Workflow proposed in this research

From an initial dataset, a subset of instances was selected and manually labeled by domain experts and two datasets were separated, one for model training and the other for evaluation. The oversampling and undersampling strategies were applied directly on the training set to consolidate the training data set.

For the new proposed strategy, both the manually labeled training set and the complete dataset were used, which was previously indexed in a general purpose search engine such as *Elasticsearch* for efficiency. In this case, the training set was further used to obtain the key terms representing each class of the problem and then retrieve unlabeled documents containing those terms to enrich the minority classes.

Once the datasets were consolidated with the class balancing strategies applied, classification models were trained and evaluated from the set reserved for this purpose. In the following sections, the most important issues related to the developed process are explained in greater detail.

3.1 Description of the dataset

For the experiments, were used a collection of 24700 e-mails generated from academic questions made by students to the administrative staff of the National University of Lujan. These questions are about procedures derived from the academic activity and the original e-mails were used without fixing any kind of typos or syntax errors. From those 24700 e-mails, 1000 were randomly selected and labeled around the question topic by a domain expert.

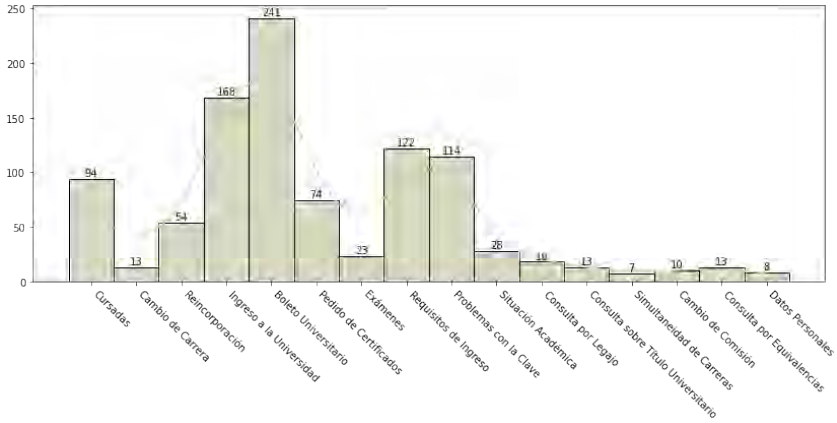


Fig. 2. Observed frequency for classes resulting from manual labeling

The 16 classes, all independent of each other, resulting from the labeling are: *Boleto Universitario, Cambio de Carrera, Cambio de Comisión, Consulta por Equivalencias, Consulta por Legajo, Consulta sobre Título Universitario, Cursadas, Datos Personales, Exámenes, Ingreso a la Universidad, Pedido de Certificados, Problemas con la Clave, Reincorporación, Requisitos de Ingreso, Simultaneidad de Carreras y Situación Académica.* The frequency distribution for each class is shown in Figure 2.

As it can be seen, the classes are highly imbalanced, an aspect that usually difficulties the classification process and that will be addressed with the process proposed in this paper.

3.2 Strategies used for the treatment of imbalanced datasets

The three strategies used for the treatment of the imbalanced data prior to the training of the automatic classification models are briefly presented below.

***Oversampling* Strategies.** The strategies implemented[3] were *RandomOverSampler*, *SMOTE*, *ADASYN* and *BorderSMOTE*. The first strategy, also known as ROSE (from the acronym for *random over sampling examples*), consists of generating new samples by random sampling with replacement of the current available samples and relies on a theoretical basis supported by the properties of kernel methods [24]. For its part, SMOTE (*Synthetic Minority Over-sampling Technique*) is one of the most recognized oversampling strategies, where, broadly speaking, the minority class is oversampled by introducing synthetic examples based on its k nearest neighbors depending on the amount of oversampling required [8]. In this sense, Borderline-SMOTE [17] is a variant of SMOTE that basically tries to determine the instances of the minority classes that are on the boundaries and generate synthetic instances from them. Finally, the essential idea of ADASYN (*Adaptive Synthetic Sampling*) [19] is to use a weighted distribution for the different examples of minority classes according to their level of learning difficulty, where more synthetic data is generated to examples of minority classes that are more difficult to learn.

***Undersampling* Strategies.** The strategies implemented were *RandomUnderSampler*, *ClusterCentroids* and *EditedNearestNeighbours*. The first strategy arbitrarily removes instances of the majority class in the training dataset [18] while in the case of the strategies based on *clustering* [22], a undersampling method based on replacing or removing instances by the centroids of the minority class instances is employed to reduce the number of majority class data samples.

On the other hand, the strategy *Edited Nearest Neighbours* [32] applies the nearest neighbor algorithm and "edits" the data set by removing samples that do not "sufficiently" match their neighborhood.

Proposed strategy. The strategy was initially presented as a semi-supervised classification strategy [16]. From an initial base with traditionally labeled mails, an extraction of the main features for each class is performed using different techniques, in this case TF-IDF and SS3 due to the results obtained in the previous work.

In the case of the TF-IDF technique [2], under this strategy, weighting per term grouped by class is used to determine which are the most important for each

[3] Implementations were performed with the **Imbalanced-learn** library for Python.

class. In the case of SS3 [4], it generates a function *gv(w, c)* that weights words relative to categories; to be more specific, *gv* takes a word w and a category c and generates a number in the interval [0,1] that represents the degree of confidence with which w belongs exclusively to c.

After retrieving the representative terms per class with both strategies, with the complete knowledge base indexed in a general purpose search engine such as *Elasticsearch*, documents from each class are retrieved based on the features detected by each technique and a new dataset is consolidated based on the instances that were retrieved by both feature selection strategies.

These instances are complemented by training dataset instances prior to the training of the classification model in order to balance it.

3.3 Generation of the Classification Models

As for classification techniques, support vector machines (SVM) were used because of their high performance for vectorized data, since vectorized data is generally required for the resampling strategies to be implemented.

SVM is a classical approach that has gained popularity over time due to some attractive features and its empirical performance. The main objective of support vector machines is to select the hyperplane which separates the training instances with a maximum distance criterion [27].

To evaluate the models, the remaining 200 manually labeled instances were reserved. Finally, the analysis of the selection of the generated models was performed based on the standard metrics *accuracy*, *precision* and *f1-score*.

4 Experiments

For the experiments[4], the training set with the 800 instances was used in all cases. Prior to training, queries were vectorized using 3-4 character grams and a TF-IDF weighting in all cases, and then class balancing strategies were applied. SVM was used in combination with a grid search alternating C (0.01, 0.1, 1), gamma (0.1, 0.01) parameters as well as kernels (rbf, linear, sigmoid), with and without class weighting.

It is important to clarify that in the case of the proposed strategy, 200 instances were retrieved for each class and feature selection technique from the database indexed in Elasticsearch, which resulted in a limitation because the number of instances resulting from the cross-linking between the instances retrieved by the two techniques meant that in some classes the amount of balance required for balancing was not reached, although the existing imbalance was reduced. This option was chosen over that of recovering a larger number of instances, with a lower coincidence *score*, in order not to introduce noise in the training set. To mitigate this situation, a variant of the proposed strategy is the definition of a smaller alternative N of instances, such as the average available per class, in order to reduce the distortion.

[4] Experiments available at `github.com/jumafernandez/imbalanced_data`

Next, classifiers were trained from the balanced datasets from the different strategies and the performance of the models was evaluated with the 200 instances reserved for this purpose, applying the accuracy, F1-score and precision metrics. The results are presented in Table 1.

Table 1. Experiments with class balancing techniques

Strategy	Accuracy	F1-Score	Precision
SVM (without class balancing)	0.810	0.80	0.82
RamdomOverSampler	0.810	0.80	0.81
SMOTE	0.805	0.79	0.81
ADASYN	0.810	0.80	0.81
BorderSMOTE	0.805	0.79	0.81
RamdomUnderSampler	0.660	0.68	0.73
ClusterCentroids	0.645	0.68	0.75
EditedNearestNeighbours	0.665	0.60	0.61
Proposed strategy	**0.820**	**0.83**	**0.85**
Proposed strategy (n = average = 115)	**0.820**	**0.83**	**0.85**

Based on the above experiments, it can be stated that none of the pre-existing techniques, either oversampling or undersampling, were able to improve the results obtained with the original dataset with the highly imbalanced classes. On the other hand, it is observed that the proposed strategy improved all the metrics in both variants equally.

In turn, another advantage of the proposed strategy, by incorporating non-synthetic instances to the training dataset, lies in the possibility of using it for new classification approaches based on neural networks, either those of deep learning as well as those based on transformers, a limitation that is observed in balancing strategies based on synthetic examples in general. The results of running the experiments in BERT (Bidirectional Encoder Representations from Transformers)[5] are transcribed below [12].

Table 2. Experiments with class balancing techniques with BERT

Strategy	Accuracy	F1-Score	Precision
BERT (without class balancing)	0.860	0.847	0.845
Proposed strategy	**0.865**	**0.865**	**0.878**
Proposed strategy (n = average = 115)	0.840	0.837	0.854

Table 2 shows that the proposed strategy is still effective but only for the conventional approach. In the case of the variant by the mean number of in-

[5] For model training, we experimented with a pre-trained model native to the Spanish language [6] and a set of hyperparameters successfully used in a previous work [15].

stances per class, the results are lower for the *accuracy* and *f1-score* metrics, between 1% and 2%, and higher in similar proportions for the *precision*.

5 Conclusions

This paper presents a novel strategy for learning from imbalanced data sets based on class oversampling by retrieving new unlabeled instances from a data repository of the same nature as the labeled data.

The fact that the instances for resampling come from real instances is presented as an advantage over strategies that generate synthetic samples. In principle, it may appear as a weakness to require an additional repository of data for experimentation. However, in full-scale problems of the real world it is normal to have a large -though unlabeled- data repository available.

Another advantage of the proposed strategy, by incorporating non-synthetic instances to the training data set, lies in the possibility of using it for new classification approaches based on neural networks, whether *deep learning* or *transformer-based*, a limitation that is observed in strategies based on synthetic examples in general.

Based on the results obtained, it can be concluded that this new strategy is competitive with respect to other resampling strategies widely used in the scientific community, either for traditional classification approaches, such as the one proposed for SVM, or for new transformer-based approaches, such as BERT.

Finally, although the present study has limited the experimentation to the domain of e-mail classification, we believe that the proposed strategy is generalizable to other domains where unlabeled text documents are available and, as future work, we propose to carry out further work applied to a more general text classification context.

References

1. Alghoul, A., Al Ajrami, S., Al Jarousha, G., Harb, G., Abu-Naser, S.S.: Email classification using artificial neural network. ACM (2018)
2. Bafna, P., Pramod, D., Vaidya, A.: Document clustering: Tf-idf approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). pp. 61–66. IEEE (2016)
3. Barua, S., Islam, M.M., Yao, X., Murase, K.: Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Transactions on knowledge and data engineering 26(2), 405–425 (2012)
4. Burdisso, S.G., Errecalde, M., Montes-y Gómez, M.: A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications 133, 182–197 (2019)
5. Cardenas, M.E., Castillo, J.J., Navarro, M., Hernández, N., Velazco, M.: Herramientas para el desarrollo de sistemas de análisis de textos no estructurados. In: XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan). (2019)

6. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
7. Chawla, N.V.: Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook pp. 875–886 (2009)
8. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357 (2002)
9. Chen, S., He, H., Garcia, E.A.: Ramoboost: Ranked minority oversampling in boosting. IEEE Transactions on Neural Networks 21(10), 1624–1642 (2010)
10. Chen, Z., Tao, R., Wu, X., Wei, Z., Luo, X.: Active learning for spam email classification. In: Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence. pp. 457–461 (2019)
11. Dalal, M.K., Zaveri, M.A.: Automatic text classification: a technical review. International Journal of Computer Applications 28(2), 37–40 (2011)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Fanny, F., Muliono, Y., Tanzil, F.: A comparison of text classification methods k-nn, naive bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT 3(2), 157–160 (2018)
14. Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F.: Learning from imbalanced data sets, vol. 10. Springer (2018)
15. Fernandez, J.M., Cavasin, N., Errecalde, M.: Classic and recent (neural) approaches to automatic text classification: a comparative study with e-mails in the spanish language. In: Short Papers of the 9th Conference on Cloud Computing, Big Data & Emerging Topics. p. 20 (2021)
16. Fernández, J.M., Errecalde, M.: Multi-class e-mail classification with a semi-supervised approach based on automatic feature selection and information retrieval. In: Conference on Cloud Computing, Big Data & Emerging Topics. pp. 75–90. Springer (2022)
17. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International conference on intelligent computing. pp. 878–887. Springer (2005)
18. Hanafy, M., Ming, R.: Improving imbalanced data classification in auto insurance by the data level approaches. International Journal of Advanced Computer Science and Applications 12(6) (2021)
19. He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). pp. 1322–1328. IEEE (2008)
20. Jung, H.J., Lease, M.: Evaluating classifiers without expert labels. arXiv preprint arXiv:1212.0960 (2012)
21. Lematre, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research 18(1), 559–563 (2017)
22. Lin, W.C., Tsai, C.F., Hu, Y.H., Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Information Sciences 409, 17–26 (2017)
23. Liu, B., Tsoumakas, G.: Dealing with class imbalance in classifier chains via random undersampling. Knowledge-Based Systems 192, 105292 (2020)
24. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery 28(1), 92–122 (2014)

25. Mohammed, R., Rawashdeh, J., Abdullah, M.: Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th international conference on information and communication systems (ICICS). pp. 243–248. IEEE (2020)
26. Shelke, M.S., Deshmukh, P.R., Shandilya, V.K.: A review on imbalanced data handling using undersampling and oversampling technique. Int. J. Recent Trends Eng. Res 3(4), 444–449 (2017)
27. Skiena, S.S.: The data science design manual. Springer (2017)
28. Tang, G., Pei, J., Luk, W.S.: Email mining: tasks, common techniques, and tools. Knowledge and Information Systems 41(1), 1–31 (2014)
29. Usai, A., Pironti, M., Mital, M., Mejri, C.A.: Knowledge discovery out of text data: a systematic review via text mining. Journal of knowledge management (2018)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
31. Wang, T., Lu, C., Ju, W., Liu, C.: Imbalanced heartbeat classification using easyensemble technique and global heartbeat information. Biomedical Signal Processing and Control 71, 103105 (2022)
32. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics (3), 408–421 (1972)

Un nuevo enfoque basado en perfiles con aprendizaje de representaciones

Dario G. Funez[1], Marcelo L. Errecalde[1],
Leticia C. Cagnina[1,2]

[1] LIDIC, Universidad Nacional de San Luis, San Luis, Argentina.
[2] Consejo Nacional de Investigaciones Científicas y Técnica (CONICET), Argentina.
{funezdario,merrecalde,lcagnina}@gmail.com

Resumen Los Enfoques Basados en Perfiles (EBP's) han mostrado muy buen comportamiento específicamente en la tarea de atribución de autoría. Este trabajo tiene como finalidad extender al EBP empleando aprendizaje de representaciones. Para ello, se utilizará la gran flexibilidad de los mecanismos de coincidencia (matching) que proveen los embeddings. La similitud entre perfiles, en este caso, ya no considerará únicamente aquellas palabras que coinciden "exactamente", sino aquellas que son lo "suficientemente similares", de acuerdo a un umbral predeterminado. Este trabajo comprende un estudio exhaustivo comparativo empleando las colecciones Enron y CIAPPA, donde quedará probada la viabilidad y efectividad de nuestra propuesta en relación a enfoques de EBP clásicos como SPI y KRD empleando escenarios con diferentes métodos de embeddings, tales como Word2Vec, Fastext y Glove.

Palabras claves: Enfoque Basados en Perfil, Aprendizaje de Representaciones, Atribución de Autoría, Perfilado de Autor, Embeddings

1. Introducción

Los Enfoques Basados en Perfiles (EBP's) han sido aplicados exitosamente para tratar problemas de atribución de autoría [Stamatatos, 2009]. En general, en este problema, se dispone de un conjunto de autores con sus respectivos documentos de autoría indiscutida y un texto de autoría desconocida es asignado a un único autor elegido de los candidatos.

Los EBP's construyen un perfil para cada autor, con información obtenida de los documentos redactados por el mismo [Escalante *et al.*, 2011]. Sus principales ventajas son la sencillez y la eficiencia en la categorización de textos y, en general, son muy competitivos respecto a los enfoques más tradicionales basados en instancias (EBI's). Es importante observar que, si bien en sus orígenes los perfiles se asociaban con los autores de un texto, los EBP's pueden ser utilizados en cualquier problema de categorización de textos donde, en lugar de autores, tenemos clases arbitrarias.

Un problema que presentan los EBP's es cuando el vocabulario del texto de prueba, no es un subconjunto del vocabulario del perfil del autor que se quiera

comparar. Es decir, existen palabras en el texto de prueba que no coinciden con la del perfil del autor. EBP en sus fórmulas de similitud/distancia emplea la cantidad de palabras que coinciden *exactamente*. En caso que se usen sinónimos o palabras similares, no se tiene en cuenta que se refieren al mismo concepto porque son palabras lexicográficamente diferentes. Este inconveniente es denominado como problema de la *no coincidencia* (o *mismatching*) en el área de recuperación de la información [Lizarralde *et al.*, 2017]. Este se presenta cuando se comparan unidades léxicas (palabras o términos) que son atómicas e indivisibles y son iguales cuando existe una coincidencia exacta entre sus componentes. Estas son disímiles en cualquier otro caso.

Algunos trabajos previos se han centrado en capturar relaciones semánticas entre palabras, a través de representaciones de palabras que contengan esa información [Mikolov *et al.*, 2013a,Bojanowski *et al.*, 2017,Devlin *et al.*, 2018]. En el área de aprendizaje de representaciones (*representation learning*) se define que a partir de datos sin ningún preprocesamiento especial, se los puede proyectar a un espacio de representaciones vectoriales con más semántica implícita. Este espacio de representaciones es conocido como *embeddings*, y se lo ha aplicado exitosamente en áreas muy diferentes del Procesamiento del Lenguaje Natural (PLN) como la traducción automática [Zou *et al.*, 2013], parsing sintáctico [Weiss *et al.*, 2015], clasificación de textos [Kim, 2014] y *question answering* [Bordes *et al.*, 2014], entre otras. El significado del *embedding* de una palabra (en inglés, *word embedding*) es una representación de una palabra por medio de un vector, cuyas componentes contienen relaciones sintácticas y semánticas [Almeida y Xexéo, 2019]. Por ejemplo, las palabras *rey* y *reina* en una representación clásica de palabras serían distintas, pero sus embeddings estarían cercanos en el espacio vectorial proyectado.

La propuesta de este trabajo surge a partir de la anomalía de los EBP's, en cuanto al cálculo de la similitud entre los perfiles. Los embeddings de palabras contienen relaciones implícitas entre los mismos que no son actualmente consideradas en los perfiles de palabras. Teniendo en cuenta esta falencia, se planteó la hipótesis de que con la incorporación de embeddings de palabras al EBP, mejoraríamos su desempeño en diversas tareas de análisis de autoría: atribución de autoría, perfilado de autor y categorización por tópicos. Para ello, se probó la hipótesis anterior en diferentes colecciones (dos de atribución de autoría, dos de perfilado de autor y una de categorización por tópicos) pero por cuestiones de espacio sólo se muestran en el estudio experimental dos de ellas. Se seleccionaron dos colecciones representativas de las tareas de atribución de autoría y perfilado de autor, que difieren en el idioma de los textos (inglés y español), poseen diferente cantidad de documentos (362 versus 196), distinto número y balance de las clases (80 versus 2, desbalanceada versus balanceada). Se utilizaron 3 enfoques de embeddings (estáticos) clásicos como Word2Vec, Fastext y Glove.

El resto de este trabajo está organizado como sigue: en la Sección 2 se describen brevemente las tareas que se abordarán en el estudio experimental; en la Sección 3 se describen los EBP's y en la Sección 4 se enumeran diferentes formas de aprendizaje de representaciones. Luego, en la Sección 5, se explica la propuesta

y en la Sección 6, se detalla la experimentación con el nuevo método en dos colecciones distintas. La Sección 7 finaliza con algunas conclusiones y trabajos futuros.

2. Tareas de Categorización de textos

En esta sección se introducirán brevemente las dos tareas de Análisis de Autoría evaluadas en este trabajo: Atribución de Autoría (AA) y Perfilado de Autor (PA).

En atribución de autoría, se dispone de un conjunto de autores con sus respectivos documentos de autoría indiscutida y un texto de autoría desconocida es asignado a un único autor elegido de los candidatos [Stamatatos, 2009]. Los componentes de AA son: el conjunto de autores candidatos, un conjunto de documentos (conjunto de entrenamiento) escrito por algún autor candidato (todos los autores deben tener algún documento) y un conjunto de documentos (conjunto de prueba) de autoría desconocida que deben ser correspondidos con los autores candidatos.

En la tarea de perfilado de autor se pretende descubrir tanto como sea posible sobre un autor desconocido, analizando sólo el texto escrito por él [Rangel, 2013]. Así, características de los autores como por ejemplo género, edad, personalidad u orientación política, pueden ser descubiertas y, mediante el PA, clasificar textos de un autor desconocido según su perfil. Al igual que en AA, en PA se cuenta con un conjunto de textos de entrenamiento de varios autores que pertenecen a una clase particular que los caracteriza (el perfil) y se pretende clasificar documentos de prueba en base a los perfiles modelados.

3. Enfoques Basados en Perfiles

Los EBP's han sidos aplicados exitosamente para resolver problemas de AA [Stamatatos, 2009], siendo en diferentes momentos el estado del arte de esta tarea. El EBP consiste en construir un perfil para cada autor, con información obtenida de los documentos redactados por el mismo [Escalante *et al.*, 2011]. Para elegir el autor sobre los K autores candidatos, se utiliza la similitud o distancia entre el perfil del texto de prueba y el de cada perfil de los autores, y se elige el que resulte con mayor similitud o menor distancia. Si bien en los orígenes de los EBP's los perfiles se asociaban con los autores de un texto, también se pueden utilizar en cualquier problema de categorización de textos donde, en lugar de autores, tenemos clases arbitrarias y los perfiles se generan con los documentos de cada clase.

Para la generación de los perfiles de autor, se extraen de los documentos del autor un conjunto de L características. Para obtener el perfil de un autor, empleando por ejemplo la característica 3-gramas de caracteres, se recuperan todos los 3-gramas de todos los documentos del autor, y luego se los ordena en forma creciente por la cantidad de ocurrencias. Los n-gramas son subcadenas de n componentes consecutivos, estos pueden ser caracteres o palabras

[Cavnar y Trenkle, 1994]. El valor L es un parámetro del EBP y solamente se utilizan, para el caso del ejemplo, los $L's$ 3-gramas más frecuentes del perfil.

En el proceso de clasificación, lo primero que se debe realizar es la obtención del perfil del documento de prueba de autoría desconocida. Luego, mediante el uso de alguna medida de distancia o similitud, se debe comparar el perfil del documento de prueba con cada perfil de los K autores candidatos [Funez *et al.*, 2013]. La respuesta del clasificador es elegir el autor cuyo perfil es el más parecido al perfil del documento de prueba. Una componente principal en los EBP's es la de determinar la similitud/distancia entre los perfiles. La mayoría de los trabajos que han propuesto mejoras a los EBP's se centran en definir medidas de similitud más complejas y eficientes. A continuación se define la terminología que se empleará después en la definición de las medidas:

a) Se asume un escenario de K autores (clases) candidatos, con P_1 .. P_K perfiles correspondientes a los K autores (clases).
b) T_j denota el j-ésimo perfil de test.
c) La notación I_j^i significa el conjunto de términos que aparecen en la intersección de los perfiles P_i y T_j.
d) $f_X(n)$ denota la frecuencia de la característica n en el perfil X.
e) La medida denotada con S significa que es una medida de similitud y la expresada con D de distancia.

Las siguientes medidas son las más aplicadas en los EBP's:

a) Distancia Relativa de Keselj (Keselj's Relative Distance): Es un medida de distancia, también conocida como N-Gramas Comunes (CNG por sus siglas en inglés), en la cual la comparación de perfiles se realiza con una frecuencia normalizada de términos, como lo expresa la Ecuación 1 y se calcula respecto a los perfiles P_i y T_j [Keselj *et al.*, 2003]:

$$D_{krd}(P_i, T_j) = \sum_{n \in P_i \cup T_j} \left(\frac{2 \cdot (f_{P_i}(n) - f_{T_j}(n))}{f_{P_i}(n) + f_{T_j}(n)} \right)^2 \quad (1)$$

b) Similitud por Intersección de Perfiles Simplificada (Simplified Profile Intersection): Es una medida de similitud dada por la Ecuación 2 que calcula la cantidad de características que son comunes a ambos perfiles, sin aplicar ninguna normalización [Frantzeskou *et al.*, 2007]. En AA ha tenido mejor desempeño SPI con respecto a KRD.

$$S_{spi}(P_i, T_j) = \mid I_j^i \mid \quad (2)$$

4. Aprendizaje de representaciones

En los últimos años, en el PLN se ha investigado cómo capturar las relaciones semánticas entre las palabras y/o frases a través de representaciones más avanzadas [Almeida y Xexéo, 2019]. En el área del aprendizaje de representaciones se proyectan datos *crudos* a espacios de representación con mayor

semántica implícita. Este espacio proyectado se conoce como *embeddings* y son vectores densos de longitud fija que se obtienen usualmente mediante dos enfoques generales [Baroni *et al.*, 2014]: basados en *predicción* y basados en *conteo*. En los primeros, se aprende a predecir la probabilidad de ocurrencia del contexto de una palabra, usualmente mediante un enfoque de red neuronal (por ejemplo Word2Vec). Los embeddings (vectores) se derivan de los pesos de la red neuronal aprendidos en la resolución de esta tarea. Los modelos basados en conteo, en cambio, emplean información global recolectando estadísticas de la colección y el conteo de co-ocurrencias de palabras (por ejemplo Glove).
Otra diferencia importante de los embeddings es si éstos son *estáticos* (por ejemplo, Word2Vec, Fastext o Glove) o *contextuales* (por ejemplo, Bert). En los primeros se aprende un embedding fijo (único) por cada palabra/término en el vocabulario de embeddings. En los segundos, se aprenden embeddings contextuales dinámicos como la popular familia de representaciones BERT, en las que el vector para cada palabra es diferente en diferentes contextos. En las siguientes subsecciones se explican los modelos (de embeddings estáticos) Word2vec, Fastext y Glove que se utilizarán en la experimentación.

4.1. Word2Vec

Word2Vec entrena una red neuronal con textos y esta permite obtener a partir de una colección de documentos, los embeddings para cada palabra de la colección [Mikolov *et al.*, 2013b]. La intuición de Word2Vec es que entrenaremos a un clasificador en una tarea de predicción binaria: "¿Es probable que la palabra w aparezca cerca de una palabra objetivo?" En realidad, no nos importa esta tarea de predicción; en su lugar, tomaremos los pesos del clasificador aprendido como embeddings de palabras. El aspecto interesante, es que el texto bajo consideración actúa como datos de entrenamiento supervisados implícitamente para dicho clasificador (auto-supervisión), evitando la necesidad de cualquier tipo de señal de supervisión etiquetada a mano.
El modelo Word2vec puede usar uno de los siguientes tipos de arquitectura para el aprendizaje de los embeddings: CBOW o SG. En las dos arquitecturas se emplean redes neuronales con una única capa oculta, y en el peor de los casos tienen una complejidad de entrenamiento logarítmica lineal. La red neuronal utiliza el algoritmo *Retropropagación* (Backpropagation) para aprender los pesos de la capa oculta que darán origen a los embeddings de las palabras. Estos embeddings, pueden servir para derivar una representación de los documentos mediante alguna forma de agregación (usualmente el promedio) o como entrada para otras arquitecturas de redes neuronales más complejas (redes neuronales recurrentes, LSTM, etc).

4.2. Fastext

Este modelo de representación de palabras es una mejora de Word2Vec que toma en cuenta su *morfología* [Bojanowski *et al.*, 2017]. Se consideran como

unidades a las subpalabras y se representan las palabras por la suma de sus n-gramas de caracteres. En Fastext cada palabra w es representada como una bolsa de n-grams de caracteres, y se le agregan a cada palabra al principio el caracter $<$ y al final $>$, para distinguir los prefijos y sufijos de las demás secuencias de caracteres. La representación de una palabra es la suma de las representación vectorial de sus n-gramas.

4.3. Glove

Glove [Pennington *et al.*, 2014] es un algoritmo de aprendizaje no supervisado de embeddings de palabras cuyo entrenamiento se realiza sobre estadísticas globales de co-ocurrencia palabra a palabra recolectando estadísticas de la colección. Los embeddings obtenidos tienen subestructuras lineales importantes del espacio vectorial de palabras.
El modelo Glove se basa en que las diferencias vectoriales de los embeddings de las palabras, capturan lo mejor posible el significado de juntar ambas palabras. El entrenamiento del modelo Glove se lleva a cabo sobre un matriz global de co-ocurrencia palabra-palabra. La meta del entrenamiento de Glove es aprender los embeddings de las palabras, ajustando el producto punto con el logaritmo de la probabilidad de co-ocurrencia de las palabras. De esta manera, se produce la asociación del logaritmo de cocientes de probabilidades de co-ocurrencia, con la diferencia de vectores en el espacio vectorial de palabras.

5. Perfilado de autor con embeddings

Los embeddings de las palabras contienen información sobre relaciones semánticas entre las mismas que permiten identificar palabras "similares". En este trabajo se propone probar el efecto de modificar los perfiles para que, en lugar de contener las palabras (atómicas/indivisibles), se las sustituya por los embeddings (vectores) de estas palabras. Esto permite, mediante métodos como la similitud coseno entre vectores, reconocer aquellas palabras que se asemejan "lo suficiente" a una palabra específica. Para ello, se define un parámetro de umbral th, que indica el mínimo nivel de similitud que dos palabras deben tener para ser consideradas "semejantes" o "similares". Es claro, que cuando $th = 1$, sólo consideraremos como similares las palabras iguales.
Dado que los perfiles de palabras contienen las entradas de palabras con su frecuencia asociada, usaremos en los perfiles con embeddings una versión "ponderada" de la SPI que, al igual que la distancia KRD toma en cuenta la frecuencia de ocurrencia de las mismas. Por otra parte, ya no se considerará únicamente aquellas palabras que coinciden "exactamente" entre los perfiles, sino aquellas que son lo suficientemente similares, de acuerdo al umbral th.
A continuación se da el pseudo-código de esta implementación:

```
funcion sim_perf_embed(perfil_emb p1,p2)
acum = 0
Para cada e1 de p1
  Para cada e2 de p2
     sim = coseno(e1,e2)
     Si sim >= th
        acum += sim * frec(e2)
retornar acum
```

Como se puede observar, si no se tomara la frecuencia de las palabras y $th = 1$, sería la SPI ya descripta previamente. Esta nueva función de similitud entre perfiles, que denominamos *sim_perf_embed* recibe como parámetros dos perfiles de embeddings, el del documento de prueba (e1) y el de la clase (e2). Para cada par de embeddings de ambos perfiles, se computa su similitud coseno y, si este valor supera el umbral th, se incrementa la variable *acum*, de acuerdo a su similitud y su frecuencia en la clase. Finalmente el valor final *acum* es retornado por la función de similitud.

6. Evaluación experimental

En esta sección se describen las colecciones utilizadas en la experimentación, los resultados obtenidos y un análisis de los mismos.

6.1. Corpus Enron

Este corpus (*https://data.mendeley.com/datasets/n77w7mygwg/2*) es obtenido a partir del *Enron Email Dataset* [Halvani y Graner, 2018]. Es muy utilizado en investigaciones de la tarea de identificación de autoría. Está compuesto de 80 autores con emails que son textos planos formales escritos en inglés. Cada documento ha sido escrito por un único autor, y es una compilación de los más recientes emails producidos por cada uno de ellos. El conjunto de *entrenamiento* lo comprenden 282 documentos (80 autores, cada uno posee entre 2 y 4 documentos cada uno). El conjunto de *prueba* lo conforman 80 autores con un documento cada uno (80 documentos en total). En la tabla 1 se muestran los resultados de medida F para distintos L's con las métricas básicas SPI y KRD, y para el caso de perfiles con embeddings se usó la implementación de la función *sim_perf_embed* descripta previamente. De esta surgieron tres variantes: sim_perf_W que emplea Word2vec[1] para obtener los embeddings del corpus, sim_perf_F que usa Fastext[2] y sim_perf_G[3] que utiliza Glove. Para sim_perf_W se le eligieron los siguientes parámetros: tamaño de embedding de 12 y th = 0,99. sim_perf_F usó

[1] https://radimrehurek.com/gensim/models/word2vec.htmlmodule-gensim.models.word2vec
[2] https://radimrehurek.com/gensim/models/fasttext.htmlgensim.models.fasttext.FastText
[3] https://nlp.stanford.edu/projects/glove/

Enfoque perfiles			Enfoque perfiles con embeddings		
			Del Corpus		
L	*KRD*	*SPI*	*sim_perf_W*	*sim_perf_F*	*sim_perf_G*
50	0,1141	0,1154	0,1429	0,1419	0,1677
100	0,2752	0,2669	0,277	0,289	0,2702
200	0,3775	0,3642	0,3489	0,3654	0,3983
230	0,3749	0,3583	0,3748	0,3445	0,3935
250	0,3398	0,3435	0,3373	0,3648	0,3881
280	0,3895	0,402	0,4016	0,3862	0,4175
300	0,3714	0,3735	0,4283	0,3808	0,4659
330	0,4227	0,4235	0,4498	0,4252	0,4625
350	0,3873	0,3787	0,4148	0,4458	0,473
400	0,4095	0,4076	0,4519	**0,4713**	0,4402
500	0,4421	0,4402	0,4787	0,3876	0,4885
700	**0,5335**	**0,5335**	**0,5848**	0,4419	**0,5737**
800	0,4858	0,4691	0,5306	0,3983	0,5239
900	0,4658	0,4617	0,5047	0,3296	0,5181
1000	0,4817	0,4817	0,4863	0,2795	0,4697
1100	0,4146	0,4146	0,4713	0,2645	0,4727
1200	0,4084	0,4084	0,4879	0,2834	0,4997
1500	0,4265	0,4149	0,5004	0,2804	0,4997
1800	0,4265	0,4149	0,5004	0,2804	0,4997
2000	0,4265	0,4149	0,5004	0,2804	0,4997
2300	0,4265	0,4149	0,5004	0,2804	0,4997
2500	0,4265	0,4149	0,5004	0,2804	0,4997
3000	0,4265	0,4149	0,5004	0,2804	0,4997

Tabla 1. Medida F para el problema identificación de autoría Colección Enron.

un tamaño de embedding de 30 y th = 0,99. sim_perf_G usó un tamaño de embedding de 5, th = 0,9995, velocidad de aprendizaje = 0,05 y 30 épocas. Los parámetros fueron seleccionados luego de realizar una búsqueda exhaustiva de valores y se eligieron los que mejor comportamiento alcanzaron en la experimentación.

En la tabla 1 se encuentran resaltados los mejores resultados para los distintos enfoques probados y se puede observar que sim_perf_W alcanza la mejor medida F para $L = 700$ con $0,5848$ y los enfoques SPI y KRD obtuvieron su mejor medida F para $L = 700$ con $0,5335$. Así, sim_perf_W supera en un 9 % a SPI y kRD. Por otra parte, se puede observar que KRD no supera a SPI para ningún L. Respecto a los valores generales con los distintos L, sim_perf_W supera en un 86 % de los casos tanto a SPI como a KRD. También se puede apreciar que para L's mayores a 300 sim_perf_W supera a SPI en todos los casos. En particular, luego de $L = 1500$ no se tienen mejoras para los casos de las tres variantes. Por último, es claro que sim_perf_F alcanzó el peor resultado de las tres variantes con embeddings y sim_perf_G obtuvo una diferencia levemente inferior, considerando el mejor valor de la medida F, con respecto a sim_perf_W.

6.2. Corpus para la Identificación de la Alineación Política de Periodistas Argentinos (CIAPPA)

Este corpus está compuesto de 196 documentos pertenecientes a 10 periodistas, 5 de estos explícitamente apoyan las acciones del gobierno argentino que gobernó (entre los años 2012 y 2015), y los otros 5 son opositores al gobierno en ese periodo [Mercado *et al.*, 2019]. El corpus se divide en dos

grupos de documentos de acuerdo a la orientación política de los periodistas. De esta manera, 98 documentos pertenecen a la clase oficialista y 98 a la clase opositora, así el corpus es balanceado en sus dos clases. En la experimentación se planteó el problema de perfilado de autor para identificar la clase de un documento, y así determinar si el autor es oficialista u opositor al gobierno. Para el corpus CIAPPA no se encuentran disponibles los conjuntos de *entrenamiento* y *prueba*, estos se obtuvieron de forma aleatoria del corpus original. El conjunto de *entrenamiento* quedó compuesto por 84 documentos oficialistas y 84 opositores quedando balanceado en las dos clases, mientras que el conjunto de *prueba* lo comprenden 14 oficialistas y 14 opositores.

En la tabla 2 se muestran los resultados de medida F para los modelos basados en perfiles considerando los enfoques SPI y KRD y perfiles con embeddings. Para sim_perf_W y sim_perf_F se utilizaron los siguientes parámetros: el tamaño del embedding de 100 y th = 0,85. Para el caso de sim_perf_G se usó un tamaño de embedding de 5, th = 0,99, velocidad de aprendizaje: 0,05 y 30 épocas. Los mejores valores se muestran resaltados en la tabla 2 y se puede observar que sim_perf_W alcanza la mejor medida F para $L = 7000$ con $0,9282$ superando a SPI y KRD que obtuvieron $0,8916$.

Enfoque perfiles			Enfoque perfiles con embeddings		
			Del Corpus		
L	*KRD*	*SPI*	*sim_perf_W*	*sim_perf_F*	*sim_perf_G*
10	0,3333	0,3333	0,317	0,3333	0,3333
20	0,3333	0,3333	0,5618	0,3333	0,3333
50	0,4285	0,392	0,3437	0,4466	0,3858
70	0,747	0,747	0,5351	0,7857	0,8212
100	0,7083	0,747	0,6428	0,7417	0,6256
120	0,7846	0,7846	0,5692	0,7857	0,6679
150	0,747	0,7128	0,6066	0,7005	0,7005
180	0,747	0,747	0,63541	0,6111	0,6679
200	0,7846	0,78461	0,6781	0,6888	0,747
400	0,6781	0,6781	0,7496	0,6111	0,733
500	0,7083	0,7083	0,74968	0,7005	0,7812
600	0,8212	0,8212	0,7142	0,7005	0,8193
700	0,7857	0,7857	0,7857	0,7417	0,8212
750	0,8212	0,7857	0,8212	0,8212	0,8212
1000	0,8564	0,8564	0,8212	0,7812	0,8564
2000	0,747	0,747	0,8564	0,7857	0,8193
2300	0,8564	0,8564	0,8564	0,7496	0,8564
2500	0,8212	0,8212	0,8916	0,641	0,8564
2800	0,8564	0,8193	0,8564	0,7496	0,8564
3000	0,8564	0,8564	0,8564	0,7496	0,8564
3500	0,8564	0,8564	0,8564	0,6428	0,8564
4000	0,8193	0,8193	0,8541	0,641	0,8541
4500	0,8564	0,8564	0,8916	0,7128	0,8916
5000	0,8564	0,8564	0,8916	0,74176	0,8541
6000	0,7754	0,8155	0,8155	0,7417	0,8155
7000	**0,8916**	**0,8916**	**0,9282**	0,7812	**0,8916**
8000	0,8541	0,85416	0,9282	0,747	0,8541
9000	0,8193	0,8193	0,9282	0,747	0,8916
10000	0,8193	0,8193	0,89161	0,7417	0,8541

Tabla 2. Medida F corpus CIAPPA problema de perfilado de autor.

7. Conclusiones y Futuras Extensiones

Los EBP's han mostrado su eficiencia y buen comportamiento, particularmente en tareas de atribución de autoría. No obstante esto, se basan en criterios de comparación exacta entre palabras que podrían ser mejorados con los nuevos enfoques de aprendizaje de representaciones (embeddings).
Este trabajo, da una propuesta de cómo llevar a cabo esta tarea, proponiendo una función de similitud entre perfiles que toma en cuenta la similitud de sus embeddings constituyentes. Hasta donde sabemos, este es el primer trabajo en el área de los EBPs donde se propone una extensión con estas características.
Nuestra propuesta, obtiene mejores resultados que enfoques clásicos de EBP como SPI y KRD en las colecciones Enron y CIAPPA que difieren en el tipo de tarea, lenguaje y número y nivel de balance de sus clases. En ambos casos, se probaron embeddings estáticos clásicos como Word2Vec, Fastext y Glove obteniendo con los embeddings Word2Vec (variante Sim_perf_W) el mejor desempeño. Si bien por razones de espacio, sólo se reportan los resultados de estas dos colecciones, los resultados obtenidos con otras colecciones clásicas (como 20NewsGroup) han mostrado la viabilidad y efectividad de nuestra propuesta.
Como trabajo futuro se propone realizar un estudio experimental más exhaustivo con diferentes formas de ponderar la frecuencia de las palabras en los perfiles, combinación de perfiles con embeddings y perfiles con n-gramas de caracteres y adaptación de embeddings contextualizados (tipo BERT) al esquema de trabajo propuesto.

Referencias

[Almeida y Xexéo, 2019] Almeida, F. y Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

[Baroni *et al.*, 2014] Baroni, M., Dinu, G., y Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. *ACL*, pp. 238–247.

[Bojanowski *et al.*, 2017] Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[Bordes *et al.*, 2014] Bordes, A., Chopra, S., y Weston, J. (2014). Question answering with subgraph embeddings. *CoRR*, abs/1406.3676.

[Cavnar y Trenkle, 1994] Cavnar, W. B. y Trenkle, J. M. (1994). N-gram-based text categorization. En *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pp. 161–175.

[Devlin *et al.*, 2018] Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Escalante *et al.*, 2011] Escalante, H. J., y Gómez, M. M., y Solorio, T. (2011). A weighted profile intersection measure for profile-based authorship attribution. En *Proceedings of MICAI 2011*, volumen 7094, pp. 232–243.

[Frantzeskou *et al.*, 2007] Frantzeskou, G., Stamatatos, E., Gritzalis, S., Chaski, C. E., y Howald, B. S. (2007). Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *International Journal of Digital Evidence*, 6(1):1–18.

[Funez *et al.*, 2013] Funez, D. G., Cagnina, L., y Errecalde, M. L. (2013). Determinación de género y edad en blogs en español mediante enfoques basados en perfil. En *XVIII Congreso Argentino de Ciencias de la Computación*.

[Halvani y Graner, 2018] Halvani, O. y Graner, L. (2018). Rethinking the evaluation methodology of authorship verification methods. En *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 40–51. Springer.

[Keselj *et al.*, 2003] Keselj, V., Peng, F., Cercone, N., y Thomas, C. (2003). N-gram-based author profiles for authorship attribution. En *Proceedings of the Pacific Association for ComputationalLinguistics*, pp. 255–264.

[Kim, 2014] Kim, Y. (2014). Convolutional neural networks for sentence classification.

[Lizarralde *et al.*, 2017] Lizarralde, I., Rodriguez, J. M., Mateos, C., y Zunino, A. (2017). Word embeddings for improving rest services discoverability. En Monteverde, H. y Santos, R., editores, *CLEI*, pp. 1–8. IEEE.

[Mercado *et al.*, 2019] Mercado, V., Villagra, A., y Errecalde, M. L. (2019). Exploratory analysis of a new corpus for political alignment identification of argentinian journalists. En *XXV Congreso Argentino de Ciencias de la Computación (CACIC)(Universidad Nacional de Río Cuarto, Córdoba, 14 al 18 de octubre de 2019)*.

[Mikolov *et al.*, 2013a] Mikolov, T., Chen, K., Corrado, G., y Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Mikolov *et al.*, 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., y Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

[Pennington *et al.*, 2014] Pennington, J., Socher, R., y Manning, C. D. (2014). Glove: Global vectors for word representation. En *EMNLP*, volumen 14, pp. 1532–1543.

[Rangel, 2013] Rangel, F. (2013). Identifying information about gender, age, emotions and beyond. En *Proceedings of the 5th BCS IRSG Symposium on Future Directions in Information Access*.

[Stamatatos, 2009] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*, 60(3):538–556.

[Weiss *et al.*, 2015] Weiss, D., Alberti, C., Collins, M., y Petrov, S. (2015). Structured training for neural network transition-based parsing. *CoRR*, abs/1506.06158.

[Zou *et al.*, 2013] Zou, W. Y., Socher, R., Cer, D., y Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. En *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1393–1398.

Análisis de deuda técnica de UX en repositorios de GitHub

Ana Liz Lubomirsky, Juan Cruz Gardey, and Alejandra Garrido

LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina
ana.lubo99@gmail.com
{jcgardey,garrido}@lifia.info.unlp.edu.ar

Resumen El concepto de deuda técnica se utiliza para denotar problemas de calidad del software que a medida que transcurre el tiempo se hacen progresivamente más dificiles de reparar. Este trabajo consiste en investigar la presencia de deuda técnica relacionada a la experiencia de usuario (UX) en las incidencias (issues) de 8 proyectos de GitHub, con el fin de fundamentar la necesidad de administrar esta deuda de UX, tal como ocurre con otros tipos de deuda técnica. Mediante diferentes análisis, se caracterizaron los issues relacionados a la UX y se identificaron aquellos que pueden contribuir a la acumulación de deuda de UX, para comparar su índice de resolución frente con otros tipos de issues, y así determinar cuanta importancia se le da al tratamiento de la deuda de UX. Los resultados muestran que los issues que generan deuda de UX con frecuencia no se resuelven, o su atención se posterga para priorizar otro tipo de issues.

Keywords: deuda técnica; experiencia de usuario; refactoring

1. Introducción

La deuda técnica es una metáfora propuesta por Ward Cunningham que describe las consecuencias de las acciones de desarrollo de software que, de forma intencionada o no, priorizan el desarrollo de la funcionalidad requerida por el cliente por encima de consideraciones de diseño e implementación y sus parámetros de calidad [1]. Conceptualmente, la deuda técnica es un análogo de la deuda financiera; la acumulación de deuda genera un interés que hace que su costo de remediación sea cada vez mayor.

Originalmente la métafora de deuda técnica fue pensada para describir cuestiones de implementación a nivel de código fuente, pero luego surgieron otros tipos de deuda como por ejemplo de arquitectura, de requerimientos, de documentación, etc [8]. Dentro de estos tipos de deuda, aparece la deuda técnica de diseño de experiencia de usuario (UX debt), a la que Kuusinen define como "diseño no del todo correcto que posponemos hacerlo bien" [6]. Si bien existen trabajos que reconocen a la UX debt como un tipo de deuda técnica, ninguno de estos provee una caracterización que permita identificarla y mantenerla bajo control.

Volviendo a la metáfora original, la presencia de *code smells* es uno de los indicadores de deuda técnica [8]. Los code smells son un conjunto de malas prácticas de codificación que pueden ser resueltos a través de *refactorings*. Los refactorings son transformaciones de código que tienen por objetivo mejorar la calidad del mismo sin modificar la funcionalidad de la aplicación. Siguiendo esta misma filosofía, en trabajos previos se ha propuesto el concepto de *UX smells* como indicadores de una mala experiencia de usuario [5]. Estos UX smells no tienen que ver con aspectos funcionales del sistema, sino con cuestiones de la interacción del usuario que pueden mejorarse a través de uno o más UX refactorings. Así como los code smells generan deuda técnica, los UX smells contribuyen a la acumulación de UX debt.

Este trabajo consiste en investigar la presencia de UX debt (en términos de UX smells) en interfaces web de proyectos en GitHub, con el objetivo de recopilar evidencia acerca del nivel de atención que reciben los UX smells. Para esto se propone analizar los *issues* de un conjuntos de proyectos, que básicamente son incidencias reportadas por los mismos colaborados del proyecto o por usuarios externos. Un issue puede ser entre otras cosas un error a corregir, una petición para añadir una nueva funcionalidad o una pregunta para aclarar alguna cuestión puntual.

A partir del análisis de los issues, se propone realizar una clasificación automática de éstos en la que primero se identifican aquellos que reportan aspectos relacionados a la experiencia de usuario y luego dentro de este grupo, se identifican los que cumplen con la noción de UX smell. Es decir, que no reflejan cuestiones funcionales del sistema (como bugs o feature requests), sino aspectos de la interfaz de usuario (UI) que pueden modificarse para mejorar la interacción. Posteriormente, se analiza la resolución de los UX smells para determinar la existencia de UX debt en los diferentes proyectos.

El artículo está organizado de la siguiente manera. La sección 2 presenta el trabajo relacionado a la deuda técnica, UX debt, y al análisis de issues en sistemas de gestión de incidencias. La sección 3 explica la clasificación de issues realizada y el posterior análisis de la resolución de los mismos. Finalmente la sección 4 describe las conclusiones y trabajos futuros.

2. Trabajo Relacionado

No existen muchos trabajos que mencionen a la UX como un tipo específico de deuda técnica. Algunos de estos utilizan una definición de "usability debt" que en general concuerda con la idea de posponer el diseño de UX, pero no proveen una definición práctica que permita identificarla [4,6]. En este trabajo se utiliza el concepto de UX smell como elementos que contribuyen a la acumulación de UX debt, así como los code smells son un componente fundamental de la deuda técnica.

Estudios anteriores han utilizado sistemas de gestión de issues para analizar la deuda técnica [7] basándose en la premisa de que los desarrolladores comúnmente documentan esta deuda a través de comentarios del código fuente, como han

observado Potdar y Shihab, en su estudio a través del análisis de más de 100K comentarios de código [9].

Además, otros estudios proponen el uso de técnicas de procesamiento de lenguaje natural para apoyar la identificación de deuda técnica admitida, que es un caso particular de deuda técnica donde los desarrolladores reconocen explícitamente sus decisiones de implementación subóptimas [2]. Maldonado et al. [3] propusieron una técnica para identificar con precisión este tipo de deuda, basada en palabras clave y frases fijas.

En este trabajo se analizan los issues con el propósito de profundizar el estudio de la presencia de UX debt, que puede ser introducida de forma intencional o no. Se realiza una clasificación diferenciando los issues genéricos de los issues de UX, y de estos se separan los issues de UX que se pueden resolver con refactorings de UX.

3. Análisis de issues en GitHub

3.1. Selección de repositorios

El primer paso del proceso de análisis fue la búsqueda de repositorios públicos que sean útiles para la búsqueda a realizar. En particular, se seleccionaron proyectos de aplicaciones web que hagan un uso intensivo de una interfaz de usuario. Adicionalmente, se establecieron tres criterios más:

- Actividad. Que sean repositorios con actividad reciente, es decir, que no sean repositorios archivados u obsoletos.
- Cantidad de issues. Se buscaron repositorios con al menos 2000 issues para que el análisis sea representativo.
- Popularidad. Se buscaron repositorios que sean populares para asegurarse que tengan cierto movimiento de issues. Como medida de popularidad se utilizó el ranking de estrellas (stars) que es la manera en que los usuarios pueden valorar los repositorios.

Para la búsqueda de los proyectos se utilizó el buscador de GitHub. Con el término "web application" se listaron todos los proyectos relacionados con aplicaciones web, y luego de ordenarlos por popularidad en forma descendente, se analizó el nombre y la descripción de cada uno para determinar si cumplía con los criterios anteriores. La tabla 1 muestra los datos de los 8 proyectos seleccionados. Se decidió limitar la cantidad de proyectos elegidos a 8 porque el análisis de cada uno requiere un tiempo considerable.

3.2. Recuperación de issues

Luego de seleccionar los repositorios a analizar, se descargaron los issues de cada uno utilizando la API de GitHub y se almacenaron en un archivo csv (valores separados por comas) para su posterior procesamiento. De todos los datos que proporciona la API, se seleccionaron los que se listan a continuación:

Tabla 1. Datos de los repositorios elegidos

Proyecto	Nombre en GitHub	#Issues	Cerrados
Jitsi Meet	jitsi/jitsi-meet	4705	66 %
Matomo	matomo-org/matomo	11823	77 %
ERPNext	frappe/erpnext	12282	75 %
KeystoneJS	keystonejs/keystone-classic	2458	88 %
Visual Studio Code	microsoft/vscode	81381	95 %
Wallabag	wallabag/wallabag	3027	82 %
Open MCT	nasa/openmct	2736	73 %
Superset	apache/superset	8059	88 %

- **Título (title):** el título del issue.
- **Descripción (body):** el contenido del issue.
- **Etiquetas (labels):** etiquetas usadas para clasificar y categorizar el issue, esto ayuda a tener una idea rápida del tipo de issue, y también poder filtrar por dichas clasificaciones.
- **Estado (state):** estado del issue, puede ser abierto o cerrado (open/closed).
- **Fecha de creación (created_at):** fecha en que se creó el issue.
- **Fecha de cierre (closed_at):** fecha en que se cerró el issue (si es que está cerrado).
- **Cantidad de comentarios (comments):** comentarios realizados por los colaboradores y usuarios del proyecto.

Además, se agregaron los siguientes atributos calculados a partir de los datos de cada issue:

- **Solucionado (verdadero/falso):** en principio se consideró resuelto a todo issue cuyo estado era cerrado. Luego analizando los issues, se identificó que en muchos casos estos se cierran asignándoles la etiqueta "wontfix", la cual se utiliza para indicar que el issue no será resuelto. Por ese motivo, se agregó este atributo cuyo valor es verdadero únicamente cuando el estado del issue es cerrado y no posee la etiqueta wontfix.
- **Tiempo de resolución:** es la diferencia en días entre la fecha de cierre del issue y la fecha de apertura.

3.3. Clasificación de issues de UX

Teniendo en cuenta que cada proyecto contiene diferentes tipos de issues, para identificar los UX smells primero fue necesario categorizar aquellos issues relacionados con la experiencia de usuario. Esta clasificación se realizó de forma automática mediante la búsqueda de ciertas palabras clave (ver apéndice[1]). Este listado de palabras clave se confeccionó inspeccionando manualmente un

[1] https://bit.ly/3SkvQqA

conjunto aleatorio de issues de cada proyecto analizado que los mismos usuarios etiquetaron como "UX".

Para cada issue, se agregó un atributo **UX** (verdadero/falso) cuyo valor es verdadero si al menos 3 de las palabras clave están contenidas en el issue. Para esta búsqueda se realizó el proceso de tokenización del issue, el cual consiste en dividir el texto en las palabras que lo conforman, considerando el título, contenido y los labels.

Es importante mencionar que este proceso de búsqueda fue iterativo. En cada iteración, se realizó la clasificación con las palabras seleccionadas hasta el momento y luego se inspeccionó un conjunto aleatorio de 20 issues de cada proyecto para refinar el listado de palabras. Esto se repitió en todos los proyectos hasta alcanzar al menos una precisión del 80 % en el conjunto de issues inspeccionado. En cuanto a la cantidad de palabras clave que un issue debía contener, se realizaron pruebas con diferentes umbrales (1, 2 y 3 valores) y la mejor precisión se obtuvo con 3.

3.4. Clasificación de UX smells

Luego de clasificar los issues de UX, el siguiente paso fue identificar dentro de este grupo, aquellos que pueden considerarse UX smells. Un UX smell se puede definir como un indicio de una experiencia de usuario deficiente [5]. Un ejemplo de UX smell es *Unformatted Input*, que ocurre cuando se utiliza un campo de texto libre en un formulario para ingresar un dato con cierto formato (como por ejemplo una fecha), lo cual es propenso a errores. Si bien ya se definió un catálogo de UX smells en [5], este no es definitivo y podría ser extendido a medida que se encuentran casos de una mala experiencia de usuario. Es por eso que el objetivo en este punto es identificar issues que cumplan con los criterios de un UX smell: que no sea un bug o un error de codificación, y que no implique agregar o quitar funcionalidad de la aplicación. A continuación se presenta un issue que se considera UX smell:

"Link preview text on hover are inconsistent. *Our link preview text on hover are inconsistent across various views, we should work to make all of them consistent."* (microsoft/vscode).

En un primer análisis de los issues se pensó que habría tres separaciones estrictamente marcadas dentro de los issues de UX: bugs, feature-requests y smells. Por este motivo, en una primera instancia únicamente se utilizaron los labels que los mismos usuarios asignan para descartar bugs ("bug") y feature-request ("feature" y "feature-request"), pero luego se decidió agregar también las palabras de la descripción, porque no siempre los issues estaban bien etiquetados.

Al igual que la clasificación anterior este también fue un proceso iterativo de dos pasos: primero se realizaba la búsqueda y después se analizaban un conjunto de 20 issues en cada proyecto para refinar el listado de palabras clave. Estos pasos se repitieron hasta alcanzar una precisión del 80 % en el conjunto de issues inspeccionado.

En las primeras iteraciones, al leer detenidamente los issues uno por uno, se observó que había muchos UX smells que estaban etiquetados por los usuarios como "feature-requests", lo que hizo tomar la decisión de no descartar desde el principio estos issues y replantear la manera de detectar cada tipo de issue. Así, se decidió descartar solamente los bugs y se trabajó en refinar las listas de palabras para identificar los smells.

Analizando los issues se observó que era muy difícil tener un sólo listado de palabras generalizado para clasificar los smells de cualquier proyecto y alcanzar la precisión buscada. Dado que cada proyecto tiene sus particularidades y palabras específicas en los issues, para los primeros tres proyectos analizados (Jitsi Meet, Matomo y ERPNext), se realizó un refinamiento exhaustivo de las palabras clave (ver apéndice[2]). Para los siguientes repositorios a tratar, se utilizó un único listado de palabras general dado que no se identificaron grandes diferencias.

3.5. Resultados

Con el objetivo de analizar cuánta atención se presta a los UX issues se calcularon las siguientes métricas:

- **Proporción de issues resueltos**: se consideran resueltos a los issues cerrados que no poseen la etiqueta "wontfix". En particular, se calculó la proporción de smells resueltos (Smells-PR) y del resto de los UX issues que no son smells (UX-PR).
- **Tiempo de resolución promedio**: es la cantidad de días entre la fecha de apertura y la fecha de cierre del issue. Se calculó tanto para los smells (Smells-TR) como para el resto de los UX issues (UX-TR).
- **Cantidad de comentarios promedio**: también se calculó para los smells (Smells-C) y para el resto de los UX issues (UX-C), con el fin de tener una medida del grado de discusión que tienen los issues y analizar si esto tiene alguna relación con su resolución.

Los resultados aparecen en la tabla 2. Se puede observar que:

- La proporción de UX issues resueltos es siempre menor que la de issues en general (ver columna "Cerrados" en la tabla 1). En algunos casos la diferencia es pequeña pero en otros, como Jitsi Meet, es casi la mitad.
- Con respecto a la proporción de smells resueltos, ERPNext es el único proyecto en el que esta es notablemente menor que la del resto de los UX issues. En los proyectos restantes (a excepción de KeystoneJS), la diferencia es mínima. Sin embargo, en todos estos casos se observa una diferencia considerable en el tiempo de resolución de los UX issues versus UX smells. El hecho que los UX smells tarden más tiempo en resolverse que los demás issues de UX, indica que la resolución de UX smells suele postergarse en favor de otros issues.

[2] https://bit.ly/3SkvQqA

Tabla 2. Resultados de la búsqueda automática. PR: proporción de issues resueltos. TR: tiempo de resolución (en días).

Repositorio	UX-PR	Smells-PR	UX-TR	Smells-TR	UX-C	Smells-C
Jitsi Meet	35 %	34 %	112,40	228,57	5,43	3,48
Matomo	64 %	58 %	166,55	348,90	6,96	5,62
ERPNext	65 %	38 %	221,82	335,40	2,05	1,66
KeystoneJS	84 %	91 %	165,21	179,60	4,22	5,24
VS Code	92 %	90 %	12,97	15,40	4,15	4,52
Wallabag	78 %	72 %	147,68	220,50	3,61	3,48
Open MCT	69 %	64 %	139,44	168,11	3,01	3,27
Superset	86 %	85 %	145,93	210,96	3,96	4,19

- En el caso específico de VS Code, los resultados muestran que los smells reciben la misma atención que el resto de los UX issues porque tanto la proporción de resolución como el tiempo de resolución muestran valores muy similares, y además son significativamente mayor y menor que los demás proyectos.
- Con respecto a la cantidad de comentarios promedio no se encontró una relación entre esta cantidad y la resolución de los issues.

3.6. Búsqueda manual

Los resultados anteriores fueron obtenidos a partir de la clasificación automática de UX issues basada en la búsqueda de palabras clave. Observando la tabla 2, resulta llamativo que haya un gran número de UX smells resueltos; en el caso de KeystoneJS y VS Code poseen una proporción del 90 %. Analizando el conjunto de UX smells resultantes de la clasificación, se encontraron muchos falsos positivos, es decir, issues que fueron clasificados como UX smells cuando en realidad no lo eran. Esto se debe a que la búsqueda a través de la palabras clave tienen sus limitaciones, ya que una misma palabra puede tener diferentes significados en distintos contextos. Además, también puede ocurrir que los usuarios utilicen alguna de las palabras usadas para identificar el smell para reportar otro tipo de issues.

Por otro lado, la única forma de determinar automáticamente si un issue está resuelto o no es a partir de su estado y sus etiquetas. Sin embargo, no siempre un issue cerrado está efectivamente resuelto. Por ejemplo, en muchos proyectos ocurre que los issues se cierran automáticamente luego de un cierto período de inactividad, o los mismos desarrolladores deciden no resolver el issue, pero no lo etiquetan como "wontfix".

Por los motivos detallados anteriormente, se realizó un análisis manual de los issues clasificados como UX smells para comprobar que estuvieran correctamente clasificados y que efectivamente estuvieran resueltos aquellos que estaban cerrados. En este análisis se incluyeron los proyectos: Jitsi Meet, Matomo, ERPNext, KeystoneJS y Wallabag.

3.6.1. Análisis de los Pull Request

Para determinar el impacto de la resolución de los UX smells en el código fuente, se recurrió al análisis de los **pull requests** (PRs) asociados a los mismos. Un pull request es una petición que hace un usuario para incorporar modificaciones al código fuente, que consisten en una colección de **commits**; cada uno de estos es una captura del estado del proyecto en un momento concreto.

Como la API de GitHub no proporciona detalles acerca de los PRs, se inspeccionaron en forma individual los smells clasificados en la web de GitHub, con los siguientes objetivos:

- Obtener el tiempo que transcurre desde la creación del issue hasta que se menciona en un PR como una medida de cuánto se pospone la resolución de un issue reportado.
- Calcular el tiempo de resolución del issue contabilizando la cantidad de días desde que se creó hasta que se cerró el PR. Esto se hizo porque se pensó que este tiempo sería más preciso que contar los días que el issue estuvo abierto.

Sin embargo, analizando los issues se observó que la proporción de smells resueltos con PRs asociados era muy baja como para calcular las métricas descriptas anteriormente. Por este motivo es que se decidió descartar el análisis de los PR, y se desarrolló una clasificación manual de los UX smells para analizar su resolución.

3.6.2. Resolución de UX smells

Inspeccionando los issues individualmente y sus discusiones generadas entre los usuarios, luego de eliminar los falsos positivos que arrojó el análisis automático, se llevó a cabo la siguiente clasificación:

- **Descartado (D)**. Un UX smell es descartado cuando los mismos usuarios luego de una discusión, determinan que la incidencia planteada no constituye un problema de experiencia de usuario. En general estos issues se cierran sin tener ningún impacto en el proyecto.
- **Ignorado (I)**. El issue no presenta ningún tipo de actividad: no tiene comentarios, ni tiene PRs asociados.
- **Resuelto (R)**. Esto ocurre cuando tiene algún PR asociado, o también cuando en la discusión generada alguno de los usuarios indica que fue resuelto en una nueva versión de la aplicación.
- **No resuelto (NR)**. A diferencia del ignorado tiene actividad, pero no hay evidencia de que haya sido resuelto. En ciertos casos también ocurre que los colaboradores deciden que resolver el smell no entra en las prioridades del proyecto y lo dejan de lado.

A cada issue se le asignó solo una de las categorías anteriores. Los resultados de la clasificación se muestran en la tabla 3. Se puede observar que:

- En casi todos los casos la cantidad de smells obtenidos en esta búsqueda manual (Smell-BM) es considerablemente más baja que la cantidad de smells

Tabla 3. Resultados de la clasificación manual de smells. Smells-BA: cantidad de smells detectados en la búsqueda automática. Smells-BM: cantidad de smells sin los falsos positivos.

Proyecto	**Smells-BA**	**Smells-BM**	**D**	**NR**	**I**	**R**	**P-R**
Jitsi Meet	37	29	8	1	13	7	33 %
Matomo	183	115	6	17	21	71	65 %
ERPNext	87	56	4	17	19	16	31 %
KeystoneJS	74	15	1	4	0	10	71 %
Wallabag	94	32	3	5	7	17	59 %

detectados automáticamente. Esto muestra que la búsqueda automática contiene un gran número de falsos positivos.

- La proporción de smells resueltos (P-R), que no considera los issues descartados, es significativamente más baja en los casos de ERPNext, KeystoneJS y Wallabag en comparación a los resultados de la tabla 2. Esto se debe a que muchos de los issues en estos proyectos están cerrados, por lo cual en un principio se consideraron resueltos, pero luego en el análisis de sus comentarios no se encontró evidencia de que esto haya sido así.
- A excepción de Matomo y KeystoneJS, el resto de los proyectos tiene por lo menos el 40 % de los smells sin resolver, lo que evidencia que los UX smells son dejados de lado. En el caso puntual de Jitsi Meet, la gran mayoría de los issues no resueltos fueron ignorados, lo cual es una muestra que el equipo de desarrollo desestimó los smells sin analizar si convenía solucionarlos o no. Los otros dos proyectos, ERPNext y Wallabag, contienen smells ignorados en una proporción similar a la de los no resueltos: issues que fueron discutidos pero finalmente no se resolvieron.

4. Conclusiones y trabajo futuro

Este trabajo presenta evidencia de la presencia de deuda técnica de UX en un conjunto de proyectos de GitHub. Esta deuda se caracteriza a partir de la identificación de UX smells: problemas con la experiencia del usuario de una aplicación que pueden resolverse sin alterar la funcionalidad de la misma (a través de UX refactorings).

Primero se desarrolló una clasificación automática de issues para identificar aquellos que cumplen con el concepto de UX smell. Esta clasificación se realizó a través de la búsqueda de palabras clave siguiendo un proceso iterativo. Los resultados muestran que en muchos casos los UX smells o no se resuelven (la mitad de los proyectos tienen al menos 35 % de issues sin solucionar), o son postergados en favor de otros issues (esto se refleja en el tiempo de resolución).

En los resultados de la clasificación anterior, se observó que muchos de los issues clasificados como UX smells en realidad no lo eran, y que a su vez en diferentes ocasiones estos se cerraban sin ser resueltos. Por este motivo, se decidió

realizar un análisis sistemático de los UX smells para eliminar los falsos positivos y determinar cuáles de ellos habían sido efectivamente resueltos. Como resultado, la mayoría de los proyectos analizados muestran una proporción menor de UX smells resueltos que la clasificación automática, lo cual reafirma la idea que los smells se postergan y esto es lo que contribuye a la acumulación de UX debt.

Como trabajo futuro, se propone trabajar en la mejora de la clasificación automática de UX smells, utilizando técnicas de minería de texto. Por otro lado, también queda pendiente estudiar otras métricas que permitan estimar con mayor precisión el costo de resolución de los issues, así poder tener una medida de cuánto trabajo implica remediar la deuda acumulada.

Referencias

1. Behutiye, W.N., Rodríguez, P., Oivo, M., Tosun, A.: Analyzing the concept of technical debt in the context of agile software development: A systematic literature review. Information and Software Technology 82 (2017)
2. Da Silva Maldonado, S.: Detecting and quantifying different types of self-admitted technical debt. In 7th International Workshop on Managing Technical Debt (MTD). 9–15. (2015)
3. Da Silva Maldonado, Shihab, T.: Using natural language processing to automatically detect self-admitted technical debt. IEEE Transactions on Software Engineering 43, 11 (2017), 1044–1062 (2017)
4. da Fonseca Lage, L., Kalinowski, M., Trevisan, D., Spinola, R.: Usability technical debt in software projects: A multi-case study. In: 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. IEEE (2019)
5. Grigera, J., Garrido, A., Rivero, J., Rossi, G.: Automatic detection of usability smells in web applications. International Journal of Human-Computer Studies 97 (2017)
6. Kuusinen, K.: Bob: a framework for organizing within-iteration ux work in agile development. In: Integrating User-Centred Design in Agile Development. Springer (2016)
7. Lage, Kalinowski, T.S.: Usability technical debt in software projects: A multi-case study. ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 2019, pp. 1-6, doi: 10.1109/ESEM.2019.8870180. (2019)
8. Li, Z., Avgeriou, P., Liang, P.: A systematic mapping study on technical debt and its management. Journal of Systems and Software 101, 193–220 (2015)
9. Potdar, S.: An exploratory study on self-admitted technical debt. In 30th International Conference on Software Maintenance and Evolution (ICSME). 91–100. (2014)

Análisis de Performance de Base de Datos Sql y NoSql aplicado a Datos de Entidades Públicas.

Mercedes Barrionuevo[1] and Mariela Rodriguez[2]

[1] Universidad Nacional de San Luis, San Luis, Argentina
[2] Universidad Nacional de Jujuy, Jujuy, Argentina
mdbarrio@unsl.edu.ar
mariela.rodriguez@fi.unju.edu.ar

Resumen En los últimos años hemos sido testigos de revoluciones tecnológicas que se suceden a un ritmo tan acelerado que parecen imperceptibles. La era del Big Data ha traído consigo una gran cantidad de datos que necesitan ser almacenados de la forma más eficiente posible y ser recuperados en un tiempo considerablemente rápido. Contar con herramientas de administración de base datos tanto relacional como no relacional es de vital importancia. Esta área es de constante interés para determinar cuál de ellas se comporta mejor u obtiene una mayor performance en un dominio de datos en particular. Finalmente, es preciso concluir si para un dominio de interés, es necesario poner énfasis en la optimización del espacio de almacenamiento o en el tiempo de respuesta de una consulta.

Palabras claves: Base de Datos Relacional. Base de Datos No Relacional. Postgress. MongoDB. ElasticSearch. DNRPA

1. Introducción

La cambiante situación económica actual del país va dejando ciertas incógnitas del poder económico de sus habitantes. Analizar la compra y los hurtos de autos nuevos o usados, es una alternativa de interés a considerar, la cual nos permitirá obtener algunos indicadores de problemas o tendencias actuales. Utilizar diversas tecnologías de preprocesamiento, almacenamiento y visualización de datos de manera conjunta nos pueden ayudar a tal fin.

La gestión de las bases de datos es fundamental para todos los trabajos de estas áreas. Un sistema de gestión de bases de datos (SGBD) es un programa que permite a uno o varios usuarios acceder a una base de datos [2]. Permite manejar los accesos diferenciados (identificación, seguridad) y permite interpretar las búsquedas para ingresar, modificar o suprimir datos. Se pueden diferenciar 2 grandes familias de SGBD: los SQL y los NoSQL. Para saber cuál tecnología elegir, en este trabajo vamos a modelar consultas y determinar su velocidad de respuesta para calcular la performance de cada una de ellas.

Las bases de datos SQL (acrónimo de Structured Query Language), también llamadas bases de datos relacionales, están constituidas por un conjunto de tablas

en las que los datos están clasificados por categorías. Ejemplo de este tipo de base de datos son: *Oracle [3], PostgreSql [4] y MySql [5].*

Por otro lado, las bases de datos NoSQL son no relacionales. Éstas no necesitan un esquema fijo y son fácilmente modulares. El objetivo es recuperar los datos de un mismo lugar sin necesidad de pasar por las relaciones entre tablas. Ejemplo de estas bases de datos son: *MongoDB [6] [7] , Elasticsearch [8] [9] y Neo4J [10].*

Distintos Ministerios, Secretarías y Organizaciones dependientes del Poder Ejecutivo Nacional Argentino han abierto sus datos. Entre ellos, el Ministerio de Justicia y Derechos Humanos [11] publica datos, actualizados mensualmente, referidos a Estadística de trámites de maquinarias, vehículos, embargos, inscripciones, bajas, transferencia, prendas, robos y recuperos de autos, entre otros. Todos estos datos publicados poseen licencia Creative Commons Attribution 4.0 y tienen una frecuencia de actualización mensual.

Por lo tanto, el objetivo planteado en este trabajo es determinar el SGBD adecuada o recomendable para gestionar datos del registro de automotores de la DNRP, para lograrlo se implementan consultas directas a las bases de datos mediante la interfaz de usuario correspondiente.

Este documento está organizado como sigue: la sección 2 describe la metodología involucrada en el desarrollo de este trabajo y los pasos realizados dentro de esta metodología. La sección 3 detalla la visualización de los resultados obtenidos en las consultas junto con la comparativa de los tiempos de ejecución de las consultas en cada uno de los SGBD. Finalmente se detallan las conclusiones y líneas futuras de trabajo.

2. Metodología

Para llevar adelante el proceso de comparación de motores de bases de datos y extraer información de forma sistemática se hará uso de la metodología CRISP-DM [15], la cual permite entender el proceso de descubrimiento de conocimiento. CRISP-DM es una metodología creada para trabajar con proyectos de minería de datos, pero de acuerdo a sus fases se adapta al actual proyecto explorando información para la concreción del objetivo propuesto. El ciclo de vida del proyecto de minería de datos, en esta metodología, consiste en seis fases: *Comprensión del negocio, Compresión de los datos, Modelado de datos, Evaluación e Implementación* [17] . Cada una de estas etapas se abordan en las siguientes secciones.

2.1. Comprensión del Negocio

En esta fase se debe comprender los objetivos generales y determinar los objetivos técnicos del proyecto. [16] El objetivo del negocio es medir la performance de las bases de datos estructurales versus las no estructurales y, en base a ello, poder recomendar la tecnología más adecuada para trabajar con datos del Registro Nacional del Automotor (DNRPA) [11].

Para concretar el objetivo del negocio, se tiene como objetivo técnico, realizar las consultas acordes y necesarias a ser ejecutadas en las diversas bases de datos mencionadas anteriormente, como así también, visualizar las consultas y determinar la existencia de posibles patrones de comportamiento mencionando las posibles causas.

2.2. Herramientas de software y hardware utilizadas

El software utilizado para la ejecución de las consultas es el siguiente:

1. **Sistema Operativo**: Windows 10 - 64 bit y Ubuntu 20.04 Linux
2. **MongoDB Server version 5.0**: Bases de datos NoSQL
3. **Studio 3T version 2022.6.1**: GUI de MongoDB con funciones de consulta visual utilizada para la exportación e importación de colecciones, vistas o consultas.
4. **PostgreSql versión 14.4**: Motor de Base de Datos Relacional.
5. **PgAdmin versión 4.0**: GUI de administración de PostgreSql.
6. **Pentaho Data Integration [14]:** Herramienta de la suite de Pentaho de las que se denomina ETL (Extract – Transform – Load).
7. **Spoon**: Spoon es una Interfaz Gráfica de Usuario (GUI), que permite diseñar transformaciones y trabajos de ETL.
8. **Logstash 7.6:** Herramienta para manejo de grandes volúmenes de archivos entre gestores de bases de datos y sistemas de archivos. Permite transformar datos y enviarlos a diversas bases de datos como MongoDB y ElasticSearch.
9. **ElasticSearch** 7.6 - Base de datos NoSQL.
10. **Power Bi 2.106**: Herramienta de visualización de Datos

Mientras que el hardware utilizado y las características de la máquina sobre la que se realizaron las consultas son:

1. Notebook: HP 15-dw2xxx
2. RAM: 8 GB
3. Disco: SSD 250 GB
4. Procesador: Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz. 10th generación.

2.3. Comprensión de los datos

La recolección de datos inicial corresponden a *"robos y recuperos de autos"* e *"inscripciones iniciales de autos"* de la DNRPA correspondientes al periodo 2018-2022. Cada archivo posee alrededor de 10.000 registros, los cuales representan datos recolectados en un mes, y 25 atributos con información relacionada al tipo de trámite, datos del registro donde se realiza el trámite, datos del vehículo y de su propietario. La cantidad total del dataset es de aproximadamente 2 millones de registros.

Los archivos se dividen en dos grupos bien diferenciados: *Incripciones de autos* [12], el cual posee toda información relacionada a las inscripciones de

vehículos y datos de su primer propietario; *Robos y recuperos de automotores* [13], contiene datos referidos a los trámites de robos y recuperos de automotores.

Los atributos de los archivos de inscripciones tanto como de robos y recuperos de automotores son: *tipo de trámite, fecha del trámite, fecha de la inscripción inicial, código de la seccional del registro, registro seccional descripcion, registro seccional provincia, automotor origen, automotor anio modelo, automotor tipo código,automotor tipo descripción, automotor marca código, automotor marca descripción, automotor modelo código, automotor modelo descripción, automotor uso código, automotor uso descripción, titular tipo persona, titular domicilio localidad, titular domicilio provincia, titular genero, titular anio nacimiento, titular país nacimiento, titular porcentaje, titular domicilio provincia id, titular país nacimiento id.*
Se consideró apropiado realizar la exploración de datos y la verificación de la calidad de datos en las fases posteriores.

2.4. Preparación de los datos

En esta fase se procede a preparar los datos para que luego se apliquen las técnicas necesarias para el proyecto. En esta sección también, se describe aspectos de la calidad de los datos. La limpieza de datos se realizó mediante una herramienta de ETL, denominada Spoon [14].

La extracción de datos se realizó desde los archivos *csv* obtenidos de la DNRP y cargados a Spoon, se cargaron 106 archivos que contienen información de Enero de 2018 a Mayo de 2022.

Las transformaciones se realizan para unificar y corregir los nombres de las marcas, tipos de vehículos, tipos de trámites y registros encontrados. Las transformaciones de limpieza más relevantes se detalla a continuación:

1. Los archivos de *robos y recuperos de autos* poseen muchos valores nulos en diversos atributos como automotor_modelo_año. En algunos casos los valores nulos son reemplazados con información encontrada en otra columna y en otros casos son ignorados por no tener valores de referencia.
2. Respecto a los tipos de trámites realizados en las seccionales de la DRNPA se identificaron 9 tipos de inscripciones distintas las cuales fueron unificadas en una única categoría.
3. Los atributos correspondientes a Marcas, Tipo de vehículos, Tipo trámites, Registro del automotor contienen datos ingresados de forma manual, la cual es necesario realizar una limpieza, a fin de unificar estos datos para obtener resultados consistentes en los pasos posteriores. En la Fig. 1 siguiente se muestra la cantidad de registros antes y después de la limpieza de datos.

La construcción de datos se realizó mediante la agregación de las variables mes y año, que representan el mes y año de realización del trámite, para ser tratados como elementos independientes y no como un valor agrupado.

Finalmente, los datos son enviados preprocesados a los motores de base de datos de PostgresSql, MongoDB y ElasticSearch. La elección de los mismos se

Limpieza	Marcas	Tipo de Vehículo	Tipo Tramite	Registro A
Inicial	1790844	1090	11	868
Limpio	468	47	3	343

Figura 1. Limpieza de Marcas, Registro, Tipo Automotor y Tipo trámite

fundamenta en sus características de código abierto, buenas posiciones en rankings de bases de datos más utilizadas, y en particular, las bases de datos no estructuradas por ser escalables y tolerantes a fallos en ambientes de datos masivos.

Las transformaciones se hicieron de acuerdo a las características propias de las base de datos estructuradas o no estructuradas. Para la carga de Postgres fue necesario realizar la disgregación y normalización de los datos. En el caso de MongoDB se generó un archivo único para la generación de la colección. El envío de datos a ElasticSearh se realizó mediante un archivo de texto (csv). Esto último es porque la herramienta Spoon permite enviar datos a diversas bases de datos como mongoDB y PostgreSql incorporando distintos drivers, pero no se encontró el driver adecuado para la base de datos ElasticSearch. Por lo tanto, para la carga de datos a la base de datos de ElasticSearch se utilizó la herramienta Logstash a través del archivo *pipeline_elasctic.conf*.

2.5. Modelado, Evaluación e Implementación de los datos

En esta fase es necesario seleccionar la técnica con la que se desarrollará el proyecto. El objetivo del proyecto es determinar la base de datos más adecuada o recomendable para gestionar datos del registro de automotores de la DNRP, para lograrlo se puede implementar consultas directas a la base de datos mediante la interfaz de usuario que cuente cada uno de ellos. En la siguiente sección se detallan como se implementaron las consultas, las respuestas obtenidas y los tiempos de ejecución de las mismas.

- **Consulta 1:** ¿Cuáles son los autos importados de empresas que sufrieron robos y han sido recuperados en el último mes?
  Esta consulta en los 3 SGBD devuelve 16 registros de autos importados que han sido robados y recuperados el último mes, tal como lo muestra la Fig. 2.
- **Consulta 2:** ¿Los compradores de automotores son mayormente hombres, mujeres o personas jurídicas?
  Esta consulta devuelve que la mayoría de los compradores son hombres con 1120667 registros.
- **Consulta 3:** ¿De qué marca y modelo (año) de auto son los más robados?
  Esta consulta devuelve que los autos mas robados son Volkswagen modelo 2011 con 1677 registros, tal como lo muestra la fig. 3.
- **Consulta 4:** ¿Cuáles fueron los meses de menor venta de autos?
  Esta consulta identifica los meses en los que la venta cayó a niveles más bajos en el periodo 2018 a 2022, siendo tales meses abril 2020, marzo 2020 y diciembre 2021.

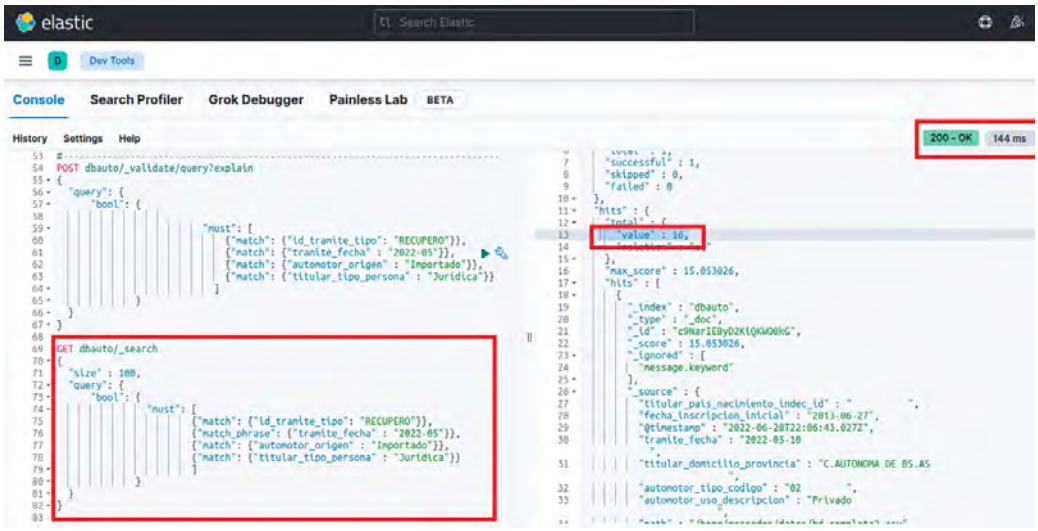


Figura 2. Consulta en ElasticSearch.

```
SELECT COUNT(anio_modelo) AS cantidad, anio_modelo, automotor_marca_descripcion
FROM tramites WHERE automotor_marca_descripcion = (
 SELECT marca.automotor_marca_descripcion AS marca_auto
   FROM tramites AS tram INNER JOIN marca_automotor AS marca
     ON tram.id_marca_automotor = marca.id_marca_automotor
   WHERE tram.id_tipo_tramite_sec IN (2, 3)
   GROUP BY marca_auto ORDER BY COUNT(tram.id_marca_automotor) DESC LIMIT 1 )
AND id_tipo_tramite_sec IN (2, 3)
GROUP BY anio_modelo, automotor_marca_descripcion
ORDER BY COUNT(anio_modelo) DESC LIMIT 1
```

Figura 3. Consulta 3 en pgAdmin 4.

- **Consulta 5:** ¿Cuál es la marca de autos mas vendidos por provincia? Esta consulta detalla por ejemplo que para la provincia de Bs. As la marca más vendida es Volkswagen, mientras que para Córdoba es Fiat, para Santa Fe es Toyota y así siguiendo para el resto de las provincias como se puede ver en la gráfica 5. La figura 4 muestra su resolución en MongoDB.

```
db.getCollection("inscripciones_robados").aggregate([
        //Etapa 1: filtro por registros de incripciones
        {"$match" : {"id_tramite_tipo" : {"$regex" : "INSC"}}},
        //Etapa 2: agrupo por marca y provincia y cuento la cantidad de registros
        {"$group" : {
                "_id" : {
                    "marca" : "$automotor_marca_descripcion",
                    "prov" : "$registro_seccional_provincia"
                },
                "suma" : {"$sum" : 1.0}
         }},
        //Etapa 3: ordeno de mayor a menor por los totales obtenidos en suma.
        {"$sort" : {"suma" : -1.0}},
        //Etapa 4:
        {"$group" : {"_id" : {"nombre" : "$_id.prov"},
                    "data" : {
                    "$push" : {
                        "provincia" : "$_id.prov",
                        "marca" : "$_id.marca",
                        "ventas" : "$suma"
                    }
         }},
        //Etapa 5: me quedo con el primer registro cuyo valor es el máximo.
        {"$group" : {
                "_id" : {
                    "p" : {"$first" : "$data.provincia"},
                    "m" : {"$first" : "$data.marca"},
                    "v" : {"$first" : "$data.ventas"}
         }}
]
```

Figura 4. Consulta 5: Marca más vendida por provincia.

3. Exploración y Visualización de los datos

De los datos recolectados se pudieron realizar diferentes gráficas para poder analizar si los resultados obtenidos con las distintas consultas se veían reflejados en dichos gráficos.

Realizadas las consultas, se hizo un análisis visual de los datos obtenidos en las diferentes consultas. La Fig. 5 visualiza los resultados de las consultas 1 a la 3. En principio se muestra los datos de los vehículos importados de las empresas que fueron recuperados en el mes de mayo de 2022. A continuación se visualiza la cantidad de vehículos comprados por género, siendo el 51 % de género masculino, 31 % de género femenino y por último 18 % corresponden a titulares jurídicos (empresas).

La Fig. 6 detalla la cantidad de ventas por mes desde el 2018 hasta la actualidad. En este gráfico se puede ver una baja significativa entre los años 2018 y 2019. Se estima que algunos de los factores de dicha caída fue el incremento de los precios y la escasa financiación, provocando la caída abrupta de ventas, sobre todo en Diciembre de 2021. Por otro lado, se puede ver que otros 2 picos mínimos de ventas ocurridos en Abril y Marzo del 2020 ocasionados por la pandemia Covid19.

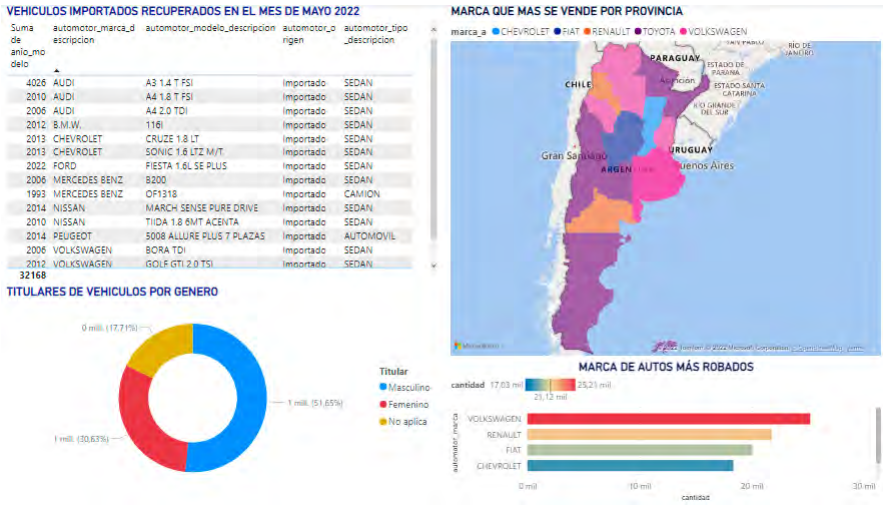


Figura 5. Vehículos recuperados, Titulares por Género y Marcas por provincia.

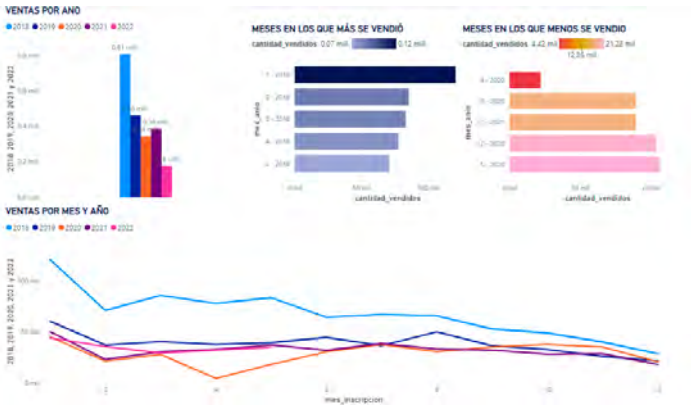


Figura 6. Histórico de Ventas de automotores.

3.1. Comparativa de Tiempos de Ejecución

En el cuadro 3.1 se muestran los tiempos de ejecución promedio de las consultas en las distintas base de datos, representadas gráficamente en la Fig. 7. Dichos valores son el resultado de 10 corridas por consulta.

_	Consulta 1	Consulta 2	Consulta 3	Consulta 4	Consulta 5
PosgreSql	1.014	0.876	1.162	0.953	0.755
MongoDB	4.788	7.988	2.750	7.033	7.486
ElasticSearch	0.144	0.667	0.243	0.425	0.732

Cuadro 1. Tiempos Promedio de Ejecución en segundos

En el cuadro 2 se muestra la varianza entre las distintas mediciones para cada consulta, siendo las más significativas los tiempos de la consulta 1 y 2 en MondoDB. Esto se debe al hecho de tener un tiempo extra de preparación propio del motor de base de datos al iniciar las ejecuciones.

_	Consulta 1	Consulta 2	Consulta 3	Consulta 4	Consulta 5
PosgreSql	0.212	0.143	0.256	0.122	0.053
MongoDB	10.03	20.96	0.011	0.055	0.063
ElasticSearch	0	0.001	0.002	0.001	0.003

Cuadro 2. Varianza de los Tiempos de Ejecución en segundos

Como se pueden observar en las gráficas anteriores los tiempos de ejecución de ElasticSearch y sus varianzas son considerablemente mejores.

4. Conclusiones

Este proyecto se encuentra enmarcado en la categoría de almacenamiento de grandes volúmenes de datos, que de acuerdo a la frecuencia de actualización de datos del Registro de la Propiedad del Automotor, irá creciendo de manera continua. Por lo tanto, decidir dónde y cómo almacenar esta información es de vital importancia.

Las base de datos SQL y NoSQL son dos tecnologías que tienen la misma finalidad: almacenar datos y ofrecer las herramientas para leer y manipular esos datos. Elegir la base de datos más adecuada es una tarea muy importante porque será la base de trabajo de todas las profesiones en el campo de la ciencia de datos. Sin embargo, esta elección no es fácil y la respuesta no siempre es evidente.

Las tareas realizadas al conjunto de datos fueron: preparación y limpieza de datos, la cual demandó la mayor parte del tiempo del proyecto, luego se

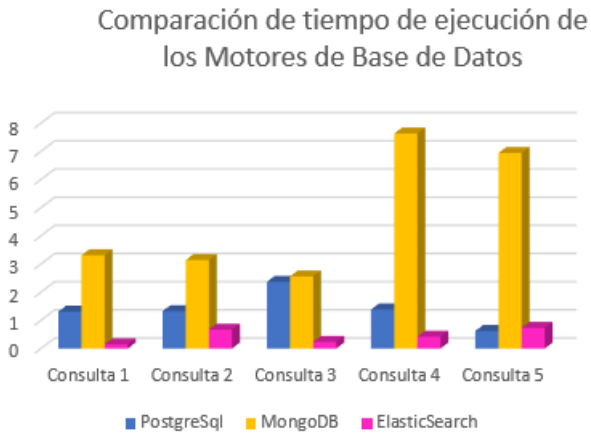


Figura 7. Resultados de consultas en distintas Base de Datos.

modelaron las consultas en PostgreSql, MongoDB y ElasticSearch, y finalmente, se obtuvieron los resultados y tiempos de respuesta.

De los tiempos de ejecución obtenidos se puede concluir que ElasticSearch tiene un mejor tiempo de respuesta con respecto a las otras dos bases de datos. Esto se debe a la naturaleza distribuida y su capacidad de manejo de datos estructurados y no estructurados. Por otra parte, se puede ver que PostgreSql tiene mejores tiempos de respuesta que MongoDB, dado que este último es más eficiente cuando se trata de datos no estructurados.

Por lo tanto, para el conjunto de datos trabajado, se recomienda el uso de Elasticsearch por permitir consultas mediante textos completos, autocompletado y realización de correcciones ortográficas, sin necesidad de realizar la tarea de preprocesado tan exhaustiva.

Como trabajo futuro, se propone optimizar las consultas para evaluar los nuevos tiempos de respuesta y verificar si ocurre alguna mejora en los mismos, como así también ejecutarlas en un hardware más potente.

Agradecimientos. Al profesor Mg. Javier Bazzocco, docente del Centro de Investigación LIFIA.

Referencias

1. Learning PostgreSQL. Salahaldin Juba, Achim Vannahme, Andrey Volkov. 2015. ISBN: 9781783989188
2. Beynon-Davies, P.: Sistema de Base de Datos pp.33-53. Editorial Reverté (2014). España.
3. Oracle, `https://www.oracle.com/ar/database/`. Consultado en Junio de 2022.
4. PostgreSql, `https://www.postgresql.org/`. Consultado en Junio de 2022.
5. MySql, `https://www.mysql.com/`. Consultado en Junio de 2022.

6. MongoDB, https://www.mongodb.com/es. Consultado en Junio de 2022.
7. Data Modeling for MongoDB: Building Well-Designed and Supportable MongoDB Databases. Steve Hoberman. 2014
8. ELK - ElasticSearch, https://www.elastic.co/es/what-is/elasticsearch
9. Martos, C., Uso y Ventajas de Elasticsearch en Bases de Datos No Relacionales (2019). Universidad Autónoma de Madrid. Escuela Politécnica Superior. España.
10. Neo4J, https://neo4j.com/. Consultado en Junio de 2022.
11. Dirección Nacional de Registros Nacionales de la Propiedad Automotor y Créditos Prendarios. Datos públicos generados, guardados y publicados por organismos de gobierno de la República Argentina, Ministerio de Justicia y Derechos Humanos. https://datos.gob.ar/dataset?tags=dnrpa Último acceso 30 de Junio 2022.
12. Dirección Nacional de Registros Nacionales de la Propiedad Automotor y Créditos Prendarios (https://datos.gob.ar/dataset/justicia-transferencias-autos). Último acceso 30 de Junio 2022.
13. Dirección Nacional de Registros Nacionales de la Propiedad Automotor y Créditos Prendarios. (https://datos.gob.ar/dataset/justicia-robos-recuperos-autos). Último acceso 30 de Junio 2022.
14. Pentaho Data Integration. https://sourceforge.net/projects/pentaho/. Consultado en Junio de 2022.
15. Chapman, P., Clinton, J., Kerber, R., Khabaza, T. y otros: CRISP-DM V 1.0. pp.10 – 12. CRISP-DM consortium: NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc and OHRA Verzekeringen (2000)
16. Galan Cortina, V., Castro Galan, E.: . Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el Entorno Universitario. Universidad Carlos III de Madrid. Escuela Politécnica Superior Ingeniería en Informática (2015)
17. Moine, J. M., Haedo, A. S., Gordillo, S.: . Estudio comparativo de metodologías para minería de datos (2011). XIII Workshop de Investigadores en Ciencias de la Computación. RedUNCI
18. Db-engines ranking of wide column stores. https://db-engines.com/en/ranking. Consultado en Agosto de 2022.
19. Duran-Cazar, Jhonatan, Tandazo-Gaona, Eduardo y cia. Rendimiento de base de datos columnares". Universidad Politécnica Salasiana. Revista de Ciencia y Tecnología. Ecuador. 2019.

A hybrid approach to boost the permutation index for similarity searching

Karina Figueroa[1], Antonio Camarena-Ibarrola[1], Nora Reyes[2], Rodrigo Paredes[3], and Braulio Ramses Hernández Martínez[1]

[1] Universidad Michoacana
{karina.figueroa,antonio.camarena}@umich.mx,1578615h@umich.mx
[2] Universidad de San Luis, Argentina nreyes@unsl.edu.ar
[3] Universidad de Talca, Chile raparede@utalca.cl

Abstract. We propose a hybrid strategy that combines three ideas, namely, a convenient way for reducing the length of the permutations, using a permutation similarity measure adjusted for these clipped permutations, and the use of the closest permutant of each object as a pivot for it. In this way, we increase the discriminability of the permutation index in order to reduce even more the number of distance computations without reducing the answer quality. The performance of our proposal is tested using two classical real-world databases: NASA and Colors which are part of the SISAP project's metric space benchmark. We reduced more than 30% of the number of distance evaluations needed to solve the queries on both databases.

Keywords: Similariy search · Permutant-based index · Permutation similarity measures

1 Introduction

Nowadays, multimedia databases are widely used, and of course, the information retrieval is a crucial task. Similarity searching is the only operation that makes sense with this kind of data because two elements are never exactly the same. The similarity is a concept that depends on the database's domain, it is modeled and defined by experts of each field, and it is frequently expensive to compute in terms of arithmetic operations, I/O events, etc. Naturally, when a query is given, the goal is to answer it as quickly as possible. One way to achieve efficiency is to reduce the number of distance computations for answering a query.

There are two kinds of similarity queries, namely, *K-Nearest Neighbor query* $NN_K(q)$ and *Range query* $R(q, r)$. The $NN_K(q)$ retrieves the K database elements that are the most similar to q. The $R(q, r)$ finds the elements of the database whose distance to q is lower than or equal to the radius r.

One way to represent the problem is by mapping it to a metric space [5]. A metric space is a pair $(\mathbb{U}, d)$, where $\mathbb{U}$ is the universe of valid objects and d is a distance function that allows us to compare any two objects from $\mathbb{U}$. Let $\mathbb{X} \subseteq \mathbb{U}$ be the database of interest and $n = |\mathbb{X}|$. As we assume that the function d

is expensive to compute, our goal is to minimize the use of d when answering queries. One issue in this kind of problems is the intrinsic dimension [10] because when it is high, the distance between any pair of different objects tends to be the same, so searching complexity rises as the intrinsic dimension increases.

Assuming we cannot establish a total order over a multimedia database, we have to resort to using a proximity index. An index is a data structure that allows us to obtain a candidate list without sequentially scanning the entire database (unthinkable for huge datasets). There are three well-known index families, namely, the ones based on *pivots*, the ones based on *compact partitions*, and recently, the ones based on *permutations*. Pivot-based and compact-partition-based indexes are *exact* proximity indexes, while permutations-based ones are *approximate*; in the sense that we may lose a few relevant objects from the query answers, but accepting this loss allows us to improve dramatically the searching time.

In this paper, we propose a hybrid method to improve the performance of the *permutations*-based indexes, combining three main ideas: the first one is to conveniently reduce the length of the permutations stored within the index, the second is adapting the permutation similarity measure for these clipped permutations, and the third one is to use the closest permutant of each object as a pivot for it. This novel strategy allows us to improve the already remarkable performance of the permutation-based index when solving similarity queries.

The performance of our proposal is tested using two classical real-world databases: NASA and Colors, which are part of the SISAP project's metric space benchmark available at [8]. We reduce more than 30% of the number of distance evaluations needed to solve the queries on both databases.

The rest of this article is organized as follows: in Section 2 we describe the related work on metric spaces and similarity search. Section 3 shows our novel hybrid index and Section 4 gives its experimental evaluation using two real world datasets from SISAP library [8]. Finally, we expose conclusions and some possible extensions for this work in Section 5.

2 Related Work

Similarity searching in metric spaces has been studied in three leading families of algorithms: pivot-based algorithms [12, 5] (for low intrinsic dimension), partition-based algorithms [4, 6] (for medium to high intrinsic dimension), and permutation-based algorithms [3, 1] (for high intrinsic dimension). As we aforementioned, the permutation-based approach is one of the best representative methods to solve approximate similarity searches. In the following we briefly describe the pivot-based and permutation-based algorithms as they are relevant for this work.

2.1 Pivot-based algorithm

Pivot-based algorithms use a subset of database objects $P = \{p_1, p_2, \ldots, p_k\} \subseteq \mathbb{X}$ to compute pseudo-coordinates. Each database object $x \in \mathbb{X}$ is represented

by a vector containing its k distances to every pivot $p_i \in P$. Let $D(u, P) = (d(x, p_1), \ldots, d(x, p_k))$ be this vector. Given a query $R(q, r)$, we first represent q in the same coordinate system as $D(q, P) = (d(q, p_1), \ldots, d(q, p_k))$. Thus, by virtue of the triangle inequality, any object $x \in \mathbb{X}$ can be discarded if $|d(p_i, x) - d(p_i, q)| > r$ for any pivot $p_i \in P$. Finally, to obtain the query answer, all the non-discarded database objects are directly compared with q and only the objects whose distance is within threshold r are reported.

2.2 Permutation-based algorithm

This kind of indexes use some distinguished elements from the database $\mathbb{X}$ as references points of view. These elements are called *permutants*. The main idea of this method was introduced in [2]. Let $\mathbb{P}$ be the permutant set, formally, $\mathbb{P} = \{p_1, \ldots, p_k\} \subseteq \mathbb{X}$. For the sake of producing the index, each $u \in \mathbb{X}$ computes $D(u, \mathbb{P}) = \{d(u, p_1), \ldots, d(u, p_k)\}$, that is, u computes its distance to every permutant. Then, each object u sorts the set $\mathbb{P}$ using the distances computed in $D(u, \mathbb{P})$ in increasing order. This ordering is called the *permutation* of u, which is denoted by Π_u. Therefore, the permutant in the first position of Π_u is the closest one, and so on. Inversely, let Π_u^{-1} be the inverse of the permutation Π_u, so we can use Π_u^{-1} to identify the position of any permutant in Π_u.

As an example, Figure 1 depicts a subset of points in $\mathbb{R}^2$, considering Euclidean distance. The set of permutants is $\mathbb{P} = \{p_1, p_2, p_3, p_4, p_5, p_6\}$; that is, $k = 6$. If we consider the object $u \in \mathbb{X}$, $D(u, \mathbb{P}) = (3, 4, 6, 3, 2, 5)$ where $d(u, p_1) = 3$ and so on; then $\Pi_u = (5, 1, 4, 2, 6, 3)$. It can be noticed that the closest permutant is p_5, because $d(u, p_5) = 2$. The inverse permutation Π_u^{-1} is $(2, 4, 6, 3, 1, 5)$. Then, $\Pi_u^{-1}(p_2) = 4$ means p_2 is in the 4th position in Π_u. We note that we need $O(nk)$ distance computations to obtain all the permutations.

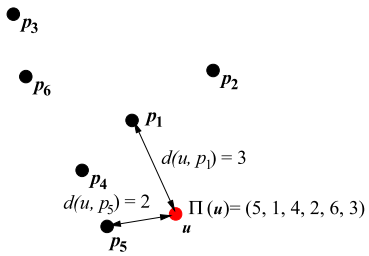


Fig. 1. Example of a permutation considering $\mathbb{P} \subset \mathbb{R}^2$ using Euclidean distance.

As two identical elements must have exactly the same permutation, we expect that two similar elements have similar permutations. Therefore, when we search for elements similar to a query q, the problem is to find objects whose permutations similar to Π_q. The advantage of this approach is that computing

the permutation similarity is cheaper than computing the *distance function* d. There are different measures to compute similarity between permutations [11]. One of the most used is the Spearman Footrule measure, defined as:

$$F(u,q)_k = F(\Pi_u, \Pi_q) = \sum_{i=1}^{k=|\mathbb{P}|} |\Pi_u^{-1}(p_i) - \Pi_q^{-1}(p_i)| \quad (1)$$

The basic method stores the whole permutation of each database object, hence it needs $O(nk)$ space.

An interesting member of this index family is the Metric Inverted File. In the next section we briefly describe it along with some of its improvements.

2.3 Metric Inverted File (MIFi)

Amato and Savino proposed using an inverted file of permutations [1], where each permutant in $\mathbb{P}$ has its respective entry in the inverted file. We call MIFi index this approach. To produce the index, they define parameter $m_i < |\mathbb{P}|$ which is used during the preprocessing time. For each permutant $p \in \mathbb{P}$ the MIFi index stores the list of the elements $u \in \mathbb{X}$ such that its permutation Π_u has the permutant p within the first m_i positions. The list for each permutant p stores pairs (u, pos), where u denotes an object in $\mathbb{X}$ and pos referes to the position of p within the permutation Π_u.

Given the query q, we need to determine Π_q. The MIFi index uses another parameter $m_s \ll m_i$ for searching. The MIFi search method only retrieves the posting lists of the first m_s permutants in Π_q and next, it unites all of them to obtain the candidate set. Finally, all the elements in the union of the lists are directly compared with q using the distance d to produce the query answer. Authors in [1] proposed a variant of the Spearman Footrule permutation similarity measure, because each permutation was clipped by the parameter m_i.

In the works [9, 7], the authors improved the performance of the MIFi index in two ways. On the one hand, each posting list stores only elements u but not the positions pos [9] and the short permutation of each element is maintained. They also proposed a new way to compute the Spearman Footrule measure. On the other hand, to reduce the candidate list size, even more, a new parameter m_{s_r} is selected according to the radius of the similarity query [7] (instead of the fixed-parameter m_s from the MIFi index).

3 Our proposal

We look to improve even more the performance of the alternative to MIFi index presented in [7]. Our proposal considers several aspects. The first one is to have smaller permutations in the index, the second one is to use one of the permutants as a pivot, and the third is to consider a modified permutation similarity measure. After explaining in detail these three main aspects of our contribution, we show how to combine them in order to produce our novel index and its respective searching algorithm.

3.1 Clipped permutations

Instead of having a maximum global length m_i, each permutation can be shortened with a different length, by considering for each $u \in \mathbb{X}$ its appropriate permutation prefix. To do so, we consider the distance to its closest permutant; that is, $d(u, p_{\Pi_u(1)}) = r_u$, and keep for u those permutants within a distance lower than or equal to $2r_u$. We ran preliminary experiments to determine that $2r_u$ performs well when solving queries, but this clearly deserves further study. We call Π'_u the clipped permutation of u.

Let m_u denote the length of the trimmed permutation for element u; that is, $m_u = |\Pi'_u|$ is the length of the prefix selected. In this case, since some objects could have only one permutant within distance $2r_u$ from u, we propose using a minimum global length $m_{\min}$ for all the permutations. Likewise, since some objects could have all the permutants within distance $2r_u$ from u, we also propose a maximum global length $m_{\max}$ for all the permutations.

In our example of Section 2.2, Figure 1, for element u we obtained $r_u = 2$. So, its clipped permutation is $\Pi'_u = (5, 1, 4, 2)$ because $d(u, p_2) = 4 \leq 2r_u$. It can be noticed that, by this way of shortening the permutations, each permutation would have a different length.

3.2 Including a single pivot

When we search for a query q, we have to compute Π_q by calculating all the distances between q and every permutant in $\mathbb{P}$. Besides, we know that if we keep the distances from the element $u \in \mathbb{X}$ to all the permutants in $\mathbb{P}$, we can use them to obtain lower bounds of the distance from u to q, as in a pivot-based algorithm. Hence, we can discard the elements whose lower bounds exceed the search threshold r. However, storing all the distances between the elements $u \in \mathbb{X}$ and the permutants $p \in \mathbb{P}$ is expensive.

We also know that a good pivot for estimating the distance from u to q is some element similar to u; so, we decided to use the permutant closest to u as its pivot. Then, we already have the pivot identifier and we only need to store the distance to it. Therefore, we only need one distance for each object $u \in \mathbb{X}$, which implies that we keep exactly n extra distances in the index, which is negligible for the index size.

Continuing with our example in Figure 1, the closest permutant of object u is p_5, so, we use it as pivot and store the distance 2.

3.3 Permutation similarity measure

Given a query q, we need to calculate all the distances between q and the permutants in $\mathbb{P}$ to obtain Π_q. At this point, we have the complete query permutation (with length k) and the distances $D(q, \mathbb{P})$. Thus, it is possible to compare any clipped permutation Π'_u with Π_q, using the same Equation 2 proposed in [7]:

$$F^*(u, q)_{m_u} = F^*(\Pi'_u, \Pi_q) = \sum_{i=1}^{m_u} |i - \Pi_q^{-1}(\Pi_u(i))| \qquad (2)$$

If all the clipped permutations had the same size, we could directly use Equation 2, as it computes a similarity measure that is fair when all the permutation prefixes have the same size. However, this is not the case, thus we have to readapt the similarity measure considering different sizes of prefixes.

This adaptation corresponds to define mechanisms to apply penalties when we find a permutant that does not belong to Π'_u and, analogously, when we miss a permutant from the prefix of Π_q. Fortunately, they also improve the discriminability of our proposal, as can be seen in Section 4.

Penalty when a permutant does not belong to Π'_u Each permutant clipped from Π'_u adds a penalty that considers how big is the displacement of the remaining permutants in Π'_u with respect to their positions in Π_q. We call the maximum of all these displacements $maxi$. So, we add $maxi \cdot (k - m_u)$ to F^*.

Note that if the permutants in Π'_u are placed in the prefix of Π_q, this penalty is very mild. Also, the penalty increases as long as displacements are bigger.

Penalty when missing a permutant from the prefix of Π_q Two permutations starting with the same permutants give a strong suggestion that the respective objects could be similar. Likewise, if some of the permutants in the prefix of Π_q does not belong to Π'_u, we have a strong indicator that object u could be irrelevant to the query q.

So, we need to establish a criterion about what is this prefix. Analogously to Π'_u, considering the query radius r and using $D(q, \mathbb{P})$, we compute how many permutants have their distances from q within threshold $2r$. This value is called m_q, so the prefix of m_q permutants is called Π'_q. Notice that, if $m_q < m_{\min}$ then m_q is set to $m_{\min}$. Otherwise, if $m_q > m_{\max}$ then m_q is set to $m_{\max}$.

Therefore, we determine how many permutants in Π'_q are missing in Π'_u. We call this value c. Finally, as this is a strong indicator that u is not relevant to q, we strongly penalize the measure with $c \cdot F^*$. Of course, if all the permutants in Π'_q occur in Π'_u, this term is zero. But, the more the number of missing permutants in Π'_u, the greater the penalty (and each increment is also very strong).

Resulting permutation similarity measure We use Equation 2 and these two penalties in order to compute the permutation similarity measure. The obtained measure is depicted in Algorithm 1.

The variable t accumulates the similarity measure, c is initialized as m_q so we start by assuming that we miss all the permutant in Π'_q, and $maxi$ is initialized as zero. Then, we compute a **for** cycle to review all the permutants in Π'_u (Lines 3 throu 9). Line 4 computes the displacement Δ_i for each permutant in Π'_u and accumulate it in t. Line 5 updates the value of $maxi$, when the displacement increases. In Line 6, we verify whether the permutant $\Pi_u(i)$ belongs to the prefix Π'_q, in whose case, we decrease c by 1 (Line 7), as we found another permutant within Π'_q. Finally, in Line 10 we apply the penalties and return the permutation similarity measure.

Algorithm 1 distanceBetweenPermutations($\Pi_q^{-1}, m_q, \Pi_u, m_u, k$)

```
1: OUTPUT: Reports modified Spearman Footrule.
2: t ← 0, c ← m_q, maxi ← 0
3: for i ← 1 to m_u do
4:    Δ_i ← |i − Π_q^{-1}(Π_u(i))|,   t ← t + Δ_i
5:    maxi ← max(Δ_i, maxi)
6:    if Π_q^{-1}(Π_u(i)) < m_q then
7:       c ← c − 1
8:    end if
9: end for
10: return  t ← t + maxi · (k − m_u) + c · t
```

3.4 Solving similarity queries

Given the dataset $\mathbb{X}$, we chose a subset of k objects at random to compute the permutations. Then, for each object $u \in \mathbb{X}$ we compute its permutation and its clipped version Π'_u, and we store both Π'_u and the distance to the closest permutant in the index.

Given a query q, we compute its permutation Π_q and its prefix Π'_q. Since we have the distance between u and its closest permutant, and we already compute $D(q, \mathbb{P})$ at querying time, then, we can calculate a lower bound of $d(u, q)$ with any permutant. Thus, for each u, $d(u, q)$ is lower bounded by $|d(u, p) - d(p, q)|$, using the closest permutant to u as a pivot. Then, if $|d(u, p) - d(p, q)| > r$ then u can be discarded from the candidate list. Only the non-discarded objects are included in the candidate list. We sort increasingly the candidate list according to our adapted permutation similarity measure. Next, a small portion of this list is traversed and compared with q using the distance function d.

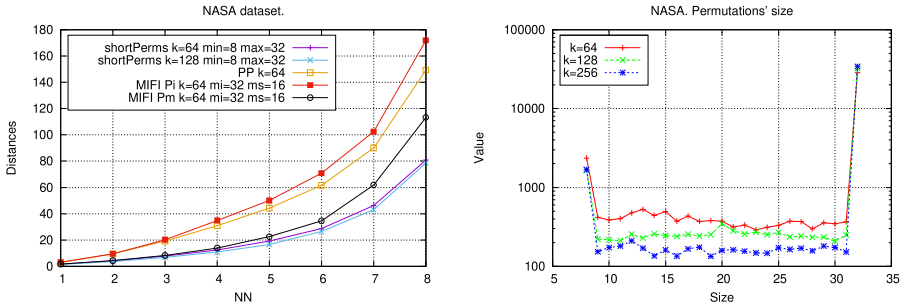
4 Experimental Results

The performance of our proposal has been tested using two classical real-world databases: NASA and Colors. These datasets are available from SISAP project's metric space benchmark set [8]. Any quadratic form can be used as a distance on these spaces, so we chose Euclidean distance as the simplest meaningful alternative for both databases.

4.1 NASA

NASA is a dataset with 20-dimensional vectors. They were generated from images downloaded from NASA[4] and there is not duplicate vectors. The total number of vectors is 40,150. The first 39,650 are indexed and the remaining 500 vectors are used as queries.

[4] Available at `http://www.dimacs.rutgers.edu/Challenges/Sixth/software.html`.



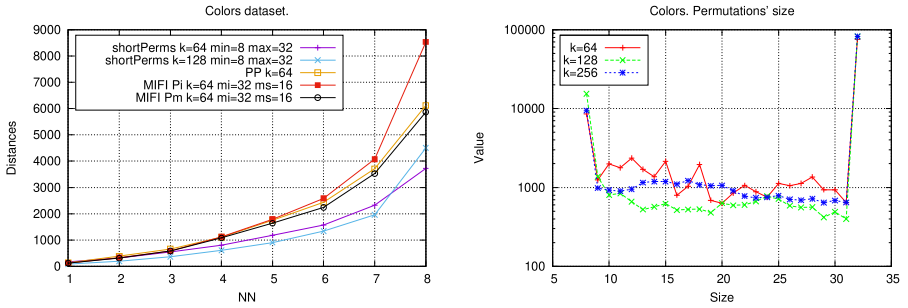
(a) $NN_K(q)$ queries. (b) Histogram of lengths of clipped permut.

Fig. 2. Performance with NASA dataset.

In Figure 2(a), we show the performance of our proposal (shortPerms) along with those of the basic permutation idea (PP, with 64 permutants), the MIFi using m_i for the missing values during the Spearman Footrule computation (MIFI Pi, using similar space of our index), and the MIFi modified as in [7] (MIFI Pm, using similar space of our index). These experiments account for how many distances are needed to obtain the true $NN_K(q)$ answer, varing $K \in [1, 8]$. Notice that our proposal makes 30% fewer distance computations than the best technique proposed in [7] for NN_8. In Figure 2(b), we show the histogram for different lengths of clipped permutations, considering that $m_{\min} = 8$ and $m_{\max} = 32$.

It is remarkable that using 64 permutants we can get clipped permutations with different m_u lengths and smaller than the original size (64), as shown in Figure 2(b). The average length of the clipped permutation is 27. We note that our index having almost half the space of PP with 64 permutants, behaves better. Moreover, when using a more extensive set of permutants (128 concerning 64), the searching costs are almost the same, since the clipped permutations have almost the same size. If we use more permutants, we increase the construction cost of the index. Therefore, it is not worth using more permutants because our proposal always leave only a few permutants that are good ones. In fact, in Figure 2(a) we omit the plot for $k = 256$ as the results are similar to those of $k = 64$ and 128.

During searches, our proposal outperforms the other variants that build the index with the same number of permutants and construction costs. This behavior may be not only due to a good clipped permutation but also to the pruning ability that gives the stored distance to the nearest permutant of each element.



(a) $NN_K(q)$ queries. (b) Histogram of lengths of clipped permut.

Fig. 3. Performance with Colors dataset.

4.2 Colors

This dataset consists of 112,682 color histograms represented as 112-dimensional feature vectors, from an image dataset [5]. Similarly with the NASA dataset, the first elements are indexed and the last 500 color histograms are used as queries.

In Figure 3(a), we show the performance of our proposal. Again, it is compared with the MIFi algorithms and the permutation-based algorithm (PP). In this dataset, our proposal needs 37% fewer distances than the best technique used in [7]. As it occurs in the NASA space, the number of permutants used does not significantly affect the search performance. On Figure 3(b), we show the histogram for each length of clipped permutations. Notice that all permutations have almost the same length for short permutations independent of the original permutation size, again with an average length of 27 permutants.

Newly, if we fix the number of permutants used to build the different alternatives of the indexes whose construction costs are the same, our proposal outperforms the others during searches.

5 Conclusions and Future Work

In this paper, we propose a new strategy for reducing the length of the permutations, which we call *clipped permutations*. We also propose a permutation similarity measure adapted for this clipping. Our approach also takes advantage of storing only one distance per database element, that is well selected, to obtain a lower bound of the distance between the element and the query. This stored distance allows for discarding many elements. This way we can use a smaller hybrid index and at the same time improving the search performance.

We have tested the performance of our proposal with two classical real-world databases: NASA and Colors, obtained from SISAP project's metric space

[5] Available at `http://www.dbs.informatik.uni-muenchen.de/~seidl/DATA/-histo112.112682.gz`.

benchmark set [8]. Our experimental results proved that our approach surpasses the other known permutation-based techniques. Therefore, the combination of these three contributions significantly improves searching performance of a permutation-based index for approximate proximity searching. As it can be noticed, we reduced more than the 30% the distance evaluations needed to solve the queries on both databases.

As future work, we will assess how scalable this method is, considering very large datasets. We will analyze how the values of $m_{\min}$ and $m_{\max}$ affect the lengths of the permutations and in consequence the impact on storage and search performance. Besides, we plan to study the actual prunning ability of the stored distances to the nearest permutant of each database element. Also, we consider investigating how the number of permutants used during the index construction affects the search performance considering other metric spaces.

References

1. Amato, G., Savino, P.: Approximate similarity search in metric spaces using inverted files. In: Proc. 3rd Intl. ICST Conf. on Scalable Information Systems, INFOSCALE 2008, Vico Equense, Italy, June 4-6, 2008. p. 28 (2008)
2. Chávez, E., Figueroa, K., Navarro, G.: Proximity searching in high dimensional spaces with a proximity preserving order. In: Proc. 4th Mexican Intl. Conf. in Artificial Intelligence (MICAI'05). pp. 405–414. LNAI 3789 (2005)
3. Chávez, E., Figueroa, K., Navarro, G.: Effective proximity retrieval by ordering permutations. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) **30**(9), 1647–1658 (2009)
4. Chávez, E., Navarro, G.: A compact space decomposition for effective metric indexing. Pattern Recognition Letters **26**(9), 1363–1376 (2005)
5. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.: Proximity searching in metric spaces. ACM Computing Surveys **33(3)**, 273–321 (2001)
6. Ciaccia, P., Patella, M., Zezula, P.: M-tree: an efficient access method for similarity search in metric spaces. In: Proc. 23rd Conf. on Very Large Databases (VLDB'97). pp. 426–435 (1997)
7. Figueroa, K., Camarena-Ibarrola, A., Reyes, N.: Shortening the candidate list for similarity searching using inverted index. In: Mexican Conf. on Pattern Recognition. vol. 12725, pp. 89–97. LNCS Springer (2021). https://doi.org/10.1007/978-3-030-77004-4_9
8. Figueroa, K., Navarro, G., Chávez, E.: Metric spaces library (2007), available at `http://www.sisap.org/Metric_Space_Library.html`
9. Figueroa, K., Reyes, N., Camarena-Ibarrola, A.: Candidate list obtained from metric inverted index for similarity searching. In: Martínez-Villaseñor, L., Herrera-Alcántara, O., Ponce, H., Castro-Espinoza, F.A. (eds.) Advances in Computational Intelligence. pp. 29–38. Springer International Publishing, Cham (2020)
10. Navarro, G., Paredes, R., Reyes, N., Bustos, C.: An empirical evaluation of intrinsic dimension estimators. Inf. Syst. **64**, 206–218 (Mar 2017). https://doi.org/10.1016/j.is.2016.06.004
11. Skala, M.: Counting distance permutations. J. of Discrete Algorithms **7**(1), 49–61 (Mar 2009). https://doi.org/10.1016/j.jda.2008.09.011
12. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach, Advances in Database Systems, vol. 32. Springer (2006)

An Efficient Dynamic Version of the Distal Spatial Approximation Trees

Edgar Chávez[1], María E. Di Genaro[2], and Nora Reyes[2]

[1] Centro de Investigación Científica y de Educación Superior de Ensenada, México
elchavez@cicese.mx
[2] Departamento de Informática, Universidad Nacional de San Luis, Argentina
{mdigena,nreyes}@unsl.edu.ar

Abstract. Metric space indices make searches for similar objects more efficient in various applications, including multimedia databases and other repositories which handle complex and unstructured objects. Although there are a plethora of indexes to speed up similarity searches, the *Distal Spatial Approximation Tree* (DiSAT) has shown to be very efficient and competitive. Nevertheless, for its construction, we need to know all the database objects beforehand, which is not necessarily possible in many real applications. The main drawback of the DiSAT is that it is a static data structure. That means, once built, it is difficult to insert new elements into it. This restriction rules it out for many exciting applications. In this paper, we overcome this weakness. We propose and study a dynamic version of DiSAT that allows handling lazy insertions and, at the same time, improves its good search performance. Therefore, our proposal provides a good tradeoff between construction cost, search cost, and space requirement. The result is a much more practical data structure that can be useful in a wide range of database applications.

Keywords: similarity search, dynamism, metric spaces, non-conventional databases

1 Introduction

The metric space approach has become popular in recent years to handle the various emerging databases of complex and unstructured objects- On these kinds of databases, it is only meaningfully searching for similar objects [4, 11, 12, 6]. Similarity searches have applications in a vast number of fields. Some examples are non–traditional databases, text searching, information retrieval, machine learning and classification, image quantization and compression, computational biology, and function prediction, among others. These problems can be mapped into a *metric space model* [4] as a metric database. In this model, there is a universe $\mathbb{U}$ of objects, and a non negative real-valued distance function $d : \mathbb{U} \times \mathbb{U} \longrightarrow \mathbb{R}^+ \cup \{0\}$ defined among them. This distance function, called also a metric, satisfies the three axioms that make the pair$\mathbb{U}, d)$ a *metric space*: *strict positiveness* ($d(x, y) \geq 0$ and $d(x, y) = 0 \Leftrightarrow x = y$), *symmetry* ($d(x, y) = d(y, x)$), and *triangle inequality* ($d(x, z) \leq d(x, y) + d(y, z)$). We have a finite *database* $\mathbb{X} \subseteq \mathbb{U}$, $|\mathbb{X}| = n$.

Thereby, "proximity" or "similarity" searching is the problem of looking for objects in a dataset $\mathbb{X}$, similar enough to a given query object $q \in \mathbb{U}$, under a specific distance function. The smaller the distance between two objects, the more "similar" they are. The database can be preprocessed to build a *metric index*; that is, a data structure to speed up similarity searches. There are two typical similarity queries: *range queries* and *k-nearest neighbors queries* [4].

There are a large number of metric indexes for metric spaces [4, 12, 11]. The *Distal Spatial Approximation Tree* (DiSAT) is an index based on dividing the search space and then approaching the query spatially. DiSAT is algorithmically interesting by itself, and it has been shown to give an attractive tradeoff between memory usage, construction time, and search performance. The DiSAT has a significant advantage over other indices because it does not require any parameter tuning. However, DiSAT is a static index; that is, the index has to be rebuilt from scratch, or it requires an expensive updating when we insert a new element into the database.

For several applications, a static scheme may be acceptable. However, many relevant ones do require dynamic capabilities. Actually, in many cases, it is sufficient to support insertions, such as in digital libraries and archival systems, versioned and historical databases, and several other scenarios where objects are never updated or deleted. The *Distal Spatial Approximation Forest* (DiSAF) [2] is a dynamic index, based on the DiSAT. It uses the *Bentley-Saxe method* (BS)[1], that allows to transform a static index into a dynamic one only if on this index the search problem is *decomposable*. However, although the DiSAF admits insertions and DiSAF and DiSAT obtain similar search performance, its construction costs are very high. Therefore, in this paper, we are focused on a new dynamic version of the DiSAT that takes advantage of all our knowledge on the DiSAT and other metric space indexes. This new version significantly reduces the construction costs and obtains better search costs than DiSAT. We are focused only on supporting insertion and range searches, and we left deletions, k-NN searches, and other improvements as future works.

The rest of this paper is organized as follows. In Section 2 we describe some basic concepts. Next, in Section 3 we detail the *Distal Spatial Approximation Trees* (DiSAT), the *Distal Spatial Approximation Forest*, and some notions of their close relatives: *Spatial Approximation Trees* (SAT) and the *Dynamic Spatial Approximation Trees* (DSAT). Section 4 introduces our new dynamic version of DiSAT. In Section 5 we experimentally evaluate the performance of our proposal. Finally, we draw some conclusions and future works in Section 6.

2 Previous Concepts

The metric space model can be formalized as follows. Let $\mathbb{X}$ be a universe of *objects*, with a nonnegative *distance function* $d : \mathbb{X} \times \mathbb{X} \longrightarrow \mathbb{R}^+$ defined among them. This distance satisfies the three axioms that make $(\mathbb{U}, d)$ a *metric space*: strict positiveness ($d(x, y) = 0 \Leftrightarrow x = y$), symmetry ($d(x, y) = d(y, x)$) and triangle inequality ($d(x, z) \leq d(x, y) + d(y, z)$). We handle a finite *dataset* $\mathbb{U} \subseteq \mathbb{X}$, which can be preprocessed (to build an index). Later, given a new object from $\mathbb{X}$ (a *query* $q \in \mathbb{X}$), we must retrieve all similar elements found in $\mathbb{U}$. There are two typical queries of this kind:

Range query: Retrieve all elements in $\mathbb{U}$ within distance r to q. That is, $\{x \in \mathbb{U}, d(x,q) \leq r\}$.

k-nearest neighbors query (k-NN): Retrieve the k closest elements to q in $\mathbb{U}$. That is, a set $A \subseteq \mathbb{U}$ such that $|A| = k$ and $\forall x \in A, y \in \mathbb{U} - A, d(x,q) \leq d(y,q)$.

The distance is assumed to be expensive to compute. Hence, it is customary to define the complexity of the search as the number of distance evaluations performed, disregarding other components such as CPU time for side computations and even I/O time. In this scenario, the goal is to preprocess the dataset such that queries can be answered with as few distance evaluations as possible. In this paper, we are devoted to the most basic type of queries; range-queries. k-nearest neighbor queries can be obtained from range queries in an optimal way [7, 8], so we can restrict our attention to range queries.

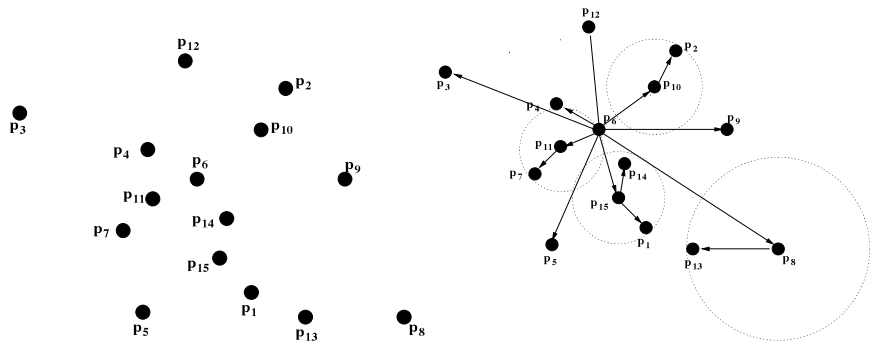
There are a plethora of indexes to speed up similarity searches [11, 12, 4] Algorithms to search in general metric spaces can be divided into two large areas: *pivot-based* and *clustering* algorithms. However, some algorithms combine ideas from both areas.

3 Distal Spatial Approximation Trees

The *Distal Spatial Approximation Tree* (DiSAT) [3] is a variant of the *Spatial Approximation Tree* (*SAT*) [9]. DiSAT and SAT are data structures that use a spatial approximation approach. They are iteratively getting closer to the query by starting at the root as navigating the tree. The DiSAT is built as follows. An element a is selected as the root, and it is connected to a set of *neighbors* $N(a)$, defined as a subset of elements $x \in \mathbb{X}$ such that x is closer to a than to any other element in $N(a)$. The other elements (not in $N(a) \cup \{a\}$) are assigned to their closest element in $N(a)$. Each element in $N(a)$ is recursively the root of a new subtree containing the elements assigned to it. For each node a the covering radius, the maximum distance $R(a)$ between a and any element in the subtree rooted at a, is stored. The starting set for neighbors of the root a, $N(a)$ is empty. Therefore we can select *any* database element as the first neighbor. Once this element is fixed, the database is split into two halves by the hyperplane defined by proximity to a and the recently selected neighbor. Any element in the a side can be selected as the second neighbor. While the root zone (those database elements closer to the root than the previous neighbors) is not empty, it is possible to continue with the subsequent neighbor selection.

The DiSAT tries to increase the separation between hyperplanes, which in turn decreases the size of the covering radius, the two parameters governing the performance of these trees. The performance improvement consists in selecting distal nodes instead of the proximal nodes selected in the original algorithm. Considering an example of a metric database illustrated in Fig. 1(a) and Fig. 1(b) shows the DiSAT obtained by selecting p_6 as the tree root. We depict the covering radii for the neighbors of the tree root. It is possible to obtain completely different trees (DiSATs) if we select different roots, and each tree may have different search costs.

Algorithm 1 gives a formal description of the construction of DiSAT. Range searching is done with the procedure described in Algorithm 2. This process is invoked as



(a) Example of a space. (b) DiSAT.

Fig. 1. Example of a metric database in $\mathbb{R}^2$, and DiSAT obtained if p_6 were the root.

`RangeSearch(`a`,`q`,`r`,`$d(a,q)$`)`, where a is the tree root, r is the radius of the search, and q is the query object. One key aspect of DiSAT is that a greedy search will find all the objects previously inserted. For a range query of q with radius r, and being c the closest element between $\{a\} \cup N(a) \cup A(a)$ and $A(a)$ the set of the ancestors of a, the same greedy search is used entering all the nodes $b \in N(a)$ such that $d(q,b) \leq d(q,c) + 2r$ because any element $x \in (q,r)_d$, can differ from q by at most r at any distance evaluation, so it could have been inserted inside any of those b nodes [12, 9]. In the process, all the nodes x founded close enough to q are reported.

3.1 Distal Dynamic Spatial Approximation Forest

The Bentley-Saxe method (BS) allows transforming a static index into a dynamic one if on this index the search problem is *decomposable*, based on the binary representation of the integers [1]. The Distal Spatial Approximation Forest (DiSAF) [2] applies the BS method to a DiSAT to transform it into a dynamic one. In this case, we use the BS method to have several subtrees T_i, particularly DiSATs. For this reason, this index is called as *Distal Dynamic Spatial Approximation Forest* (DiSAF), because we have a *forest* of DiSATs. Each subtree T_i into the DiSAF is a DiSAT in the forest that will have 2^i elements.

Considering the example illustrated in Fig. 1(a), the Fig. 2 illustrates the DiSAF obtained by inserting the objects $p_1, \cdots, p_{15}$ one by one, in this order. As we have 15 elements, DiSAF will build four DiSATs: T_0, T_1, T_2, and T_3. The final situation will have: T_0 with the dataset $\{p_{15}\}$, T_1 with $\{p_{13}, p_{14}\}$, T_2 with $\{p_9, \ldots, p_{12}\}$, and T_3 with $\{p_1, \ldots p_8\}$. We depict the covering radii for the neighbors of the tree roots; some

Algorithm 1 Process to build a DiSAT for $\mathbb{U} \cup \{a\}$ with root a.

```
BuildTree (Node a, Set of nodes U)
1.  N(a) ← ∅     /* neighbors of a */
2.  R(a) ← 0     /* covering radius */
3.  For v ∈ U in increasing distance to a Do
4.     R(a) ← max(R(a), d(v, a))
5.     If ∀b ∈ N(a),  d(v, a) < d(v, b) Then
6.        N(a) ← N(a) ∪ {v}
7.  For b ∈ N(a) Do S(b) ← ∅
8.  For v ∈ U − N(a) Do
9.     c ← argmin_{b∈N(a)} d(v, b)
10.    S(c) ← S(c) ∪ {v}
11. For b ∈ N(a) Do BuildTree (b, S(b))
```

Algorithm 2 Searching of q with radius r in a DiSAT with root a.

```
RangeSearch (Node a, Query q, Radius r, Distance d_min)
1. If d(a, q) ≤ R(a) + r Then
2.    If d(a, q) ≤ r Then Report a
3.    d_min ← min {d(c, q),  c ∈ N(a)} ∪ {d_min}
4.    For b ∈ N(a) Do
5.       If d(b, q) ≤ d_min + 2r Then
6.          RangeSearch (b, q, r, d_min)
```

covering radii are equal to zero. As the DiSAF has not any parameter, the only way to obtain different forests is by considering different insertion orders.

Dynamic Spatial Approximation Tree

The *Dynamic Spatial Approximation Tree* (DSAT) [10] is an online version of the SAT. It is designed to allow dynamic insertions and deletions without increasing the construction cost for the SAT. An astounding and unintended feature of the DSAT is boosting the searching performance. The DSAT is faster in searching even if, at construction, it has less information than the static version of the index. For the DSAT, the database is unknown beforehand, and the objects arrive at the index at random and the queries. A dynamic data structure cannot make strong assumptions about the database and will not have statistics about all of the database.

4 Dynamic Distal Spatial Approximation Trees

As we mentioned, the DiSAT is a static index that must be rebuilt from scratch or requires an expensive updating when we insert a new element into the database. On the other hand, DiSAF allows to insert elements and obtains a similar search performance as DiSAT, but its construction costs are very high because each insertion has to rebuild some subtrees. Therefore, using our deep knowledge of DiSAT and its relatives

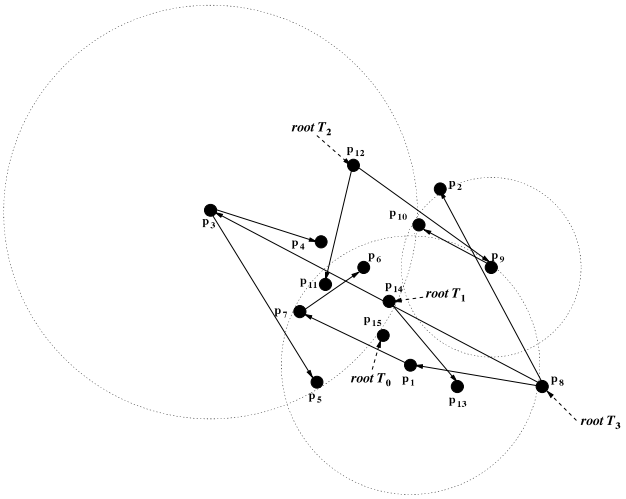


Fig. 2. Example of the DiSAF, inserting from p_1 to p_{15}.

and also taking advantage of storing one distance per element, we propose a new dynamic version of DiSAT that can be built by inserting the elements individually. The *Dynamic Distal Spatial Approximation Tree* (DDiSAT) reduces the construction costs significantly with respect DiSAF and obtains better search performance than DiSAT.

We want to avoid reconstruction at each insertion to reduce construction costs. Therefore, we consider using lazy insertions; we need to ensure that several insertions do not need to do any rebuilding and that only some of them require rebuilding the index. Each DDiSAT node can store an element a, its covering radius $rc(a)$, its set of neighbors $N(a)$, and a bag $B(a)$ of pairs of (element, distance), that are new elements into the database and the distance is its distance from a. The main idea is only to rebuild the DDiSAT when the new insertion in a bag makes the number of elements in the bags (pending insertion in the DiSAT) equal to the number of nodes in the DDiSAT. The above means, the DDiSAT reaches twice of the original elements inside its nodes. We have two cases to consider during insertions into the DDiSAT:

- If the DDiSAT has i nodes and less to i elements into their bags, we insert the new element x into a node bag and do not need to rebuild the DDiSAT.
- Otherwise, we retrieve all the elements into the DDiSAT (in nodes and bags), and we rebuild the tree as a DiSAT.

Therefore, most of the insertions will proceed as follows. When we insert a new element x into the DDiSAT, we search its insertion point. This search begins at the tree root. At any DDiSAT node, let be b its element, if b is closer to x than any neighbors in $N(b)$ we insert the pair $(x, distance(b, x))$ into the bag $B(b)$ of this node. Otherwise, we go down by the node of the nearest element to x in $N(b)$. As the new element x insertion goes down through the tree nodes, we have to update the covering radii. This

Algorithm 3 Searching of q with radius r in a DDiSAT with root a.

```
RangeSearch (Node a, Query q, Radius r, Distance d_min)
1. If d(a,q) ≤ R(a) + r Then
2.    If d(a,q) ≤ r Then Report a
3.    For any pair (x, d_x) ∈ B(a)
4.       If |d(a,q) − d_x| ≤ r Then
5.          If (d(x,q) ≤ r Then Report x
6.    d_min ← min {d(c,q), c ∈ N(a)} ∪ {d_min}
7.    For b ∈ N(a) Do
8.       If d(b,q) ≤ d_min + 2r Then
9.          RangeSearch (b, q, r, d_min)
```

way, we avoid several rebuilding of the tree and ensure to do not degrade the search performance. As it can be observed, as the DDiSAT grows in elements, the number of insertions needed to double the number of elements also increases. Thus the reconstructions will be more sporadic. However, they will involve more elements.

During searches, we take advantage of all the information from the tree. As in a search on a DiSAT (Algorithm 2), we also use the distances stored in the buckets. The Algorithm3 illustrates the new search process. This process is invoked as `RangeSearch` $(a, q, r, d(a,q))$, where a is the tree root, r is the radius of the search, and q is the query object.

5 Experimental Results

For the empirical evaluation of the indices, we consider three widely different metric spaces from the SISAP Metric Library (`www.sisap.org`) [5].

Dictionary: a dictionary of 69,069 English words. The distance is the *edit distance*, the minimum number of character insertions, deletions, and substitutions needed to equal two strings. This distance is useful in text retrieval to cope with spelling, typing, and optical character recognition (OCR) errors.

Color Histograms: a set of 112,682 8-D color histograms (112-dimensional vectors) from an image database[3]. Any quadratic form can be used as a distance; we chose Euclidean as the simplest meaningful distance.

NASA images: a set of 40,700 20-dimensional feature vectors, generated from images downloaded from NASA[4]. The Euclidean distance is used.

When we evaluate construction costs, we build the index with the complete database. If the index is dynamic, the construction is made by inserting, one by one, the objects. Otherwise, the index knows all the elements beforehand. To evaluate the search performance of the indexes, we build the index with the 90% of the database elements and we use the remaining 10%, randomly selected, as queries. So, the elements used as query

[3] At `http://www.dbs.informatik.uni-muenchen.de/~seidl/DATA/histo112.112682.gz`

[4] At `http://www.dimacs.rutgers.edu/Challenges/Sixth/software.html`

objects are not in the index. We average the search costs of all these queries. All results are averaged over 10 index constructions with different datasets permutations.

We consider range queries retrieving on average 0.01%, 0.1%, and 1% of the dataset. This corresponds to radii 0.051768, 0.082514 and 0.131163 for the Color Histograms; and 0.605740, 0.780000 and 1.009000 for the NASA images. The Dictionary has a discrete distance, so we used radii 1 to 4, which retrieved on average 0.00003%, 0.00037%, 0.00326%, and 0.01757% of the dataset, respectively. The same queries were used for all the experiments on the same datasets. As we mentioned previously, given the existence of range-optimal algorithms for k-nearest neighbor searching [7, 8], we have not considered these search experiments separately.

We show the comparison between our dynamic DDiSAT, the DiSAF and the DSAT, and the static alternatives SAT and DiSAT. The source code of the different SAT versions (SAT and DSAT) is available at`www.sisap.org`. A final note in the experimental part is the arity parameter of the *DSAT* which is tunable and is the maximum number of neighbors of each tree node. In our experiments, we used the arity suggested by authors in [10]: the best arity for the NASA images and for Color histograms is of 4, and arity 32 for the Dictionary. Figure 3 illustrates the construction costs of all indices on the three metric spaces. As it can be seen, DDiSAT surpasses DiSAF on construction costs. On the other hand, DSAT does not make any reconstruction while it builds the tree via insertions. It must be considered that SAT and DiSAT are built with all the elements known simultaneously, not dynamically.

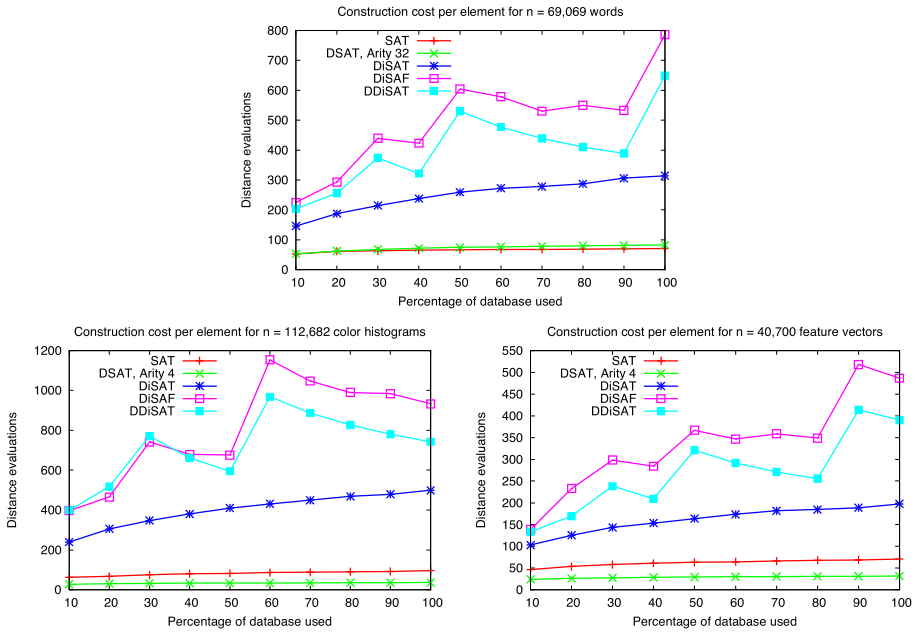


Fig. 3. Construction costs for the three metric spaces considered.

We analyze search costs in Figure 4. As can be noticed, DDiSAT surpasses the dynamic indexes DiSAF and DSAT in all the spaces. Moreover, DDiSAT obtains the best search performance concerning the other four indexes (static and dynamic ones). Therefore, we can affirm that the heuristic of construction of DiSAT allows surpassing in searches the other strategies used in SAT and DSAT, and combining it with the bags into the nodes that store new elements near them, it is possible to obtain even better results. Besides, we have obtained a dynamic index that overcomes DiSAT at searches. Moreover, DDiSAT has the advantage over DSAT, which does not have any parameters to tune.

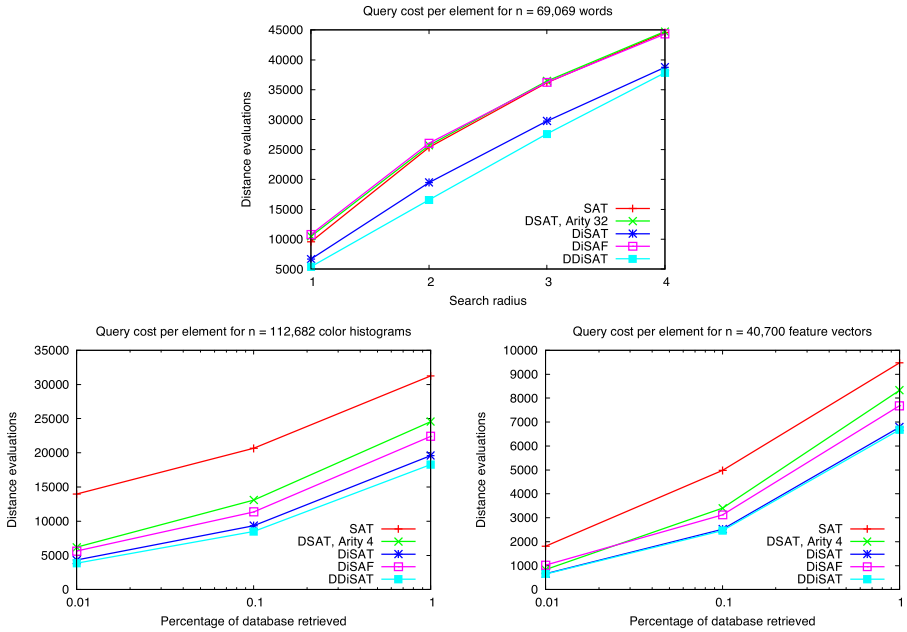


Fig. 4. Search costs for the three metric spaces considered.

6 Conclusions

We have presented a new dynamic version of the DiSAT, which at this time can handle insertions and improve its search quality. As we mentioned, there are few data structures for searching metric spaces that are dynamic and efficient. Furthermore, we have shown that we can take advantage of the heuristic used in DiSAT even more. As the distal nodes produce more compact subtrees, which in turn give more locality to the underlying partitions implicitly defined by the subtrees, we can use these partitions over the metric space to assign each new element to its closest object in the tree while it is waiting to be actually inserted as a DiSAT node.

The DiSAT was a promising data structure for metric space searching, with several drawbacks that prevented it from being practical: high construction cost and inability to accommodate insertions and deletions. We have addressed some of these weaknesses. We have obtained reasonable construction costs, and it is still possible to improve it. For example by providing a bulk-loading algorithm to initially create the DDiSAT if we know aforehand a subset of elements, avoiding some unnecessary rebuildings when we insert elements and combining with *lazy insertion* that do not always rebuild trees.

In future works, we consider making the DDiSAT fully dynamic; that is, supporting deletions and designing an efficient bulk-loading algorithm, which allows for reducing more the insertion costs. We also consider to design an efficient alternative of k-NN search that applies a smart solution by taking advantage of all distances calculated in order to shrink, as soon as possible, the radius from q that encloses k elements.

References

1. Jon L. Bentley and James B. Saxe. Decomposable searching problems i. static-to-dynamic transformation. *Journal of Algorithms*, 1(4):301–358, 1980.
2. Edgar Chávez, María E. Di Genaro, Nora Reyes, and Patricia Roggero. Decomposability of disat for index dynamization. *Journal of Computer Science and Technology*, pages 110–116, 2017.
3. Edgar Chávez, Verónica Ludeña, Nora Reyes, and Patricia Roggero. Faster proximity searching with the distal sat. *Information Systems*, 2016.
4. Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.
5. Karina Figueroa, Gonzalo Navarro, and Edgar Chávez. Metric spaces library, 2007. Available at `http://www.sisap.org/Metric_Space_Library.html`.
6. Magnus Hetland. The basic principles of metric indexing. In Carlos Coello, Satchidananda Dehuri, and Susmita Ghosh, editors, *Swarm Intelligence for Multi-objective Problems in Data Mining*, volume 242 of *Studies in Computational Intelligence*, pages 199–232. Springer Berlin / Heidelberg, 2009.
7. Gísli R. Hjaltason and Hanan Samet. *Incremental Similarity Search in Multimedia Databases*. Number CS-TR-4199 in Computer science technical report series. Computer Vision Laboratory, Center for Automation Research, University of Maryland, 2000.
8. Gísli R. Hjaltason and Hanan Samet. Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems*, 28(4):517–580, 2003.
9. Gonzalo Navarro. Searching in metric spaces by spatial approximation. *The Very Large Databases Journal (VLDBJ)*, 11(1):28–46, 2002.
10. Gonzalo Navarro and Nora Reyes. Dynamic spatial approximation trees. *Journal of Experimental Algorithmics*, 12:1.5:1–1.5:68, June 2008.
11. Hanan Samet. *Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.
12. Pavel Zezula, Giussepe Amato, Vlatislav Dohnal, and Michal Batko. *Similarity Search: The Metric Space Approach*, volume 32 of *Advances in Database Systems*. Springer, 2006.

XVII Workshop Arquitectura, Redes y Sistemas Operativos **(WARSO)**

Coordinadores

Carlos Buckle (UNPSJB)

Marcelo Arroyo (UNRC)

Jorge Ardenghi (UNS)

Estudio Experimental del Comportamiento de Métricas de QoS y QoE de Streamings de Video Multicast IPTV

Santiago Pérez[1], Higinio Facchini[1], Pablo Varela[1], Bruno Roberti[1], Alejandro Dantiacq[1], Fabián Hidalgo[1], María Stefanoni[1], Matilde Césari[1]

[1] CeReCoN – Departamento de Electrónica – Facultad Regional Mendoza – UTN
Rodríguez 273, Ciudad Mendoza
CP (M5502AJE) República Argentina
{santiagocp, higiniofac}@frm.utn.edu.ar

Abstract. Digital television is the most important advance in television technology. IPTV describes a mechanism for transporting a stream of video content over a network that uses the IP network protocol. IP Television must be a totally personalized experience that must guarantee Quality of Service (QoS), and QoE (Quality of Experience) within the organization, including the LAN Networks of a Television Channel, or the traditional LAN Networks. In the present research work, the behavior of IPTV traffic was analyzed in an experimental LAN network with controlled traffic. Different codecs were used for contrast, and detailed quantitative results of QoS metrics and, from them, indicative QoE values were established. The findings provide guidance on suitable software and network topology configurations for managing similar networks and provide detailed values for simulation analysts.

Keywords: IPTV, Multicast traffic, Codecs, QoS, QoE

1 Introducción

En términos generales, IPTV es una definición que aplica a la entrega de canales de televisión tradicional, entre otras cosas películas y contenido de video a demanda, a través de una red de tipo privada. Desde la perspectiva de un proveedor de servicio, IPTV abarca la adquisición, procesamiento y entrega segura de contenido de video, a través de una infraestructura de red basada en el protocolo IP. Desde la perspectiva del usuario final, IPTV se ve y funciona como un servicio estándar de televisión pago.

La definición oficial aprobada por la Unión Internacional de Telecomunicaciones sobre IPTV (ITU-T FG IPTV) es la siguiente: IPTV se define como servicios multimedia como televisión, video, audio, texto, gráficos y datos entregados a través de redes basadas en IP, gestionados para proporcionar el nivel de calidad requerido de servicio y experiencia, seguridad, interactividad y confiabilidad. En la transmisión del formato audiovisual IPTV debe entenderse que los receptores serán dispositivos tales como tabletas, notebooks, smartphones, computadoras, televisores, etc. Y que hay que diferenciar este servicio de otros en línea, tales como canales de internet gratuitos, o de tipo "Youtube", en los que los videos se pueden recargar y mirar sin garantías de calidad.

En este trabajo se describe un experimento de tráfico de video IPTV multicast en una red LAN de laboratorio real como test bed, emulando una Red LAN IPTV de un canal de televisión, o de Campus LAN. Se usó una topología de red experimental, de tráfico controlado, con clientes alámbricos e inalámbricos, usando el Software FFmpeg Server, como servidor de video, y el analizador de tráfico WireShark. El tráfico de video IPTV se codificó en H.264, H.265, VP8 y Theora, para contrastar y comprender el impacto de distintos codecs sobre la QoS y QoE del tráfico de la red. Los experimentos se realizaron usando un tráiler de video de la película Star Trek. Este trabajo es una continuación de experimentaciones realizadas sobre redes cableadas Ethernet y Wi-Fi, para tráfico de video general (no IPTV) con codecs H261, H263 y H264 en IPv4.

El resto de este documento se estructura de la siguiente manera: la sección 2 Protocolos y Codecs describe las principales características de estos componentes utilizados en el estudio experimental; la sección 3 Escenarios y Recursos Experimentales describe la topología, y los dispositivos hardware y software del ensayo; la sección 4 Resultados de la experimentación describe cuantitativamente los valores obtenidos de las diferentes métricas de QoS y QoE; y, finalmente, en la sección 5 se plantean las conclusiones del presente trabajo.

2 Protocolos y Codecs

En este apartado se plantea en forma resumida las principales características de los protocolos y codecs utilizados.

2.1 Multidifusión IP

La multidifusión IP (multicast IP) es una tecnología de conservación de ancho de banda que reduce el tráfico, porque entrega simultáneamente una sola secuencia de información a los millares de destinatarios corporativos y a los hogares.

Algunas de las ventajas de una solución multicast de distribución de contenidos son:

- Escalabilidad: ya que los requisitos de ancho de banda ya no son proporcionales al número de receptores.
- Rendimiento: debido a que el procesamiento de un único flujo de datos por fuente siempre será más eficiente que procesar un flujo por receptor.
- Menor gasto de capital: debido a la relajación en las especificaciones de los elementos de red necesarios para proporcionar el servicio.

Los routers (enrutadores) emplean protocolos multidifusión que construyen árboles de distribución para transmitir el contenido multidifusión, que aseguran la mayor eficiencia para el envío de datos a múltiples receptores. En IP se utilizan protocolos como PIM-SM, PIM-SSM u otros. Para nuestro trabajo se utilizó PIM-SM (Protocol Independent Multicast – Sparse Mode).

2.2 Codec de video

La codificación de video se refiere al proceso de convertir video sin formato a un formato digital que sea compatible con muchos dispositivos. Para reducir un video a un tamaño más manejable, los distribuidores de contenido utilizan una tecnología de compresión de video llamada codec. Esto se realiza mediante métodos sofisticados con pérdida, durante los cuales se descartan los datos innecesarios. Un codec actúa sobre el video, tanto en la fuente para comprimirlo, como antes de la reproducción para descomprimirlo.

Para la experimentación se seleccionó una combinación equilibrada de estándares ya establecidos y algunos más novedosos. Los utilizados en este trabajo son:

- H.264/MPEG-4 AVC: Es una norma promovida conjuntamente por la UIT y la ISO, que ofrece un gran avance significativo en la eficiencia de compresión para lograr una reducción de alrededor de 2 veces en la velocidad de bits, en comparación con MPEG-2 y MPEG-4 de perfil simple.
- H.265/ MPEG-H Parte2/ High Efficiency Video Coding (HEVC): Define un formato de compresión de video, sucesor de H.264/MPEG-4 AVC, desarrollado conjuntamente por la ISO/IEC Moving Picture Experts Group (MPEG) y ITU-T Video Coding Experts Group (VCEG), como ISO/IEC CD 23008-2 High Efficiency Video Coding. Este estándar puede utilizarse para proporcionar mejor calidad de videos de bajo bitrate con la misma tasa de datos. Es compatible con la televisión en ultra-alta definición y resoluciones hasta 8192x4320.
- VP8: Es un códec de video de On2 Technologies anunciado el 13 de septiembre de 2008. El 19 de mayo de 2010, Google, que adquirió On2 Technologies en 2009, liberó el códec VP8 como código abierto (bajo una licencia permisiva similar a la licencia BSD).
- Theora: Es un códec de vídeo libre desarrollado por la Fundación Xiph.Org, como parte de su proyecto Ogg. Basado en el códec VP3. Google, en 2010, empezó a financiar parte del proyecto de Ogg Theora Vorbis. Theora es un códec de vídeo de propósito general con bajo consumo de CPU.

3 Escenarios y recursos experimentales

En la comunidad científica-tecnológica, existe diversas líneas de investigación dedicadas al estudio del tráfico de video IPTV, basadas en trabajos experimentales realizados sobre redes reales, que muestran el comportamiento de cada caso [1-11]. Los análisis se realizan con la captura de tráfico en escenarios de tráfico de video real y/o sintético. Lamentablemente, los trabajos de investigación, relacionados con la temática de IPTV LAN para Canales de Televisión o Redes LAN tradicionales, muestran, en general, una falta de uniformidad de los escenarios de experimentación, en la cantidad y tipos de codecs, en los videos utilizados, etc. Estos y otros aspectos complican, en su conjunto, los contrastes entre trabajos contemporáneos entre sí, y los realizados previamente. Esta ha sido la principal motivación para proponer un nuevo escenario y, en consecuencia, una nueva metodología de experimentación para la captura del tráfico de video IPTV.

3.1 Escenario de experimentación

La topología utilizada es una computadora funcionando como servidor de streaming, y computadoras de escritorio (PCs) como clientes, todas conectadas en los extremos de una red mixta conformada por routers y switchs, con distintos tipos de enlaces interconectando a los mismos. En esta topología, los enlaces son del tipo FastEthernet, con una velocidad de transmisión de 100 Mbps. Para el funcionamiento entre los routers R1 a R6 se configuró el protocolo de enrutamiento OSPF v2. Para los mismos routers se configuró el protocolo de enrutamiento multicast PIM de modo SM. Los enlaces redundantes existentes se plantearon para aproximar una red real, pero se configuró el protocolo de ruteo, para que el tráfico entre el servidor y cada cliente siga siempre una única ruta. La Figura 1 muestra la topología de trabajo.

- Como servidor: una computadora de escritorio, con CPU Intel Core I5, con 8 GB de RAM, y sistema operativo Linux Ubuntu.
- Como clientes: computadoras de escritorio CPU AMD Athlon(tm) II X2 250, a 3 GHz, con 4 GB de RAM, y sistema operativo Windows 10 de 64 bits,
- Los routers R1, R2, R3 y R4 fueron modelo Cisco 2811, y los routers R5 y R6 fueron resueltos con switchs multicapa Cisco WS-CS3750.
- Finalmente, para la conexión de los routers a las PCs se usaron switchs Cisco Layer 2 Catalyst Model WS-2950-24.

El software utilizado como servidor de streaming es FFmpeg [12]. El framework consta de una base de componentes, que interactúan con la aplicación a través de comandos ffmpeg, para completar los procesos de streaming de manera correcta.

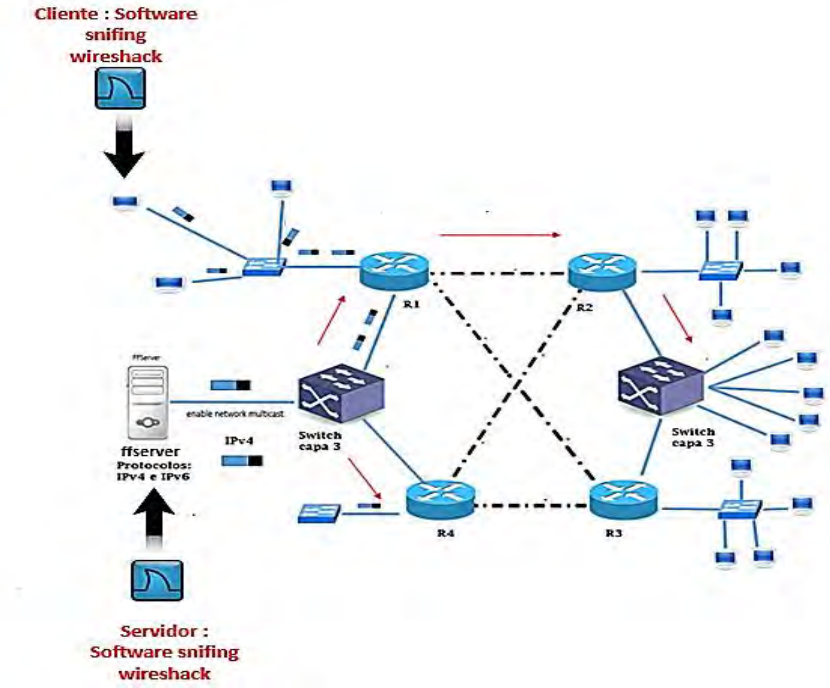


Fig. 1. Topología de red utilizada.

3.2 Video y sniffer utilizado

Se utilizó un archivo de video codificado con los codecs mencionados previamente. Se trata de un tráiler de la película Star Trek [13] de 2 minutos 11 segundos de duración. La Tabla 1 muestra la comparación de las características propias de cada códec para este video.

Wireshark [14] fue la aplicación utilizada para la medición y captura de tráfico. Este software comprende una serie de características que sirven para analizar cada uno de los paquetes de datos, y así mismo, hacer una evaluación de conjunto de paquetes correspondientes a la transmisión del video. Para el trabajo de experimentación se evaluaron ciertos parámetros de tráfico de red, como el retardo, la cantidad de paquetes transmitidos, etc.

A partir del escenario de experimentación, se realizaron una serie de ensayos con las siguientes consideraciones:

a) Se configuró el archivo de video codificados en uno de los 4 formatos mencionados en el servidor de streaming,
b) Antes de comenzar las mediciones, se sincronizaron en tiempo todos los equipos involucrados en la topología, usando un servidor NTP local.
c) Desde el servidor se envió el archivo del video en el codec particular a la red en formato multidifusión (multicast).
d) Posteriormente, se cambió el formato de video a los otros tres codecs de video, repitiendo el paso c.

Tabla 1. Propiedades del Trailer Star Trek.

Video	**H.264**	**H.265**	**Theora**	**VP8**
Format	MPEG-4	MPEG-4	Ogg	WebM v2
File size	79,9 MiB	72,3 MiB	83,3 MiB	78,6 MiB
Duration	2 min 11 s	2 min 11 s	2 min 11 s	2 min 11 s
Bit rate mode	Variable	Variable	Variable	Variable
Bit rate	5 109 kb/s	4 620 kb/s	5 329 kb/s	5 028 kb/s
Video				
Format	AVC	HEVC	Theora	VP8
Bit rate	5 011 kb/s	4 514 kb/s	5 010 kb/s	4 721 kb/s
Width [pixeles]	1 280 pixeles	1 280 pixeles	1 280 pixeles	1 280 pixeles
High [pixeles]	528 pixeles	528 pixeles	528 pixeles	528 pixeles
Aspect ratio	2,4:1	2,4:1	2,4:1	2,4:1
Frame rate mode	constant	constant	constant	constant
Frame rate [fps]	23,976 FPS	23,976 FPS	23,976 FPS	23,976 FPS
Bits/(pixel*frame)	0.309	0.279	0.309	0.291
Audio				
Format	AAC LC	AAC LC	Vorbis	Vorbis
Bit rate mode	Variable	constant	Variable	Variable
Bit rate	98,7 kb/s	99,7 kb/s	98,7 kb/s	98,7 kb/s
Maximum bit rate	167 kb/s	167 kb/s	167 kb/s	167 kb/s
Channel	2 canales	2 canales	2 canales	2 canales
Sampling rate	44,1 kHz	44,1 kHz	44,1 kHz	44,1 kHz
Track size	1,54 MiB	1,56 MiB	1,54 MiB	1,54 MiB

Durante cada ensayo, se realizaron capturas de tráfico en el servidor y en cada uno de los clientes, usando el software sniffer Wireshark. Los paquetes se filtraron por tipo de paquetes RTCP o UDP. Para la transferencia del video se necesitaron 59.644, 54.202, 46.502 y 61.986 tramas, para los codecs H.264, H.265, Theora y VP8, respectivamente. Las capturas se exportaron desde archivos tipo .pcap de Wireshark a archivos tipo .csv. Un archivo en el formato .csv es, básicamente, un archivo de texto, que permite guardar la captura como un vector de tramas, para ser analizados bajo Excel o un programa a media (en nuestro caso se utilizó el lenguaje Python).

4 Resultados de la experimentación

4.1 Análisis de QoS

La Calidad de Servicio puede considerarse en base al comportamiento de ciertos parámetros de red, de un tráfico determinado, en una arquitectura de red dada. Los parámetros de calidad de servicio estandarizados son: el retardo, la diferencia de retardos (jitter), y la pérdida de paquetes. Los valores de referencia para estas métricas de red son 100 ms, 30 ms y 0,1%, respectivamente.

En la Figura 2 se pueden apreciar, a los fines comparativos, la métrica de retardo para cada uno de los codecs analizados. Se observa que los valores satisfacen holgadamente los requerimientos de QoS, y qué en la banda de 1 a 2 ms, los codecs presentan un comportamiento claramente diferenciado, a excepción de los codecs H.264 y H.265.5. Estos últimos codecs concentran sus retardos en el orden de los 2 mseg, mientras que el códec Theora loa hace en el orden de los 1,5 mseg. Finalmente, el códec VP8 concentra sus retardos en el orden de 1 mseg. De la misma forma, en la misma Figura 3 se presenta el comportamiento de los distintos codecs bajo estudio para la métrica jitter. Se observa que los valores satisfacen holgadamente los requerimientos de QoS, y que en la banda de los -0,010 a los + 0,020 ms, los codecs presentan un comportamiento claramente diferenciado para esta métrica.

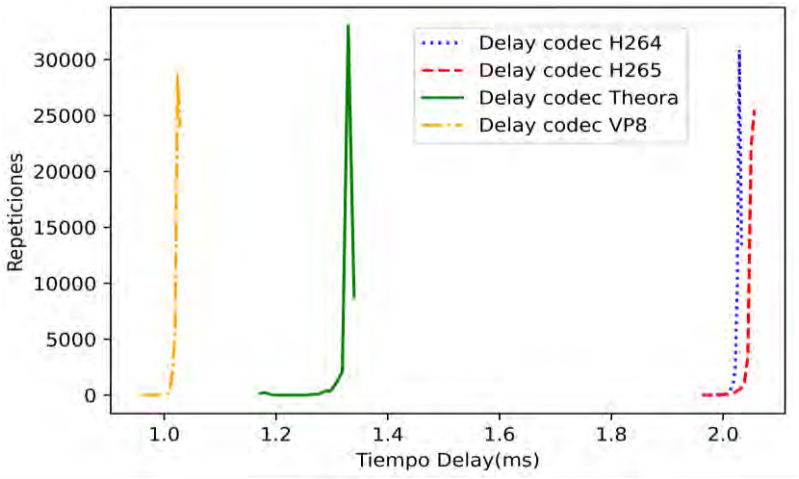


Fig. 2. Se observa el comportamiento solapado de los codecs para la métrica retardo.

Cada códec tiene una respuesta en la forma de triángulo. Sin embargo, se destaca que los códec tienen bases y alturas distintas, lo que implica la amplitud de banda de valores de jitter, y el nivel de repetición con que se concentran la mayor cantidad de valores

En la red LAN de laboratorio, que permitía un análisis de tráfico controlado, no se observaron pérdidas de tramas.

4.2 Análisis de QoE

Los procesos de QoS, por si solos, no son totalmente adecuados para proporcionar una garantía de rendimiento, debido a que no tienen en cuenta la percepción del usuario sobre el comportamiento de la red. Esto condujo a la disciplina emergente de Calidad de Experiencia (QoE, Quality of Experience). La QoE, desde una visión teórica, es el grado de satisfacción o molestia del usuario de una aplicación o servicio. Resulta del cumplimiento de sus expectativas, con respecto a la utilidad/disfrute de la aplicación o servicio, a la luz de la personalidad y el estado actual del usuario.

Para evaluar o medir la calidad de la experiencia percibida por el usuario se han propuesto tres métodos: (i) métodos subjetivos, (ii) métodos objetivos y (iii) métodos indirectos. Los métodos subjetivos están relacionados con la utilización de personas, para evaluar la calidad del video en un ambiente controlado, mediante el uso de encuestas. Los métodos objetivos son algoritmos, que utilizan una señal de referencia completa, parcial o sin utilizar señal de referencia, para medir calidad del video. Por último, están los métodos indirectos, que mediante un modelo matemático evalúan la calidad de experiencia asociada al video. Este modelo matemático es generado básicamente teniendo en cuenta la variación de las métricas de QoS.

Teniendo en cuenta lo mencionado, y en una primera aproximación, la QoE de la red se evaluó por un método indirecto. Con esos datos, y como una continuación de la presente investigación, se perfeccionarán los resultados de QoE con métodos que aporten la subjetividad del usuario.

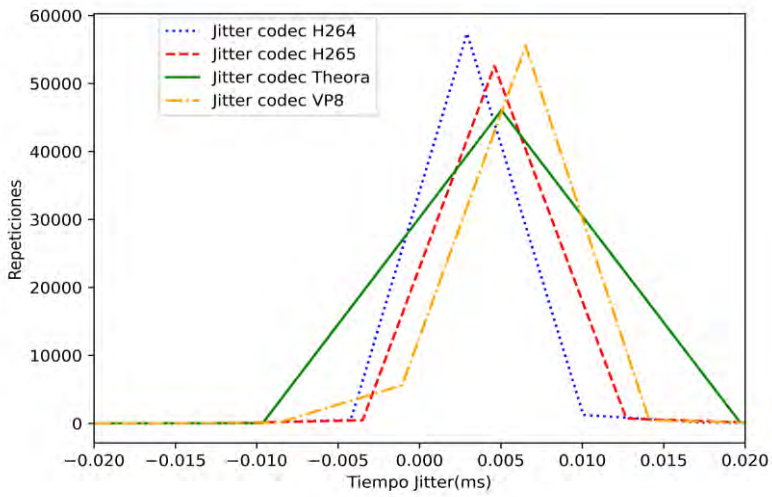


Fig. 3. Se observa el comportamiento solapado de los codecs para la métrica jitter.

Por ello, se adoptó el modelo matemático propuesto en [8], a través de la siguiente expresión:

$$QoE = 1 / (retardo + K * Jitter) * e^{\wedge} (pérdida\ de\ paquetes) \quad (1)$$

donde K permite balancear el impacto del parámetro jitter, con respecto al retardo, para el cálculo de la QoE de los usuarios. Ninguno de los parámetros utilizados en la expresión podría ser negativo. Para la evaluación se utilizó el valor de K=2.

En la Figura 4 se muestra, a los fines comparativos, la QoE para cada uno de los codecs analizados, en función del retardo y usando como parámetros los valores promedio y máximo de jitter, para cada caso. Se observa qué en la banda de 1 a 2 ms, los codecs presentan un comportamiento claramente diferenciado, a excepción de los codecs H.264 y H.265. Estos últimos codecs tienen una peor respuesta para la QoE. El códec VP8 presenta un mejor comportamiento, mientras que el codec Theora se ubica en una respuesta intermedia.

Y en la Figura 5 se presenta, a los fines comparativos, la QoE para cada uno de los codecs analizados, en función del jitter, y usando como parámetros los valores promedio y máximo del retardo, para cada caso. Se observa que en la banda de -0,10 mseg a +0,15 mseg, los codecs presentan un comportamiento claramente diferenciado, a excepción de los codecs H.264 y H.265. Estos últimos codecs tienen una peor respuesta para la QoE. Nuevamente, el códec VP8 presenta un mejor comportamiento, mientras que el codec Theora se ubica en una respuesta intermedia.

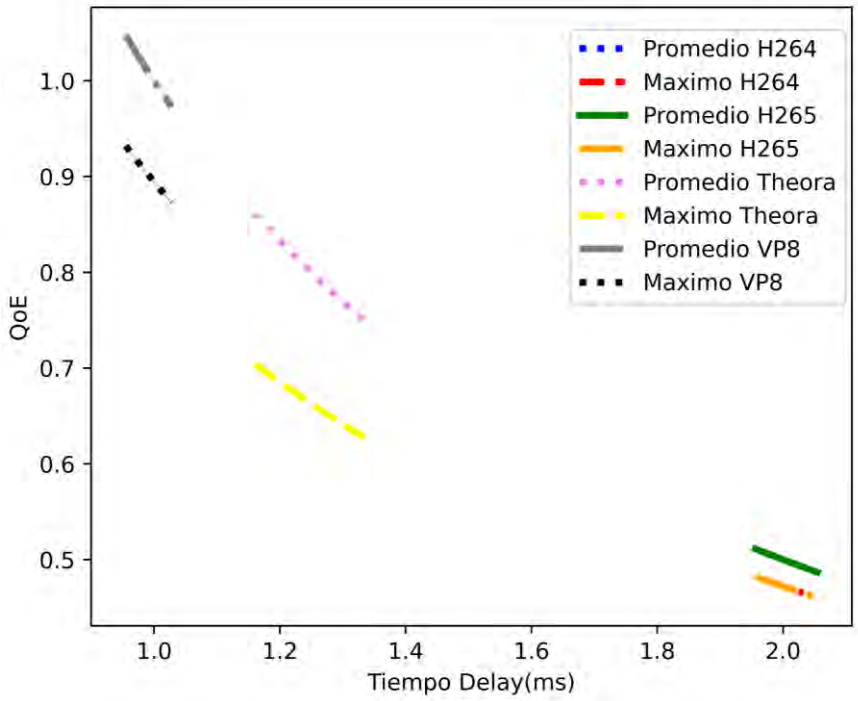


Fig. 4. Comparación de la QoE en función del retardo, para los valores promedio y máximo del jitter para los 4 codecs.

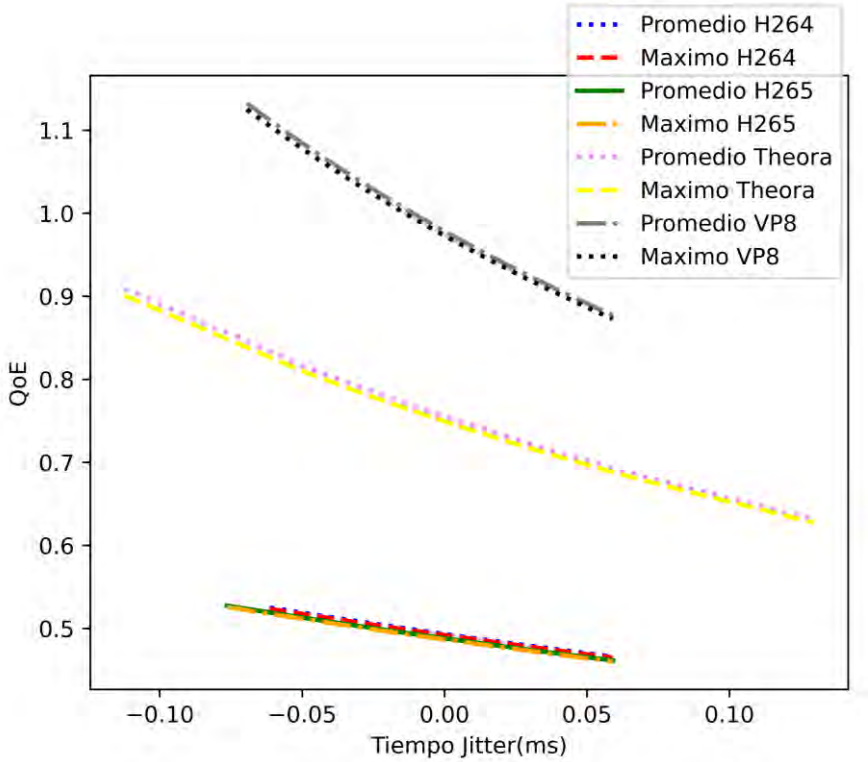


Fig. 5. Comparación de la QoE en función del jitter, para los valores promedio y máximo del retardo para los 4 codecs.

5 Conclusiones

Se analizaron detalladamente los alcances de la QoS y de QoE para el tráfico IPTV sobre una red LAN experimental. Se utilizó una topología de red real de tráfico controlado, un servidor y clientes de video IPTV, un trailer de Star Trek, con 4 subescenarios o casos particulares, por cada uno de 4 codecs: H.264, H.265, Theora y VP8. El tráfico fue capturado para análisis, en todos los casos, usando un sniffer.

Se demostró experimentalmente que:

- La métrica de retardo para tráfico IPTV, en la red experimental, satisface holgadamente los requerimientos de QoS, y presenta un comportamiento claramente diferenciado, a excepción de los codecs H.264 y H.265. Estos últimos codecs concentran sus retardos en el orden de los 2 mseg, mientras que el códec Theora lo hace en el orden de los 1,5 mseg. Finalmente, el códec VP8 concentra sus retardos en el orden de 1 mseg.
- La métrica jitter para tráfico IPTV, en la red experimental, satisface holgadamente los requerimientos de QoS, y presenta un comportamiento claramente diferenciado. Cada códec tiene una respuesta en la forma de triángulo, cuando se

representa la cantidad de valores, en repetición, en función del jitter. Los códec tienen bases y alturas distintas. El codec H.264 (Theora) presenta la mejor (peor) respuesta general.

- La respuesta de QoE, en base a la expresión matemática utiliza, usando el retardo (jitter) como variable independiente, y el jitter (retardo) como parámetro, muestra que los H.264 y H.265 tienen el peor comportamiento. El códec VP8 presenta el mejor comportamiento, mientras que el codec Theora se ubica en una respuesta intermedia.

Las conclusiones podrán utilizarse como referencias a contextos similares, y por analistas de simulación para la parametrización de los simuladores de tráfico IPTV. La experiencia reunida con los estudios realizados sobre la QoS y QoE del tráfico IPTV en una red LAN, utilizando diferentes codecs abrió una serie de alternativas de profundización de la línea de investigación, con el objeto de avanzar el perfeccionamiento de los resultados de QoE, incluyendo algunas consideraciones de subjetividad del usuario.

Referencias

1. Driscoll G. “Next Generation IPTV Services and Technologies”. (1ª Ed.)”. Editorial: Wiley-Interscience, Canada, 2018.
2. J. Lloret, A Canovas, J. J. P. C. Rodriguez, K. Lin. “A network algorithm for 3D/2D IPTV distribution using WIMAX and WLAN technologies” Springer Science-Business. 2011.
3. J. C. C. Valencia, W. C. Muñoz, G. E. C Golondrino. “Análisis de QoS para IPTV en un entorno de redes definidas por software”. Revista Ingenierías Univer. de Medellín. 2019.
4. J. Cuellar, J. Arciniegas, J. Ortiz. “Modelo para la medición de QoE en IPTV”. Editorial: Universidad Ecesi. Colombia, 2018.
5. J. M. J. Herranz., J. L. Mauri. "Estudio de la variación de QoE en Televisión IP cuando varían los parámetros de QoS". Tesis Master. Univer. Politecnica de Valencia, Gandia 2014.
6. J. C. Cuellar, S. Acosta, J. L. Arciniegas. “QoE/QoS Mapping Models to measure Quality of Experience to IPTV Service”. Conferencia Paper. Publicación: ResearchGate. Octubre 2018.
7. G. Baltoglou, E. Karapistoli, P. Chatzimisios. “IPTV QoS and QoE Measurements in Wired and Wireless Networks”. Publicación: Globecom. Editor IEEE. 23 de Abril de 2013.
8. A. C. Solbes, D. J. L. Mauri, D. J. T. Gironés. “Diseño y Desarrollo de un Sistema de Gestión Inteligente integrado de servicios de IPTV estándar, estereoscópico y HD basado en QoE”. Tesis, Universidad Politecnica de Valencia, Gandia 9 de Septiembre de 2013.
9. H. A. Facchini, S. C. Perez, A. Dantiacq, F .Hidalgo. ““Estudio Experimental de Tráfico de Video en Redes IPv6 Multicast IEEE 802.11ac”, CACIC 2019, CeReCoN. Universidad Tecnológica Nacional, UTN Mendoza, Octubre 2019.
10. H. A. Facchini, S. C. Perez, F .Hidalgo, P. Varela. “Análisis, simulación y estudio experimental del comportamiento de métricas de QoS y QoE de streamings de video multicast IPTV. WICC 2020. CeReCoN. UTN Mendoza, Mayo 2020.
11. S. C. Perez, G. Q. Salomón, H. Facchini. “Comparación del comportamiento de los códecs de video en el entorno WI-FI IEEE 802.11ac”. Argencon 2020. CeReCoN. Universidad Tecnológica Nacional y Universidad Nacional de Chilecito. Chaco, Diciembre 2020.
12. Ffmpeg, Available: https://www.ffmpeg.org/
13. Video Star Trek, Available: https://www.youtube.com/watch?v=g5lWao2gVpc
14. Wireshark, Available: https://www.wireshark.org/

Emuladores de sistemas embebidos dentro de contenedores

Esteban Carnuccio[1], Waldo Valiente[1], Mariano Volker[1], Matías Adagio[1], Micaela Antelo[1]

[1] Universidad Nacional de La Matanza,
Departamento de Ingeniería e Investigaciones Tecnológicas
Florencio Varela 1903 - San Justo, Argentina
{ecarnuccio, wvaliente, mvolker, maadagio, mantelo}@unlam.edu.ar
www.unlam.edu.ar

Resumen. En la educación de sistemas embebidos, es necesario que el estudiante interactúe con ellos, con el fin de poder completar su aprendizaje. Para esto, una manera es a través del uso de emuladores. Los cuales permiten el contacto con el embebido de forma similar a su empleo físico. En este sentido el presente artículo, expone los trabajos que se realizaron para plantear las bases que permitan emular mediante QEMU, las placas de desarrollo: Raspberry PI, ESP32 y STM32F103C8T6. Estas se ejecutan dentro de contenedores Docker. De manera tal, que los contenedores de las imágenes de los sistemas embebidos permitan realizar pruebas en un entorno estandarizado. De forma, que las aplicaciones del embebido puedan funcionar en un entorno virtual. Así se les podrá ofrecer a los estudiantes una herramienta con la cual puedan realizar sus trabajos sin tener la necesidad de incurrir en costos.

Palabras Clave. Emuladores, Docker, Sistemas Embebidos, QEMU

1 Introducción

En la actualidad existen varias iniciativas y proyectos educativos, que buscan enseñar las tecnologías que componen los Sistemas Ciber Físicos (CPS), término del inglés *Cyber-Physical Systems*. El objetivo de los CPS se basa principalmente en las interacciones dadas por sus componentes y el entorno, junto con las funciones de control y los mecanismos de comunicación entre ellos. Es por ello por lo que en esta investigación se orienta en la emulación de sistemas embebidos (SE), con conexión al exterior. Los destinatarios de los SE son los estudiantes universitarios y de postgrado. La fundamentación de lo antes descripto es recapitulada en el estudio sobre este tema [1]. En ella se diferencian los principales enfoques sobre los tópicos de gestión de proyecto, el diseño del sistema o las técnicas de redes. Además, resalta el uso de una plataforma flexible que ejecute un sobre SE con conectividad de red. Ya que, la conexión es una característica vital, debido a que el SE utiliza diferentes esquemas, que le permite comunicarse con otros dispositivos. Para esto se utilizan mecanismos tales como Wifi, Bluetooth,

Ethernet, entre otros. De esta manera, el SE puede intercambiar datos con servidores u otros dispositivos. Para así formar una topología de computación en la nube [2] [3].

Para gestar esta idea, los contenedores Docker son una de las herramientas que ayudan en la construcción, gestión y pruebas de las complejas topologías que componen estos sistemas. Gracias a que los contenedores, brindan un ambiente aislado que permiten trabajar con paquetes de software sin utilizar virtualización del hardware.

1.1 Docker

Para explicar el funcionamiento de Docker se utilizará la metáfora explicada en [4]: "Antes los trabajadores encargados de mover mercancías comerciales dentro y fuera de los barcos en el puerto (estibadores) requerían de habilidades especiales. Las mercancías eran de diferentes tamaños y formas. Los experimentados estibadores eran apreciados por su capacidad para acomodar los distintos tipos de mercancías dentro de los barcos. Contratar a estas personas no era económico, pero hacían un trabajo realmente eficiente. Como alternativa, surgieron los contenedores marítimos, que son cubos paralelepípedos de iguales proporciones, que permiten simplificar la carga y descarga de los barcos". Dichos contenedores admiten que esas mercancías, que poseen formas irregulares, se guarden desde el origen antes de llegar al puerto. Como los contenedores poseen un tamaño estándar, los barcos cargan y descargan mucho más rápido, incluso en forma automatizada. Esta metáfora es conocida en el ámbito de proyectos de software, porque se invierte mucho tiempo y energía en conseguir software heterogéneo (mercancías). Estos se integran de formas complejas en el sistema (barco). Gracias a Docker, permite que los diferentes sistemas, involucrados en el proceso de desarrollo, hablen un mismo idioma. Haciendo de esta manera, que trabajar en la integración de diferentes sistemas sea más sencillo. Ya que cada imagen Docker, es generada y mantenida como una caja negra.

Internamente Docker se compone de imágenes. Cuando se instancia esta, se asocia a un nuevo contenedor. Por lo que se pueden generar varios contenedores a partir de la misma imagen. En la figura (Fig. 1) se muestra la relación entre dos contenedores, que provienen de la misma imagen, con el sistema operativo (SO) en donde ejecutan.

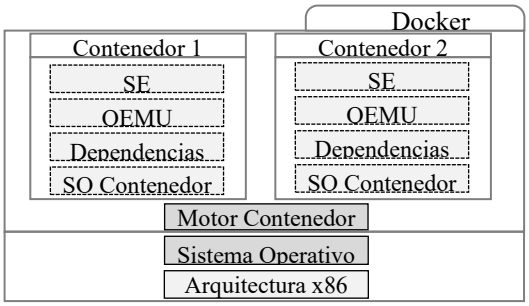


***Fig. 1** – Docker con el emulador del SE.*

Como se observa en la figura anterior, en la base de la pila se encuentra la arquitectura de hardware, que pertenece a la computadora. Sobre este funciona el Sistema Operativo Anfitrión y por encima del mismo el motor de los contenedores Docker. Dentro del contenedor se apilan las capas que forman la imagen. En la parte inferior se encuentra instalado el *kernel* mínimo del SO del contenedor (por ejemplo, Ubuntu). Luego se instalan las dependencias necesarias. Debido a que los SE suelen poseer una arquitectura hardware diferente a la que utiliza una computadora, que es del tipo x86. Para ejecutarlos se necesita de un programa que permita su funcionamiento. En esta investigación se eligió al emulador QEMU. Por eso la capa siguiente, de la imagen del contenedor, se instala QEMU, que se encuentra ya preparado y configurado para que emule los SE. Finalmente, en la capa superior se instala las herramientas propias para la construcción y uso del SE.

1.2 QEMU

QEMU (*Quick Emulator*) fue publicado inicialmente por Fabrice Bellard en 2003 [5]. En esta publicación se detalla su objetivo inicial: "*QEMU logra una emulación rápida del espacio de usuario en Linux en x86 y PowerPC, mediante la traducción dinámica. Su objetivo principal es lograr ejecutar el proyecto Wine en arquitecturas que no sean x86*". Lo anterior explica en forma simple y concisa su funcionamiento. Actualmente QEMU mantiene la idea inicial, con el agregado de diferentes tipos de arquitecturas. Es gratuito y de código abierto, pudiendo alcanzar un rendimiento nativo. Ya que el código se traduce a medida que se procesa. En la figura (Fig. 2), se ilustra que QEMU ejecuta como un programa desde la plataforma Anfitrión o *host*, por lo general, una máquina x86. Por el otro lado, el SE ejecuta dentro de la emulación que brinda QEMU, desde una arquitectura destino o *target*. Para unir estos dos mundos, QEMU usa el módulo TGC (*Tiny Code Generator*). El TCG permite la traducción dinámica del conjunto de instrucciones emuladas desde *target*, para que sean ejecutadas por las instrucciones que entiende el *host*. Esta traducción consiste en buscar las secuencias cortas de código de la arquitectura de origen, los traduce a la arquitectura de destino y captura las secuencias resultantes [6]. Lo antes dicho, es solo una de las muchas funciones que brinda el emulador. Ya que, la ejecución del SE es más complejo, intervienen lo referente al acceso a las regiones de memoria, el acceso al mapeo de los dispositivos de E/S, los temporizadores, los controladores, las interrupciones, depuración y estadísticas.

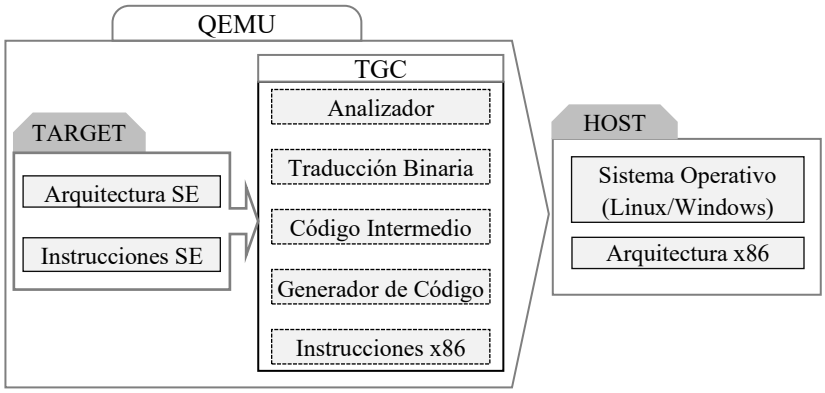


Fig. 2 - Emulación de SE con QEMU.

A continuación, se describirán los principales trabajos de SE previos a esta investigación, como así también de las emulaciones que realizan.

2 Trabajos relacionados

En [7] se desarrolló un laboratorio virtual usando al emulador QEMU dentro de contendedores Docker. El cual permitía emular Raspberry pi y ESP32. Ese trabajo de investigación se centró en realizar la comunicación entre varias placas ESP32 y Raspberry pi emuladas. Para realizar su conexión virtual se utilizó el protocolo REST[1] y MQTT[2]. No obstante, la comunicación solo fue realizada en forma interna, dentro de la red virtual, por lo que no se generan comunicación con servidores externos. Al mismo tiempo, este trabajo se encuentra acotado, debido a que las placas ESP32 que genera QEMU no permite emular sensores y actuadores. Según el autor, esto se debe a que utilizó la versión de Qemu que generó el proyecto de Espressif.

Por otro lado, en [8] se diseñó e implementó un laboratorio virtual en la nube, el cual consiste en dos partes: Una parte cliente y otro servidor. En donde la primera, está compuesta por Raspberry Pi físicas. Mientras que la segunda se encuentra en servidores en la nube, los cuales ejecutan contenedores Docker con el emulador QEMU instalado. Allí el estudiante puede ejecutar determinados sistemas emulados remotamente. Pero este trabajo no contempla la utilización en la comunicación de los dispositivos emulados con elementos externos, tales como Smartphones.

Por ese motivo en este trabajo de investigación, se plantean las tareas iniciales que se llevaron a cabo para construir las bases de una plataforma de sistemas embebidos emulados dentro de contenedores Docker. Los cuales permitirán emplear a diferentes SE, tales como ESP32, STM32 y Raspberry, ejecutándose sobre QEMU. Para que posteriormente, en próximos trabajos, estas puedan comunicarse con el exterior, y de esta forma permitan intercambiar datos con aplicaciones móviles.

[1] Interfaz de comunicación entre cliente y servidor

2 Protocolo de comunicación enfocado a la conectividad Machine-to-Machine (M2M)

En este sentido el presente documento se divide en las siguientes secciones: Primero, se describen las diferentes características que presentan los SE de la investigación. Luego se explica el proceso de configuración, instalación y creación, que se ha seguido para construir las imágenes Docker. Finalmente se explica un ejemplo de comunicación externa utilizando la Raspberry Pi emulada.

3 Desarrollo

Cuando se trata de elegir una plataforma de sistemas embebido, existen una amplia variedad de SE compactos de bajo presupuesto. Con estos se pueden realizar proyectos con fines educativos, comercial, por entretenimiento a la electrónica o la programación. Todas ellas son diseñadas para diferentes mercados. No obstante, su elección depende de las especificaciones técnicas, que se dividen en características o cantidad de conexiones que posean.

3.1 Especificaciones técnicas de los sistemas embebidos

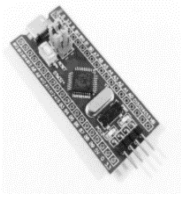
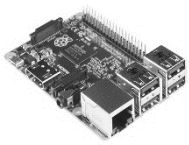
	ESP32	**STM32-*Blue Pill***	**Raspberry Pi 2**
Imagen			
Características			
SoC	Tensilica Xtensa	STM32F103C8T6	Broadcom BCM2836
Procesador	Tensilica Xtensa LX6	ARM Cortex-M3	ARM Cortex-A7
Núcleos	2	1	4
Dimensiones [mm]	48 x 25,5	53 x 23	86 x 56
Voltaje de trabajo [v]	3.3	3 - 5.5	5
Frecuencia del reloj [Hz]	240 M	72 M	900 M
Memoria interna [bytes]	512 K	128 K	1 G
Conexiones			
Conectores GPIO	34	37	40
I^2C	2	2	6
USB	☑	☑	☑
Bluetooth	☑	☒	☑
Ethernet	☑	☒	☑
Wifi	☑	☒	☑
Salida a monitor	☒	☒	☑

Tabla 1 - Comparación de Sistemas Embebidos utilizados.

3.1.1 Características de ESP32

Salió al mercado a fines del 2016 [9], está constituido en arquitectura System On Chip (SOC) diseñado por Espressif Systems. El microcontrolador más difundido entre los modelos, incluido en la placa de desarrollo, es el ESP-WROOM-32. También cuenta con el chip CP2102N, que permite la transferencia por USB. Además, posee dos pulsadores de *reset* y *boot* [10].

3.1.2 Características de STM32

También conocida como *BluePill* [11], es el apodo que se le da a la placa de desarrollo STM32F103. Esta posee un microcontrolador ARM Cortex-M3 de 32 bits. La placa posee 40 pines, un botón de *reset*, dos LED (uno de estado y otro de encendido), 2 conectores puentes (para configurar el modo de trabajo). Como conexión externa, solo posee puerto USB.

3.1.3 Características de Raspberry Pi 2

Tiene su origen en 2012, se diseñó como una minicomputadora de bajo costo. De forma tal, que le sea fácil de transportar a los estudiantes. En ella funciona un Sistema Operativo tipo Linux, pueden ejecutar aplicaciones avanzadas como suites ofimáticas, editores de fotos, así como aplicaciones de servidor web Apache. Es por ello, que se la puede conectar al monitor, teclado y mouse [12]. El nuevo modelo Raspberry Pi 4 modelo B, salió en junio de 2019. Posee dos puertos micro HDMI para conectar hasta dos pantallas. También incluye los puertos USB para conectar periféricos. Por el lado de la conectividad al exterior, incorpora un puerto Gigabit Ethernet, un módulo WiFi y otro de Bluetooth. No obstante, la versión que incluye la emulación es para la Raspberry Pi 2, las principales diferencias con el modelo actual es que soporta un solo puerto HDMI. Además, posee una menor frecuencia de reloj y memoria. Que son recursos más que suficientes para los proyectos incluidos.

3.2 Proceso de creación

La construcción del entorno para los SE, difiere ya que cada uno posee sus propias herramientas de compilación. Si bien todos usan QEMU, este también tiene diferentes versiones, ya que surgen como proyectos alternativos a la línea base del emulador. En la Fig. 3 se detalla las capas que componen las tres imágenes Docker generadas en esta investigación.

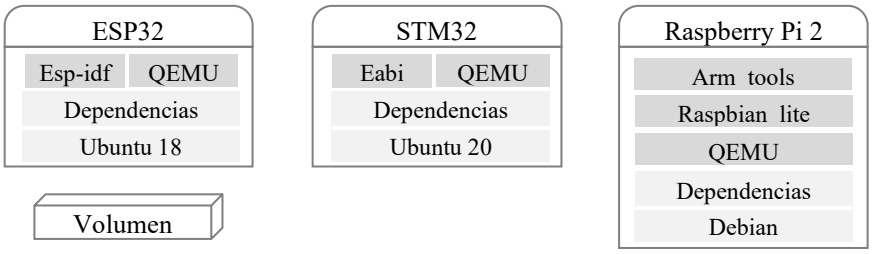


Fig. 3 - Composición de las imágenes ESP32 (a), STM32 (b) y Raspberry Pi 2(c)

3.2.1 Imagen de ESP32

Para armar el ambiente del ESP32, se implementó una imagen de Docker que posee las herramientas de compilación y el emulador QEMU. La selección de utilizar una versión anterior del Sistema Operativo base del contenedor, se debe a compatibilidad, requerida por el proyecto de QEMU que se construyó con dicha versión. Incluso se debió mantener como dependencias las dos versiones del lenguaje Pyhton (2.3 y la 3). Para el conjunto de herramientas de compilación y despliegue se utiliza el proyecto esp-idf [13]. Mientras que el emulador QEMU (versión 2.7.0), es un proyecto independiente del usuario Ebiroll [14]. Para funcionar, el emulador ejecuta directamente el grupo de ROMs que forman el bootloader, las funciones núcleo y la lógica del programa principal. Para trabajar en el proyecto se utiliza la técnica de volumen de Docker. En ella se permite trabajar con una estructura de archivos, que permanecerán consistente, aunque el contenedor se destruya, incluso permite usar el mismo volumen entre distintos contenedores. Para construir los proyectos con esp-idf para ESP32, se debe realizar la configuración inicial, desde la interfaz *menuconfig* y luego compilarlo. El resultado de la etapa de compilación es un binario con la lógica del programa. Este binario, se lo puede grabar directamente en el dispositivo físico o, como en nuestro caso, se lo ejecuta desde el emulador.

3.2.2 Imagen de STM32

Esta investigación se basó en el trabajo realizado por [15], para poder implementar la emulación del SE *BluePill* sobre QEMU. Como ese proyecto se encuentra desarrollado para emular otra placa de la misma familia, pero de distinto tipo al STM32F10C8T6 utilizado, fue necesario realizar adaptaciones en su código fuente. Por lo tanto, debido a las modificaciones realizadas y para facilitar el uso de la emulación de esta placa de desarrollo, se creó un repositorio GitHub propio, con las adaptaciones que permiten formar a la imagen Docker para la *BluePill* [16]. De esta manera, dentro del contenedor asociado a la misma, se encuentra todo el entorno de trabajo configurado para poder emular esta placa fácilmente. A su vez, contiene código fuente de distintos ejemplos. Los cuales le permiten al usuario probar los diferentes componentes que ofrece QEMU, durante la emulación de la *BluePill*.

3.2.3 Imagen de Raspberry Pi 2

Para implementar la emulación de la Raspberry Pi, se hizo uso de la imagen de Docker creado por [18]. El cual posee instalado una versión QEMU, configurada para poder ejecutar una emulación de Raspberry Pi 2. Dentro del emulador se ejecuta una versión minimalista del Sistema Operativo oficial de este embebido, denominado *Raspbian Lite.* Esta versión no posee instalada su interfaz gráfica (GUI), sino que solamente se puede acceder a su línea de comandos. Desde allí se puede controlar la Raspberry emulada.

Por otro lado, dentro de un contenedor Docker no se pueden ejecutar aplicaciones gráficas. Esto se debe a que no está preparado por defecto para ello. Sin embargo, la imagen creada por [18], incorpora los cambios en la configuración para poder realizarlo. De esta forma, se puede ejecutar una versión adaptada de Debian en forma gráfica. Dentro de la cual, a su vez se ejecuta la versión de QEMU antes mencionado. Esto se puede visualizar en la (Fig. 3-c). Esta característica es de mucha utilidad, para poder ejecutar herramientas gráficas dentro del contenedor Docker.

Cuando se crea el contenedor, este automáticamente crea un servidor que permite acceder a través de un navegador web a la interfaz gráfica del SO Debian. De esta forma, por medio del localhost, se puede controlar el S.O gráfico y al emulador Qemu, en donde se ejecuta Raspbian.

Otra característica de este contenedor es que permite acceder al sistema de archivos del Raspbian Lite vía ssh. Lo que facilita el trabajo en el emulador desde el host anfitrión.

3.3. Ejemplo de GPIO y API REST en Raspberry Pi 2

Empleando la imagen de la Raspberry Pi, anteriormente mencionada, se realizaron pruebas iniciales para ejecutar ejemplos de script de Python. Estos programas tienen como funcionalidad probar desde la emulación la comunicación con servidores externos, empleando el protocolo REST. En dichos scripts se realizaron peticiones GET y POST a un servidor en la nube. Lo que genero muy buenos resultados.

Por otra parte, también se realizó la prueba de la emulación de los puertos GPIO de la Raspberry Pi. Para ello se hizo uso de las bibliotecas provistas por [19] , que permiten emular los puertos. No obstante, no se pudo hacer funcionar esta biblioteca dentro de QEMU. Por lo que se probó su funcionamiento dentro del contenedor, pero sobre el S.O Debian. Esto se realizó para poder aprovechar las pantallas GUI que genera dicho código. La cual permite visualizar en forma gráfica los pines GPIO. De esta manera se puede interactuar con los mismos. En la siguiente figura se muestra dicha pantalla.

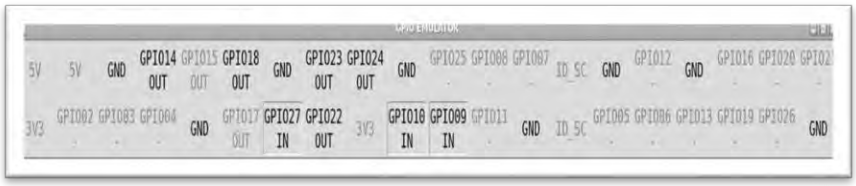


Fig. 4 - Interfaz gráfica GPIO.

4 Trabajos futuros

Se plantean tres líneas de investigaciones a desarrollar:

- La primera, en la cual se estima implementar la conectividad de los tres sistemas embebidos a un Gateway para lograr una topología compleja de internet de las cosas.
- De los tres sistemas embebidos listados en la presente investigación, el STM32 BluePill no presenta conectividad al exterior. Por lo tanto, se está analizando la factibilidad de implementar algún módulo que permita la conectividad mediante un puerto serial a través del SO anfitrión.
- Mientras que en la tercera se pretende realizar interfaces gráficas que faciliten su uso.

5 Conclusiones

En este trabajo se empezó a construir un entorno de emuladores de distintas placas de desarrollo. La cual se sustenta en el objetivo de facilitar las actividades de los estudiantes durante su formación académica. Para ello se está haciendo uso del emulador QEMU, dentro de contenedores Docker. De forma tal que los estudiantes puedan realizar prácticas con sistemas embebidos ESP32, STM32 (BluePill) y Raspberry Pi 2. A pesar de que su configuración es diferente, gracias al uso de los contenedores, se pretende que posea una rápida utilización para el usuario, de forma tal que lo opere fácilmente. A si mismo este trabajo sienta las bases, para en futuros trabajos realizar la interconexión con dispositivos externos, tales como Smartphone.

6 Referencias

[1] S. Martin, Teaching and Learning Advances on Sensors for IoT, Basel, Suiza: MDPI, 2021, pp. 10-27.

[2] P. Waher, Learning Internet of Things, Packt Publishing, 2015, pp. 2-5.

[3] Sachan, Internet de las cosas (IoT) y sus aplicaciones, Amazon Digital, 2020, p. 15.

[4] G. Sébastien, Docker Cookbook: Solutions and Examples for Building Distributed Applications, O'Reilly, 2015.

[5] F. Bellard, «QEMU x86 emulator version 0.1,» 2003. [En línea]. Available: https://www.winehq.org/pipermail/wine-devel/2003-March/015577.html.

[6] A. Höller, A. Krieg, T. Rauter, J. Iber y C. Kreiner, «QEMU-Based Fault Injection for a System-Level Analysis of Software Countermeasures Against Fault Attacks,» 2015 Euromicro Conference on Digital System Design (DSD), Madeira, Portugal, 2015, pp 4.

[7] J.-P. Ernits, «VIRTUAL IOT LAB FOR EMBEDDED SOFTWARE DEVELOPMENT FOR ESP32 AND RASPBERRY PI BASED DEVICES,» Tallin University Of Technology, 2020.

[8] G. Song, Y. Nie, G. Chen y Y. Tong, «Design and Implementation of virtual simulation experiment platform for computer specialized courses,» de *Journal of Physics: Conference Series*, 2020., pp. 1-7

[9] Á. B. Herranz, «Desarrollo de aplicaciones para IoT con el módulo ESP32,» Universidad de Alcalá, Alcalá de Henares, 2019, pp 15-20.

[10] Espressif Systems, «ESP32 Series Datasheet,» Espressif Systems, China, 2022, pp. 48

[11] STMicroelectronics, «STM32F103x8 DataSheet,» STMicroelectronics, 2015.

[12] Raspberry Pi, «Raspberry Pi 4 Model B DATASHEET,» Raspberry Pi, Pencoed, England, 2019.

[13] Espressif, «GitHub - espressif / esp-idf: Espressif IoT Development Framework,» 2022. [En línea]. https://github.com/espressif/esp-idf.

[14] Ebilroll, «GitHub - Ebiroll: Add tensilica esp32 cpu and a board to qemu and dump the rom to learn more about esp-idf,» 2021. [En línea]. Available: https://github.com/Ebiroll/qemu_esp32.

[15] Beckus, «beckus-qemu stm32» 2018. [En línea]. Available: http://beckus.github.io/qemu_stm32/.

[16] SOAUnlam, «soa-emulador-bluepill» 2022. [En línea]. Available: https://github.com/soaunlam2021/emulador-stm32-Bluepill.

[17] W. Gay, Beginning STM32: Developing with FreeRTOS, libopencm3 and GCC 1st ed. Edición, Berkley, Estados Unidos: Apress, 2018.

[18] M. Ambass, «deskctopcontainers raspberrypi,» 2017. [En línea]. Available: https://hub.docker.com/r/desktopcontainers/raspberrypi.

[19] nosix, «nosix-emulator» 2021. [En línea]. Available: https://github.com/nosix/raspberry-gpio-emulator.

Sistema de Archivos Paralelos con Aplicaciones de Machine Learning

Nicolás Benquerenca[1], Román Bond[1], Martín Morales[1,2], Diego Encinas[1,3]

[1]SimHPC-TICAPPS. Universidad Nacional Arturo Jauretche. Florencio Varela, 1888, Argentina.
[2]Centro CodApli. FRLP. Universidad Tecnológica Nacional. La Plata, 1900, Argentina.
[3]Instituto de Investigación en Informática (III-LIDI). Facultad de Informática, Universidad Nacional de La Plata - Centro Asociado CIC. La Plata, 1900, Argentina.

{nbenquerenca, rbond, martin.morales, dencinas}@unaj.edu.ar

Resumen. Se propone la investigación, análisis y evaluación del impacto de aplicaciones del tipo Machine Learning en un sistema de archivos paralelos, a nivel de rendimiento y uso de recursos. Para tal motivo se plantea el estudio del sistema de archivos paralelo BeeGFS, como infraestructura, y el uso de aplicaciones de Machine Learning como herramienta de benchmark para obtener los resultados necesarios y posterior análisis. Los sistemas de archivos paralelos nos permiten incrementar el rendimiento de los "File Servers" que requieren de mayor capacidad de respuesta a operaciones de lectura y escritura por accesos recurrentes y concurrentes a datos, donde los sistemas de archivos convencionales como "Network File System" no pueden satisfacer esta capacidad, entre otras grandes ventajas.

Palabras clave: Sistemas de Archivos, BeeGFS, Benchmarks, Cloud Computing, Machine Learning.

1 Introducción

La era digital atraviesa año tras año nuevos desafíos tecnológicos, por lo que indefectiblemente aparecen nuevas barreras a superar tanto a nivel de hardware como de software. Hace unos años, con el crecimiento continuo de aplicaciones y usuarios como por ejemplo en las redes sociales, los sistemas tradicionales como Network File System (NFS) comenzaron a mostrar sus falencias respecto a soluciones que demandaban alto rendimiento. Así ocurrió en otros ámbitos tecnológicos como la migración de Asymmetric Digital Subscriber List (ADSL) a fibra óptica, los sistemas de archivos se enfrentaron también a un mismo factor común: el aumento de demanda de recursos. La época de la centralización, escalabilidad vertical y las soluciones unificadas en arquitecturas de hardware empezaron a mostrar sus limitaciones y se inició la migración a soluciones descentralizadas. Por eso, fue necesario, por un lado, la innovación tecnológica y, por otro, los sistemas de archivos paralelos entraron en auge. Con características como escalamiento horizontal, alta disponibilidad, duplicidad de información, acceso concurrente, la barrera se fue superando.

Hoy el desafío es distinto. No sólo por el simple hecho de que día a día aumenta la demanda, sino porque ésta actúa de modo distinto. Las aplicaciones de hace cinco años ya no se comportan igual, y los usuarios tienen nuevas necesidades. Estos últimos, ya no sólo no les alcanza con tener acceso a los programas sino que se está en un periodo en donde se necesita que sean inteligentes. Es decir, se espera que los GPS den la mejor ruta de un camino no sólo basada en kilómetros, si la frase que se está escribiendo en un procesador de texto tiene sentido, se creen subtítulos en un

video en vivo, que las recomendaciones estén basadas en el estado de ánimo del usuario o hasta que detecten patrones para la detección temprana de enfermedades. Estas son algunas de las funcionalidades con las que ya se está interactuando y que vienen a dar sentido a esta nueva era digital inteligente: la de la Inteligencia Artificial (IA).

Debido a que esta transformación avanza rápidamente, es necesario empezar a supervisar los sistemas, analizar la información recolectada y desarrollar herramientas de monitoreo y benchmark específicas que permitan- desde el lado del hardware- determinar si las actuales soluciones disponibles de sistemas de archivos paralelos están preparadas para el cambio que se avecina. Para contribuir a actuales y futuras investigaciones sobre esta cuestión, en el presente trabajo se buscará aportar información mediante la monitorización, sobre el impacto y comportamiento de aplicaciones del tipo Machine Learning en sistemas de archivos paralelos, como BeeGFS.

2 Sistema de archivos paralelo

Se da este nombre a un tipo de sistema de archivos distribuido. A su vez, se diferencian de los convencionales -como NFS- porque almacenan datos a través de varios servidores conectados a la red, denominados comúnmente como "nodos", que utilizan la técnica de stripes y storage targets, que se explicará en posteriores apartados.

Este método de trabajo permite acceso simultáneo y de alto rendimiento a los datos almacenados en los servidores, a través de múltiples canales mediante procesos informáticos que generan las aplicaciones. Al ser escalado fácilmente, permiten trabajar enormes volúmenes de datos sin inconvenientes.

Una característica que comparten todas las distribuciones de sistemas de archivos paralelos es la utilización de diversos servicios dependiendo de la función que tengan, en diferentes servidores. Por ejemplo, los de metadata y almacenamiento son comunes entre ellos; algunos softwares incorporan servicios de supervisión, gestión y administración. Esto último, sumado a la utilización de múltiples canales, permite a los clientes acceder por caminos independientes tanto a los servidores de metadatos como a los que almacenan los datos, a diferencia de como ocurre en los sistemas tradicionales donde todos acceden a un mismo nodo.

3 BeeGFS

Es un sistema de archivos paralelo POSIX (Portable Operating System Interface), independiente del hardware, enfocado en el rendimiento y diseñado para facilitar el uso, la instalación y la gestión. Se encuentra diseñado para trabajar en una variedad de entornos, entre los que se destacan los orientados al rendimiento, como HPC, IA y Deep Learning [1], entre varios más.

BeeGFS distribuye de manera automática los datos que carga el usuario, entre distintos servidores. Esto permite escalar fácilmente en rendimiento y capacidad del cluster, simplemente añadiendo servidores a la infraestructura de acuerdo a las necesidades que se presenten [2].

3.1 Arquitectura

En la figura 1 se visualiza la arquitectura típica de BeeGFS:

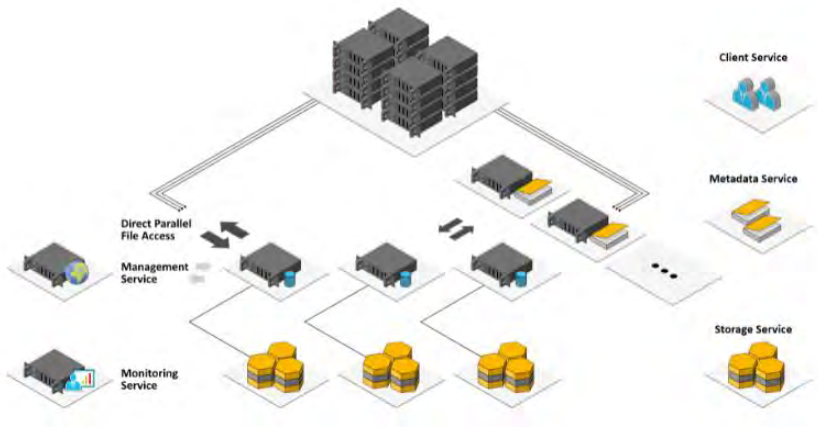


Fig. 1.. Infraestructura BeeGFS típica [3]

Se utilizarán tres tipos de arquitectura para exponer los resultados luego de su respectiva evaluación, de acuerdo a los siguientes esquemas:

- Arquitectura 1: Compuesta por 1 servidor de administración, 1 servidor de metadata, 1 servidor de almacenamiento y 1 cliente.
- Arquitectura 2: Compuesta por 1 servidor de administración, 1 servidor de metadata, 2 servidores de almacenamiento y 1 cliente.
- Arquitectura 3: Compuesta por 1 servidor de administración, 2 servidores de metadata, 2 servidores de almacenamiento y 1 cliente.

El despliegue de las arquitecturas se implementó de forma virtual utilizando VirtualBox [4]. Cada máquina virtual cuenta con la siguiente configuración de hardware:

- Servidor metadata/administración/almacenamiento: disco de 60 GB, 1 GB de memoria, 2 cpu y el sistema operativo Ubuntu 20.4.
- Cliente: disco de 60 GB, 2 GB de memoria, 2 cpu y el sistema operativo Ubuntu 20.4.

3.2 Parametrización BeeGFS

BeeGFS, así como gran parte de los sistemas archivos paralelos, permite definir opcionalmente una serie de parámetros, como por ejemplo número de storage targets, tamaño del stripe, caché o entre varios tipos de algoritmos de planificación dependiendo del entorno de producción y las necesidades particulares de la situación, junto a varias opciones más [5].

Se detallarán los parámetros que resultan relevantes y sobre los cuales se realizarán modificaciones para visualizar el cambio en los comportamientos, en caso de existir.

3.2.1 Stripe

Este parámetro hace referencia a los “fragmentos” entre los cuales el sistema divide un directorio o archivo y en el cual se define el tamaño del bloque de los mismos.

En la figura 2 se visualizan los stripes correspondientes a los datos dentro de los nodos de almacenamiento. Cada uno de estos “fragmentos” en su conjunto forman el dato en sí.

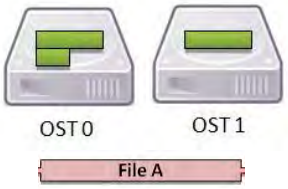


Fig. 2. Stripe [6]

3.2.2 Cache

En informática, hace referencia a la utilización de un espacio para el almacenamiento de datos de forma temporal. Es una técnica comúnmente utilizada para agilizar procesos o reducir exigencias hacia servidores, disminuyendo operaciones de I/O, entre otras funcionalidades.

BeeGFS permite customizar dos técnicas de caché en el servicio del cliente [7]:

- Buffered: Opción seteada por defecto. Utiliza un pequeño grupo de búferes estáticos para lectura y escritura. Como máximo almacena unos cientos de kilobytes de un archivo.
- Native: Es una alternativa opcional a Buffered. Aún se encuentra en estado de experimentación por lo que no se recomienda su utilización en entornos de producción que se requiere de alta disponibilidad. Permite el almacenamiento en memoria de varios gigabytes, dependiendo de la capacidad de la RAM del cliente.

3.2.3 Parametrizaciones alternativas

Con el objetivo de visualizar diferencias en el comportamiento de las gráficas respecto de las parametrizaciones estándar del sistema de archivos, se realizaron pruebas en la arquitectura 3 editando los parámetros de tamaño de stripe en servidores, tipo de caché en cliente y cantidad de memoria RAM en cliente. Los cambios son los siguientes:

- Stripe: BeeGFS establece de manera predeterminada el uso de un stripe de 512K. Se aumentó a 2 MB. El aumento del tamaño del stripe es una técnica que se suele utilizar para reducir la sobrecarga de mensajes de parte del cliente hacia los servidores.
- Caché: Se modificó el tipo de caché en el cliente, pasando de Buffered a Native. Esto permitirá un mayor almacenamiento del lado del cliente, en caso de que la aplicación lo permita. La finalidad es detectar si varía el comportamiento en los servidores durante la ejecución de la aplicación.
- Memoria RAM: Se aumentó la cantidad de memoria en el cliente de 2 GB a 3 GB.

4 Machine Learning

A partir de la lectura de la bibliografía actual [8], se la define como el campo que se ocupa de las cuestiones de cómo construir programas de computadora que mejoran automáticamente con la experiencia. Para contribuir a esta definición, también es posible describirla como una evolución en la rama de la tecnología del desarrollo de software que permite diseñar algoritmos capaces de simular la inteligencia humana.

Esto lo realiza mediante un proceso de aprendizaje y entrenamiento de modelos predictivos.

4.1 Dataset

Es una colección de piezas de datos que una computadora maneja como una unidad, con fines analíticos y predictivos. Los datos deben ser comprensibles y uniformes para ser entendidos por la máquina, que los usará para entrenar un algoritmo con el simple objetivo de encontrar patrones predecibles de dicho conjunto de datos.
La gran mayoría de aplicaciones de machine learning necesitan de datos previos para poder entrenar sus modelos. Dependiendo del tipo de problema, el tamaño del dataset puede ser un factor crítico en cuanto operaciones de lectura y escritura para los sistemas de archivos paralelos.

Comúnmente estos tienen una estructura "normalizada" y compuesta por imágenes clasificadas en distintos tipos: las de entrenamiento y las de pruebas, entre ellas completando un dataset. Un 70-80% del dataset total es utilizado para set de entrenamiento mientras que el 30-20% restante es para set de pruebas. A su vez, estos también son utilizados para evaluar los modelos en etapas intermedias de ejecución de la aplicación. Por ejemplo, si se calcula sobre el set de entrenamiento, su finalidad es entender qué tan bien está aprendiendo el modelo. Por el contrario, si se hace sobre el set de pruebas, obtendremos una idea de que tan bien el modelo se está volviendo capaz de generalizar.

Existen distintas webs para obtener datasets gratuitos, donde muchos de ellos vienen preparados ya con su aplicación y que son ideales para realizar pruebas de comportamientos en laboratorios. En este trabajo en particular se utilizó la web de Kaggle [9].

5 Benchmark

Benchmark es una técnica basada en una prueba realizada sobre un determinado sistema o componente, con el fin de medir el rendimiento de dicho sistema o componente [10]. En el presente trabajo, se utilizará la siguiente aplicación de Machine Learning para medir el rendimiento:

- FRUITS 360: Es un dataset de imágenes que contiene imágenes de frutas y vegetales de 100x100 píxeles en 131 clases [11]. Está compuesto por 90483 en total, donde de ellas 67692 son de entrenamiento y 22688 de prueba. Un peso aproximado de 1,2 GB.

6 Herramientas útiles

6.1 Collectl

Es una herramienta All In One de monitoreo de rendimiento. A través de la misma podemos realizar una supervisión y recopilación de datos en tiempo real de un amplio conjunto de subsistemas [12][13]. Esto es útil para verificar el estado general de un sistema, determinar qué estaba haciendo el mismo en un punto determinado o simplemente para realizar múltiples evaluaciones necesarias.

6.2 Darshan

Es un software diseñado para capturar de forma precisa el comportamiento de entrada/salida de una aplicación, incluyendo propiedades como patrones de acceso a archivos [14][15].

7 Resultados

De los datos recopilados de las ejecuciones correspondientes, se destacan los siguientes tópicos:

- Tiempo de ejecución arquitectura 1: 2 minutos y 11 segundos aproximadamente.
- Tiempo de ejecución arquitectura 2: 1 minuto y 59 segundos aproximadamente.
- Tiempo de ejecución arquitectura 3: 2 minutos y 5 segundos aproximadamente.

7.1 Resultados metadata

7.1.1 Arquitectura 1

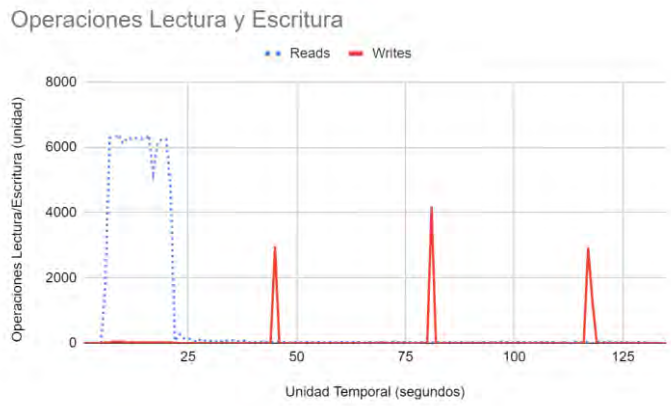


Fig. 3. Operaciones lectura y escritura metadata arq. 1

7.1.2 Arquitectura 3

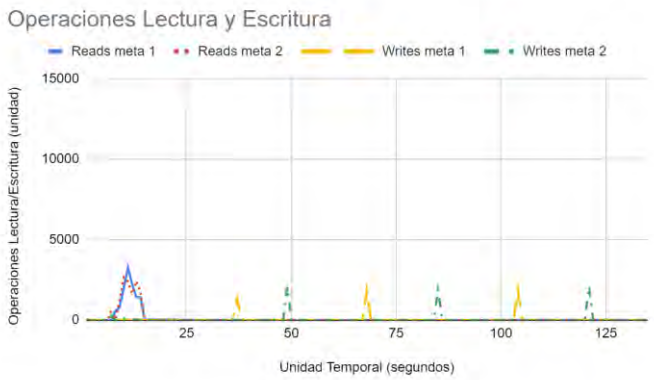


Fig. 4. Operaciones lectura y escritura metadata arq. 3

En la arquitectura 2 no se reflejaron diferencias significativas respecto de la Fig.3.

7.2 Resultados storage

7.2.1 Arquitectura 1

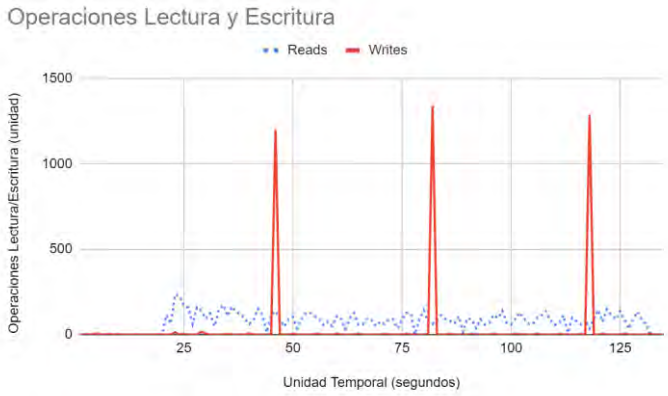


Fig. 5. Operaciones lectura y escritura storage arq. 1

7.2.2 Arquitectura 2

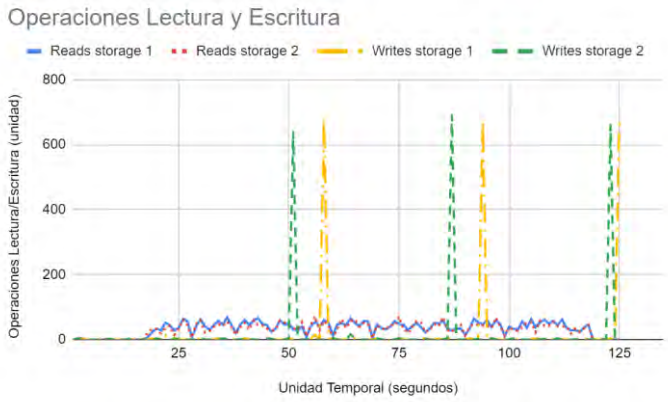


Fig. 6. Operaciones lectura y escritura storage arq. 2

En la arquitectura 3 no se reflejaron diferencias significativas respecto de la figura 6.

7.3 Resultados cliente

7.3.1 Arquitectura 1

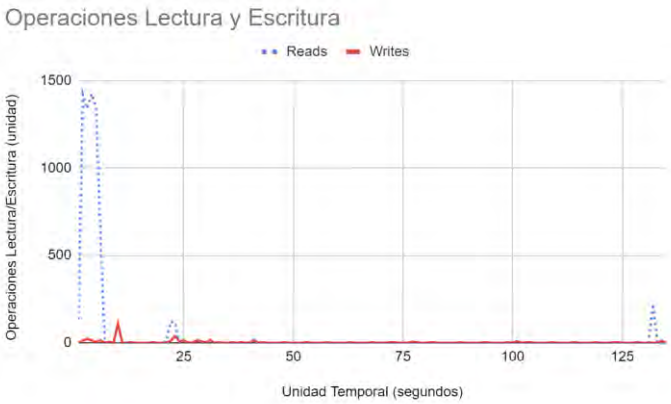


Fig. 7. Operaciones lectura y escritura cliente arq. 1.

En las arquitecturas 2 y 3 no se reflejaron diferencias significativas respecto de la figura 7.

8 Consistencia de datos

Para esta prueba se utilizó la herramienta Darshan. El objetivo de la misma es verificar si efectivamente la aplicación de FRUITS-360 está trabajando sobre el sistema de archivos paralelo y no sobre una unidad local del nodo.

Esta resulta muy interesante debido a que permite constatar la consistencia de la información obtenida tanto del lado de los servidores, con collectl, como del lado del cliente, mediante Darshan.

```
# ***************************************************
# DXT_POSIX module data
# ***************************************************

# DXT, file_id: 7540834710928794733, file_name: /mnt/beegfs/fruits-360/fruit-classification.py
# DXT, rank: 0, hostname: cliente1
# DXT, write_count: 0, read_count: 9
# DXT, mnt_pt: /mnt/beegfs, fs_type: beegfs
# Module    Rank  Wt/Rd  Segment          Offset       Length    Start(s)      End(s)
 X_POSIX      0   read        0            4691           22      0.0197      0.0206
 X_POSIX      0   read        1               0         4713      0.0213      0.0219
 X_POSIX      0   read        2            4713            0      0.0219      0.0222
 X_POSIX      0   read        3               0         8736      0.0290      0.0290
 X_POSIX      0   read        4            8736            0      0.0290      0.0290
 X_POSIX      0   read        5               0         2677      0.0293      0.0293
 X_POSIX      0   read        6            2677            0      0.0293      0.0293
 X_POSIX      0   read        7               0         2501      0.0294      0.0294
 X_POSIX      0   read        8            2501            0      0.0294      0.0294
```

Fig. 8. Logs darshan.

De la figura 8 se comprueba que realmente la aplicación que se encuentra ejecutando en el cliente está realizando las operaciones de lectura sobre el sistema de archivos analizado.

9 Conclusiones

Se observa en los resultados obtenidos, que el patrón de trabajo de las aplicaciones de Machine Learning es contraria a los de aquellas que no involucren IA. En programación tradicional, es normal ver gráficos en donde predominen los picos de operaciones de escritura. Por el contrario, este tipo de soluciones presentan grandes volúmenes de operaciones de lectura, principalmente de mayor estrés en el inicio pero continuas durante la ejecución.

Respecto de los tres tipos de arquitecturas propuestas, todas las ejecuciones se realizaron bajo las mismas condiciones de hardware y software, con la aplicación corriendo a 50 épocas. De los resultados se distingue que el incremento tanto de servicios de Metadata como de Storage no aparejó una reducción porcentual significativa en lo que respecta al tiempo de ejecución de la aplicación, de acuerdo a la siguiente tabla:

Tabla 1. Tiempos de ejecución

Ejecución (Segundos)	Ejecución (Segundos)	Ejecución (Segundos)	Reducción 1-2	Reducción 1-3
113	119	125	9%	5%

En donde sí se contempla el escalamiento horizontal es en el uso de recursos por parte de los nodos de Metadata. BeeGFS reparte las operaciones entre los nodos disponibles. Se observa en los resultados una reducción de la carga de trabajo del

orden de promedio 50% en la arquitectura 3. Verificando en Metadata, en la arquitectura 1 se obtuvo un pico de operaciones de lectura de 6384 en 1 segundo, mientras que en la arquitectura 3, con la adición de un segundo nodo de metadata, el pico de operaciones de lectura se registró en el nodo Metadata1 con 3274 operaciones, resultando en una reducción de 48,72%.

Tabla 2. Picos operaciones de lectura Metadata

Max. reads arq.1	Max. reads arq.3	Promedio reducción
6834	3274	48,72%

Verificando los resultados, se entiende por qué las distintas distribuciones de sistemas de archivos paralelos están enfocando la facilidad de adhesión y remoción de nodos de Metadata a entornos de producción con un simple par de líneas en consola. Esto permite reducir el degradamiento a los que son sometidos los nodos de metadata en aplicaciones de Machine Learning y, además, aumentar el poder de procesamiento de operaciones de ese servicio en aplicaciones que así lo requieran. En este caso en particular, BeeGFS posibilita adherir un nuevo servicio de metadata a la infraestructura simplemente instalando los paquetes correspondientes y las configuraciones básicas. Lo anterior, junto a un reinicio del servicio de metadata, el nodo ya se encuentra operativo para recibir solicitudes de los clientes y dividir las cargas de trabajo con los otros nodos que brinden la misma función, casi como un Plug & Play. De igual forma se realiza lo mismo con los otros servicios involucrados, como dato adicional.

Respecto a los storages, a partir de lo visto en los gráficos de los resultados, son sometidos a patrones de lectura de poca exigencia para los nodos, con picos bajos y continuos. En operaciones de escritura - y al igual que sucede en metadata- se advierten picos de operaciones con periodicidades casi idénticas, referentes a checkpoints que realiza la aplicación para salvaguardar la información en caso de detención de la ejecución.

Referencias

1. Chowdhury, F., Zhu, Y., Heer, T., Paredes, S., Moody, A., Goldstone, R., Mohror, K., Yu, W.: I/O Characterization and Performance Evaluation of BeeGFS for Deep Learning (2019).
2. Heichler, J.: An Introduction to BeeGFS (2014).
3. BeeGFS Arquitecture, https://doc.beegfs.io/latest/architecture/overview.html
4. Mergen, M., Uhlig, V., Krieger, O., Xenidis, J.: Virtualization of High-Performance Computing (2006).
5. BeeGFS Documentation, https://doc.beegfs.io/latest/index.html
6. High Performance & Scientific Computing, https://oit.utk.edu/hpsc/isaac-open/lustre-user-guide/
7. Client Side Caching Modes, https://doc.beegfs.io/latest/advanced_topics/client_caching.html
8. Yang. Q., Liu, Y., Chen, T., Tong, Y.: Federated Machine Learning: Concept and Applications (2019).
9. Kaggle, https://www.kaggle.com/
10. Benquerenca, N.; Bond, R.; Morales, M.; Encinas, D.: Rendimiento de sistema de archivos en arquitecturas distribuidas y paralelas (2020).
11. Fruits 360, https://www.kaggle.com/datasets/moltean/fruits
12. Collectl, http://collectl.sourceforge.net/

13. Kunkel, J. M., Betke, E., Bryson, M., Carns, P., Francis, R., Frings, W., Laifer, R., Mendez, S.: Tools for Analyzing Parallel I/O (2019)
14. Welcome to the Darshan project, https://www.mcs.anl.gov/research/projects/darshan/
15. Lindi, B.: I/O-profiling with Darshan (2012).

Una experiencia de implementación de infraestructura informática: recorriendo el camino desde lo académico hasta la instalación y puesta en funcionamiento

Damián Ferrara, Leopoldo Nahuel, Lourdes Di Santo, Antonella Basalo, ,
Augusto Bertuzzi Gaspari, Emanuel Rodriguez

GIDAS - Grupo de I&D Aplicado a Sistemas informáticos y computacionales
Universidad Tecnológica Nacional (UTN) - Facultad Regional de La Plata (FRLP)
Av. 60 esq. 124 s / nº CP 1900, La Plata, Buenos Aires, Argentina
{dferrara,augustobg,errodriguez,
lnahuel,ldisanto,abasalo}@frlp.utn.edu.ar

Resumen. El propósito de esta investigación surge de la necesidad de instalación de un servidor adquirido por el laboratorio GIDAS. La instalación de esté, nos permitirá disponer de nuestra propia infraestructura IT para dar servicios y centralizar información de los diferentes equipos de investigación. Nos centraremos en el recorrido desde lo académico a las experiencias de instalación y puesta en funcionamiento. La manera propuesta para alcanzar el objetivo planteado y fomentar capacidades para la resolución de problemas, transferencia a las prácticas, trabajo cooperativo, habilidades y la creación de conocimientos por parte de los becarios, es una estrategia de enseñanza denominada método de proyecto. Al ser una estrategia transdisciplinaria tiene relación con una amplia gama de técnicas de enseñanza y de aprendizaje, como lo son el estudio de casos, el debate, el aprendizaje basado en problemas, etc. Se describe detalladamente todos los pasos seguidos en la instalación, así como los problemas que surgieron para llegar a la puesta en funcionamiento.

Palabras Claves: Servidor, Infraestructura IT, Estrategia de Enseñanza.

1 Introducción

El Grupo de Investigación y Desarrollo Aplicado a Sistemas informáticos y Computacionales (GIDAS), ha adquirido un servidor Lenovo ThinkServer SR530 [1] que necesita ser instalado en la sala de servidores de nuestra casa de estudios, para poder virtualizar los servicios requeridos por los distintos grupos de investigación dentro del laboratorio. Se busca, de esta manera, contar con un servidor propio que evite los problemas de los servicios “en la nube”, como pueden ser sus altos costos, restricciones en las versiones gratuitas, limitación en las herramientas disponibles (se circunscribe a un “ecosistema”) y que permita ejercer la propiedad de los datos allí volcados, tanto por motivos de seguridad como de privacidad; pero a su vez manteniendo la posibilidad del trabajo en equipo a distancia, almacenamiento, sistematización y centralización de los datos.

Buscamos documentar la implementación de esta solución de hardware y software en un entorno académico, que nos permita consolidar los conocimientos adquiridos en la carrera; experimentar, probar y generar soluciones para nuestro equipo de trabajo.
Nos centraremos en la instalación física y puesta en marcha del servidor y la instalación de la solución Proxmox Virtual Environment 6.4 para controlar ese hardware. Podemos listar las distintas ventajas y desventajas de contar con un servidor propio contra la utilización de servicios en la nube.

Tabla 1. Comparación entre nubes públicas y privadas.

Servicio	Servidor privado	Servidor público
Privacidad	Nosotros decidimos quienes tienen acceso a la información y cuándo	Si bien es posible configurar los accesos y resguardos de la información, desconocemos el uso que la empresa pueda estar haciendo con nuestros datos (punto importante si se manejan datos sensibles)
Seguridad	100% dependiente del manejo propio	100% dependiente de la empresa a la que le confiamos el servicio
Soporte	Es necesario procurar el propio soporte o tercerizarlo.	Depende del plan contratado, pero usualmente es de fácil contacto, aunque puede llevar un costo extra.
Almacenamiento	Restringido a los límites del hardware. Fácilmente expandible. Costo por única vez.	Restringido a los límites del plan contratado. Fácilmente expandible. Costo por cantidad.
Riesgos	Estar preparado para eventualidades "black swan" (tiene un costo muy alto)	Datos replicados en distintos servidores del mundo hacen que los riesgos de perder información sean muy bajos, pero aún sujetos a la arbitrariedad de la empresa
Costos	Gran costo inicial, bajo costo de mantenimiento. No hay limitación en cantidad de usuarios	Costo relativamente pequeño, pero a lo largo de todo el proyecto. Muchas veces depende de la cantidad de accesos, más allá de las características del servicio

2 Implantación de Infraestructura: pasos hacia el objetivo

Poner en funcionamiento el servidor implicó una serie de capacitaciones técnicas, investigaciones y planificaciones.
Primeramente, fue necesario estudiar las características del servidor como modelo y sus prestaciones de servicios. Este equipo cuenta con cuatro discos de estado sólido de 1 terabyte; los cuales no incluidos, que fue necesario instalar. Para su colocación, el servidor cuenta con espacio para ocho unidades SAS/SATA Hot-Swap de 2,5" o 4 de 3,5", o bien cuatro unidades SATA Simple-Swap de 3,5"; hasta 2 M.2. Una vez comprendido el mecanismo de estas bahías Hot-Swap, las cuatro unidades de disco fueron instaladas.

Por otro lado, el servidor contaba con un módulo DIMM de memoria RAM de 16 gigabytes TruDDR4 a 2666 MHz / 2933 MHz de fábrica. Se instaló un módulo más al original, lo que resultó en 32 gigabytes.
El servidor Lenovo XClarity Controller es el motor de gestión integrado en los servidores ThinkSystem diseñado para estandarizar, simplificar y automatizar las tareas básicas de gestión de servidores. Con esta interfaz gráfica inicializamos y realizamos la configuración RAID de los discos. La elección de esta configuración nos llevó a una pequeña evaluación de las ventajas y desventajas, quedando seleccionada finalmente la configuración del RAID en nivel 5.
La secretaría de TICs de La UTN-FRLP brindó la capacitación que nos permitió comprender cómo es la topología del centro de cómputos, los pasos de instalación, hasta quedar en funcionamiento y configurar los accesos remotos. El servidor fue puesto en un rack, con sus correspondientes guías telescópicas y bastidores sobre el chasis.
Finalmente, aunque la puesta en marcha fue concretada, en esta etapa se nos presentó un verdadero dilema. En primer lugar, se pensó en instalar un sistema operativo Linux Debian con algún programa de virtualización, lo cual fue considerado poco óptimo. Otras posibilidades que se evaluaron fueron VMware, vSphere y Citrix XenServer. Finalmente se decidió instalar Proxmox Virtual Environment 6.4 para controlar el hardware.
Proxmox es una plataforma completa de gestión de servidores de código abierto para la virtualización empresarial. Integra estrechamente el hipervisor KVM y los contenedores de Linux (LXC), la funcionalidad de redes y almacenamiento definido por software, en una sola plataforma. Con la interfaz de usuario integrada basada en web, puede administrar máquinas virtuales y contenedores, clústeres para alta disponibilidad y recuperación con facilidad de las herramientas integradas ante desastres [2].

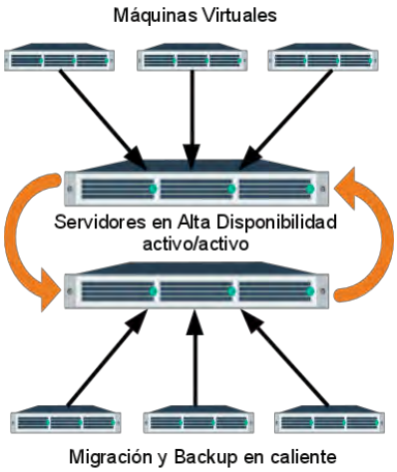


Fig. 1. Virtualización servidor Proxmox

3 Experiencias del trabajo realizado: aciertos y desaciertos

Se comenzó con la adquisición de un servidor y se requirió su instalación a finales del 2021. Dado que el laboratorio no cuenta con un espacio físico adecuado para la instalación de un servidor, se trabajó bajo la coordinación con la secretaria de TICs de la FRLP, que facilitó el lugar en el centro de cómputo de la Facultad y capacitaciones prácticas que complementaron la investigación previa teórica.

En un primer momento, se había optado por la instalación de un Sistema operativo Linux Debian, que no se lo encontró adecuado para los propósitos de prestar servicios y máquinas virtuales para el laboratorio; dado que implicaba tener una capa más entre la máquina virtual y el sistema operativo. Este fue remplazado por Proxmox que permite crear y gestionar no solo máquinas virtuales, sino también contenedores de manera sencilla.

4 Trabajo Futuro

Una de las principales funciones que se prevé para el servidor es la de brindar servicios a los distintos grupos de investigación del laboratorio.
Si bien existen numerosas herramientas con distintos propósitos específicos para este fin, decidimos comenzar el proceso con la instalación de Redmine, plataforma que, entre varias otras funcionalidades, permite la creación y seguimiento de incidentes.
El siguiente objetivo a partir de la puesta en marcha de Redmine [3] es poder utilizarla para generar un registro de demanda de necesidades que existen en el laboratorio y a partir de ello poder darles un orden de prioridad. De esta manera, se podrá relevar cuáles son las necesidades que tienen los diferentes equipos, jerarquizarlas, buscar las soluciones correspondientes y ponerlas en práctica. Se intenta que sea de una forma transparente, tanto para quienes enviaron el ticket, para los que trabajan en resolverlo e incluso para otros integrantes del laboratorio que deseen visualizarlos.

5 Manos a la obra: Experiencia pedagógica y didáctica

A raíz de la adquisición del servidor, se planteó un programa de formación y obtención de conocimientos. El cual consiste en:
Los propósitos que se persiguieron fueron generar una formación teórica y práctica, en todos los niveles referentes a la instalación, configuración, seguridad y mantenimiento de servidores. Los objetivos buscaban que los integrantes del grupo TDO pudieran instalar y administrar un servidor, generando la documentación correspondiente. Además, fomentar capacidades para la resolución de problemas, transferencia a las prácticas, trabajo cooperativo y habilidades.
La estrategia de enseñanza utilizada para alcanzar los objetivos fue el método de proyecto. Consiste en una estrategia en la que el producto del proceso de aprendizaje es un proyecto o programa de intervención profesional, en torno al cual se articulan todas las actividades formativas [2].
El método de proyectos al ser una estrategia transdisciplinaria tiene relación con una amplia gama de técnicas de enseñanza y de aprendizaje, como lo son el estudio de casos, el debate, el aprendizaje basado en problemas, entre otros. El trabajar una o más de estas técnicas en conjunto con el método de proyectos crea un ambiente altamente propicio para la adquisición y el desarrollo de conocimientos, habilidades y actitudes en todos los participantes. Además de los conocimientos propios de cada disciplina, los estudiantes adquieren y desarrollan:

- Herramientas cognitivas y ambientes de aprendizaje que motiven a los participantes a representar sus ideas.
- ·La formación sus propias representaciones de tópicos y cuestiones complejas.
- Aprendizajes de ideas y habilidades complejas en escenarios realistas.
- La aplicación de sus habilidades a una variedad de contextos.

• La construcción de su propio conocimiento, de manera que sea más fácil para los participantes transferir y retener información.
• Las habilidades sociales relacionadas con el trabajo en grupo y la negociación [4].

El presente proyecto se pensó en tres etapas: la primera fue un trabajo de campo, observación y análisis de parte del grupo. Con las conclusiones de la etapa anterior se planificó una instancia de búsqueda de espacio y ubicación del servidor; finalmente, la tercera se trató de la instalación del sistema Proxmox en el servidor.

Dado que para ubicar el servidor no implicó ningún tipo de inconveniente, se pasó de la primer a la tercera etapa sin inconvenientes. Y lo que fue una temporalización inicial pensada en seis horas, se redujo a cinco, ya que la segunda etapa se realizó junto a la final.

6 Conclusiones

Esta investigación permitió al grupo de investigación TDO instalar un servidor en un ambiente de producción real para prestar servicios de infraestructura IT al resto del laboratorio GIDAS.

Para lograr el objetivo planteado en el párrafo anterior recurrimos a una estrategia didáctica conocida como el método de proyecto, el cual es una estrategia transdisciplinaria que tiene relación con una amplia gama de técnicas de enseñanza y de aprendizaje, que permitió crear un ambiente altamente propicio para la adquisición y el desarrollo de conocimientos, habilidades y actitudes en todos los participantes.

Al finalizar la instalación se creó un manual donde se documentaron las actividades realizadas y la resolución de problemas que surgieron, lo que permitió ordenar los conocimientos adquiridos.

Los pasos a seguir serán comenzar a prestar servicios de Infraestructura IT; se instalará Redmine y conocer sobre las funciones y beneficios de esta última.

Referencias

1. ThinkSystem SR530 Setup Guide. (2020). Lenovofiles.com. https://thinksystem.lenovofiles.com/help/topic/7X07/setup_guide.pdf
2. Proxmox, https://www.proxmox.com/en/proxmox-ve, ultimo acceso 2020/08/21.
3. Guide - redmine. (s/f). Redmine.org. https://www.redmine.org/guide
4. Maria Cristina Davini,: Método de enseñanza. 1a ed. Santillana, Buenos Aires (2008)
5. Edgar Olguín Guzmán, Jorge Martín Hernández Mendoza: El método de proyectos como estrategia didáctica, Publicación semestral, Vol. 10, No. 19 (2021) 43-45.

XV Workshop Innovación en Sistemas de Software **(WISS)**

Coordinadores

Pablo Fillottrani (UNS)

Marcelo Estayno (UNLZ)

Alicia Mon (ITBA)

Dante Zanarini (UNR)

A Wizard for Composing SPARQL Queries in the GF Framework for Ontology-Based Data Access

Sergio Alejandro Gómez[1,2] and Pablo Rubén Fillottrani[1,2]

[1]Laboratorio de I+D en Ingeniería de Software y Sistemas de Información (LISSI)
Departmento de Ciencias e Ingenería en Computación
Universidad Nacional del Sur
San Andrés 800, (8000) Bahía Blanca, Argentina
Email: {sag,prf}@cs.uns.edu.ar
[2]Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

Abstract. Ontology-Based Data Access is a methodology concerned with bridging the gap between legacy data sources and semantic web technologies by providing protocols and tools for translating old data into ontologies. Querying modern ontologies represented as networks of objects interlinked by relations and properties and stored in OWL/RDF text files requires writing SPARQL queries, an activity requiring technical proficiency that is not usually in the hands of lay users. We extend our prototype of OBDA called GF to include the functionality of executing arbitrary SPARQL queries posed against OWL/RDF ontologies obtained by OBDA from H2 relational databases as well as Excel and CSV spreadsheets. To help naive users with less technical programming skills perform queries on such ontologies, we introduce a wizard for visually expressing a subset of SPARQL queries in a Query-By-Example approach.

Keywords. Ontologies, Ontology-Based Data Access, SPARQL, Knowledge Representation

1 Introduction

The Semantic Web (SW) is a version of the web where data resources have a precise meaning given in terms of conceptualizations known as ontologies that allow software agents to reason about such meaning automatically [1]. Ontology-Based Data Access (OBDA) is a discipline concerned with bridging the gap between legacy data sources and SW technologies by providing protocols and tools for translating old data into ontologies [2]. Querying ontologies provides many benefits for querying relational data as it allows the usage of open-world semantics in contrast to closed-world semantics and also allows to make explicit implicit conclusions hidden in the non-trivial subclass and composition relations that describe the underlying application domain modeled by the queried ontologies.

One of the advantages of OBDA is that old, legacy data can be then combined with new ontological data. Legacy data include tabular data as relational

database, Excel spreadsheets and CSV text files. Modern ontological data in contrast is represented as networks of objects interlinked by relations and properties and stored as OWL/RDF text files distributed in the SW. Querying modern ontologies requires writing SPARQL queries [3], an activity that requires technical proficiency that quite normally is not in the hands of lay users.

In this work, we extend a prototype of OBDA called GF [4] that we have been developing in the last years to include the functionality of executing arbitrary SPARQL queries posed against OWL/RDF ontologies obtained by OBDA from H2 relational databases as well as Excel and CSV spreadsheets. Also to help naive users with less technical programming skills to perform queries on such ontologies, we introduce a wizard for expressing a subset of SPARQL queries visually based on a Query-By-Example (QBE) approach [5]. Our solution provides a concrete way of writing SPARQL queries over legacy data without requiring the user to know explicitly SPARQL syntax. For reproducibility of the reported results, an executable file along with the files of the examples presented in this paper and its results can be checked online at `http://cs.uns.edu.ar/~sag/gf-v4.3`.

The rest of the work is structured as follows. In Sect. 2, we review the subset of SPARQL queries that our wizard can generate. In Sect. 3, we present the wizard to build the queries discussed previously. In Sect. 4, we review related work. Finally, in Sect. 5, we conclude and foresee future work.

2 Queries in SPARQL

SPARQL is the standard query language and protocol for Linked Open Data and RDF databases that can efficiently extract information hidden in non-uniform data and stored in various formats and sources, such as the web or RDF triple-stores. The distributed nature of SW data, unlike relational databases, helps users to write queries based on what they want to know instead of how the data is organized. In contrast to the SQL query language for relational databases, SPARQL queries are not constrained to working within one database – federated queries can access multiple data stores (or endpoints) because SPARQL is also an HTTP-based transport protocol, where any endpoint can be accessed via a standardized transport layer. RDF results can be returned in several data-interchange formats and RDF entities, classes, and properties are identified by IRIs such as `<http://example.org/Person/name>`, which are difficult to remember even knowing SPARQL and the underlying structure of the data source.

As mentioned in the introduction, we propose a wizard for visually composing SPARQL queries posed against a data source expressed as an OWL/RDF ontology. We now present the subset of queries that we solve with our implementation. We present a running example with which we present some prototypical queries and in Sect. 3 we show how these queries can be solved by using the wizard that we defined. We present a relational database schema for which the GF system produces an ontology automatically. Then we show some SPARQL queries posed against the ontology. We will see that writing those queries from scratch present an important challenge even for experienced users and that the proposed wizard

can help in easing such task by allowing the composition of queries by a Query-By-Example methodology (i.e., visually and abstracting from some of the inner details of the query structure and the queried dataset).

Example 1. In Fig. 1, we define the schema of a very simple relational database and show how its translation to an OWL Description Logic (DL)[1] ontology should be and then propose some iconic SPARQL queries. There are two tables: *Person* and *Phone*. A person has a unique identifier, a name, a weight in kilograms, a sex that is false if the person is female and true if the person is male, also a person has a birth date. A phone has a unique identifier, a number, a price, and its owner. There is an implicit one-to-many relation from Person to Phone, meaning that a person can have 0, 1, or more phones and a phone can belong to 0 or at most 1 person.

Notice that in this work, we have added extra functionality to the direct mapping specification programmed in previous versions of GF (see [4] and references therein for details) in order to simulate the natural joins between tables and be able to retrieve that characteristic from SPARQL. Thus, the person now knows his phones and vice versa.

Person (*personID*, *name*, *weight*, *sex*, *birthDate*)
Phone (*phoneID*, *number*, *price*, *owner*)

Person

personID	*name*	*weight*	*sex*	*birthDate*
1	John	120.0	true	2001-01-01
2	Paul	110.0	true	2002-01-01
3	Mary	60.0	false	2001-01-01

Phone

phoneID	*number*	*price*	*owner*
1	555-1234	200.00	1
2	555-1235	220.00	1
3	555-1236	230.00	2

Fig. 1. Relational schema and instance for tables *Person* and *Phone*

Example 2 (Continues Ex. 1). In Fig. 2, the UML design of the classes Person and Phone can be seen. In Fig. 3, the instances of classes Person and Phone are shown. There are three people, two males named John and Paul, and one female of name Mary. John has two phones (viz., 1 and 2), Paul has only one (viz., 3) but Mary has none. The class HeavyYoungMan is defined as a subclass of Person according to the SQL filter: `select "personID" from "Person" where "sex"=true and "birthDate">='2001-01-01' and "weight">=100.0`.

We now explore several paradigmatic query cases in SPARQL. The choice of the particular syntax of some queries is due to that they are presented in the exact way that they are composed by our tool employing the visual specification that we present in Sect. 3. We solve a very specific subset of queries and categorize its cases as: queries over a single class, queries over a simple hierarchy of classes, and queries over an association.

[1] In this context, we see a DL ontology as a mathematical conceptualization of an equivalent OWL/RDF file, which is understood as the serialization of such ontology. We refer the reader to [6].

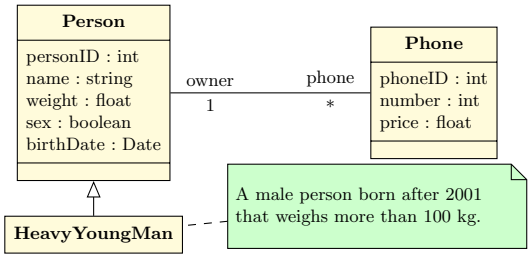


Fig. 2. UML class diagram for people and their phones obtained via OBDA from Fig. 1

```
Person(p1).                personID(p1,1)          name(p1, john).        weight(p1,120.0).      sex(p1,true).
birthDate(p1,2001-01-01).  phone(p1,t1)            phone(p1,t2).          HeavyYoungMan(p1).
Person(p2).                personID(p2,2).         name(p2, paul).        weight(p2,110.0).      sex(p2,true).
birthDate(p2,2002-01-01).  phone(p2,t3).           HeavyYoungMan(p2).
Person(p3).                personID(p3,2).         name(p3, mary).        weight(p3,60.0).       sex(p3,false).
birthDate(p3,2001-01-01).
Phone(t1).                 number(t1, 555-1234).   price(t1, 200.0).      owner(t1, p1).
Phone(t2).                 number(t2, 555-1235).   price(t2, 220.0).      owner(t1, p1).
Phone(t3).                 number(t3, 555-1236).   price(t3, 230.0).      owner(t1, p2).
```

Fig. 3. Assertional knowledge about people and their phones for UML diagram in Fig. 2 obtained from the relational instance in Fig. 1

Example 3 (Continues Ex. 2). We start with a *selection query having several conditions over a single class*: Select the portion of the data that comprise all the females that were born in 2001 that weigh less than 70 kilos, and her name optionally starts with an *M*, contains an *r*, and ends with a *y*. When relevant in all queries we ask the query processor to show at most 10 results starting with the first result. The text of the SPARQL query can be seen in Listing 1.1. The result of the query is computed in tabular form:

id	*name*	*isMale*	*bd*	*weight*
3	Mary	false	2001-01-01	60.0

Example 4 (Continues Ex. 2). We now show a *totalization query*: select the average weight of the men. The source code for the query is shown in Listing 1.2. The result of the query is:

averageWeight
115.0

Example 5 (Continues Ex. 2). We now show a query that works by *grouping similar data according to the value of a field*: Categorize people by sex and compute the average and maximum weight, least birthdate, person count, and sum of weights. The code of the query can be read in Listing 1.3. The result of the query is:

isMale	*averageWeight*	*maximumWeight*	*leastBirthDate*	*personCount*	*weightSum*
false	60.0	60.0	2001-01-01T00:00:00	1	60.0
true	115.0	120.0	2001-01-01T00:00:00	2	230.0

Example 6 (Continues Ex. 2). We now show a *query over a simple hierarchy of classes*, in particular showing the case of inheritance of attributes in subclassing: Find the name of all the young heavy weighted men. The source code is in Listing 1.4. The result of the query reads as:

personID	*name*
1	John
2	Paul

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?id ?name ?isMale ?bd ?weight
WHERE
{
        ?x rdf:type <http://example.org/Person> .
        ?x <http://example.org/Person/personID> ?id .
        ?x <http://example.org/Person/name> ?name .
        ?x <http://example.org/Person/sex> ?isMale .
        ?x <http://example.org/Person/birthDate> ?bd .
        ?x <http://example.org/Person/birthDate> ?bd .
        ?x <http://example.org/Person/weight> ?weight .
        ?x <http://example.org/Person/name> ?name .
        ?x <http://example.org/Person/name> ?name .
        FILTER ( strstarts(str(?name), 'M') && ?isMale = false && ?bd >= '2001-01-01T00:00:00'^^xsd:dateTime
        && ?bd <= '2001-12-31T00:00:00'^^xsd:dateTime && ?weight < 70
        && regex(str(?name), 'r', "i") && strends(str(?name), 'y') )
}
ORDER BY DESC(?name)
LIMIT 10
OFFSET 0
```

Listing 1.1. SPARQL query for all the females that were born in 2001 that weigh less than 70 kilos, and her name optionally starts with an M, contains an r and ends with a y

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (AVG(?weight) AS ?averageWeight)
WHERE
{
        ?x rdf:type <http://example.org/Person> .
        ?x <http://example.org/Person/weight> ?weight .
        ?x <http://example.org/Person/sex> ?isMale .
        FILTER ( ?isMale = true )
}
```

Listing 1.2. SPARQL query for the average weight of the men

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?isMale (AVG(?weight) AS ?averageWeight) (MAX(?weight) AS ?maximumWeight) (MIN(?bd) AS ?leastBirthDate)
(COUNT(?id) AS ?personCount) (SUM(?weight) AS ?weightSum)
WHERE
{
        ?x rdf:type <http://example.org/Person> .
        ?x <http://example.org/Person/sex> ?isMale .
        ?x <http://example.org/Person/weight> ?weight .
        ?x <http://example.org/Person/weight> ?weight .
        ?x <http://example.org/Person/birthDate> ?bd .
        ?x <http://example.org/Person/personID> ?id .
        ?x <http://example.org/Person/weight> ?weight .
}
GROUP BY ?isMale
```

Listing 1.3. SPARQL query for categorizing people by sex and computing the average and maximum weight, least birthdate, person count, and sum of weights

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?personID ?name
WHERE
{
        ?x rdf:type <http://example.org/HeavyYoungMan> .
        ?x <http://example.org/Person/personID> ?personID .
        ?x <http://example.org/Person/name> ?name .
}
LIMIT 10
OFFSET 0
```

Listing 1.4. SPARQL query for finding the name of all the young heavy weighted men

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?personID ?name ?p ?phoneID ?phoneNumber ?phonePrice
WHERE
{
        ?x rdf:type <http://example.org/HeavyYoungMan> .
        ?x <http://example.org/Person/personID> ?personID .
        ?x <http://example.org/Person/name> ?name .
        ?x <http://example.org/Person/ref-phone> ?p .
        ?p rdf:type <http://example.org/Phone> .
        ?p <http://example.org/Phone/phoneID> ?phoneID .
        ?p <http://example.org/Phone/number> ?phoneNumber .
        ?p <http://example.org/Phone/price> ?phonePrice .
        FILTER (strstarts(str(?name), 'John') && regex(str(?phoneNumber), '555', "i") && ?phonePrice >= 100)
}
LIMIT 10
OFFSET 0
```

Listing 1.5. SPARQL query for finding all the heavy men named John that have a phone containing 555 in its number and with a price of at least 200 dollars

Example 7 (Continues Ex. 2). We now show a *query over an association*: Find all the heavy men named John that have a phone containing 555 in its number and with a price of at least 200 dollars. The source code of the query is presented in Listing 1.5 and its result is:

personID	*name*	*p*	*phoneID*	*phoneNumber*	*phonePrice*
1	John	http://example.org/Phone/phoneID=1	1	555-1234	200.0
1	John	http://example.org/Phone/phoneID=2	2	555-1235	220.0

Example 8 (Continues Ex. 2). As our last example, we present a *totalization query over a composition*: Find the average price of phones whose owner weighs between 110 and 120 kg. The source code of the query is in Listing 1.6. The result of the query is:

averagePrice
216.66666666666666

3 A Wizard for Writing SPARQL Queries

Now we present a wizard for writing the queries presented previously in a visual way. We based our approach on the Query-By-Example (QBE) paradigm where queries are specified by giving symbolic examples of the information to be retrieved. As in most QBE solutions, our program uses the usual form called QBE grid to indicate the subject, predicate, and object of the triples involved in the

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT (AVG(?phonePrice) AS ?averagePrice)
WHERE
{
        ?phone rdf:type <http://example.org/Phone> .
        ?phone <http://example.org/Phone/price> ?phonePrice .
        ?phone <http://example.org/Phone/ref-owner> ?phoneOwner .
        ?phoneOwner rdf:type <http://example.org/Person> .
        ?phoneOwner <http://example.org/Person/weight> ?ownersWeight .
        ?phoneOwner <http://example.org/Person/weight> ?ownersWeight .
        FILTER (?ownersWeight >= 110 && ?ownersWeight <= 120)
}
```

Listing 1.6. SPARQL query to find the average price of phones whose owner weighs between 110 and 120 kg

query, the conditions they have to satisfy if a totalization or grouping is involved and aliases for results. The names of properties and concepts are presented synthetically to avoid the information overload associated with full IRIs. As in all QBE environments, there is a parser that can convert the user's actions into statements expressed in a manipulation language, in this case, SPARQL. Behind the scenes, it is this statement that is executed. A suitably comprehensive front-end can minimize the burden on the user to remember the finer details of SPARQL, and it is easier and more productive for end-users (and even programmers) to select concepts and properties by selecting them rather than typing in their names.

We now address a brief description of the wizard. The ontology to be queried has to be loaded into the system. The limitations of the current status of the system include that only one ontology can be queried at a time. The ontology that is queried cannot reference other ontologies except the one that defines the basic datatypes. Our implementation addresses the visual specification employing a form, then generates automatically the source code of the equivalent SPARQL query and this query is evaluated against the ontology using the RDF4J library (see `https://rdf4j.org/`) and then generates a web page showing the result of the query (see accompanying online documentation).

For space reasons, we will only discuss how the queries of Sect. 2 are expressed in our tool. In Fig. 4, we can see how the SPARQL query presented in Lst. 1.1 is visually codified. The user has to name the subject of the triples (viz., x), then establish the concept the subject belongs to (viz., Person), and then for each property that the user desires a column in the result, has to assign an alias and establish a condition, that can be deemed as invisible and/or optional if desired (viz., property sex with alias *isMale* and value equal to `false`). Notice how the user interface hides the low-level details of IRIs from the user.

In Fig. 5, we can see visual specification of the SPARQL query of Listing 1.2. In this case, as this totalization query must compute a single number (i.e., the average weight of the men), only one field has to be made visible and the result column for this property has to be named (viz., *averageWeight*). More importantly, in this kind of query a totalization function has to be selected (viz., *Average*).

Subject	Concept	Property	Alias	Order	Visible	Function	Operator	Value	Optional	Result
x	Person	Person/personID	id	<none>	☑	<none>	<none>		☐	
x	Person	Person/name	name	descending	☑	<none>	starts with	M	☐	
x	Person	Person/sex	isMale	<none>	☑	<none>	=	false	☐	
x	Person	Person/birthDate	bd	<none>	☑	<none>	>=	2001-01-01	☐	
x	Person	Person/birthDate	bd	<none>	☐	<none>	<=	2001-12-31	☐	
x	Person	Person/weight	weight	<none>	☑	<none>	<	70	☐	
x	Person	Person/name	name	<none>	☐	<none>	contains	r	☐	
x	Person	Person/name	name	<none>	☐	<none>	ends with	y	☐	

Fig. 4. Querying people with several conditions

Subject	Concept	Property	Alias	Order	Visible	Function	Operator	Value	Optional	Result
x	Person	Person/weight	weight	<none>	☑	Average	<none>		☐	averageWeight
x	Person	Person/sex	isMale	<none>	☐	<none>	=	true	☐	

Fig. 5. Finding the average weight of the men

In Fig. 6, we can see the visual specification of the SPARQL query of Listing 1.3. This kind of query shows how to partition a set of individuals using the values of a property (in this case *sex*). As the *sex* property is of Boolean type, the set of people is partitioned into two disjoint subsets (assuming that the sex for all people is determined), this is done by using the *Group* function. For each sex, the usage of several totalization functions are shown: *Average*, *Max*, *Min*, *Count*, and *Sum* for computing the average and maximum weight, least date of birth, the number of people and the sum of their weights. Notice that variables for the results must be defined (viz., *averageWeight*, *maximumWeight*, *leastBirthDate*, *personCount*, and *weightSum*.

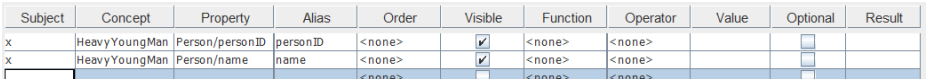
Subject	Concept	Property	Alias	Order	Visible	Function	Operator	Value	Optional	Result
x	Person	Person/sex	isMale	<none>	☑	Group by	<none>		☐	
x	Person	Person/weight	weight	<none>	☑	Average	<none>		☐	averageWeight
x	Person	Person/weight	weight	<none>	☑	Max	<none>		☐	maximumWeight
x	Person	Person/birthDate	bd	<none>	☑	Min	<none>		☐	leastBirthDate
x	Person	Person/personID	id	<none>	☑	Count	<none>		☐	personCount
x	Person	Person/weight	weight	<none>	☑	Sum	<none>		☐	weightSum

Fig. 6. Totalizing functions according to sex

In Fig. 7, we see that querying a hierarchy of classes is straightforward as the inheritance of properties (attributes) is computed seamlessly. In this case, it is shown how the names and identifiers of people can be used for the class HeavyYoungMan which is a subclass (sub-concept) of Person. Notice that in particular, this is the visual presentation of the SPARQL query of Listing 1.4.

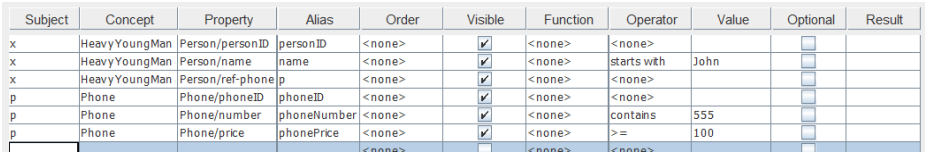
In Fig. 8, we see how an association between classes can be queried (this is the visualization of the SPARQL query in Listing 1.5). In particular, two variables for the subjects have to be defined: x for people and p for phones. Notice in the third row how x is associated with p by means of the Person/ref-phone property.

Finally, in Fig. 9, we can observe the visual expression of the SPARQL query in Listing. 1.6 showing how to perform a totalization over an association. Notice



Subject	Concept	Property	Alias	Order	Visible	Function	Operator	Value	Optional	Result
x	HeavyYoungMan	Person/personID	personID	<none>	☑	<none>	<none>		☐	
x	HeavyYoungMan	Person/name	name	<none>	☑	<none>	<none>		☐	

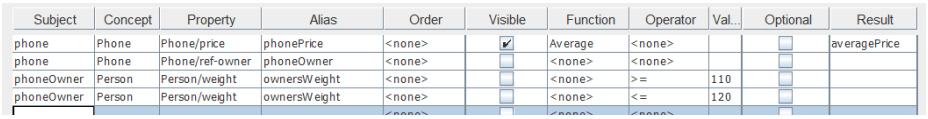
Fig. 7. Querying a hierarchy: Find the name of heavy men



Subject	Concept	Property	Alias	Order	Visible	Function	Operator	Value	Optional	Result
x	HeavyYoungMan	Person/personID	personID	<none>	☑	<none>	<none>		☐	
x	HeavyYoungMan	Person/name	name	<none>	☑	<none>	starts with	John	☐	
x	HeavyYoungMan	Person/ref-phone	p	<none>	☑	<none>	<none>		☐	
p	Phone	Phone/phoneID	phoneID	<none>	☑	<none>	<none>		☐	
p	Phone	Phone/number	phoneNumber	<none>	☑	<none>	contains	555	☐	
p	Phone	Phone/price	phonePrice	<none>	☑	<none>	>=	100	☐	

Fig. 8. Querying an association: Find the heavy men with their phones

how again two different variables for the subject have to be defined for indicating the association between subject and objects in RDF triples and also how the *averagePrice* variable in the result column has to be declared.



Subject	Concept	Property	Alias	Order	Visible	Function	Operator	Val...	Optional	Result
phone	Phone	Phone/price	phonePrice	<none>	☑	Average	<none>		☐	averagePrice
phone	Phone	Phone/ref-owner	phoneOwner	<none>	☐	<none>	<none>		☐	
phoneOwner	Person	Person/weight	ownersWeight	<none>	☐	<none>	>=	110	☐	
phoneOwner	Person	Person/weight	ownersWeight	<none>	☐	<none>	<=	120	☐	

Fig. 9. Querying an association: Find the average price of phones of people weighing between 110 and 120 kg

4 Related Work

Swipe [7] implements a search-by-example approach to query Wikipedia where naive users can enter query conditions directly on the Infobox of a Wikipedia page, and then Swipe uses these conditions to generate equivalent SPARQL queries and execute them on DBPedia. As Swipe, our system makes querying ontologies user-friendly but our system is more general as it is not limited to DBPedia. Our system could do something similar by, given a Wikipedia page, first downloading the associated DBPedia OWL ontology and loading it in GF, then expressing the query on the GF wizard and executing it. Like DBPedia, iSparQL end-point [8], our system allows also us to enter a SPARQL query in text form to be submitted against the current ontology loaded in the program. Diaz et al. [9] present SPARQLByE (for SPARQL by Example) which is a front-end for DBPedia where a naive user can input positive and negative examples of what he desires, and then the system uses a reverse engineering heuristic to induce a SPARQL query. As our system, SPARQLByE abstracts full IRIs and works with joins and optional statements. Horridge and Musen [10] present Snap-SPARQL, a Java framework for working with SPARQL and OWL, that includes a parser, axiom template API, SPARQL algebra implementation, and graphical user interface components for reading, processing, and executing SPARQL

queries. Our system does this by using an auxiliary library and provides a visual interface for the composition of queries. In brief, our solution provides a concrete way of writing SPARQL queries over legacy data expressed as an OWL ontology without requiring the user to know explicitly SPARQL syntax and it is available as a downloadable standalone application unlike many of the solutions reviewed here that are custom built for specific ontologies. However, referring to external ontologies is not supported in the current version of GF's implementation.

5 Conclusions and Future Work

We presented an extension for the GF framework for ontology integration to allow a naive user to build SPARQL queries visually by using a Query-By-Example approach. We presented several examples of how the approach works. The limitations of our approach include that in its current state it is only capable of working with a single data source comprised of an OWL ontology loaded into memory. Then it does not allow to make use of several data sources at the same time nor make the query refer to other data sources. We have not tested our implementation with naive users to account for its usability in real cases. Part of our current research is focused on solving these matters.

Acknowledgments. This work was supported by Secretaría General de Ciencia y Técnica, Universidad Nacional del Sur, Argentina, and by Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC-PBA).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284**(5) (2001) 34–43
2. Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., Zakharyaschev, M.: Ontology-Based Data Access – A Survey. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18). (2018) 5511–5519
3. Harris, S., Seaborne, A.: SPARQL 1.1 Query Language for RDF W3C recommendation 21 march 2013 (2013) https://www.w3.org/TR/rdf-sparql-query/.
4. Gómez, S.A., Fillottrani, P.R.: Ontology Metrics and Evolution in the GF Framework for Ontology-Based Data Access. In: Computer Science – CACIC 2021. Springer International (2022)
5. Zloof, M.M.: Query by Example. In: NCC (proceedings). Volume 44. Anaheim, California: AFIPS (May 1975)
6. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017)
7. Atzori, M., Zaniolo, C.: Swipe: searching wikipedia by example. In: Proceedings of the 21st International Conference on World Wide Web. (2012) 309–312
8. Grobe, M.: RDF, Jena, SparQL and the Semantic Web. In: SIGUCCS '09: Proceedings of the 37th annual ACM SIGUCCS fall conference: communication and collaboration. (oct 2009) 131–138
9. Diaz, G., Arenas, M., Benedikt, M.: SPARQLByE: querying RDF data by example. Proceedings of the VLDB Endowment **9** (09 2016) 1533–1536
10. Horridge, M., Musen, M.: Snap-SPARQL: A Java Framework for Working with SPARQL and OWL. In: International Experiences and Directions Workshop on OWL. (04 2016) 154–165

Modelado Conceptual en Industria 4.0: *Mapeo sistemático de la literatura*

Ayelén Zapata[1], Marcelo Fransoy[1], Salvador Soto[1], Martín Di Felice[1], Marisa Panizzi[1]

[1] Programa de Maestría en Ingeniería en Sistemas de Información. Escuela de Posgrado. Universidad Tecnológica Nacional. Regional Buenos Aires (UTN-FRBA). Medrano 951. (C1179AAQ). CABA, Argentina.
aye.zapata@gmail.com, marcelo.fransoy@gmail.com, salvasoman@gmail.com, mdifelice@live.com.ar, marisapanizzi@outlook.com

Resumen. El concepto de Industria 4.0 refiere a una nueva manera de producir mediante la adopción de tecnologías 4.0 basadas en soluciones enfocadas en la interconectividad, la automatización y los datos en tiempo real. Este trabajo presenta los resultados de un mapeo sistemático de literatura sobre el modelado conceptual de la industria 4.0. Se realizó una búsqueda en las bibliotecas digitales *Scopus*, *IEEE Xplore* y *ACM DL* desde enero de 2017 hasta mayo del 2022. De un total de 61 artículos se analizaron 30 estudios primarios. Se encontraron que solo dos artículos usan DSML (Lenguajes de Modelado de Dominio Específico*)* y UML (en inglés, *Unified Modeling Language*). El 63,33% de los estudios proponen una solución a través de un modelo, mientras que un 13,34% lo hacen a través de herramientas, métodos y procesos. Por último, un 23.33% presentan el estado de la cuestión.

Palabras clave: Modelado conceptual, industria 4.0, mapeo sistemático de la literatura.

1 Introducción

De acuerdo a Zenón en [1] en la solución de problemas la elaboración del modelo conceptual es aquella representación gráfica, escrita o mental elaborada por el analista y que emplea como marco de apoyo para situar y ordenar de sus percepciones, para con ello fijar la estructura del problema, delimitar el área de interés y decidir qué aspectos son relevantes y cuáles no.

Según Sokolowski *et al.*[2] un sistema es una idea generalizada de uno o de un grupo de componentes que interactúan y su funcionalidad deseada es articulada por medios gráficos y textuales. El nivel de generalización es lo que distingue a un modelo conceptual de otros modelos, siendo típicamente más informal en términos de detalles y certeza, enfocándose en una comunicación rápida de las características principales del sistema objetivo.

De acuerdo al Ministerio de Desarrollo Productivo en [3] la industria 4.0 se refiere a una nueva manera de producir mediante la adopción de tecnologías 4.0, es decir, de soluciones enfocadas en la interconectividad, la automatización y los datos en tiempo real.

Su primera mención formal con esta connotación data del año 2011, en la Feria de Hannover, Alemania, en la presentación del artículo Industria 4.0: Con el internet de las cosas camino de la 4° revolución industrial [4], donde se expuso cómo Alemania

podría ser el próximo líder y proveedor del nuevo mercado en 2020 gracias al internet de las cosas en el entorno industrial.

De acuerdo a Rainer Drath y Alexander Horch en [5], las hipótesis o fundamentos que deben ocurrir para que se den las condiciones para el desarrollo de la industria 4.0 sostienen que la infraestructura de comunicación en los sistemas de producción será más asequible y por tanto será parte de todo; los dispositivos en el campo, máquinas, plantas y fábricas (incluso productos individuales) estarán más conectados a una red (la Internet o una red privada del fabricante); y que los dispositivos en el campo, máquinas, plantas y fábricas serán capaces de almacenar documentos y conocimiento acerca de sí mismos fuera de su corporeidad en la red.

El mapeo sistemático de literatura, también conocido y enunciado en el presente trabajo como SMS (*Systematic Mapping Study*), nace en el campo de investigación de ingeniería de software como ingeniería en software basada en evidencia[6], que apunta a un enfoque basado en evidencia a la investigación teórica y práctica de la ingeniería de software. Este enfoque basado en evidencia surge a su vez de la medicina, ya que las investigaciones reflejaban la opinión de expertos para dar consejos médicos y estos no eran confiables dada la no acumulación de evidencia científica que lo sustente. El propósito de la búsqueda de evidencia es el de proveer los medios por los cuales la mejor evidencia actual de la investigación puede ser integrada con experiencia práctica y valores humanos en el proceso de toma de decisiones relacionado al desarrollo y mantenimiento de software [6].

Dentro de los trabajos relacionados podemos mencionar la investigación de Dreyfus *et al.*[7] en la cual realizan una Revisión Sistemática de la Literatura (en inglés, *Systematic Literature Review* o SLR), con el objetivo presentar un análisis detallado de los 199 artículos que identifica y generar un modelo conceptual para ello. Posteriormente asignan estos documentos a categorías y destacan las deficiencias. Finalmente, discuten el uso de Metrología Virtual en varios campos industriales, subrayando su potencial para todas las industrias manufactureras.

Por su parte, en el estudio Dornelles *et al.*[8], realizan también una SLR para construir un marco conceptual para consolidar una visión común sobre este tema creciente pero fragmentado mediante la integración de una amplia gama de hallazgos de la literatura. El estudio sistematiza dicho conocimiento en una perspectiva singular y consolidada sobre las tecnologías y el trabajo de la Industria 4.0.

Wankhede *et al.*[9] realiza también una SLR para generar un marco conceptual con los lineamientos y la estrategia para implementar las tecnologías de la industria 4.0 en la industria automotriz.

Mientras que Ding *et al.* [10] realiza la primera revisión sistemática de la literatura que vincula la Industria 4.0 con la fabricación ágil y esbelta, proponiendo un marco conceptual sobre sus relaciones. Finalmente, Machado *et al.*[11] realiza una revisión sistemática de la literatura teniendo en cuenta la manufactura sustentable.

Este artículo se desarrolla en el marco del Seminario de Modelado Conceptual de la Maestría en Ingeniería de Sistemas de Información de la Universidad Tecnológica Nacional, Regional Buenos Aires con el propósito internalizar la importancia de la conceptualización para entender la parte del mundo que queremos representar en un computador, y compartir, comunicar y asentar ese conocimiento y que Solo partiendo de un modelo conceptual es posible diseñar e implementar un proceso correcto, efectivo y eficiente de análisis de datos.

En este artículo se presenta un mapeo sistemático de la literatura (SMS) para analizar el estado del arte respecto al modelado conceptual para la industria 4.0 Para realizar el SMS se siguieron los lineamientos propuestos por Kitchenham *et al.* [12].

El artículo se estructura de la siguiente manera: en la Sección 2 se describe la planificación del SMS, en la Sección 3 se describe su ejecución. Los resultados se presentan en la Sección 4. En la Sección 5 se presenta un análisis de las amenazas a la validez y, finalmente, en la Sección 6 se exponen las conclusiones y trabajos futuros.

2 Planificación del SMS

En esta sección se presenta la definición del protocolo de revisión del SMS compuesto por las preguntas de investigación (PI), estrategia de búsqueda, selección de los estudios, criterios y proceso de selección, formulario de extracción y el proceso de síntesis de los datos.

El objetivo de este SMS es responder la siguiente pregunta de investigación (PI): *¿Cuál es el estado del arte respecto al modelado conceptual de la industria 4.0 en las fábricas actualmente?*

Esta pregunta principal se descompone en un conjunto de subpreguntas (PI1-6), las cuales se presentan en la Tabla 1 junto con su motivación.

Tabla 1. Preguntas de investigación (PI) y motivación.

Pregunta de investigación (PI)	Motivación
PI1: ¿Qué contribuciones existen respecto al modelado conceptual en las fábricas en el contexto de la industria 4.0?	Conocer las contribuciones de modelado conceptual en el contexto de la industria 4.0.
PI2: ¿Qué lenguaje de modelado se utiliza en la industria 4.0?	Determinar el lenguaje que se utiliza para el modelado en la industria 4.0.
PI3: ¿Qué diagramas se consideran para el modelado en la industria 4.0?	Identificar qué diagramas se utilizan en el modelado de la industria 4.0.
PI4: ¿En qué tipos de industrias se llevan a cabo los estudios?	Identificar si corresponde a la industria automotriz, aeroespacial, financiera, farmacéutica, etc.
PI5: ¿A qué pilar de la industria 4.0 contribuye?	Identificar en qué tipo de tecnologías se enfoca el modelo.
PI6: ¿Cuál es el tipo de investigación?	Identificar los tipos de investigación de los estudios de acuerdo con la clasificación propuesta por Wieringa *et al.*[13].

Se realiza una búsqueda automática en las librerías y plataformas digitales *Scopus* e *IEEE Xplore* y *ACM DL* por tratarse de las bibliotecas más utilizadas en el campo de las ciencias de la computación. Se consideran artículos de congresos y artículos de revistas publicados en el período comprendido entre enero del año 2017 hasta mayo del año 2022.

La cadena de búsqueda resultante es:

("industry 4.0" AND "conceptual model" AND "factory") OR ("industry 4.0" AND "conceptual model*" AND "manufactur*") OR ("industry 4.0" AND "conceptual framework" AND "factory") OR ("industry 4.0" AND "conceptual framework" AND "manufactur*")*

Los criterios de inclusión y exclusión utilizados para el proceso de selección de artículos se presentan en la Tabla 2.

Tabla 2. Criterios de inclusión y exclusión.

Criterios de inclusión	Criterios de exclusión
I1. Artículos en inglés	E1. Mapeos Sistemáticos de la Literatura (SMS) y Revisiones Sistemáticas de la Literatura (SLR)
I2. Para artículos del mismo autor y enfocadas en la misma investigación, se toma el más reciente y completo.	E2. No accesibles
I3. Artículos publicados entre enero de 2017 y mayo de 2022.	E3. Literatura gris
I4. Artículos que contengan cadenas candidatas en el título, palabras clave y/o en el resumen.	E4. Artículos cuyo contenido no se enfoquen en el modelado conceptual.

El proceso de selección de los estudios consiste en los siguientes pasos: 1) realizar la búsqueda en las fuentes definidas aplicando la cadena en el título, palabras clave y/o en el resumen, 2) eliminar los artículos duplicados, 3) aplicar los criterios de inclusión y exclusión en el título, resumen y palabras clave, 4) aplicar los criterios de inclusión y exclusión al texto completo. Este proceso permitió la selección de los estudios primarios (EP) que se analizaron para dar respuesta a las preguntas de investigación (PI) formuladas.

Para dar respuesta a cada una de las preguntas de investigación (PI) se definió un esquema de clasificación, que por restricciones de espacio se presenta en un apéndice [14], junto con el formulario de extracción de datos.

3 Ejecución del SMS

En esta sección, se presenta la búsqueda realizada en las librerías y plataformas digitales, la selección de estudios primarios de acuerdo con lo definido en el protocolo de revisión del SMS. De un total de 58 artículos encontrados, se analizaron 30 estudios primarios. El listado de los estudios analizados se presenta en el apéndice en [14].

4 Resultados del SMS

En la Tabla 3 se presenta una síntesis de los resultados del análisis de los estudios primarios en base a lo establecido en el esquema de clasificación definido en el protocolo de revisión. A continuación, se pretende dar respuesta a las preguntas de investigación (PI) en base al material analizado.

Tabla 3. Síntesis de los resultados obtenidos.

Estudio	[PI1] Contribución	[PI2] Lenguaje	[PI3] Diagrama*	[PI4] Industria	[PI5] Pilar	[PI6] Tipo
[EP1]	Modelo	No especifica	Diagrama de actividad	Fabricación de material de transporte	IoT CC	Propuesta de solución

Estudio	[PI1] Contribución	[PI2] Lenguaje	[PI3] Diagrama*	[PI4] Industria	[PI5] Pilar	[PI6] Tipo
[EP2]	Modelo	No especifica	Otros	Manufactura	AI AR	Propuesta de solución
[EP3]	Método	UML	Diagrama de clases	Fabricación de material de transporte	No especifica	Propuesta de solución
[EP4]	Modelo	No especifica	Diagrama de actividad	Manufactura	AI IoT SA	Propuesta de solución
[EP5]	Modelo	No especifica	Otros	No especifica	BD	Propuesta de solución
[EP6]	Modelo	No especifica	Otros	Manufactura	No especifica	Propuesta de solución
[EP7]	Modelo	No especifica	Diagrama de actividad	Manufactura	AR	Propuesta de solución
[EP8]	Modelo	No especifica	Otros	Fabricación de material de transporte	SA AR IoT	Propuesta de solución
[EP9]	Modelo	No especifica	Otros	Manufactura	AM, CB, VR, AR, IoT, BD, CC	Propuesta de solución
[EP10]	Buenas prácticas	No especifica	Otros	Manufactura	IoT CS BD	Evaluación
[EP11]	Modelo	No especifica	Otros	Manufactura	IoT BD AI CC SA	Propuesta de solución
[EP12]	Modelo	No especifica	Otros	Manufactura	SA	Propuesta de solución
[EP13]	Modelo	No especifica	Otros	Manufactura	SA	Propuesta de solución
[EP14]	Modelo	No especifica	Otros	Manufactura	CS	Evaluación
[EP15]	Modelo	No especifica	No especifica	Manufactura	IoT CC CS	Artículo de opinión
[EP16]	Modelo	No especifica	No especifica	Manufactura	SA	Propuesta de solución
[EP17]	Modelo	No especifica	No especifica	Manufactura	IoT AI CC	Validación
[EP18]	Modelo	No especifica	No especifica	Manufactura	AR SA	Propuesta de solución
[EP19]	Modelo	No especifica	No especifica	No especifica	AI	Evaluación

Estudio	[PI1] Contribución	[PI2] Lenguaje	[PI3] Diagrama*	[PI4] Industria	[PI5] Pilar	[PI6] Tipo
[EP20]	Modelo	No especifica	No especifica	No especifica	CS	Evaluación
[EP21]	Proceso	No especifica	No especifica	Manufactura	IoT CC BD	Propuesta de solución
[EP22]	Modelo	No especifica	No especifica	No especifica	No especifica	Propuesta de solución
[EP23]	Proceso	No especifica	No especifica	Fabricación de material de transporte	RFID SA	Propuesta de solución
[EP24]	Modelo	No especifica	No especifica	No especifica	No especifica	Evaluación
[EP25]	Modelo	No especifica	No especifica	Manufactura	IoT CS BD	Evaluación
[EP26]	Herramienta	DSML	Diagrama de secuencia	Agricultura	AR AI	Propuesta de solución
[EP27]	Modelo	No especifica	No especifica	Manufactura	No especifica	Propuesta de solución
[EP28]	Modelo	No especifica	No especifica	Servicio de envío de paquetes	IoT	Evaluación
[EP29]	Modelo	No especifica	No especifica	Manufactura	CS	Propuesta de solución
[EP30]	Modelo	No especifica	No especifica	Manufactura	IoT CS BD	Evaluación

PI1: ¿Qué contribuciones existen respecto al modelado conceptual en las fábricas en el contexto de la industria 4.0?

Elnagar *et al*. [EP1] propone un modelo conceptual para que las empresas adopten aprendizaje profundo para IoT, aplicando el enfoque de aprendizaje profundo federado (FDL) y presentando un marco para su aplicación en una planta de fabricación de automóviles industria 4.0.

Bennulf M. *et al.* [EP2] presenta un modelo conceptual de comunicación y negociación entre agentes para un sistema Plug&Produce en la industria manufacturera, que es más flexible y aumenta la velocidad de adaptación para incorporar nuevos productos y recursos.

Polacsek T. *et al.* [EP3] propone un método basado en modelos para desarrollar en conjunto el diseño y la producción de un producto, permitiendo evaluar su factibilidad de fabricación, aplicado en la industria 4.0.

Por otra parte, Serrano-Ruiz J.C. *et al.* [EP4], Gupta S. et al [EP5], Saboor A. et al [EP7], Kim T.H. *et al.* [EP8], Rahamaddulla *et al.* [EP9], Le *et al.* [EP11], Onaji *et al.* [EP12], Eirinakis *et al.* [EP13], Taifa *et al.* [EP16], Oluyisola et al [EP17], Nick et al [EP18], Frank *et al.* [EP22], Boucher *et al.* [EP27], Manavalan *et al.* [EP28], Kunath *et al.* [EP29] y Zhong *et al.* [EP30] proponen una solución para el desembarco de las

tecnologías que componen la Industria 4.0 dentro las organizaciones, también en forma de modelo o marco conceptual.

Reyes J. *et al.* [EP6] propone un modelo conceptual que fusiona las tecnologías de la industria 4.0 con herramientas de manufactura esbelta para reducir el desperdicio y minimizar costos, en el contexto de la planificación de la cadena de suministro ajustada.

Doyle-Kent *et al.* [EP14], Culot *et al.* [EP15], Pere*s et al.* [EP19], Rojas *et al.* [EP20], Boukerika *et al.* [EP24] y Hubert Backhaus *et al.* [EP25] buscan presentar un estado de la cuestión con respecto al modelado conceptual y la industria 4.0, y para ello se basan en la presentación de un marco conceptual.

Por otro lado, Frank *et al.* [EP21] y Raharno *et al.* [EP23] también proponen soluciones pero en forma de procesos. Mientras que Chen *et al.* [EP26] lo hace en forma de herramienta.

Finalmente, Cañas *et al.* [EP10] busca presentar el estado de la cuestión, poniendo en evidencia las buenas prácticas y principios a la hora de implementar una transformación hacia la industria 4.0.

PI2: ¿Qué lenguaje de modelado se utiliza en la Industria 4.0?

Chen *et al.* [EP26] propone una herramienta para la aplicación de una solución específica con tecnologías de la Industria 4.0 y para eso se vale de un lenguaje de modelado de dominio específico (DSML), mientras que Polacsek T. *et al*. [EP3] se vale de UML.

El resto de los estudios no hace referencia a un lenguaje específico de modelado.

PI3: ¿Qué diagramas se consideran para el modelado en la Industria 4.0?

Elnagar *et al.* [EP1] presenta un diagrama de actividad para representar el funcionamiento del enfoque de aprendizaje profundo federado (FDL) que presenta en su modelo, al igual que Serrano-Ruiz J.C. *et al.* [EP4] para representar su modelo de programación de producción inteligente y que Saboor A. *et al*. [EP7] para su modelo de sistema de fabricación sin intervención humana.

Bennulf M. *et al.* [EP2], Gupta S. et al [EP5], Reyes J. *et al.* [EP6], Kim T.H. *et al*. [EP8], Rahamaddulla *et al.* [EP9], Cañas *et al. [*EP10], Le *et al.* [EP11], Onaji *et al*. [EP12], Eirinakis *et al.* [EP13] y Doyle-Kent *et al.* [EP14] utilizan un diagrama para representar su modelo, pero que no se corresponde con ninguna de las categorías de diagramas definidas en el protocolo de revisión.

Polacsek T. *et al*. [EP3] utiliza un diagrama de clases para representar su modelo y las conexiones entre producto y producción.

Chen *et al.* [EP26], para su herramienta, utiliza un diagrama de secuencias con el fin de especificar el funcionamiento de esta.

El resto de los estudios no utilizan diagramas para la especificación de su modelo.

PI4: ¿En qué tipos de industrias se llevan a cabo los estudios?

Elnagar *et al.* [EP1], Bennulf M. *et al.* [EP2], Serrano-Ruiz J.C. *et al.* [EP4], Reyes J. et al [EP6], Saboor A. *et al.* [EP7], Kim T.H. *et al.* [EP8], Rahamaddulla *et al.* [EP9], Cañas *et al.* [EP10], Le *et al.* [EP11], Onaji *et al.* [EP12], Eirinakis *et al.* [EP13], Doyle-Kent *et al.* [EP14], Culot *et al*. [EP15], Taifa *et al.* [EP16], Oluyisola *et al*. [EP17], Nick *et al*. [EP18], Frank *et al.* [EP21], Hubert Backhaus *et al.* [EP25], Boucher *et al.* [EP27], Kunath *et al.* [EP29] y Zhong *et al*. [EP30] enfocan sus trabajos dentro de la industria manufacturera. Elnagar *et al.* [EP1], Polacsek T. *et al.* [EP3] y Raharno *et al.*

[EP23], lo hace en la industria de fabricación de material de transporte. Chen *et al.* [EP26] propone su solución dentro de la industria de la agricultura. Finalmente, Manavalan *et al.* [EP28] lo hace dentro de la industria del servicio de envío de paquetes.

El resto de los estudios no hace referencia a una industria específica.

PI5: ¿A qué pilar de la industria 4.0 contribuye?

Bennulf M. *et al.* [EP2], Serrano-Ruiz J.C. *et al.* [EP4], Le *et al.* [EP11], Oluyisola et al [EP17], Peres *et al.* [EP19] y Chen *et al.* [EP26] se enfocan en la implementación de inteligencia artificial. Cañas *et al.* [EP10], Doyle-Kent *et al.* [EP14], Culot *et al.* [EP15]; Rojas *et al.* [EP20], Hubert Backhaus *et al.* [EP25], Kunath *et al.* [EP29] y Zhong *et al.* [EP30] hacen referencia a la aplicación de sistemas cobóticos.

Elnagar *et al.* [EP1], Rahamaddulla *et al.* [EP9], Cañas *et al.* [EP10], Le *et al.* [EP11], Culot G. *et al*. [EP15], Oluyisola et al [EP17], Frank *et al.* [EP21], Manavalan *et al.* [EP28] y Zhong *et al.* [EP30] mencionan a Internet de las cosas.

La computación en la nube es abordada por Elnagar *et al.* [EP1], Rahamaddulla *et al.* [EP9], Le *et al.* [EP11], Culot *et al.* [EP15], Oluyisola et al. [EP17], Frank *et al.* [EP21], mientras que Big Data y analítica también es mencionada por Gupta S. *et al.* [EP5], Rahamaddulla *et al.* [EP9], Cañas *et al.* [EP10], Le *et al.* [EP11], Frank *et al.* [EP21], Hubert Backhaus *et al.* [EP25] y Zhong *et al.* [EP30].

RFID es incluido por Kim T.H. *et al.* [EP8] y Raharno *et al.* [EP23], quien a su vez también aborda sobre sensores y actuadores, al igual que Le *et al.* [EP11], Onaji *et al.* [EP12], Eirinakis *et al.* [EP13], Taifa *et al*. [EP16] y Nick *et al.* [EP18].

Finalmente, Bennulf M. *et al.* [EP2], Saboor A. *et al*. [EP7], Nick *et al*. [EP18] y Chen *et al.* [EP26] tratan sobre robots autónomos en sus propuestas.

Hay estudios que no hacen referencia a ningún pilar en particular, así como también es común observar estudios que mencionan varios de ellos en su investigación.

PI6: ¿Cuál es el tipo de investigación?

De acuerdo a los criterios propuesta por Wieringa *et al.*[12] para la clasificación de artículos, hemos encontrado investigación de evaluación en los artículos de Cañas *et al.* [EP10], Doyle-Kent *et al.* [EP14], Peres *et al.* [EP19], Rojas *et al.* [EP20], Boukerika *et al.* [EP24], Hubert Backhaus *et al.* [EP25], Manavalan *et al.* [EP28] y Zhong *et al.* [EP30].

Los estudios de Elnagar *et al.* [EP1], Bennulf M. *et al.* [EP2], Polacsek T. et al [EP3], Serrano-Ruiz J.C. *et al.* [EP4], Gupta S. *et al*. [EP5], Reyes J. *et al*. [EP6], Saboor A. *et al*. [EP7], Kim T.H. *et al*. [EP8], Rahamaddulla *et al.* [EP9], Le *et al.* [EP11], Onaji *et al.* [EP12], Eirinakis *et al.* [EP13], Taifa *et al*. [EP16], Nick *et al*. [EP18], Frank *et al.* [EP21], Fran*k et al.* [EP22], Raharno *et al.* [EP23], Chen *et al.* /EP26], Boucher *et al.* [EP27] y Kunath *et al.* [EP29] son considerados como propuestas de solución.

Culot *et al.* [EP15] es el único estudio que se considera un artículo de opinión, mientras que Oluyisola *et al*. [EP17] es el único clasificado como artículo de validación.

5 Amenazas a la validez

A continuación se presentan las acciones tomadas para mitigar las amenazas a la validez categorizadas por Petersen *et al.* [15]:

- Descriptiva: busca asegurar que las observaciones se describan de manera objetiva y precisa. Se ha estructurado la información a recolectar por medio de un formulario de extracción de datos para responder las PIs, presentado en el apéndice en [14], para apoyar un registro uniforme de datos y objetivar el proceso de extracción de datos.
- Teórica: depende de la capacidad de obtener la información que se pretende captar. Se comenzó con una cadena de búsqueda (sección 2) adaptada a las tres bibliotecas digitales más populares sobre ciencias de la computación. Se definió un conjunto de criterios de inclusión y exclusión (sección 2, Tabla 2) para objetivar el proceso de selección.
- Generalización: es la capacidad de generalizar los resultados a todo el dominio. El conjunto de PIs es lo suficientemente general para identificar y clasificar los hallazgos sobre modelos conceptuales de la Industria 4.0, independientemente de los casos específicos, el tipo de industria, etc.
- Interpretativa: se logra cuando las conclusiones son razonables dados los datos. El proceso ha sido realizado por dos grupos de manera paralela y validado con la docente del seminario (última autora) para resolver las discrepancias en el momento de la inclusión y exclusión de los artículos. Todos los integrantes del grupo validaron las conclusiones.
- Repetibilidad: el proceso de investigación debe ser lo suficientemente detallado para garantizar que pueda repetirse de forma exhaustiva. Se ha diseñado un protocolo de revisión para el SMS lo suficientemente detallado como para permitir que otros investigadores puedan repetir el proceso.

6 Conclusiones

- La mayoría de los trabajos presentan un modelo cuya contribución es una propuesta de solución, generalmente para la introducción de las tecnologías de la industria 4.0 dentro de las organizaciones.
- No hay una utilización formal de lenguajes o diagramas para la aplicación de estos modelos.
- La industria más abordada por los diferentes estudios es la industria de la manufactura.
- La mayoría de los estudios mencionan diversos pilares en sus investigaciones, siendo los más mencionados: Internet de las Cosas (21,4%), Sensores y Actuadores (14,3%), Big Data y Analítica (12,5%) y Sistemas Cobóticos (12,5%).
- En cuanto al tipo de investigación, la mayoría de los artículos se encuadran en propuestas de solución e investigaciones de evaluación, siendo el primer tipo el más encontrado.

Los futuros trabajos para desarrollar son: a) cubrir el área de vacancia de modelado conceptual para los diferentes tipos de industria 4.0 y b) realizar el modelado conceptual con los diagramas existentes, contrastar los resultados y evaluar cual es el de mayor precisión.

Referencias

1 Fuentes-Zenón, “4 El enfoque de Sistemas en la Solucion de Problemas La Elaboracion

del Modelo Conceptual", Consultado: el 20 de junio de 2022. Disponible en: https://www.academia.edu/4090548/4_El_enfoque_de_Sistemas_en_la_Solucion_de_Problemas_La_Elaboracion_del_Modelo_Conceptual.

2 M. Banks y J. A. Sokolowski, *Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains*. Hoboken, N.J, 2010.

3 "¿Qué es la Industria 4.0?", *Argentina.gob.ar*, el 7 de abril de 2021. https://www.argentina.gob.ar/produccion/planargentina40/industria-4-0 (consultado el 28 de junio de 2022).

4 "Industrie 4.0: Mit dem Internet der Dinge auf dem Weg zur 4. industriellen Revolution - ingenieur.de", *ingenieur.de - Jobbörse und Nachrichtenportal für Ingenieure*, el 1 de abril de 2011. https://www.ingenieur.de/technik/fachbereiche/produktion/industrie-40-mit-internet-dinge-weg-4-industriellen-revolution/ (consultado el 20 de junio de 2022).

5 R. Drath y A. Horch, "Industrie 4.0: Hit or Hype? [Industry Forum]", *IEEE Ind. Electron. Mag.*, vol. 8, núm. 2, pp. 56–58, jun. 2014, doi: 10.1109/MIE.2014.2312079.

6 A. Kitchenham, T. Dybå, y M. Jørgensen, "Evidence-based Software Engineering", 2004, pp. 273–281.

7 P.-A. Dreyfus, F. Psarommatis, G. May, y D. Kiritsis, "Virtual metrology as an approach for product quality estimation in Industry 4.0: a systematic review and integrative conceptual framework", *Int. J. Prod. Res.*, vol. 60, núm. 2, pp. 742–765, ene. 2022, doi: 10.1080/00207543.2021.1976433.

8 J. de A. Dornelles, N. F. Ayala, y A. G. Frank, "Smart Working in Industry 4.0: How digital technologies enhance manufacturing workers' activities", *Comput. Ind. Eng.*, vol. 163, p. 107804, ene. 2022, doi: 10.1016/j.cie.2021.107804.

9 V. A. Wankhede y S. Vinodh, "State of the art review on Industry 4.0 in manufacturing with the focus on automotive sector", *Int. J. Lean Six Sigma*, vol. 13, núm. 3, pp. 692–732, ene. 2021, doi: 10.1108/IJLSS-05-2021-0101.

10 B. Ding, X. Ferràs Hernández, y N. Agell Jané, "Combining lean and agile manufacturing competitive advantages through Industry 4.0 technologies: an integrative approach", *Prod. Plan. Control*, vol. 0, núm. 0, pp. 1–17, jun. 2021, doi: 10.1080/09537287.2021.1934587.

11 G. Machado, M. P. Winroth, y E. H. D. Ribeiro da Silva, "Sustainable manufacturing in Industry 4.0: an emerging research agenda", *Int. J. Prod. Res.*, vol. 58, núm. 5, pp. 1462–1484, mar. 2020, doi: 10.1080/00207543.2019.1652777.

12 Kitchenham, B. y Charters, S.: Guidelines for Performing Systematic Literature Reviews in Software Engineering. Citeseer (2007).

13 R. Wieringa, N. Maiden, N. Mead, y C. Rolland, "Requirements engineering paper classification and evaluation criteria: A proposal and a discussion", *Requir Eng*, vol. 11, pp. 102–107, mar. 2006, doi: 10.1007/s00766-005-0021-6.

14 Zapata A., Fransoy M., Soto S., Di Felice M., Panizzi M. Apéndice - Modelado Conceptual en Industria 4.0: *Mapeo sistemático de la literatura.* (2022). Disponible en: https://doi.org/10.6084/m9.figshare.20341500.v1

15 K. Petersen, S. Vakkalanka, y L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update", *Inf. Softw. Technol.*, vol. 64, pp. 1–18, ago. 2015, doi: 10.1016/j.infsof.2015.03.007.

AIS-Signal Detector. Control de balizas de acceso a puertos.

Martin Andres Cachile[1], Facundo Ferro[2], Marcelo Taruschio[2], Rodolfo Bertone[3]

[1] Alumno de Grado FACEI UCALP
[2] Profesor Lic. en Sistemas FACEI UCALP
[3] Director de Carrera FACEI UCALP
Martin.cachile@gmail.com, facundoferro@gmail.com, Marcelo_taruschio@ucalp.edu.ar, rodolfo.bertone@ucalp.edu.ar

Abstract. En la labor diaria portuaria, se ha observado la existencia de grandes dificultades para la adjudicación de responsabilidad sobre los daños ocasionados a las instalaciones del puerto por parte de los buques o artefactos navales que se encuentran en la zona de influencia. Esta problemática se evidencia al momento de la reparación de dichas instalaciones, como por ejemplo boyas, muelles, balizas etc., que implican grandes erogaciones por parte de la autoridad portuaria.

En este contexto y conforme la problemática planteada, se presenta el desarrollo de un sistema denominado AIS-Signal Detector que permite determinar con exactitud, mediante la utilización de tecnología AIS (Sistema de identificación automática), el buque responsable del daño ocasionado.

Keywords: Sistemas de identificación automática, Signal Detector, GPS, daños, puertos.

1 Introducción

En el ámbito portuario y de la navegación, la tecnología del Sistema de Identificación Automática (AIS por sus siglas en inglés) ha mejorado notablemente los aspectos de navegación concernientes a la seguridad, la eficiencia, la protección del medioambiente y la identificación de los buques.

Actualmente el AIS es utilizado ampliamente en todos los puertos del mundo incluyendo los de Argentina, así como en los buques y artefactos navales. Cada buque tiene un sistema de posicionamiento global (GPS) el cual permite conocer su ubicación en tiempo real, y esta información es transmitida y administrada por el AIS. Asimismo, a nivel portuario cada boya que demarca acceso y salidas del puerto tiene su ubicación geográfica determinada por un GPS y dicha posición es conocida también por el sistema de identificación automática. Para la comunicación en tiempo real los AIS, utilizan antenas VHF que están diseñadas para obtener la máxima ganancia centradas en el rango de frecuencias del canal 16 (156,8 MHZ). [1]

1.1 AIS

La ley 20094 promulgada por el gobierno argentino, define puerto como el ámbito espacial que comprende el agua, los diques, dársenas, muelles, fondeaderos, escolleras y canales de acceso y derivación; y por tierra el conjunto de instalaciones, edificios, terrenos, vías de comunicación, indispensables para la normal actividad y desarrollo de la navegación.[2]

Para la conservación de los puertos, de acuerdo con la definición anterior, los AIS se ha constituido como un elemento indispensable para el normal desarrollo no solo de la navegación, sino de la salvaguarda de los elementos portuarios. Para ello, ofrece numerosos beneficios como proporcionar un medio de identificación confiable para los buques, transmitir con precisión la posición de boyas y, además indicar posible desplazamientos indeseados de las mismas. Como complemento difunde mensajes con información específica de datos meteorológicos, de mareas y del estado del mar, marcando o delinea rutas, áreas y límites.

AIS es una de las tecnologías de seguridad de navegación más utilizadas e importantes desde la introducción del radar. Desde 2002 su utilización es obligatoria para buques de más de 500GT y buques de pasajeros. En 2005 se agregan más tipos d de barcos que deben cumplir la normativa, que van de 15 a 45 mts de eslora [3]. Utiliza un sistema de comunicaciones a partir de cuatro canales mundiales en la banda móvil marítima VHF, para el intercambio de datos de navegación. Existen numerosos dispositivos AIS, conocidos como estaciones, que se identifican mediante una identidad única de servicio móvil marítimo (MMSI).

Una característica relevante es que las estaciones AIS están diseñadas para funcionar de forma autónoma sin la interacción del personal del buque o de tierra y también pueden recibir instrucciones para transmitir con un formato diferente.

El estándar AIS comprende varios subestándares denominados "tipos" que especifican tipos de productos individuales. La descripción para cada tipo proporciona una especificación técnica detallada que garantiza la integridad general del sistema AIS global, dentro del cual deben operar todos los tipos de productos. Los principales tipos de productos (la figura 1 representa un esquema de los tipos de productos individuales, [4]):

1. Clase A: Transceptor AIS utilizado en grandes embarcaciones comerciales, envía información continuamente y el alcance puede rondar 50 millas náuticas.
2. Clase B: Destinados a mercados de buques comerciales y de ocio más pequeños, aprovecha tecnología de identificación de barcos, tiene menores prestaciones y requisitos tecnológicos. Envía información cada 3 minutos, con un alcance promedio de 12 millas náuticas.
3. Estación base: diseñado para las autoridades de navegación para la transmisión de información entre buque y costa, y costa y buque. Las Estaciones Base AIS en red pueden ayudar a proporcionar reconocimientos globales en todo el dominio marítimo.
4. Ayudas a la navegación: Ayudas a la Navegación (AtoN) transceivers: transmiten la posición y estado de boyas, junto con información meteorológica y de estado del mar.

5. AIS receiver: Sólo recibe las señales AIS y no tiene un trasmisor para enviar señales AIS. Este producto le conviene a embarcaciones de recreo que no necesariamente quieran compartir toda la información de su barco.
6. AIS SART: Transmisor de Búsqueda y Rescate que utiliza AIS y puede usarse como ayuda para determinar la posición de un barco en situación SOS. Normalmente se usa en balsas salvavidas. AIS en aviones de Búsqueda y Rescate (SAR): se usa en aviones y helicópteros como ayuda en operaciones de búsqueda y rescate.

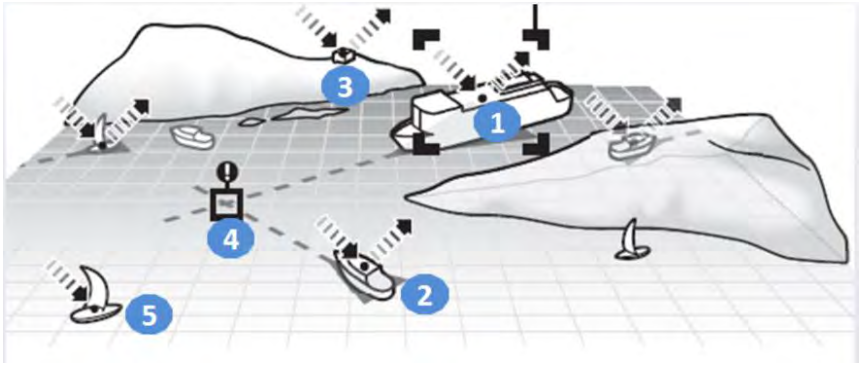


Fig. 1. Sistema AIDS, las referencias son los tipos de producto.

2 Estudio de Caso

El objetivo de este trabajo consistió en analizar y diseñar un sistema para la detección de embarcaciones responsables de los posibles daños ocasionados por las maniobras realizadas en acceso o salida del puerto a los elementos del mismo. El AIS proporciona información en particular de las embarcaciones que navegan o navegaron cerca de un puerto. Las señales recibidas son analizadas por el sistema que se desarrolla, AIS Signal Detector (AIS SD), para detectar posibles colisiones y/o destrucciones de las ayudas para la navegación (AtoN).

Para desarrollar AIS Signal Detector, se analizaron los siguientes aspectos:

- Determinar la información relevante administrada por el AIS, para conocer el trayecto del buque.
- Análisis de la información procesada por los AIS, sus prioridades y sus esquemas de acceso utilizados para transmitir los datos.
- Protocolo de actuación frente a las alertas emitidas por el sistema, como por ejemplo colisión o posible colisión con una boya, dada la trayectoria del buque.
- Crear trazados de buques o artefactos navales.
- Crear zonas de interés sobre la superficie del agua.
- Alertas de buques o artefactos navales cercanos a puntos de interés.

A través del AIS SD se lograr identificar con precisión el buque o artefacto naval que ocasione el daño. De esta forma, el sistema permite que los gastos erogados por daños ocasionados a las instalaciones portuarias eventualmente se puedan reclamar al buque involucrado.

La Convención de Bruselas de 1910 sobre abordajes, la cual unifica reglas en materia de este tema, prohíbe las presunciones sobre responsabilidades. En los casos de Alisión (colisión de una embarcación contra un objeto estacionado o fijo) las mismas están fuera de la normativa internacional de la Navegación sobre abordajes y de la normativa Argentina, dado que no son colisiones entre buques. En este sentido, como en la República Argentina no se encuentra regulada la responsabilidad ante una Alisión, la obligación de resarcir los daños causados por un buque se rigen por el principio de la responsabilidad subjetiva e indirecta del armador, fundada en la culpa del capitán y tripulación conforme el artículo 174 de la ley de Navegación Argentina (ley 20094) [2].

A partir del desarrollo de AIS SD y como el mismo procesa la información relativa a la navegación, se administran los dispositivos sensibles del puerto. Así, con este sistema se puede prevenir un daño de las instalaciones o detectar al responsable de haberlo ocasionado.

3 Solución propuesta

Como se mencionó anteriormente, los AIS clase A y B transmiten la localización del buque utilizando canales VHF. Esta información es analizada y procesada por AIS-SD. Se procesan datos estáticos y dinámicos. Ejemplos de los primeros son: número de identificación del buque (IMO), nombre del mismo, eslora. Para los datos dinámicos se pueden mencionar posición, tiempo de viaje, curso, velocidad, velocidad de giro, destino, hora de llegada, tipo de carga, etc. Para AIS-SD es importante conocer y procesar identificación del buque, posición, rumbo y velocidad, porqué a partir de estos datos se puede prever un posible acontecimiento o determinar el responsable del mismo si ya ocurrió. En caso de detectar posibles colisiones, el sistema busca información complementaria en otros AIS para confirmar la veracidad de la información procesada. De esta forma, con la corroboración de los datos es posible imputar los gastos producidos a los responsables sin el beneficio de la duda.

3.1 Software base utilizado

El sistema se fue desarrollado en Java para de esta forma, poder generar código para cualquier sistema operativo que permita instalar la máquina virtual respectiva. Además, se utilizó el framework Springboot para facilitar el desarrollo, con el patrón de diseño MVC (Model View Controler) como base. El sistema propuesto, es una API REST, que permite crear una interfaz de software de aplicación (API) que utiliza solicitudes https para acceder y manipular datos.

Springboot es un framework de código abierto que proporciona a los desarrolladores Java una estructura para comenzar una aplicación Spring con grado de Producción totalmente autoconfigurable. La principal ventaja es poder crear API propias y

desplegar servicios REST para que sea fácil la interacción con otros servicios, como por ejemplo aplicaciones móviles, webs o cualquier otro tipo de cliente que trabaje bajo el protocolo https.[5]

3.2 Funcionalidad de AIS-SD

La principal función del AIS-SD consiste en el análisis de la información almacenada en la estación de trabajo AIS instalada en el puerto donde se instala la aplicación desarrollada. El caso de prueba fue el puerto de la Ciudad de La Plata.

AIS-SD filtra las señales recibidas, primero por el rango especificado en la configuración del sistema y segundo por una diferencia en metros con la señal anterior. De esta forma si dos señales están a menos de 5 metros solo se procesará una porque el trayecto es muy pequeño.

Los filtros previstos son:

- Por perímetro
- Por puntos de interés

En el caso de filtro por perímetro, la función recibe como parámetro dos valores, el primero es una coordenada (latitud, longitud) y el segundo un lista de coordenadas que permiten definir un perímetro. La función retorna verdadero o falso en caso de determinar si la coordenada recibida está o no dentro del perímetro.

Para el filtro de puntos de interés la función opera de manera diferente. Un punto de interés se define como una zona próxima a una coordenada. Esta zona está delimita por un radio (en metros) a partir de dicha coordenada. La función recibe tres parámetros: el primero de ellos es la latitud y longitud de dispositivo (ejemplo la boya), el radio de acción a considerar es el segundo parámetro y el tercero la coordenada a evaluar. El resultado será nuevamente verdadero o falso en función de la distancia entre ambas coordenadas respecto del radio.

El problema aquí es que la distancia no puede calcularse como si fuera un plano o un mapa 2d. Como se tratan de distancias terrestres y se debe tener en cuenta la curvatura de la Tierra, y para ello se utiliza la formula del semiverseno. Esta fórmula es una ecuación para la navegación astronómica, que permite el cálculo de la distancia de círculo máximo entre dos puntos de un globo sabiendo su longitud y su latitud.

$$\operatorname{semiversin}\left(\frac{d}{R}\right) = \operatorname{semiversin}(\varphi_1 - \varphi_2) + \cos(\varphi_1)\cos(\varphi_2)\ \operatorname{semiversin}(\Delta\lambda).$$

Donde, R es el radio de la esfera (en este caso el radio del planeta Tierra y es aproximado), φ 1 es la latitud del punto 1, φ 2 es la latitud del punto 2, y Δ λ es la diferencia de longitudes. Si bien la función pierde precisión a medida que las distancias aumentan, en los casos evaluados por AIS-SE se tratan siempre de pequeñas distancias.

3.3 AIS-SD para el puerto La Plata

La figura 2 presenta la interfaz inicial del sistema desarrollado e instalado en el puerto de la ciudad de La Plata. Cuando la aplicación se instala para una zona portuaria es fundamental realizar la configuración de la misma. Dentro del área de influencia del

puerto se encuentran definidas todas las señales AID provenientes de las ayudas de navegación que serán resguardadas por la aplicación y por ende están en su zona de influencia. Para determinar la zona de influencia del puerto a analizar, basta con seleccionar un punto (latitud y longitud) dentro del mapa. Este punto debería ser aproximadamente el centro geográfico del puerto. Establecida esa coordenada, se define el radio que delimita las señales AID a recibir. En la figura 2, dicha zona está delimitada por el circulo de color verde.

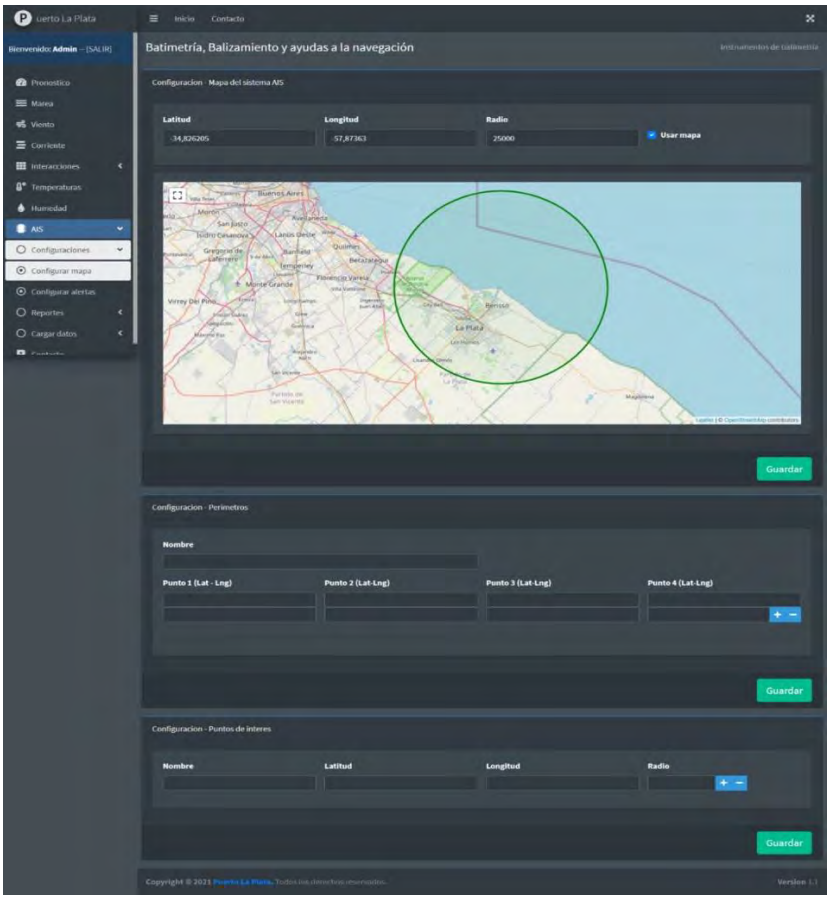


Fig. 2. Sistema AIDS-SD puerto La Plata.

Una vez establecida la zona de influencia se debe constituir al menos un perímetro del puerto. Cada perímetro queda delimitado por un conjunto de cuatro coordenadas cartográficas que delimitan el contorno de la superficie y al cual se le asigna un nombre único. Una vez delimitado los perímetros, se podrá controlar y alertar por el paso de un buque por dicho espacio.

Por último, en la configuración se deben indicar los puntos de interés. La figura 3 presenta una parte del mapa de la aplicación, donde además de observarse la zona de influencia del puerto, se muestra un punto de interés (en este caso una boya). El punto

se marca como la posición GPS del elemento en el río. Además, cada punto de interés esta nombrado de manera unívoca y tiene definido un radio de control, que permite determinar si una embarcación o artefacto de navegación se acerca o viola su zona de control. El radio de control de un punto de interés es mucho menor que el radio de la zona de influencia el cuerpo.

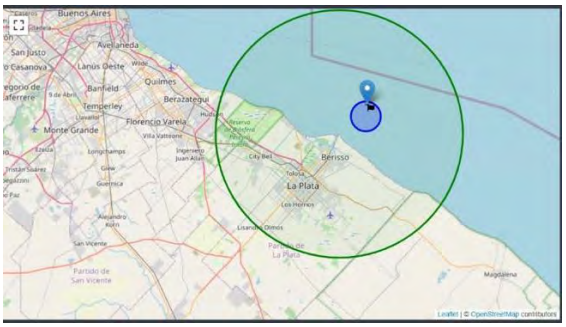


Fig. 3. Puntos de Interés.

AIS SD debe monitorear y almacenar todas las señales de su zona de influencia. Para ello se crean tareas a fin de procesar los mensajes. Una tarea es una acción que lleva a cabo un proceso. Esta tarea va a realizar el procesamiento de datos a fin de decodificar los mensajes que fueron recibidos a través del AIS, los que serán almacenados en una BD.

Luego de creada una tarea, se la identifica en forma univoca y se le asocia un archivo comprimido el cual puede constar de un conjunto de archivos con mensajes AIS los cuales posteriormente serán decodificados. Para el seguimiento de la tarea se cuenta con una tabla la cual especifica su nombre, su fecha de inicio y finalización y el estado de la misma. Estos estados puede ser: creado, en proceso, finalizado o error; en este último caso junto con un mensaje que informe el tipo de error producido.

La figura 4 presenta otra interfaz del sistema con información relativa a las embarcaciones. Esta información es la recopilada respecto de un buque debidamente identificado en un período de tiempo determinado. Se tiene: Fecha en la que se reciben señales AIS, identificación del buque (MMSI) y total de datos procesado en dicha fecha para dicho buque.

3.3 Generador de reportes

A partir de la información de configuración definida y el almacenamiento de los movimientos de buques en la zona de influencia es posible generar reportes que permitan identificar con seguridad el buque responsable de un evento determinado. Como ejemplo, si una ayuda de navegación (boya) dejara de transmitir o se detectara que su señal proviene de un ubicación GPS diferente a la que tiene definida puede significar que la misma fue colisionada por un buque. Para determinar el posible responsable, se genera un reporte de navegación por el área en cuestión, utilizando la información almacenada en la BD.

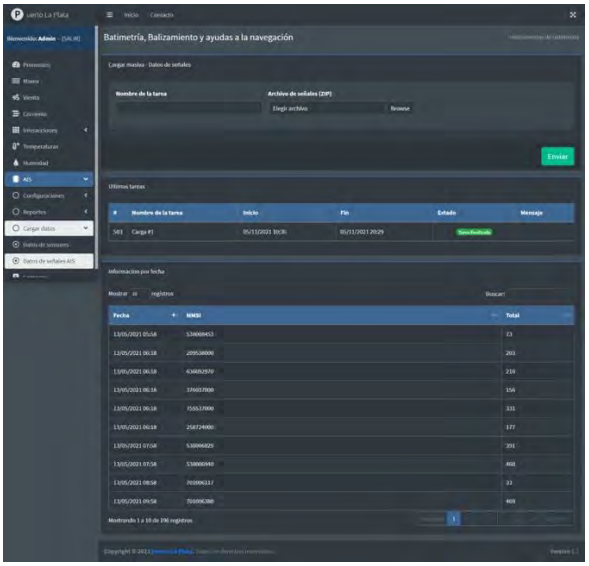


Fig. 4. Buques y posiciones.

El esquema de la figura 5 muestra el proceso de generación de reportes. Como la generación de reporte es un proceso del sistema, el mismo se define como una tarea más. Se indica un rango de fechas a evaluar y se determina con cada mensaje guardado dentro del rango de fecha especificado, si la embarcación se encontraba en el perímetro o punto de interés.

Fig. 4 Generación de Reportes.

4 Resultados Obtenidos

Al momento de definición e implantación del sistema, el puerto de La Plata carecía de una aplicación que permitiera monitorizar las incidencias producidas. Por dicha falencia, los gastos producidos en la reparación o reemplazo de ayudas de navegación eran, en general, responsabilidad económica del autoridad del puerto. Reparaciones o reemplazados de estos elementos resultan muy costosos y además, al tratarse de productos que no se fabrican en el país, contar con los mismos es lento y tedioso.

Desde que la aplicación fue puesta en funcionamiento se han detectado un número relativamente importante de eventos y en todos los casos, a partir de la generación de los reportes del sistema, se ha podido determinar las responsabilidades en cada caso.

Por ende, los costos de reparación han dejado de ser un problema para el presupuesto del puerto, dado que los mismos son imputados a los empresas propietarias de los buques.

El sistema se encuentra operativo desde hace 6 meses, considerando un tiempo muy limitado para poder extrapolar resultados mas representativos. Además, los datos de eventos producidos son considerados reservados por la autoridad portuaria.

5 Conclusiones

AIS tiene numerosas funciones que ayudan a la mejora de la navegación. Algunas de estas mejoras tienen que ver con la seguridad y eficiencia de dicha navegación, la protección del ambiente marino, facilitando la identificación de los buques. Esta última particularidad es de gran importancia para el presente trabajo, debido a que la tecnología de identificación de buque permite conocer el camino seguido por el mismo. A partir de la traza de navegación de cada barco, es posible crear un mapa virtual y detectar cuando se afecta la zona delimitada por un perímetro o punto de interés.

Los mensajes emitidos por AIS son utilizados, en su mayoría, en forma preventiva proporcionando información al navegante para la toma decisiones que eviten colisiones entre buques.

AIS Signal Detector utiliza la información emitida por AIS para determinar el responsable de generar un daño a una o varias instalaciones portuarias. El sistema procesa la BD generada a partir de los mensajes de datos proporcionados por AIS y se comparan con los perímetros y puntos de interés definidos en el sistema. Cuando se traspasa una determinada zona se crean una alerta.

AIS SD presenta una solución a un problema recurrente que es el costo de reparación de ayudas de la navegación. A partir de su utilización la información generada por AIS ya no solo funciona para evitar colisiones entre buques o como ayudas a la navegación, sino que también permite encontrar el responsable en caso de daños ocasionados a las instalaciones portuarias.

A partir de la utilización de la aplicación desarrollada, se puede decir que el puerto de La Plata es primer puerto de Argentina en contar con un sistema de esta naturaleza.

Referencias

1. https://www.argentina.gob.ar/prefecturanaval/ais
2. https://www.argentina.gob.ar/normativa/nacional/ley-20094-43550
3. 2005/53/CE, Directiva de Europa para Navegación. Enero 2005.
4. https://base-ais.com.ar/teoria-ais/
5. https://spring.io/projects/spring-boot

HERA: una Herramienta para la Evaluación de Recursos Académicos

Juan Francisco Porto[1], Enzo Rucci[2][0000−0001−6736−7358] ✉, and Gonzalo Villarreal[3][0000−0002−3602−8211]

[1] Facultad de Informática, UNLP. La Plata (1900), Bs As, Argentina
toto.lp77gmail.com
[2] III-LIDI, Facultad de Informática, UNLP – CIC.
La Plata (1900), Bs As, Argentina
erucci@lidi.info.unlp.edu.ar
[3] PREBI-SEDICI Universidad Nacional de La Plata CESGI - CIC
La Plata (1900), Bs As, Argentina
gonzalo@prebi.unlp.edu.ar

Resumen En la actualidad, determinar la calidad y el impacto de un recurso académico (revista o artículo científico) representa un verdadero desafío para un investigador. Al problema inicial de la existencia de múltiples indicadores y métricas de diversa índole para cada uno de ellos, se ha sumado el crecimiento exponencial en la cantidad de recursos para analizar gracias al desarrollo tecnológico. En este artículo se presenta el diseño y desarrollo de una herramienta web que busca dar respuesta a esta problemática. HERA es una herramienta que apunta a simplificar, agilizar y apoyar el proceso de determinar la calidad y el impacto de un recurso académico. Para ello, HERA consulta múltiples fuentes en tiempo real para luego ofrecer información de un recurso como metadatos, pertenencia a índices y bases de datos, medidas de citas y menciones, e información de la publicación donde figura dicho recurso en caso que corresponda.

Keywords: Bibliometría · Cienciometría · Evaluación bibliográfica · Recuperación de información

1. Introducción

La rapidez con la que la cantidad y disponibilidad de información aumenta ha tenido fuerte impacto en el modo en que los investigadores la utilizan. Es habitual que un investigador lea, acceda y utilice información extraída de artículos de otros colegas. Hasta hace algunas décadas atrás, el requisito de originalidad y el proceso de revisión por pares eran elementos suficientes para certificar la calidad del material con el que se trabajaba. Sin embargo, esto cambió de manera radical con el surgimiento de la Internet y de sistemas (más) automatizados para la publicación y evaluación de documentos científicos [15]. Un estudio realizado a finales del 2018 sugiere que al menos 3 millones de artículos fueron publicados en ese año, en aproximadamente 33100 revistas científicas [8].

Una característica muy buscada y asociada estrechamente con la calidad tanto de una revista como de un artículo es su impacto. De hecho, es un error común confundir calidad con impacto y tomarlos como sinónimos. La calidad de una revista se asocia a los procesos que esta sigue y a los requisitos que impone para que un artículo sea finalmente publicado. Por su parte, el impacto se asocia a la repercusión que la revista o el artículo tienen en la comunidad científica, usualmente medido en número de citas [14].

La necesidad de determinar calidad e impacto de recursos científicos se traduce en lo que hoy vemos como sistemas de evaluación, indicadores y métricas de revistas y artículos. Si bien no son una herramienta automatizada en su totalidad, ni mucho menos estandarizada, estos elementos permiten tener valores concretos o numéricos que pueden dar (o aproximar) respuestas a preguntas como: este recurso [4] ¿es de buena calidad?, ¿es de alto impacto?, ¿es útil?, ¿es válido?

Lamentablemente, los sistemas, métricas e indicadores mencionados anteriormente no están exentos de subjetividad ni de sesgos. En consecuencia, actualmente los miembros de la comunidad académica y científica se enfrentan a una tarea ardua y engorrosa cuando deben determinar la calidad y el impacto de una publicación científica. Esta situación se debe a dos factores. En primer lugar, al crecimiento exponencial en la disponibilidad de recursos que tuvo el área debido al desarrollo tecnológico. En segundo lugar, la ausencia de estándares y la consecuente existencia de múltiples métricas y sistemas de evaluación que, aunque comparten objetivos, no siempre lo llevan a cabo de la misma manera.

En este artículo se presenta HERA, una herramienta web que apunta a simplificar, agilizar y apoyar el proceso de determinar la calidad y el impacto de un recurso académico. Para ello, recopila información proveniente de diferentes bases de datos académicas (principalmente indicadores/métricas de calidad e impacto) en forma rápida, para luego ser exhibidas de forma integrada y amigable al usuario. Hasta donde llega el conocimiento de los autores, HERA es la primera herramienta de su clase que trabaja a nivel de artículo.

El resto del artículo se organiza de la siguiente forma. La Sección 2 introduce el marco referencial para este trabajo. Luego, la Sección 3 describe la herramienta propuesta. A continuación, la Sección 4 muestra su funcionamiento mientras que la Sección 5 resume las conclusiones junto al trabajo futuro.

2. Marco Referencial

2.1. Métricas e Indicadores de Calidad e Impacto Bibliográfico

En el caso de las revistas, la calidad suele determinarse a partir de su presencia en determinados sistemas de evaluación, conocidos usualmente como bases de datos bibliográficas. Cabe destacar que, en rigor de verdad, esto no se trata de una métrica por no poseer un valor numérico, con lo cual la interpretación del peso de esta

[4] Se entiende como recurso a cualquier publicación individual o seriada que resulte habitual en la ciencia.

información la hace individualmente cada investigador. Cada una de estas bases de datos tiene sus propios propósitos y criterios de inclusión, y existen en la actualidad una multiplicidad de ellas. Por su parte, la calidad de los artículos habitualmente se asocia a la revista en que fue publicada, por la necesidad de cumplir sus requisitos de publicación.

Como se mencionó anteriormente, el impacto se suele medir en número de citas. Desde el surgimiento del primer indicador en aproximadamente 1960 [1], se han desarrollado una variedad de métricas e indicadores basados en citar para determinar el impacto de una revista o artículo. Si bien este trabajo se enfoca principalmente en citas a documentos tradicionales (por ej. artículos científicos y tesis), el universo de objetos citables es cada vez más amplio, incluyendo a conjuntos de datos [2], software [9] y registros de propiedad intelectual [7], y está generando nuevos debates en la comunidad científica respecto a cómo medir su impacto, la validez de las citas, técnicas de recolección de métricas, etc [5].

Generalmente, es aceptado que un número alto de citas representa una buena medida de la calidad de un artículo. Esto es porque se puede interpretar, como se piensa comúnmente [10], que si un artículo es muy citado, significa que tiene una amplia contribución al campo de investigación que trata. Muchos autores lo habrían considerado *útil* o *válido*, y por ende lo utilizaron. Sin embargo, resulta importante aclarar que también existen múltiples controversias sobre el uso de la cantidad de citas para medir su calidad o impacto. Algunas de ellas son las siguientes:

- Un número de citas alto puede pertenecer tanto a un artículo bueno como a uno malo. Las citas pueden provenir no sólo de aquellos que valoran y aprecian la labor de un científico y la reutilizan, sino que puede darse el caso en que investigadores realizan críticas o refutan al trabajo de otro [6].

- El tipo de publicación de una investigación puede afectar directamente a la cantidad de citas. Tal situación se ve reflejada en la comunidad de Ciencias de la Computación [4], donde se suelen publicar trabajos completos en congresos y no sólo en revistas. Por esta costumbre, tienden a tener números de citas más bajos que aquellos que publican en revistas de manera tradicional.

- Diversos estudios como [3] muestran que publicar en abierto hace a los artículos más "citables". Esto no se relaciona con que sean de mejor calidad, sino por el hecho de que tienen mayor disponibilidad que los que se publican en cerrado. Considerando que muchas revistas sólo cobran si publican en abierto, esto se vuelve una dificultad adicional para los investigadores de menor cantidad de recursos económicos o antecedentes en el área.

- Lamentablemente, también se han detectado maneras de manipular las citas de aquellos sitios "inteligentes" que indexan contenido académico mediante mecanismos automatizados, como por ejemplo Google Scholar [11].

Para concluir, el número de citas podrá ser un valor objetivo pero es una medida que nace de la apreciación subjetiva de la que el autor citante no se puede desprender. Es importante tenerlo en cuenta y, de manera imperativa, complementarlo con las

demás métricas [12], y aspectos para tener un panorama amplio a la hora de evaluar la calidad de un recurso académico.

2.2. Bases de Datos Bibliográficas

En la actualidad, las dos bases de datos de revistas más reconocidas son Scopus [5] y Web of Science (WoS) [6], cuyos dueños y gestores son grandes empresas. En ambas, la indexación e inclusión de artículos y revistas es de carácter selectivo. Existen juntas directivas de científicos o editores de campos específicos que se aseguran que el material a incluir mantenga ciertos niveles de calidad que ellos consideran adecuados. Luego, ambos ofrecen otras métricas con rankings y percentiles por categoría, las conocidas citas e incluso cálculos de performance esperada. El objetivo de sitios como éstos es garantizar siempre que la información que se indexa es de carácter científico y relevante.

Para dar respuesta a las limitaciones de Scopus y WoS, surgieron diversas bases de datos provenientes de consorcios académicos, con una estrecha relación con la promoción del movimiento de acceso abierto. Un ejemplo a nivel mundial es el Directory of Open Access Journals (DOAJ) [7], que evalúa las políticas de acceso abierto para la inclusión de artículos o revistas en su base de datos. Un artículo o revista será indexado fácilmente si no tiene restricciones de acceso, y podrá incluso acceder a un sello que lo destacará, si cumple aún con requisitos más específicos que demuestren su calidad bajo los criterios del sitio. A nivel de Iberoamérica, se pueden mencionar Latindex [8] y la Red Iberoamericana de Innovación y Conocimiento Científico (REDIB) [9]. Latindex se propone recopilar publicaciones científicas producidas en iberoamérica. Las revistas indexadas aparecerán en lo que denominan Directorio, pero además existe una sección llamada Catálogo donde sólo aparecen aquellas que se consideran de la más alta calidad. Por su parte, REDIB es una plataforma que recopila contenidos científicos y académicos en formato electrónico, producidos en el ámbito iberoamericano. Está dirigida tanto a la comunidad editora y científico-académica como a la sociedad en general y al sector empresarial e industrial.

El avance tecnológico posibilitó la indexación automática de contenido en Internet por parte de motores de búsquedas como Google o Bing!. Así es como surgen bases de datos como las de Google Scholar [10], Microsoft Academic [11] y Semantic Scholar [12], que acumulan la información de lo que ellos consideren contenido académico, siendo notorio que se puede encontrar casi cualquier documento que cumpla su forma. Los algoritmos de aprendizaje automático de éstos son muy potentes, e incluso capaces de indexar hasta videoconferencias grabadas. Al ser más laxos en cuanto a criterios

[5] www.scopus.com
[6] clarivate.com/webofsciencegroup/solutions/web-of-science/
[7] doaj.org/
[8] www.latindex.org
[9] redib.org
[10] scholar.google.com
[11] www.microsoft.com/en-us/research/project/academic/
[12] www.semanticscholar.org

de inclusión, sus números de citas suelen ser más elevados que el de bases de datos curadas y más conservadoras, como Scopus y WoS.

Por último, encontramos servicios como Altmetric [13] o Dimensions [14], que generan aplicaciones que intentan enriquecer el proceso de evaluación de revistas y/o artículos proveyendo de una visión más integradora de la cuestión. Estos servicios presentan usualmente versiones ampliadas de las citas tradicionales, integrando menciones en redes sociales, videos, y conferencias, y en cálculos de diversas índoles. Por ejemplo, en base a estadísticas de acceso a los recursos, comparaciones con publicaciones de los mismos campos, y observaciones de citas en intervalos de tiempo específicos, otorgan de manera sintética indicadores de tendencias y predicciones que permiten evaluar rápidamente el impacto y contribución que un recurso logra en el mundo académico.

2.3. Estado del Arte

Existen algunas aplicaciones que intentan dar solución a la problemática descrita, ofreciendo a los investigadores una recopilación de información que permite la evaluación de los recursos académicos con métodos variados. Por ejemplo, el sitio web ¿Dónde lo Publico? (DLP) [15] recopila información de revistas de ciencias sociales y humanidades en español o portugués. Una opción superadora es la Matriz de Información para el Análisis de Revistas (MIAR) [16], que da una vista amplia, detallada e integral de métricas de los revistas que posee en su base de datos. Cabe aquí mencionar que MIAR no provee información a nivel de artículo y que el proyecto DLP ha sido discontinuado.

3. Propuesta

3.1. Propósito

La herramienta desarrollada apunta a simplificar, agilizar y dar apoyo al proceso de evaluación de recursos académicos:

- La simplificación está dada por dos factores: 1) ante la multiplicidad y diversidad de bases de datos académicas existentes, HERA realiza una selección basada en criterio de expertos, logrando que el usuario final tenga una visión representativa del recurso a bajo costo; y 2) el funcionamiento no requiere más que ingresar el identificador del recurso en un campo de búsqueda y hacer clic en un botón.
- La agilidad proviene del proceso de recopilación automático de información que realiza HERA para un recurso determinado, lo que exime al usuario de la ardua tarea de realizar búsquedas en sitios externos de forma individual.
- El apoyo al proceso de evaluación se basa en la recopilación y presentación conjunta de métricas e indicadores diversos asociado al recurso de interés. Rápidamente se podrán conocer números de citas (tradicionales y alternativas), factores

[13] `www.altmetric.com`
[14] `www.dimensions.ai`
[15] `www.dondelopulbico.com`

de impacto, información sobre licencias y acceso abierto, e inclusión en bases de datos reconocidas. De esta manera, HERA provee una visión integral del recurso para que el usuario pueda realizar un análisis propio según sus objetivos e intereses.

3.2. Análisis y Decisiones de Diseño

Identificadores de los Recursos Para recopilar los datos acerca de un recurso académico, primero es necesario buscarlo en diversas bases de datos, de las cuales es posible extraer los metadatos que describen cada recurso. En una primera instancia fue considerada la alternativa de buscar un recurso en base a su título. Esta opción probó ser de poca utilidad, ya que no sólo pocas bases de datos ofrecen la posibilidad de buscar un recurso por este medio, sino que las que sí otorgan esta posibilidad devuelven información en base a similitud. De manera alternativa, y dando solución a estos inconvenientes, se tomó la decisión de utilizar identificadores capaces de localizar de manera unívoca a los recursos académicos: DOI para artículos e ISSN para revistas/publicaciones seriadas. Todas las bases de datos seleccionadas permiten realizar búsquedas utilizando estos identificadores (según corresponda), y además evita tener que verificar que los datos devueltos sean los correctos. Colateralmente, la herramienta es incapaz de encontrar un recurso que no esté identificado por alguno de estos valores, y se introduce la leve complejidad de buscar por estos identificadores, siendo que resulta más sencillo utilizar algo natural como lo es un título.

Modalidad de Acceso a la Información de las Bases de Datos La posibilidad de almacenar la información de los recursos en una base de datos fue descartada rápidamente por el volumen de información a manejar y la necesidad de actualizaciones periódicas para mantenerla coherente. En su lugar se optó por obtener la información en tiempo real aprovechando la disponibilidad de servicios en línea que proveen la mayoría de las bases de datos. Esta alternativa no carece de desventajas, ya que introduce una pequeña carga de trabajo y demora para la visualización de la información, además de la dependencia de los servicios que se consultan. Sin embargo, se ve compensada por la omisión de la base de datos antes mencionada y del hecho de proveer siempre información actualizada.

Medios de Acceso a las Bases de Datos Inicialmente, se consideró el protocolo OAI-PMH para la obtención de información de los recursos. Sin embargo, luego de su análisis, se lo descartó por implicar diferentes inconvenientes en su uso (extracción dificultosa de un registro, carga elevada de procesamiento, inconsistencia de datos, entre otros). Afortunadamente, la mayoría de las bases de datos cuenta con una API disponible, que permite obtener información de un recurso mediante su identificador. Este proceso resulta efectivo, además de ser relativamente sencillo de usar. Para aquellos sitios que no ofrecían una API pero resultaban de gran interés para extraer datos, se optó por utilizar la técnica de *web scrapping*. Este proceso no es viable a gran escala ya que implica desarrollar un algoritmo específico para cada base de datos, que puede requerir ser actualizado ante cambios en su código HTML.

Sin embargo, es la única opción viable ante la ausencia de servicios en línea que permitan obtener la información.

Selección de Bases de Datos y Métricas Para tener una visión amplia e integral de los recursos, se consideraron las siguientes bases de datos: Scopus, WoS, DOAJ, Latindex, REDIB, Crossref, Scimago, Google Scholar, Microsoft Academic, Semantic Scholar, Altmetric y Dimensions. De este conjunto no fue posible incluir a Latindex y a Google Scholar ya que no ofrecen API a las cuales poder consultar y que la generación del código HTML de sus sitios emplean métodos para su visualización que imposibilitan su extracción. Finalmente, las bases de datos que ofrecen API y que pudieron ser aprovechadas son Scopus, WoS, DOAJ, Crossref, Microsoft Academic, Semantic Scholar, Altmetric y Dimensions. En el caso de REDID y Scimago, por su peso específico en el área, se empleó web scrapping. Más detalles se pueden consultar en [13].

3.3. Diseño y Desarrollo

HERA consiste de una aplicación web, que está dividida en una aplicación *frontend*, desarrollada con ReactJS, y en una aplicación *backend*, desarrollada con NodeJS, como se muestra en la Fig. 1. La aplicación backend tiene como objetivo servir de puente entre la aplicación frontend y los diversos servidores que proveen las métricas de los recursos académicos, concentrando de esta forma la mayor carga de trabajo en sí misma. La aplicación frontend realiza peticiones a la de backend mediante una API REST, recopilando los resultados, y es la encargada de generar las visualizaciones de los mismos en un portal web. A continuación, se resume su información pero más detalles pueden ser consultados en [13].

Diseño de Aplicación Backend Fue desarrollada con NodeJS y utilizando el framework express [16] para definir los endpoints de la API. También incorpora el módulo node-fetch para ser capaz de realizar peticiones por HTTP a las diferentes bases de datos y procesar las respuestas. Consta de dos funcionalidades principales: 1) Almacenamiento de constantes de acceso, requeridas para acceder a las API de Scopus y Microsoft Academic; y 2) Exposición de una API que brinde la funcionalidad necesaria para que la aplicación de frontend pueda delegar las solicitudes de datos a las diferentes bases de datos.

Diseño de Aplicación Frontend Fue desarrollada con ReactJS, con el objetivo de elaborar una aplicación web en la que se visualicen las métricas recopiladas. Se encarga de interpretar el pedido de un usuario y comunicarse con la aplicación de backend para obtener datos de métricas relevantes. Por ser una librería de JavaScript, la parte lógica y funcional de la aplicación puede ser programada con JavaScript puro o incorporar cualquier otra librería/framework que se desee. Para el caso de esta aplicación, en la parte funcional el único framework utilizado es la Web API Fetch [17], similar al uso de XMLHttpRequest.

[16] `https://expressjs.com/`
[17] Fetch API `https://developer.mozilla.org/en-US/docs/Web/API/Fetch_API`

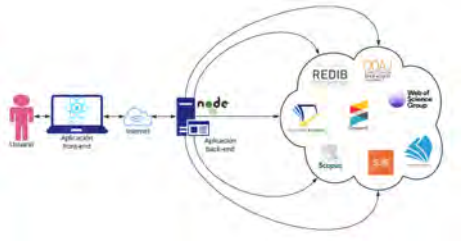


Figura 1: Esquema de comunicaciones en el funcionamiento de HERA

Figura 2: Interfaz gráfica de usuario de HERA

3.4. Interfaz Gráfica de Usuario

Se cuenta con una única interfaz de usuario, la cual se muestra en la Fig. 2. En ella, el usuario debe seleccionar el tipo de búsqueda, que puede ser por DOI o por ISSN, y luego colocar en la barra el identificador correspondiente. Finalmente, hacer clic en "Buscar".

4. Pruebas y Resultados

HERA se encuentra disponible en `http://hera.sedici.unlp.edu.ar`. A continuación se muestra su funcionamiento tanto para recursos identificados con ISSN como con DOI. Por cuestiones de espacio, sólo se muestran dos casos de prueba pero se pueden consultar el proceso de verificación completo en [13].

4.1. Búsqueda por ISSN

Al introducir un ISSN en la barra de búsqueda y presionar el botón Buscar, se realiza la búsqueda de ese recurso. De este se obtiene primero la información sobre la editorial y enlace al recurso, así como la vista del resumen de métricas recopiladas de los diferentes sitios, visualizado en la Fig. 3. Adicionalmente, se cuenta con un botón para expandir la información y así visualizar la totalidad de su contenido, como se muestra en las Fig. 4, 5 y 6. En este enlace `https://youtu.be/eERFzl0m3KU` se muestra el proceso de búsqueda con mayor detalle.

4.2. Búsqueda por DOI

Al introducir un DOI en la barra de búsqueda y presionar el botón Buscar, las métricas son recopiladas de las diferentes bases de datos y se renderizan en pantalla, como se muestra en la Fig. 7. Entre los resultados se pueden apreciar algunos metadatos de los recursos, como ser título, tipo de recurso, resumen, entre otros, así como el nombre de la revista/publicación seriada donde se publicó el artículo y un enlace para dirigirnos a la página original del mismo. También, de ser posible, la aplicación intentará recolectar métricas del la revista donde se encuentra el artículo para mostrarlas en conjunto, como se ve en la Fig. 8. Para ello, se extrae de los metadatos

Figura 3: Resultados generales de la búsqueda de un recurso ISSN.

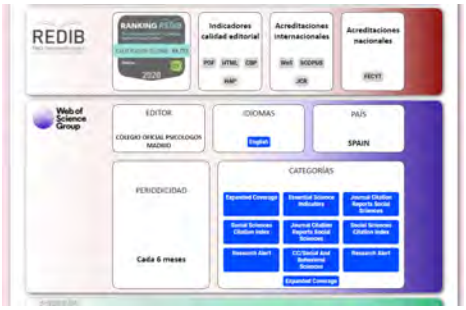


Figura 4: Vista expandida de métricas de un recurso ISSN (1).

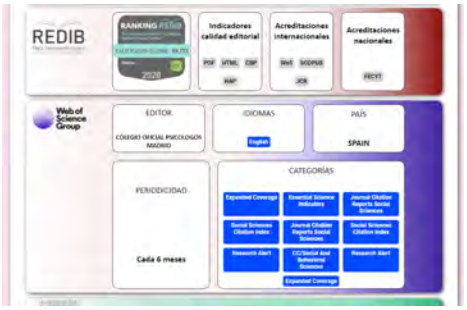


Figura 5: Vista expandida de métricas de un recurso ISSN (2)

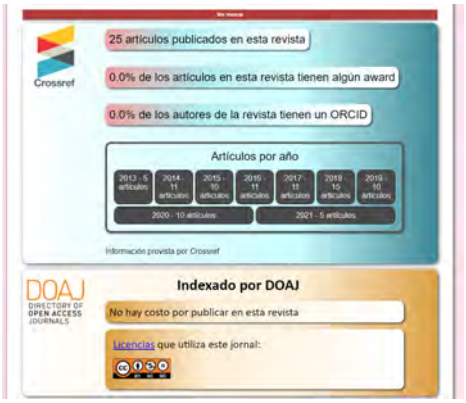


Figura 6: Vista expandida de métricas de un recurso ISSN (3).

recuperados del ISSN de la revista y se ejecuta de manera automática y en segundo plano una búsqueda por este recurso. De esta forma se amplía el contexto de evaluación de un artículo, al poder visualizar sus métricas de forma aislada y también tener acceso a las de la revista y así entender qué influencia podría tener en las del primero. En general, la calidad de un artículo está asociada a la de la revista en la que está publicado. Al presionar el botón "Ver más" que se ve en esta última imagen, se genera una vista expandida de las métricas en la cual se exhibe información más detallada (en aquellos casos donde las bases de datos nos provean algún dato adicional), ilustrado en la Fig. 9. En este enlace `https://youtu.be/aLTm9ht3LtI` se muestra el proceso de búsqueda con mayor detalle.

5. Conclusiones y Trabajo Futuro

Determinar la calidad y el impacto de una revista o un artículo científico se torna un desafío cada vez más dificultoso debido a 2 factores: (1) el crecimiento exponencial en la cantidad de recursos para analizar gracias al desarrollo tecnológico; y (2) la

Figura 7: Resultados generales de la búsqueda por DOI (vista resumida)

Figura 8: Métricas correspondientes a la publicación seriada donde se encuentra el recurso DOI (vista resumida)

existencia de múltiples indicadores y métricas de diversa índole para cada clase de recurso. En este contexto, HERA representa una herramienta que de manera ágil permite integrar y visualizar métricas de bases de datos relevantes y reconocidas en un único sitio. Esto es de vital importancia a la hora de evaluar contenido, no sólo por eximir a los investigadores de la tediosa tarea de buscar métricas en muchas páginas web, sino también por facilitar la visualización de métricas para un análisis integral de estas. Considerando las características de HERA, esperamos que los miembros de la comunidad académico-científica la encuentren útil para evaluar la calidad y el impacto de los recursos académicos y que contribuya a facilitar y acelerar dicha tarea.

Figura 9: Vista expandida para recurso DOI

Entre las líneas de trabajo futuro se encuentran:

- Expandir el banco de bases de datos académicas: Considerar índices regionales, como por ejemplo Redalyc, y también bases reconocidas pero de disciplinas más específicas como el caso de PubMed.
- Implementación de una API REST: La herramienta podría exponer parte de su funcionalidad mediante el desarrollo de una API REST para así desarrollar nuevas herramientas que se nutran de los datos brindados por HERA.
- Optimización de tiempo de búsquedas: Sería de utilidad poseer mecanismos de caché para optimizar los tiempos en caso de búsquedas recurrentes.

Referencias

1. History of citation indexing, `https://clarivate.com/webofsciencegroup/essays/history-of-citation-indexing/`
2. Berez-Kroeker, A.L., Gawne, L., Kung, S.S., Kelly, B.F., Heston, T., Holton, G., Pulsifer, P., Beaver, D.I., Chelliah, S., Dubinsky, S., Meier, R.P., Thieberger, N., Rice, K., Woodbury, A.C.: Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics **56**(1), 1–18 (Jan 2018). https://doi.org/10.1515/ling-2017-0032
3. Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., Harnad, S.: Self-Selected or Mandated, Open Access Increases Citation Im-

pact for Higher Quality Research. PLOS ONE **5**(10), e13636 (Oct 2010). https://doi.org/10.1371/journal.pone.0013636
4. Godoy, D., Zunino, A., Mateos, C.: Publication practices in the Argentinian Computer Science community: a bibliometric perspective. Scientometrics **102**(2), 1795–1814 (Feb 2015). https://doi.org/10.1007/s11192-014-1450-0
5. Groth, P., Cousijn, H., Clark, T., Goble, C.: FAIR Data Reuse – the Path through Data Citation. Data Intelligence **2**(1-2), 78–86 (Jan 2020). https://doi.org/10.1162/dint$_{a0}$0030
6. Gruber, T.: Academic sell-out: how an obsession with metrics and rankings is damaging academia. Journal of Marketing for Higher Education **24**(2), 165–177 (Jul 2014). https://doi.org/10.1080/08841241.2014.970248
7. Guerrero-Bote, V.P., Sánchez-Jiménez, R., De-Moya-Anegón, F.: The citation from patents to scientific output revisited: a new approach to the matching Patstat / Scopus. Profesional de la información **28**(4) (Jun 2019). https://doi.org/10.3145/epi.2019.jul.01
8. Johnson, R., Watkinson, A., Mabe, M.: The stm report. an overview of scientific and scholarly publishing. Tech. rep., The International Association of Scientific, Technical and Medical Publishers (2018)
9. Katz, D.S., Chue Hong, N.P., Clark, T., Muench, A., Stall, S., Bouquin, D., Cannon, M., Edmunds, S., Faez, T., Feeney, P., Fenner, M., Friedman, M., Grenier, G., Harrison, M., Heber, J., Leary, A., MacCallum, C., Murray, H., Pastrana, E., Perry, K., Schuster, D., Stockhause, M., Yeston, J.: Recognizing the value of software: a software citation guide. F1000Research **9**, 1257 (Jan 2021). https://doi.org/10.12688/f1000research.26932.2
10. Lindsey, D.: Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid. Scientometrics **15**(3), 189–203 (Mar 1989). https://doi.org/10.1007/BF02017198
11. Lopez-Cozar, E.D., Robinson-Garcia, N., Torres-Salinas, D.: Manipulating Google Scholar Citations and Google Scholar Metrics: simple, easy and tempting (Feb 2013). https://doi.org/10.48550/arXiv.1212.0638
12. Meyer, B., Choppy, C., Staunstrup, J., van Leeuwen, J.: Viewpoint Research evaluation for computer science. Communications of the ACM **52**(4), 31–34 (Apr 2009). https://doi.org/10.1145/1498765.1498780
13. Porto, J.F.: HERA: Herramienta para Enriquecimiento de Recursos Académicos. Tesina de Licenciatura en Sistemas, Universidad Nacional de La Plata (Dec 2021), `http://sedici.unlp.edu.ar/handle/10915/129874`
14. Repiso, R.: Cómo identificar una revista de calidad. Cardiocore (2015), `https://www.redalyc.org/articulo.oa?id=277041630002`
15. Rozemblum, C., Unzurrunzaga, C., Banzato, G., Pucacco, C.: Calidad editorial y calidad científica en los parámetros para inclusión de revistas científicas en bases de datos en acceso abierto y comerciales **4**, 64–80 (abr 2015), `https://www.palabraclave.fahce.unlp.edu.ar/article/view/PCv4n2a01`
16. Urbano, C., Somoza-Fernández, M., Rodríguez-Gairín, J.M., Ardanuy, J., Guardiola, E., Pons, A., Borrego, , Brucart, J.M., Cosculluela, A.: MIAR : una base de datos para la identificación y la evaluación de la difusión secundaria de revistas de humanidades y ciencias sociales (2005), `http://eprints.rclis.org/6267/`

Interacción Humano Robot en el Contexto de la Computación Afectiva
Asociando estados emocionales al comportamiento de un Robot.

Alan Roldan, Fernando Yapura, Jorge Ierache, Iris Sattolo., Fernando Elkfury, Gabriela Chapperon.

Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias
Laboratorio de Sistemas Inteligentes y Enseñanza Experimental de la Robótica
Secretaria de Ciencia y Tecnología

Cabildo 134, Buenos Aires, Argentina
[jierache ,aroldan,fyapura,isattolo,felkfury,gchapperon]@unimoron.edu.ar

Resumen: Se presentan los resultados preliminares del desarrollo de un framework de interacción emocional Humano-Robot que contribuye con la configuración de estados emocionales, del tipo de robot (físicos o virtual) y sus acciones asociadas en respuesta al estado emocional. Para este proyecto, se trabajó en la integración de distintos sistemas entre ellos se destaca el software Emotion Detection Asset, que se encargará de reconocer emociones a través de expresiones faciales, capturadas por medio de una webcam o de una imagen importada desde un archivo; interfase de usuario por cual se puede realizar diferentes configuraciones; robots físicos (Roboreptile) y/o virtuales, para la representación o ejecución de acciones en respuesta a las emociones capturadas del humano, finalmente se realizan pruebas con software de reconocimiento de emociones propietario. En la primera sección "introducción" se presentan las características generales del área de computación afectiva, enfoque categórico de emociones, modelos multimodales y unimodales, emociones, finalmente se presenta una síntesis comparativa de los trabajos específicos de emociones y robots. En la segunda sección se presenta sintéticamente el problema, en la tercera sección se plantea la solución desarrollada, en la cuarta sección se presentan las pruebas preliminares, finalmente en la quinta sección se enuncian las conclusiones y futuras líneas de trabajo.

. Keywords: Reconocimiento facial – robots – emociones – computación afectiva - expresiones faciales – framework

1. Introducción

Como parte de la nueva era digital, el reconocimiento facial se está convirtiendo en una tecnología con gran potencial. Esta puede ser aplicada en distintos ámbitos y combinada con distintos dispositivos, para desbloquear un celular, identificar personas, reconocer distintas emociones predominantes en un individuo, etcétera. En este trabajo se utilizó para detectar expresiones de felicidad, tristeza, sorpresa, asco, enojo, miedo y neutralidad, a través de gestos faciales. Se desarrolló un framework que ante una entrada responda con una determinada salida, integrando un software de reconocimiento facial con un robot. El primero, capturará la o las emociones a través de una imagen, esta puede ser una foto o una captura de una cámara en tiempo real y, a su vez, transmitir la emoción detectada, a un robot para que el mismo pueda expresarla de forma sintética y así poder simular la empatía del autómata. Si bien el framework acepta distintas entradas (software de reconocimiento facial, ingreso de valores manuales, API), éstas no funcionan simultáneamente, es por eso por lo que no es considerado multimodal.

1.1 Computación Afectiva

La computación afectiva es una de las ramas más modernas de la ciencia de la computación. Tuvo sus orígenes en un grupo de investigación del Massachusetts Institute of Technology (MIT) y fue definida por Rosalind Picard en el año 1995 como "la informática que se relaciona con las emociones, no sólo con las consideradas más importantes, como la alegría o la tristeza, sino también con el interés, el aburrimiento o la frustración, que son las que se dan en relación con los ordenadores." [1]. La computación afectiva representa uno de los desafíos actuales y emergentes en el campo de los sistemas y tecnologías de la información. Ésta se enfoca en el estudio y el desarrollo de sistemas y dispositivos que pueden reconocer, interpretar, procesar y estimular las emociones humanas. Rosalind Picard define que la computación afectiva es "la informática que se relaciona con las emociones, no sólo con las consideradas más importantes, como la alegría o la tristeza, sino también con el interés, el aburrimiento o la frustración, que son las que se dan en relación con los ordenadores" [2]. Su rápido crecimiento se ha visto reflejado en distintas ramas como seguridad, salud, marketing, robótica y educación, entre otras. Actualmente su objetivo es desarrollar dispositivos y sistemas que puedan reconocer, interpretar, procesar y/o simular las emociones humanas para mejorar la interacción entre el usuario y la computadora. Estos sistemas "afectivos" [3], por lo tanto, deben ser capaces de: 1) capturar y reconocer los estados emocionales del usuario a través de mediciones sobre señales generadas en la cara, la voz, el cuerpo, o cualquier otro reflejo del proceso emocional que se esté llevando a cabo; 2) procesar esa información clasificando, gestionando, y aprendiendo por medio de algoritmos que se encargan de recoger y comparar gran cantidad de casos, y que tienen en cuenta los estados emocionales del usuario y, en su caso, las determinadas por el ordenador; y, por último, 3) generar las respuestas y las emociones correspondientes, que pueden expresarse a través de diferentes canales: colores, sonidos, robots, o personajes virtuales dotados de expresiones faciales, gestos, voz, etc. [3].

1.2 Enfoque categórico

Paul Ekman, considerado uno de los psicólogos más destacados del siglo XX, ha sido un pionero en el estudio de las emociones humanas, y su relación con las expresiones faciales. En su libro "The repertoire of nonverbal behavior: categories, origins, usage and coding" [4] plantea la existencia de 6 expresiones faciales universales que trascienden el idioma y las diferencias regionales, culturales y étnicas, a las que relaciona con 6 emociones basales: 1) Enojo, 2) Asco, 3) Miedo, 4) Felicidad ,5) Tristeza, 6) Sorpresa. Años más tarde, en su trabajo [5] "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion", adiciona una séptima expresión facial: Desprecio. El conjunto de las emociones mencionadas anteriormente conforma el enfoque categórico. (figura 1)

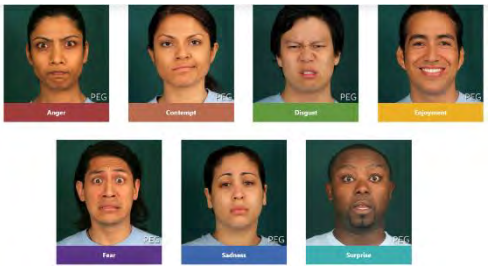


Figura N°. 1 Paul Ekman conjunto de siete emociones básicas y universales
Fuente de Imagen: https://www.paulekman.com/universal-emotions/

1.3 Enfoques unimodales y multimodales.

De acuerdo con los tipos de datos que se utilizan en los sistemas propuestos pueden clasificarse en sistemas unimodales, los que exploran una sola fuente de datos, y multimodales, los que combinan dos o más fuentes de datos. Para deducir el estado emocional de un individuo en un contexto multimodal se tiene que registrar, simultáneamente, diversa información biométrica.

Como trabajos unimodales se pueden citar como ejemplo a:[6] [7]y [8] en los cuales se captura el rostro a través de videos.

Trabajos como [9] agregan a la captura del rostro la posición de la cabeza. Como multimodales en [10] utilizan cuestionarios, tecleo y micrófono; [11] utilizan movimientos del mouse y teclado y [12] captura de patrones de tecleo. Se enuncian algunos ejemplos de trabajos multimodales que integran información fisiológica: [13] EMG Presión del volumen, Sanguíneo Conductividad de la piel Respiración, HR ECG, Volumen de respiración, Temperatura de la piel, [14] EEG HR Presión arterial GSR Respiración, [15] ECG (HRV) EEG.

En contextos multimodales existen interfaces que permiten sensar parámetros biométricos. En la actualidad se experimenta con el control de computadoras, dispositivos, robots, drones, juegos, etc. a través de Brain Control Interface (BCI o también conocidas como BMI, Brain Machine Interface), en la mayoría de los casos correspondiente a sistemas específicos e integrados a estos. En este orden podemos mencionar diversos desarrollos de control de Robots con el empleo de BCI orientados

a la navegación domótica [16] [17], otros empleos del BCI orientados al control de artefactos en un contexto de domótica [18].

1.4 Robots y Emociones

Se presenta en la tabla 1 una síntesis comparativa de trabajos en los cuales intervienen robots y transmisión de emociones, sumando la propuesta del presente artículo. Se seleccionaron las siguientes características para la comparación:

- Método de entrada: esta característica hace referencia a cómo el sistema adquiere la o las emociones. Entre ellos puede ser imagen, sonido, video, texto, etc.
- Representación a través de robot físico: esta característica hace referencia a si el sistema utiliza un robot real (Hardware) para representar la o las emociones capturadas.
- Representación a través de robot virtual: esta característica hace referencia a si el sistema utiliza un robot virtual (imágenes, avatar, etc) para representar la o las emociones capturadas.
- Robot utilizado: esta característica hace referencia al nombre del robot empleado en las pruebas.
- Acoplamiento: esta característica hace referencia a que tan dependiente es el sistema de sus distintos componentes. El acoplamiento puede ser bajo, medio o alto.
- Emociones percibidas: esta característica hace referencia a aquellas emociones que el sistema puede parametrizar y posteriormente representar.

Tabla N° 1 síntesis comparativa: Robots y Emociones

	Ierache et al [19]	Kishi[20]	Takato Horii [21]	**Propuesta**
Método de entrada	Captura de rostro (cámara). Neurosky Emotiv. Sensores fisiológicos.	Detección de la cercanía de un objeto a través de la visión del robot (cámara).	Visión del robot (cámara) Sonido (micrófonos)	Software de reconocimiento facial (cámara o imágenes). Ingreso manual de valores de emociones (teclado). Código abierto para futuras líneas de trabajo.
Rep. Robot físico	si	si	Si	Si
Rep. Robot virtual	no	no	no	Si
Robot	Bípedo Robosapiens V1 Wow Wee Robotics	KOBIAN-R humanoid robot	Robot iCub	Roboreptile Avatar virtual

	Robot móvil Lego NXT			
Acoplamiento	Bajo	Alto	Alto	Bajo
emociones	Circunflejo de Russel	Neutral, felicidad, sorpresa, asco	Neutral, Felicidad, Enojo, Tristeza, Valencia emocional	Felicidad, Tristeza, Sorpresa, Asco, Enojo, Miedo, Desprecio, Neutral (extra)

Particularmente la propuesta articula con un bajo acoplamiento, permitiendo integrar distintos robots, físicos y virtuales

2. Problema

Si bien existen múltiples proyectos similares, la mayoría de ellos sólo se centran en las expresiones y el reconocimiento de las emociones. Además, éstos se encuentran fuertemente acoplados (dedicados a un robot específico), con lo cual realizar modificaciones o agregar nuevas funcionalidades implica un gran esfuerzo, y hasta se hace imposible. Por el contrario, este trabajo se basa en el desarrollo de un Framework, el cual permite distintas entradas, así como múltiples salidas con un bajo acoplamiento entre sus partes, Es unimodal y de código abierto.

3. Solución desarrollada

Se desarrolló un framework que permite al usuario la configuración de estados emocionales desacoplados del tipo de robot (físico o virtual) que los representa. Dicho framework permite la interacción e integración entre distintos sistemas, sin romper su premisa de bajo acoplamiento. Para este proyecto, se trabajó en la integración de distintos sistemas: software Emotion Detection Asset, que se encargará de reconocer emociones a través de expresiones faciales, capturadas por medio de una webcam o de una imagen importada desde un archivo; UI por la cual se puede realizar diferentes configuraciones; robots físicos (Roboreptile) y/o virtuales, para la representación de las emociones capturadas o ingresadas manualmente; interfaz de transmisión (USB UIRT o pantalla); API, que abre el framework para ser consumido a través de la red. La figura N° 2 representa el modelo conceptual del sistema propuesto

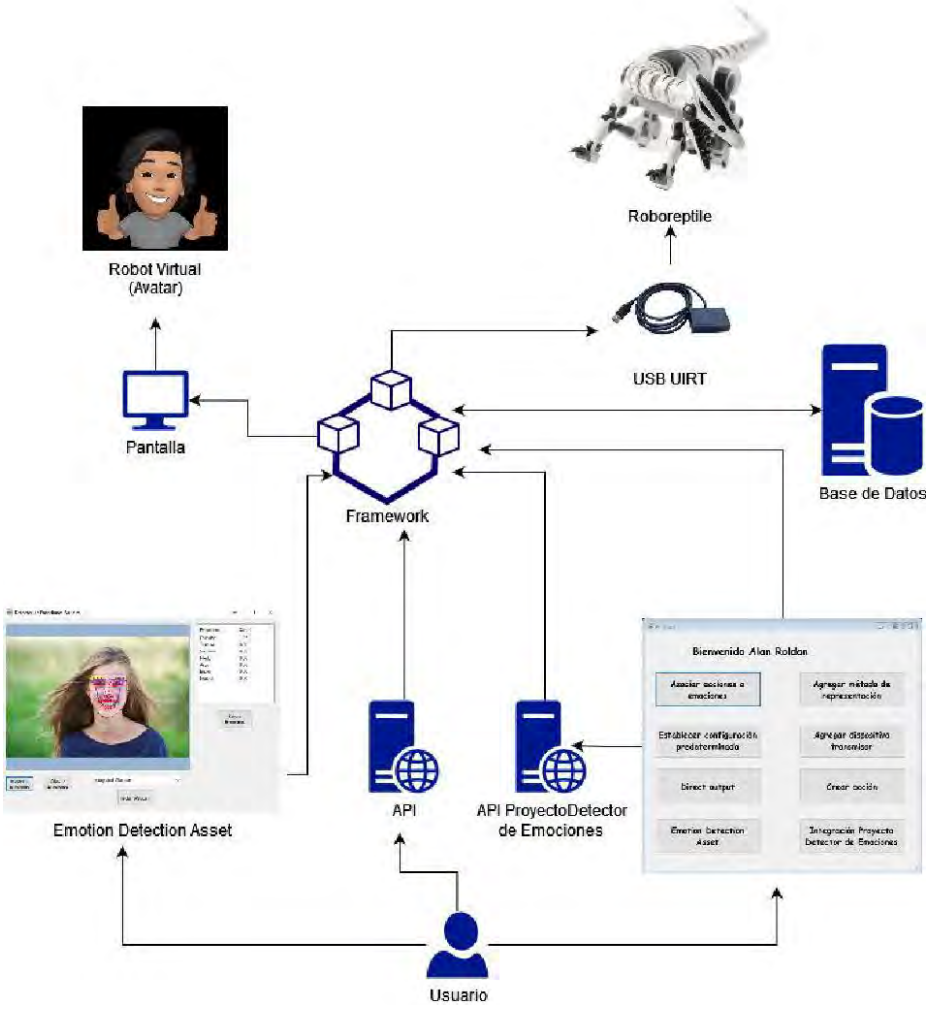


Figura 2 modelo conceptual

El framework propuesto (Figura 3) presenta en particular la asociación de emociones a acciones del robot en respuesta a las emociones detectadas del humano, la posibilidad de seleccionar el método de representación, física (robot) o virtual (avatar), el dispositivo de transmisión al robot: monitor en caso de avatar, en robots físicos (wi fi, USB UIR (IR), Emociones habilitadas, y acciones asociadas a partir de movimientos configurados previamente desde la función de carga de configuración predeterminada.

Figura 3 Framework Asociar acciones a emociones

4. Pruebas

Las pruebas tienen como objetivo verificar y validar el correcto funcionamiento de las distintas partes que componen el sistema, para demostrar que el framework pueda funcionar de manera independiente, así como también de manera conjunta con los otros componentes. En total se realizaron tres pruebas, las cuales se abordarán a continuación. Para el desarrollo de estas se utilizó un sujeto de prueba masculino, otro femenino y se verificó el funcionamiento de la API. Esta prueba fue realizada por miembros del equipo, con lo cual se conocía cómo configurar el sistema. Se esperaba que la API responda efectivamente o con un mensaje de error al request realizado a través de Postman, y así también transmitir al dispositivo de representación la emoción especificada.

4.1 Prueba general

Las pruebas con avatar pueden ser visualizadas [22] y con robot físico pueden ser visualizadas en [23], los resultados obtenidos fueron alentadores, el sistema es lo suficientemente intuitivo para un usuario promedio. También se observó la correcta interacción entre la interfaz y el framework, ya que al momento de transmitir una emoción al dispositivo de representación todo funcionó según lo esperado.

4.2 Prueba de API

Al utilizar Postman[24] para realizar esta prueba, se pudo observar fácil y rápidamente el comportamiento de la API, ya que esta responde con la representación de la emoción a través del dispositivo configurado, y además envía un mensaje a quien genere el post request.

4.3 Pruebas de Integración con Proyecto "Reconocimiento de Emociones Mediante Expresiones Faciales a Través de Regresión Logística"

Mediante esta integración se busca demostrar que es posible y relativamente sencillo agregar distintos softwares de reconocimiento facial a nuestro framework. En este caso, se integrará con el proyecto desarrollado por Carlos Barrionuevo. [25] La integración consiste en contactar la API de reconocimiento facial, enviándole una imagen para que esta sea analizada y luego devuelva la emoción predominante en la misma. Una vez instaladas las librerías necesarias para que el sistema de reconocimiento funcione, se procedió a desarrollar una nueva interfaz para poder demostrar de manera sencilla la conexión entre ambos. Para el usuario que analiza la imagen, es transparente, por lo cual hay que visualizar el código para observar las llamadas a la API. (Figura N° 4). En la línea 50 "IRestResponse response = client.Execute(request);" está realizando una llamada a la API (definido en la línea 46), a través del método POST (definido en la línea 48). A continuación, guardamos la respuesta de la API en la variable OUTPUT, la cual contiene un string, sin formato, con el contenido de la emoción predominante. Ya parseado el resultado, se genera una instancia de la clase Salida, la cual contiene las distintas funciones que transmiten la emoción de acuerdo con la configuración almacenada en la base de datos, para poder representarla, ya sea a través del robot o del avatar.

```
44      private void button2_Click(object sender, EventArgs e)
45      {
46          var client = new RestClient("http://127.0.0.1:8000/upload");
47          client.Timeout = -1;
48          var request = new RestRequest(Method.POST);
49          request.AddFile("archivo", path);
50          IRestResponse response = client.Execute(request);
51          string output = response.Content;
52          var salida = output.Split(new string[] { "emocion_predicha&#34;: &#34;" }, StringSplitOptions.None)[1];
53          salida = salida.Split(new string[] { "&#34;}" }, StringSplitOptions.None)[0];
54          button2.Enabled = false;
55          Salida emocionSalida = new Salida();
56          emocionSalida.transmitirEmocion(salida.ToString());
57      }
```

Figura 4. Código conexión contra API

4.4 Discusión de los resultados

El desarrollo experimental alcanzado permite, en comparación a los trabajos indicados en el estado del arte (Tabla N° 1 Síntesis comparativa: Robots y Emociones), alcanzar un bajo acoplamiento permitiendo la integración abierta a distintos robots físico o virtuales con sus formas de comunicación, como así también en forma independiente generar los métodos de comportamiento del robot y asociarlos a una o más emociones identificadas.

6. Conclusión y futuras líneas de trabajo

El desarrollo experimental realizado permitió integrar exitosamente el registro de emociones humanas bajo el enfoque categórico con la respuesta de acciones por parte del robot físico o virtuales, facilitando la integración de distintas fuentes de registro

emocionales, como así también robots. Futuras líneas de trabajo se orientarán bajo enfoques de registro de emociones multimodales, como voz, variación de ritmo cardiaco, conductancia de piel, EEG, entre otras, todas estas fusionadas bajo un enfoque dimensional con valores de excitación, valencia.

El presente trabajo se desarrolló en el marco PICTO-UM-2019-00005 – "Influencias del estado biométrico-emocional de personas interactuando en contextos de entornos simulados, reales e interactivos con robots".

Referencias

[1] R. Picard, Affective Computing, 1995.
[2] Picard, R. (2000). Affective Computing. En T. M. Press, Affective Computing (pp. 4-8). Cambridge Massachusetts: The Mit Press.
[3] S. Baldasarri, «Computación Afectiva: tecnología y emociones para mejorar la experiencia de usuario.,» Revista Institucional de la Facultad de Informática | UNLP, 2016.
[4] P. Ekman, The repertoire of nonverbal behavior: categories, origins, usage and coding, 1969.
[5] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer y M. Tomita, «Universal and cultural differences in the judgments of facial expressions of emotion» 1987.
[6] D. T. van der Haar, «Student Emotion Recognition» de International Conference on Human-Computer interaction, 2019.
[7] R. Zatarain Cabada, M. L. Barron Estrada, G. Halor-Hernandez y C. A. Reyes-García, «Emotion Recognition in Intelligent Tutoring Systems» de Mexican International Conference on Artificial Intelligence, México, 2014.
[8] Z. Wei-Long y L. Bao-Liang, «Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks» de IEEE TRANSACTIONS ON AUTONOMOUS MENTAL DEVELOPMENT, VOL. 7, NO. 3, SEPTEMBER 2015, IEEE, 2015, pp. 162-175.
[9] R. Xu, J. Chen, J. Han, L. Tan y L. Xu, «Towards emotion-sensitive learning cognitive state analysis of big data in education deep learning-based» de Computing, Austria, Springer Viena, 2019, pp. 1-16.
[10] E. Alepis y M. Virvou, «User Modeling: An Empirical Study for Affect Perception Through Keyboard and Speech in a Bi-modal User Interface,» de International Conference on Adaptive Hypermedia and adaptive Web-Based Systems, Berlin, Heidelberg, 2006.
[11] S. Salmeron-Majadas, O. Santos y J. Boticario, «An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low-cost affective modeling in an educational context» de 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014, 2014.
[12] E. Calot, J. Ierache y W. Hasperué, «Robustness of keystroke dynamics identification algorithms against brain-wave variations associated with emotional

variations» de Intelligent Systems and applications, Londres, Springer Cham, 2019, pp. 194-211.
[13] M. Wiew y Z. Lachiri, «Emotion Classificationin Arousal Valence Model using MAHNOB-HCI Database,» international Journal of Advanced Computer Science and Applications, pp. 1-6, 2017.
[14] J. Healey, R. Picard y E. Vyzas, «Toward Machine Emotional Intelligence: Analysis of Affective Physiological State» de Pattern Analysis & Machine Intelligence, IEEE, 2001, pp. 1175-1191.
[15] J. Marín-Morales, J. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. Scilingo, M. Alcañiz y G. Valenza, «Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors» Scientifics Reports, 2018.
[16] J. Ierache, G. Pereira y J. Iribarren, «Navigation Control of a Robot from a Remote Location via the Internet Using Brain-Machine Interface» de Robot Intelligence Technology and application, Springer Cham, 2014, pp. 297-310.
[17] J. Ierache, G. Pereira, J. Iribarren y I. Sattolo, «Robot Control on the Basis of Bioelectrical Signals",» de Robot Intelligence Technology and Applications 2012, Korea, Springer, 2012, pp. 337-346.
[18] J. Ierache, F. Nervo, G. Pereira y J. Iribarren, «Estado Emocional Centrado en Estímulos, Aplicando Interfase Cerebro-Maquina,» de XX Congreso Argentino de Ciencias de la Computación (Buenos Aires, 2014), Buenos Aires, 2014.
[19] J. S. Ierache, R. Nicolosi, G. Ponce, C. Cervino y E. Eszter, «Influencias del estado biométrico emocional de personas interactuando en contextos de entornos simulados, reales e interactivos con robots,» XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste). ISBN: 978-987-3619-27-4, Páginas: 785-789
[20] Kishi. M. I. T., «Impression Survey of the Emotion Expression Humanoid Robot with Mental Model based Dynamic Emotions» 2013.
[21] Takato Horii Y. N. a. M. A. *, «Imitation of human expressions based on emotion estimation by mental simulation» 2016.
[22] Prueba Avatar : https://drive.google.com/file/d/1L2xlGDrKfcLWsXjmn5d1N1m7RO0ArzkP/view?usp=sharing
[23] Prueba Robot : https://drive.google.com/file/d/1udy2xy8XL1QqsjtrxuCPXHlKaDKVVYQq/view?usp=sharing
[24] https://www.postman.com/ vigente Julio 2022
[25] C. Barrionuevo, J. Ierache e I. Sattolo, Reconocimiento de Emociones Mediante Expresiones Faciales a Través de Regresión Logística, XXVI Congreso Argentino de Ciencias de la Computación (CACIC) (Modalidad virtual, 5 al 9 de octubre de 2020) , ISBN: 978-987-4417-90-9, Páginas: 491-5002021.

Aplicando PageRank en Registros de Actividades Criminales: una Aproximación a la Detección de Bandas Delictivas

Sebastián P. WAHLER[1,2], Martín L. LARREA[3], and Diego C. MARTÍNEZ[3]

[1] Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco. Trelew, ARGENTINA.
http://www.ing.unp.edu.ar/dpto-informatica.html
[2] Departamento de Informática, Procuración General, Ministerio Público Fiscal, Poder Judicial de la Provincia del Chubut. Rawson, ARGENTINA.
https://www.mpfchubut.gov.ar
[3] Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur. Bahía Blanca, ARGENTINA.
https://cs.uns.edu.ar/
(e-mail: spwahler@ing.unp.edu.ar, dcm@cs.uns.edu.ar, mll@cs.uns.edu.ar)

Resumen Se presenta el resultado del estudio de las técnicas y metodologías actuales de análisis inteligente de datos y visualización para la asistencia en la investigación criminal, a partir de los registros de actividades delictivas, sus autores y las relaciones de datos que puedan derivarse a partir de ellas. Es de especial interés la identificación de bandas delictivas o criminales para propender a una persecución penal inteligente. Se discute el desarrollo de un componente de software para la visualización, incorporando la utilización del algoritmo de PageRank y detección de comunidades.

Keywords: Investigación Criminal & Análisis Inteligente de Datos & Redes Sociales & Visualización & PageRank

1. Introducción

En la actualidad las actividades criminales habituales en una ciudad o región van desde hurtos y robos de poca importancia, hasta otros de mayor gravedad como abusos sexuales y homicidios. Todos ellos son registrados de diferentes formas por las fuerzas de la ley, con datos de variada precisión que incluyen usualmente la tipificación del delito, los datos en tiempo y espacio, y en muchas ocasiones los autores correspondientes. Toda esta información respalda los procesos de investigación judicial de cada caso, pero con el transcurso del tiempo constituyen una extensa base de conocimiento sobre la cual es posible extraer valiosa información para la prevención del delito y la búsqueda de la justicia. Por ejemplo, es posible inferir relaciones de amistad o conveniencia entre diversos autores de actividades criminales a partir de los registros delictivos y es de extrema relevancia para la prevención del delito y la resolución de casos inconclusos.

Las organizaciones criminales son grupos que operan fuera de la ley, realizando actividades ilegales en beneficio propio y en detrimento de otros individuos o grupos sociales [5]. Pueden ser de diverso tamaño y cubrir áreas geográficas variadas, en muchos casos en conflicto con otras organizaciones similares. Una de las características particulares de este tipo de organizaciones es que, al estar enfocadas en actividades ilegales perseguidas por los organismos de seguridad pública, el anonimato y/o la discreción de sus miembros es de vital importancia. Esto requiere estudios de la información existente con el fin de identificar los criminales y realizar acciones apropiadas para la prevención del delito. Los miembros de las organizaciones criminales tienen a su vez diversos grados de compromiso con cada una de ellas. En muchos casos los hechos son cometidos por individuos de baja jerarquía y responsabilidad en el grupo. Por otro lado, existen otros individuos de mayor jerarquía y responsabilidad en la organización criminal, que ostentan cualidades de liderazgo. En tal sentido, con el objetivo de ayudar en la identificación de las bandas delictivas, sus integrantes y el grado de importancia de cada uno dentro de ellas, son de interés dos áreas de las Ciencias de la Computación: el área de Visualización de Información, en particular la Visualización de Grandes Conjuntos de Datos, que busca asistir a los usuarios en la adecuada comprensión de la información, y el Análisis de Redes Sociales, en donde se emplean técnicas y formalismos para la comprensión de las estructuras de las redes y sus nodos. La aplicación de técnicas visuales para la representación de este tipo de información es importante [15] [4] [12], así como el estudio de las tareas e interacciones que la visualización debe soportar [2], ya que son estas interacciones las que facilitan la exploración de la visualización de información.

En esta línea de investigación estudiamos la aplicación de estas técnicas y tecnologías, contribuyendo además al desarrollo de componentes de software para la visualización y análisis inteligente de los datos, incorporando nociones de analítica de grafos. En particular en este trabajo nos interesa la consideración del algoritmo de *PageRank*, aportando a la detección de delincuentes de relevancia entre las *comunidades de individuos*, de la misma forma que se puede discriminar la importancia de un conjunto de páginas web a través de sus vínculos. Para esto se cuenta con los registros de actividades criminales a través de la colaboración del Ministerio Público Fiscal de la provincia del Chubut, que provee la base de conocimientos para inducir el grafo de relaciones como se detalla en la siguiente sección.

2. Marco de trabajo - MPF Chubut

Desde hace pocos años, el Análisis de Redes Sociales (o SNA por sus siglas en inglés de Social Network Analysis) ha contribuído a las investigaciones criminales y a las actividades de inteligencia relacionadas. Actualmente los organismos estatales encargados de la Justicia y la prevención del delito cuentan con registros informatizados de las actividades criminales detectadas, así como de las etapas y eventos del subsecuente proceso penal. Esta información constituye en esencia, una forma de *red social*. Para este trabajo es de especial interés la información producida por las fuerzas policiales de la Provincia del Chubut y su

Poder Judicial de la mano del Ministerio Público Fiscal (MPF [6]), registradas en el sistema `Coirón`. Ésta es una herramienta que permite registrar, comunicar y gestionar las actividades, trámites y actuaciones que se realizan para un caso penal, desde la denuncia hasta su finalización. También es una herramienta de administración de información, flujo de casos, planificación, organización, coordinación y control. Su progreso, mantenimiento y mejora continua está a cargo del Equipo de Desarrollo del Departamento de Informática del Área de Planificación y Control de Gestión de la Procuración General. Entre otras funcionalidades, es de interés la incorporación de herramientas de visualización de información, potenciando el análisis que realizarán luego los especialistas de análisis criminal. En este trabajo nos enfocamos en las bandas delictivas y la *importancia relativa* de sus integrantes. Para ello es central el concepto de "*Grupo de Pertenencia*", como se denomina en el Sistema `Coirón` a la relación directa que existe entre un individuo dentro del universo de personas cargadas como actores de delitos (roles: denunciado, sospechoso o imputado) y otros individuos del mismo universo, con los cuales existan uno o más casos penales en común. El objetivo es contar con componentes de software que muestren gráficamente las relaciones entre las personas involucradas en los casos penales, enriquecida con información proveniente de la analítica de grafos.

En estos grafos un nodo es una persona (con los roles ya mencionados) si está involucrada en dos o más casos penales [4]. El tamaño del nodo posee una relación directa con la cantidad de casos penales en los que se encuentre involucrada la persona. Cuanto mayor sea el tamaño del nodo, mayor es la cantidad de casos penales en las que está involucrado. Los arcos entre pares de nodos vinculan a las personas entre sí y representan el o los casos que tienen en común. El grosor de la vinculación será directamente proporcional a la cantidad de casos en común entre un par de personas. Hay nodos que se encontrarán aislados en el grafo, y esto no significa que no estén efectivamente involucrados en casos, sino que quizás no existan relaciones para el filtro de búsqueda que se utilice en esa vista en particular.

Supongamos que una persona A se encuentra asociada a 8 casos penales, una persona B a 4 y una persona C a 2 casos. Las personas A y B se encuentran relacionadas entre sí, por estar en 3 casos en común (casos 1, 2 y 3). Por otro lado las personas A y C también se encuentran relacionadas, por tener un caso en común (caso 4). Una representación gráfica de dicha situación se muestra en la Figura 1, y puede observarse el doble de tamaño entre el nodo A y el nodo B, representando justamente la diferencia de casos entre ambos nodos (8 y 4 casos). También se ve a simple vista el grosor del enlace entre A y B tres veces más grande que el enlace entre A y C (3 casos en común entre el primer par de nodos, y sólo un caso para el último par de nodos mencionado).

Esta es, en primera instancia, una caracterización de la importancia de los individuos en la red. Sin embargo, analíticas mas complejas podrían aplicarse.

[4] Existe un gran cúmulo de personas en el sistema con sólo un caso con rol de *denunciado*, por esa razón se los excluye del universo a analizar. Serían parte del dataset a visualizar si se encuentran relacionados con otros nodos del primer grupo.

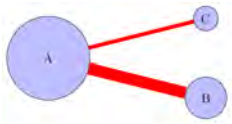


Figura 1. Ejemplo de relación entre tres personas.

3. Descripción general de los datos

Nuestro conjunto de datos consta de casos penales, actuaciones (bitácora de eventos del proceso penal), delitos, personas, elementos (denunciados y secuestrados), todos ellos relacionados, registrados entre octubre de 2006 y mayo de 2022 en la Circunscripción Judicial de Trelew - Chubut. Este conjunto de datos incluye lugares relativos a personas y a hechos delictivos, fechas, estados procesales de los casos y las personas, como así también los vínculos entre todos los conjuntos mencionados. Son 105586 casos penales, 183348 personas involucradas en casos, un universo de 132950 personas en total y 113010 delitos cargados. En relación al conjunto de datos para la visualización, constituyen: 33178 nodos, 16964 enlaces y 60513 relaciones Nodos/Enlaces.

A partir de los datos de los casos penales, pudimos construir la red de *Grupos de Pertenencia*. En esta red se eliminan los nodos de aquellas personas cuyos roles no sean referidos a actores delictivos, como ser denunciantes, víctimas o damnificados. Una visualización de los datos puede verse en la Figura 2. Al analizar la composición de la red obtenida podemos observar las relaciones que existen entre los nodos y como se "equilibra" el grafo, haciendo que aquellos nodos con pocas o nulas relaciones queden en la periferia de la gráfica. También es apreciable cierta medida de *centralidad* de aquellos nodos que son rodeados por sus relacionados, denotando cierta importancia. En la Figura 2 se visualizan sólo las 10 personas con más Casos y sus grupos de pertenencia. Claramente esos 10 nodos principales quedan rodeados de sus grupos de pertenencia y se pueden observar transitividades entre ellos a través de nodos que conforman parte del grupo de pertenencia de más de un nodo principal.

Para llevar a cabo la visualización del conjunto de datos obtenidos del análisis anteriormente descripto, se utilizó `Vis.js` [5], una biblioteca o librería de visualización dinámica basada en lenguaje Javascript. Esta librería está diseñada para manejar grandes cantidades de datos dinámicos y permitir la manipulación y la interacción con los datos. Aplicamos además algoritmos de diseño forzados "*force-directed graph drawing*", que intentan posicionar los nodos considerando las fuerzas entre dos nodos (atractivos si están conectados, repulsivos de lo contrario). Generalmente son iterativos y mueven los nodos uno por uno hasta que ya no es posible mejorar o se alcanza el número máximo de iteraciones. Los enlaces tienen más o menos la misma longitud y el menor número posible de enlaces cruzados. Los nodos conectados se juntan más mientras que los nodos aislados se alejan hacia los lados.

[5] https://visjs.org/

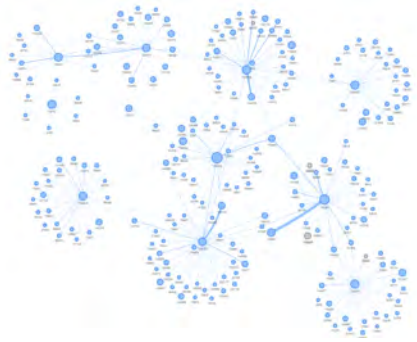


Figura 2. 10 personas con más casos en Coirón, con sus relaciones

4. Identificación de posibles bandas delictivas

Recordemos que en este trabajo nuestro principal interés es la identificación asistida de bandas delictivas y sus cualidades. Las personas nos movemos habitualmente entre lugares conocidos (hogar, trabajo, supermercado, restaurante) y con frecuencia por las mismas calles o rutas. La teoría sugiere que muchos delitos ocurren cuando se cruzan delincuentes y víctimas dentro de algunas de estas zonas de actividad. Un delincuente tenderá a cometer un delito en algún lugar que se encuentre dentro o cerca del recorrido que realiza diariamente para trasladarse o su zona de movimiento habitual. La naturaleza de los vínculos de los integrantes de una banda delictiva es una variable que aporta información sobre las características y similitudes de los miembros del grupo, atendiendo a criterios concretos: vínculo familiar, cultural, de proximidad (provienen del mismo barrio), coincidencia en prisión, especialización (habilidades delictivas), la experiencia y otras capacidades y tipos de vínculo.

Existen antecedentes reales que justifican la importancia de esta línea de trabajo. En el año 2019 fue necesaria una investigación criminal sobre reiterados robos de televisores LCD en domicilios [9], como así también una serie de hechos consecutivos vinculados al robo de cajas fuertes en empresas del parque industrial de la ciudad de Trelew. La UAC (Unidad de Análisis Criminal), organismo auxiliar perteneciente al MPF, sirvió como equipo de apoyo en la investigación de ambos modus operandi, haciendo uso de toda la información de los legajos fiscales, consultas generales y específicas contenidas en el Sistema Coirón. Fue de vital uso la información referida a los *grupos de pertenencia* de cada persona, pero devino en un arduo trabajo entrecruzando información de personas, para dar con las supuestas bandas delictivas detrás de estos hechos. Esta necesidad ha activado la línea de trabajo actual, utilizando la información ya contenida en el sistema de gestión penal, buscando proveer de una forma más directa y visual la base para el apoyo a la toma de decisiones en las investigaciones de bandas delictivas. Esto ayuda a los especialistas a detectar triangulaciones, transitividades y por supuesto *centralidades* e importancias internas en la Red. Para esto

es necesaria la consideración de técnicas que permitan distinguir la *importancia* de un nodo en un grafo particular, como explicamos en la siguiente sección.

5. PageRank y detecciones comunitarias

Como parte simplificada de la estructura de una comunidad de nodos en una red social, cada uno representa a un individuo y la red tiene una segmentación multitudinaria [11]. Algunas personas son centrales en la comunidad, algunas están al margen, establecen menos relaciones con otros y, por lo tanto, tienen una influencia menor. En esta sección, presentamos un nuevo enfoque de descubrimiento comunitario basado en el algoritmo PageRank para encontrar a estos delincuentes "importantes" ó con "mayor influencia" en nuestro grafo, con el fin de analizar supuestas bandas delictivas. Recordemos que un grafo es un par $G = (N, A, g)$ donde N es un conjunto finito no vacío de elementos denominados *nodos* (vértices), A es un conjunto de arcos y g es una función que asocia a cada arco a perteneciente a A con un par no ordenado (x, y), siendo x e y nodos pertenecientes a N. Se dice que a es un arco con vértices extremos x e y [3].

PageRank (PR) es un método que fue implementado a través de un algoritmo originalmente utilizado por Google que asigna a cada página web de un conjunto dado, un puntaje que refleja su importancia dentro del conjunto. A este puntaje se lo denomina *valor de PageRank*. Ante una consulta, el buscador utiliza estos puntajes para determinar el nivel de relevancia de las páginas, y retorna en primer lugar aquellas con un puntaje más alto. Para calcular los puntajes, PageRank utiliza la estructura de enlaces de la web [1]. Una página web tiene un valor de PageRank alto si es apuntada por muchas otras páginas, o bien si es apuntada por páginas con puntajes altos [14]. PageRank tiene una base intuitiva en el concepto de *random walks* sobre grafos [8]: supongamos que un navegante aleatorio empieza a navegar la web desde una página cualquiera. El navegante puede hacer clic en forma aleatoria sobre alguno de los enlaces presentes en la página en la que se encuentra actualmente con una probabilidad d a la que se denomina *damping factor*, o bien con probabilidad $1 - d$ accede aleatoriamente a cualquier otra página web. Este proceso se repite indefinidamente. Luego, el valor de PageRank de una página P puede ser interpretado como la probabilidad de que el navegante aleatorio se encuentre en P al finalizar el proceso. PageRank es definido formalmente de la siguiente manera [7]. Sean q_i el número de enlaces salientes que posee la página i , n el número total de páginas web, d el *damping factor* que por lo general adquiere el valor 0.85, π un vector columna denominado *vector PageRank*, y $H = (h_{ij})$ una matriz cuadrada de tamaño n tal que $h_{ij} = 1/q_i$ si existe un enlace desde la página i a la página j , y $h_{ij} = 0$ en caso contrario. El valor h_{ij} corresponde a la probabilidad de acceder a la página j desde la página i en un paso, a partir de hacer clic en alguno de los enlaces que aparecen en esta última. El valor de PageRank correspondiente a la página j es π_j, y se define recursivamente como se muestra en la ecuación 1 [10].

$$\pi_j = \frac{1-d}{n} + d\sum_{i=1}^{n} \pi_i h_{ij} \tag{1}$$

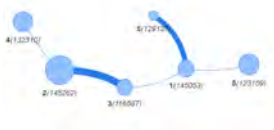
Aplicación de PageRank para bandas delictivas Nuestro dataset descrito anteriormente se obtiene a partir de consultas SQL a la Base de Datos de `Coirón`. Para hacer uso del algoritmo de PageRank se decidió incorporarlo dentro de esas consultas SQL de modo de obtener un resultado que pueda ser utilizado para la visualización. Dentro de la consulta original se genera una tabla para los nodos y otra tabla para las relaciones, de esta manera el software realizado para la visualización obtiene dichos datasets y renderiza el grafo. Para incorporar el cálculo de PageRank, inicialmente se adecuaron ambas tablas para la utilización de la fórmula, y se necesitó de ciertas tablas temporales para el cálculo. Por un lado se computó el grado de salida de cada nodo *(Out Degree)*, es decir el número de enlaces que lo conectan con otros nodos. Luego se declara el *damping factor*, en nuestro caso 0.85, luego el conteo total de nodos, y se calcula el *PageRank inicial* de cada nodo, para después comenzar la iteración buscando cumplir con la sumatoria de la fórmula. El *damping factor* corresponde a un valor probabilístico que, aplicado al escenario de paginas web, pretende capturar la posibilidad de que un usuario continúe haciendo click en los links de una página en una sesión de navegación continua. Aquí este factor tienen un significado diferente, reintrepretado como el factor en el que se diluye la importancia de un individuo entre sus pares a través de una cadena de arcos. Actualmente estamos estudiando un valor apropiado en función de los datos existentes, puesto que el damping factor es esencialmente un valor empírico. En este trabajo optamos por usar el valor propio de la propuesta original del PageRank.

```
INSERT INTO #OutDegree
SELECT #Node.id, COUNT(#Edge.src)
FROM #Node
LEFT OUTER JOIN #Edge ON #Node.id = #Edge.src
GROUP BY #Node.id
DECLARE @dampingFactor float = 0.85
DECLARE @Node_Num int
SELECT @Node_Num = COUNT(*) FROM #Node
INSERT INTO #PageRank
SELECT #Node.id, rank = ((1 - @dampingFactor) / @Node_Num)
FROM #Node
INNER JOIN #OutDegree ON #Node.id = #OutDegree.id
DECLARE @Iteration int = 0
```

```
WHILE @Iteration < 50
BEGIN
--Iteration Style
SET @Iteration = @Iteration + 1
INSERT INTO #TmpRank
SELECT #Edge.dst, rank = ((1 - @dampingFactor) / @Node_Num)
+ (@dampingFactor * SUM(#PageRank.rank / #OutDegree.degree))
FROM #PageRank
INNER JOIN #Edge ON #PageRank.id = #Edge.src
INNER JOIN #OutDegree ON #PageRank.id = #OutDegree.id
GROUP BY #Edge.dst
END
```

Una vez finalizado el desarrollo de la fórmula, se procedió a realizar pruebas que corroboren el buen funcionamiento del código. Se realizaron ejemplos para pocos nodos con pocas relaciones, de manera tal que sea sencilla la verificación. Se muestra a continuación la Figura 3 que refleja la visualización de la ejecución de PageRank para 6 nodos. En cada nodo se muestra su posición de PageRank, y entre paréntesis el identificador de cada nodo. Como se puede observar el nodo central por el PageRank calculado es el referido al ID *145053*, cuyas relaciones con 3 nodos pesan sobre las relaciones que poseen el resto de los nodos visualizados en el grafo. Es interesante ver que éste individuo no es el que posee necesariamente la mayor cantidad de casos penales, pero es el más importante entre sus pares *de su propia red social* de contactos relacionados.

Se ejecutó el algoritmo también para casos de estudio real resueltos en el MPF, donde las bandas y sus líderes han sido identificados, y de esta manera validar la relevancia de la implementación. Como ejemplo a continuación se



IdNodo	PageRank
145053	0,286845042094944
145262	0,199512347112651
116587	0,1910604814502
132310	0,109790233522352
123109	0,106270247927848
129137	0,106270247927848

Figura 3. Resultado de ejecución de PageRank para 6 nodos.

muestra la Figura 4, sobre el caso real de robo de LCDs mencionado en el capítulo anterior. Pudo validarse que todos los integrantes de la banda se encuentran en el centro del subgrafo, altamente relacionados con los nodos de colores, que reflejan a los de más valor de PageRank.

Figura 4. Resultado de ejecución de PageRank para el caso real de robo de LCDs.

Detección de comunidades Una comunidad puede ser definida como un conjunto de nodos que están más densamente conectados entre ellos que con el resto de la red. La importancia de este planteamiento radica en que se espera que los nodos que están contenidos dentro de una misma comunidad compartan atributos, características comunes o relaciones funcionales [11]. En este trabajo se aprovecha el algoritmo SQL descrito con anterioridad en el cual se dividen nodos y relaciones para ser luego visualizadas por la herramienta `Vis.JS`. En este sentido, nuestro enfoque se basa en la búsqueda de posibles personas que participen en bandas delictivas. Del origen de datos surge que para cada nodo pueden conocerse todas sus relaciones, de modo que podemos asignar a cada uno de esos nodos referentes un identificador de grupo ó cluster. Es posible entonces verificar para cada par de nodos, si pertenecen a un grupo en particular (uno de ellos será "referente" y podremos identificarlo), y de esta manera asignar a cada relación también un grupo determinado. Con ambas tablas (nodos y relaciones) actualizadas, es posible desde la herramienta de visualización, asignar colores a cada cluster, y así generar un grafo aún más práctico a la vista.

Si bien para la visualización se utiliza lenguaje JavaScript de la mano de la librería anteriormente mencionada `Vis.JS`, y los dataset se obtienen desde la Base de Datos del Sistema `Coirón` a través de consultas SQL, el software de estudio que toma los datos del dataset y los procesa para luego llamar a la librería de visualización, se encuentra desarrollado en lenguaje C Sharp de .Net

Framework. A continuación se exhibe en la Figura 5 la misma visualización que se ha presentado con anterioridad en la Figura 2, pero ahora con la detección de grupos por color. Se puede observar que cada grupo o cluster de nodos comparte el mismo color para los enlaces internos.

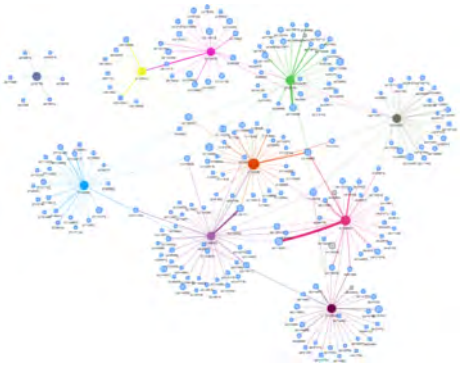


Figura 5. 10 personas con más casos en Coirón, con sus relaciones. Se agrega detección de comunidades por color.

6. Conclusiones y trabajos futuros

Se ha presentado una propuesta de estudio de las técnicas y metodologías actuales de análisis inteligente de datos y visualización para la asistencia en la investigación criminal. Todo ello a partir de los registros de actividades delictivas, sus autores y las relaciones de datos que pueden derivarse a partir de ellas. Fue de especial interés la identificación de redes ilegales, tales como bandas delictivas o criminales para propender a una persecución penal inteligente.

Del estudio propuesto, como se explicó en los capítulos anteriores, se desarrolló un módulo de software como herramienta gráfica para visualizar la red de grupos de pertenencia de los actores delictuales, incorporando un algoritmo de PageRank para reflejar aquellas personas importantes y Detección de comunidades asignandole colores a cada grupo/cluster de nodos. El desarrollo de software se implementó en el mismo Ministerio Público Fiscal, del cual se tomaron los datos para generar los datasets de pruebas. De esta manera se ha logrado no sólo estudiar las técnicas y metodologías expuestas, sino también alcanzar la puesta en producción y uso de la herramienta, por los propios actores de la investigación.

Las primeras impresiones de aquellos especialistas de investigaciones penales han sido muy satisfactorias y permiten evaluar a este trabajo como el inicio de futuros desarrollos visuales para el apoyo a la toma de decisiones en la investigación penal. Actualmente se continúa trabajando en el desarrollo de la aplicación de visualización. Se pretende modificar la fórmula original de PageRank, enriqueciéndola con información adicional según los registros judiciales. Por ejemplo, la posibilidad de darle mayor ranking inicial a aquellos nodos que tengan un peso mayor que otros según los registros. También sería interesante hacer lo mismo con el peso que poseen los enlaces entre nodos, ya que no es lo

mismo una relación de 2 casos penales en común entre dos personas, que una de 18 casos en común. De esta manera lograríamos darle más ranking también a aquellos nodos que estén relacionados con otros, en mayor cantidad de casos penales. Esto es sin duda relevante para la investigación criminal basada en antecedentes penales. También se buscará profundizar sobre diversos algoritmos de centralidad. Existen algunas nociones que son de relevancia para la identificación de la importancia de una persona en la red inducida por las causas penales. Por ejemplo, *betweenness centrality*, que modela la medida en que un nodo en particular se encuentra entre otros nodos en una red, o *closeness centrality*, que es la inversa de la suma de los caminos más cortos (geodésicas) que conectan un nodo particular con todos los demás nodos de una red [13]. De manera similar, *eigenvector centrality*, es otra forma de asignar la centralidad a un actor de la red basado en la idea de que si un nodo tiene muchos vecinos centrales, también debería ser central.

Referencias

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30(1-7), 107–117 (1998)
2. Chen, H., Atabakhsh, H., Tseng, C., Marshall, B., Kaza, S., Eggers, S., Gowda, H., Shah, A., Petersen, T., Violette, C.: Visualization in law enforcement. In: CHI'05 extended abstracts on Human factors in computing systems. pp. 1268–1271 (2005)
3. Dubinsky, E.: Mathematical structures for computer science. by judith l. gersting. The American Mathematical Monthly 91(6), 379–381 (1984)
4. Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., Liu, Q.: Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access 7, 106111–106123 (2019)
5. Finckenauer, J.O.: Problems of definition: what is organized crime? Trends in organized crime 8(3), 63–83 (2005), https://doi.org/10.1007/s12117-005-1038-4
6. Fiscal, M.P.: Página web. https://www.mpfchubut.gov.ar/
7. Franceschet, M.: Pagerank: Standing on the shoulders of giants. Communications of the ACM 54(6), 92–101 (2011)
8. Göbel, F., Jagers, A.: Random walks on graphs. Stochastic processes and their applications 2(4), 311–336 (1974)
9. Jornada, D.: Caso de estudio real. https://www.diariojornada.com.ar/57375/policiales/Como_era_el_trabajo_de_la_banda_de_los_LCD_que_fue_desbaratada_esta_semana_en_Trelew
10. Lin, J., Dyer, C.: Data-intensive text processing with mapreduce. Synthesis Lectures on Human Language Technologies 3(1), 1–177 (2010)
11. Ma, X., et al.: Exploring sharing patterns for video recommendation on youtube-like social media. Multimedia Systems 20(6), 675–691 (2014)
12. Mathew, A., Mary Jose, A., Sabu, C., Raj, A., et al.: Criminal networks mining and visualization for crime investigation. ICICNIS (2021)
13. Newman, M.E.: A measure of betweenness centrality based on random walks. Social networks 27(1), 39–54 (2005)
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
15. Xu, J., Chen, H.: Criminal network analysis and visualization. Communications of the ACM 48(6), 100–107 (2005)

Extractor de noticias para el análisis integral del impacto de la pandemia en la provincia del Chubut

Emanuel Balcazar and Leo Ordinez

Laboratorio de Investigación en Informática (LINVI), FI - UNPSJB
Bvd. Brown 3051, Puerto Madryn, Argentina
{emanuelbalcazar13,leo.ordinez}@gmail.com

Resumen En el marco de la pandemia por COVID-19, con el objetivo de analizar la situación social que se estaba atravesando, se propone la extracción y análisis de información periodística, de forma automática, para su tratamiento y explotación. Para ello, se desarrolló una herramienta que recupera notas periodísticas de los principales medios de la Provincia del Chubut, generando un dataset de alta riqueza en términos de su potencial impacto.

Keywords: extracción automática de información, medios digitales, NLP

1. Introducción

En el marco de la COVID-19, la vida de las personas y el curso de las actividades y servicios no esenciales sufrieron cambios drásticos a partir del Decreto DECNU-2020-297-APN-PTE y sus normativas anexas de nivel provincial y/o municipal en Chubut. Las medidas de Aislamiento Social, Preventivo y Obligatorio (ASPO) y Distanciamiento Social, Preventivo y Obligatorio (DISPO) generaron desencadenantes de impacto social y económico que necesitan ser identificados y monitoreados para proveer información a los formuladores de políticas públicas.

En este trabajo se propone la construcción de conocimiento a partir de diferentes estrategias y herramientas de relevamiento de información y datos, que favorezcan un análisis, procesamiento y ponderación de la situación social dada en una circunstancia sanitaria como la generada por el COVID-19. En particular se considerará la región delimitada por los límites geográficos de la provincia del Chubut, ajustando la escala territorial a nivel de ciudades, pueblos y comunas rurales. La construcción de conocimiento, se hará a partir de la extracción, procesamiento y análisis automático de información periodística publicada en la prensa provincial, la cual luego será presentada de manera acorde, a fin de poder evaluarse de manera indirecta la evolución de diferentes temáticas, que impactan en la sociedad.

Los resultados serán obtenidos mediante la aplicación de técnicas de Procesamiento de Lenguaje Natural (NLP) y extracción de datos de sitios web (web

scraping), para luego ser presentados en una aplicación web que permitirá su análisis y/o difusión, así como su explotación más profunda. En el transcurso de la pandemia se han elaborado propuestas similares, con distintos objetivos [2,3,4,5,10,12]. En todos los casos, se busca la construcción de conocimiento nuevo a partir de la minería de información.

2. Materiales

En base a experiencias anteriores de trabajos similares [1,8], muchas de las estrategias utilizadas sirvieron de insumo para este trabajo.

Como primera aproximación a la solución presentada, se realizaron una serie de experimentos utilizando el buscador de Google con la intención de obtener artículos periodísticos que, en un principio, tocaran el tópico del COVID-19 en su contenido. Seguido a lo anterior, se procedió a seleccionar cuales serian las fuentes de datos (sitios webs), que resultaban más acordes para el tipo de artículos que se esperaba. Asimismo, esta decisión se vio influenciada por la intención de solo obtener los artículos que cubran la provincia de Chubut. Luego, se hicieron una serie de extracciones de prueba analizando el contenido de los sitios web periodísticos para conocer la estructura que poseían y cómo esto afectaba a la forma de extracción del contenido. Esto terminó de definir la decisión respecto a utilizar el buscador de Google como *hub* para las búsquedas y aprovechar su parametrización.

A fin de realizar búsquedas por sitios específicos y fechas particulares, se optó por incluir en la ecuación de búsqueda una fecha en el formato utilizado por el sitio. De esta manera, cada búsqueda, dependiendo de en qué sitio se está realizando, tiene en su contenido una fecha y así Google al ver que la fecha se encuentra dentro del artículo, lo devuelva como un resultado de búsqueda. Este experimento dio buenos resultados por lo que fue la alternativa elegida.

Una vez analizada la factibilidad de utilizar el buscador de Google como *hub* para la extracción de artículos, se pasó a la versión prográmatica de dicho buscador. Esta aplicación, denominada *Google Custom Search Engine (CSE)*, ofrece una API para la realización de búsquedas, mediante ecuaciones. A la vez, posee limitaciones de acuerdo a la versión de uso gratuito y la paga, las cuales determinaron ciertas decisiones arquitectónicas.

2.1. Selección de las fuentes

En este paso, se realizó un análisis de todos los sitios web de noticias disponibles de los cuales se buscó donde extraer la información. Dicho análisis fue limitado a solo los sitios que brindaran noticias informativas de cualquier tema dentro de la provincia de Chubut dado a que esta restricción forma parte de los requerimientos y limitantes del proyecto.

Para la selección de las fuentes se tuvo en cuenta un criterio de territorialidad y uno de volumen. El primero tiene que ver con las características geográficas de la provincia del Chubut, la cual tiene una gran extensión territorial y una

alta concentración en tres zonas principalmente. La primera de ellas alrededor de las ciudades de Rada Tilly, Comodoro Rivadavia y Sarmiento, al sur de la provincia. Una segunda, al noreste, conformada por la región del Valle Inferior del Río Chubut (VIRCH) y Península Valdés, donde se destacan las ciudades de Puerto Madryn, Trelew, Rawson, Gaiman y localidades menores. Finalmente una tercera zona, ubicada en el oeste cordillerano, en la que se destacan las ciudades de Esquel y Trevelin. En cuanto a volumen, se buscó que los medios tuvieran un caudal considerable de artículos, a fin de que el análisis sea lo mas amplio y detallado posible. Los medios seleccionados fueron:

- `https://diariocronica.com.ar` (sur)
- `https://www.eldiarioweb.com` (noreste)
- `https://www.diariojornada.com.ar` (noreste)
- `https://www.elpatagonico.com` (sur)
- `https://www.elchubut.com.ar` (noreste)
- `https://radio3cadenapatagonia.com.ar` (noreste)
- `https://diariolaportada.com.ar` (oeste)
- `https://www.red43.com.ar` (oeste)

Una vez seleccionado los sitios webs, se procedió a analizar cómo se estructuraban las noticias, puntualmente el análisis se realizó investigando el HTML resultante con el fin de identificar por cada sitio el formato utilizado y las etiquetas disponibles para poder extraer la información requerida, determinando los selectores CSS involucrados.

3. Métodos

Como se mencionó antes, para el diseño e implementación de la solución, se tuvieron en cuenta experiencias previas de proyectos similares. En términos metodológicos, el presente se basó en conceptos de enfoques mayores [11,9,6,7], pero adaptados a la escala de este trabajo y, sobre todo, a su urgencia. Las herramientas utilizadas para este desarrollo fueron: *NodeJS 14.16.0*, *NPM 6.14.4*, *Python 3.9.4*, *AdonisJS 4.0.13*, *VueJS 4.3.1*, *RabbitMQ 3.8.2*, control de versiones: *Github*, editor de código: *Visual Studio Code*, *PostgresQL 12.4*, *DBeaver 6.2.0*, *Docker 20.10.5* y sistemas operativos: *Ubuntu desktop 20.4* 64-bits y *Windows 11*

3.1. Arquitectura

La arquitectura se diseñó basada en microservicios con el fin de poder separar en módulos más pequeños cada parte del procesamiento. En la Figura 1 se muestra el diseño general de la misma.

El componente **cliente** es la parte visual del sistema. Se encarga de mostrar al usuario una interfaz web con el fin de brindarle la información recopilada de manera sencilla mediante gráficos y diagramas. Además permite el acceso a los artículos extraídos y normalizados.

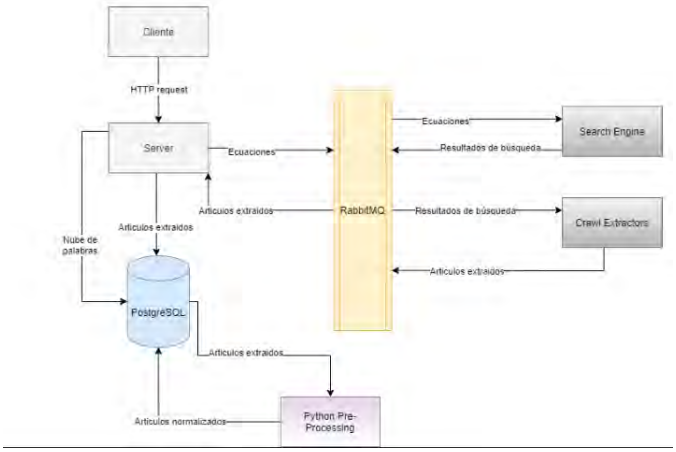


Figura 1: Arquitectura de la aplicación basada en microservicios

El componente **servidor** es el intermediario entre el componente cliente, la base de datos y la cola de mensajes. Se encarga de recibir las peticiones del cliente y obtener los datos solicitados desde la base de datos. También se conecta a la cola de mensajería para recibir los artículos que fueron extraídos y así persistirlos en la base de datos. Asimismo, el servidor tiene dos *planificadores* funcionando en simultáneo: uno se encarga de ejecutar las extracciones cada cierto período configurable; y el otro administra la construcción de nubes de palabras, a partir de los artículos procesados. El primer planificador se incorporó a la arquitectura por las restricciones de pago y correspondiente uso del servicio Google CSE. Se optó por utilizar una **base de datos relacional** para la persistencia de los datos requeridos.

La **cola de mensajería** es el componente central en el sistema, debido a que gracias a ella se comunican todos los demás componentes. La decisión de haber utilizado una cola de mensajería se justifica en la posibilidad de implementar un esquema de productor-consumidor. Esto permite, además la paralelización de los procesamientos, dado a que a través de la cola de mensajería transitan los artículos, logs del sistema, ecuaciones de búsqueda, instrucciones a otros componentes, entre otros.

El componente **motor de búsqueda** se encarga de recibir a través de la cola de mensajería una ecuación de búsqueda para luego hacer la llamada a la API haciendo uso de Google CSE. Los resultados de búsquedas se vuelven a colocar en la cola de mensajería para ser tomados por los extractores.

El componente **extractores** recibe a través de la cola de mensajería los resultados de las búsquedas y los procesa extrayendo de cada resultado el artículo en formato HTML. Luego se le aplican los selectores correspondientes al sitio y por último se envía a la cola de mensajería el articulo extraído, con datos adicionales como la ecuación utilizada, selectores, links originales y demás *metadatos*.

Por último, el componente desarrollado en **Python** se encarga de pre-procesar los artículos para permitir el posterior armado de las nubes de palabras. En este componente se normalizan los artículos extraídos, aplicando una serie de funciones de *stemming*, eliminación de caracteres inválidos, eliminación de preposiciones y demás, con el fin de obtener una versión del artículo que resume en términos neutrales el contenido del mismo. De esta manera, se puede utilizar para el armado de las nubes de palabras ya que sintetizan en conceptos y frecuencias las temáticas tratadas.

Para realizar las búsquedas, se optó por crear un componente independiente que obtenga las ecuaciones de búsqueda entrantes, ejecute la búsqueda y devuelva los resultados de la misma.El componente se divide en una serie de sub-módulos que se encargan de realizar el flujo de datos desde que llega una ecuación de búsqueda hasta que se devuelven los resultados. Para esto, se cuenta con *workers* que reciben una ecuación cada uno, donde cada worker recibirá la ecuación asociada al tópico al cual está suscrito.

Al iniciar el componente, se levantan tantos workers como sitios web configurados haya, cada worker se encargará de recibir de RabbitMQ las ecuaciones de búsqueda específicos del sitio que le corresponda, cada worker trabaja con un sitio web en particular, esto se logra utilizando el ruteo disponible de RabbitMQ el cual permite asignar un consumidor a un tópico en particular (en este caso, el tópico es el sitio web).

Una vez finalizado el proceso se obtiene un texto reducido y normalizado, que se utiliza como base para poder crear las nubes de palabras.

4. Resultados

La aplicación desarrollada dispone de un tablero de control general, que muestra estadísticas del sistema. En la Figura 2 se muestran, al mes de julio del 2022, los datos correspondientes a la recuperación de artículos desde el 1 de enero de 2020.

Un total de 87.684 artículos extraídos y normalizados, compuestos por 146.390 palabras obtenidas (muchas de las cuales son palabras de baja frecuencia que no poseen ningún significado).

Como puede apreciarse, el sitio *elchubut.com.ar* es el que más artículos publica por día, con casi 50.000, en todo el período; seguido por *diariojornada.com.ar* con 15.883 de a diferencia de los demás sitios contemplados, debido a que abarca una mayor área incluyendo noticias de otras regiones de Argentina.

4.1. Palabras más frecuentes

En la Figura 3, se presentan de las cinco palabras más frecuentes, agrupadas por la cantidad de veces que aparecen en cada sitio:

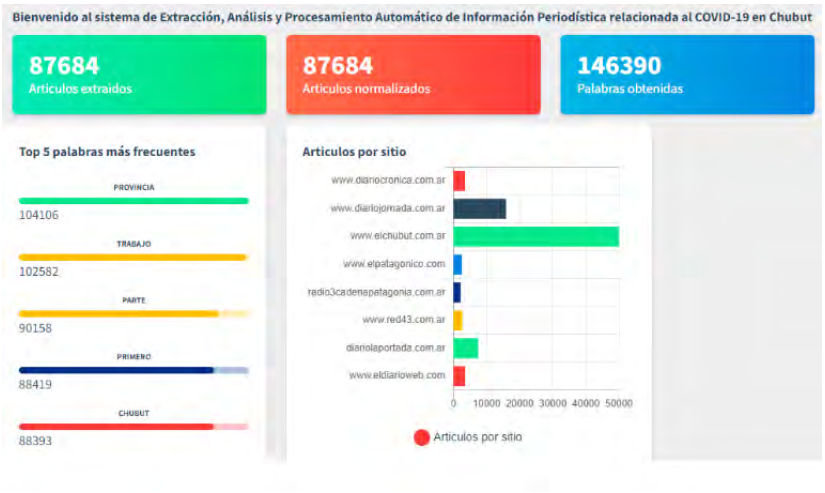


Figura 2: Tablero de control implementado.

4.2. Nube de palabras por sitio

En la Figura 4 se muestran nubes de palabras, donde la frecuencia de las palabras de un sitio particular se ve representada por el tamaño de la palabra en comparación a las demás. Esto permite ver rápidamente cuáles son las palabras obtenidas con mayor frecuencia según el sitio del que se obtuvieron. Las palabras abarcan todos los artículos de su respectivo medio (*dataset* completo).

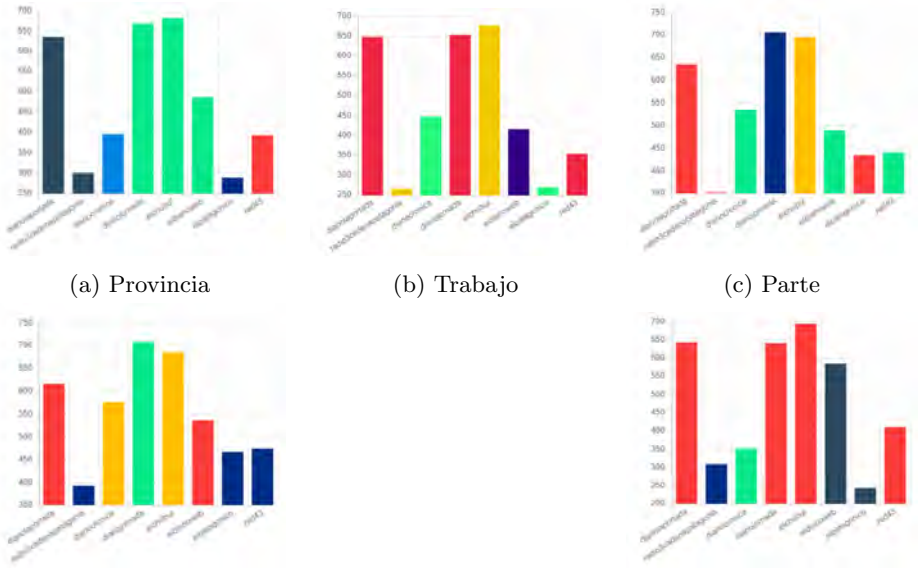
4.3. Nubes de palabras por fecha

Otro tipo de gráfico que se incluyó fueron las nubes de palabras clasificadas por rangos de fechas, el cual permite tener una visión general de todas las palabras y como va evolucionando la frecuencia de las temáticas a lo largo del tiempo.

5. Conclusiones y Trabajos Futuros

En este trabajo se presentó el desarrollo de un sistema para la recuperación, clasificación y análisis de notas periodísticas publicadas en medios digitales. El trabajo se fundamentó en la necesidad urgente, surgida por la pandemia, de contar con información acerca de "lo que ocurría en la sociedad". El desarrollo técnico involucró un análisis de factibilidad tecnológica y del entorno en el que se pensaba operar. En este sentido, la Provincia del Chubut ofició de marco y por ello se buscaron fuentes que tengan representatividad territorial y, a la vez, un volumen considerable de producción.

El desarrollo de este proyecto se dio en el marco del convocatoria COVID Federal del MINCYT. En ese contexto, el proyecto formaba parte de uno mayor,



(a) Provincia (b) Trabajo (c) Parte

(d) Primero (e) Chubut

Figura 3: Frecuencias de palabras por sitio.

que involucraba distintas aristas en el abordaje situacional de la Provincia del Chubut, durante la pandemia. Esa multidimensionalidad y multiescalaridad, se manifestaron en un trabajo interdisciplinario, que marcó los requerimientos de negocio del trabajo aquí presentado. El desarrollo informático posibilitó materializar y explorar las potencialidades de un análisis de este tipo, lo cual definió nuevos requerimientos y demandas.

En línea con los anterior, lo producido en este trabajo será utilizado como base para un futuro trabajo. En particular, el dataset obtenido de más de dos años de notas periodísticas de ocho medios de comunicación de la Provincia del Chubut, será explotado mediante técnicas de Procesamiento de Lenguaje Natural. Especificamente se trabajará sobre modelado de tópicos y modelado dinámico de tópicos, así como técnicas basadas en grafos, a fin de analizar la evolución temporal de distintos temas de interés para los medios de comunicación y, por consiguiente, para la ciudadanía.

Reconocimiento: Los autores quieren expresar un sentido recordatorio y reconocer la labor de la Dra. Florencia del Castillo, quien falleció al tiempo de publicar este artículo, en la dirección del proyecto que posibilitó este trabajo.

(a) Diario Jornada (b) Diario el Chubut (c) Diario el Patagónico

(d) Portal Radio 3 Cadena Patagonia (e) Diario Crónica (f) Diario Red43

(g) Diario La Portada (h) El Diario Web

Figura 4: Nubes de palabras de cada sitio.

Referencias

1. Balcazar, E., Bobadilla, L., Fusiman, W., Ordinez, L.: Geoperfil profesional: una herramienta automática de información sobre profesionales. In: XXVI Congreso Argentino de Ciencias de la Computación (CACIC)(Modalidad virtual, 5 al 9 de octubre de 2020) (2020)
2. Chipidza, W., Akbaripourdibazar, E., Gwanzura, T., Gatto, N.M.: A topic analysis of traditional and social media news coverage of the early covid-19 pandemic and implications for public health communication. medRxiv (2020), `https://www.medrxiv.org/content/early/2020/07/07/2020.07.05.20146894`
3. Chukwusa, E., Johnson, H., Gao, W.: An exploratory analysis of public opinion and sentiments towards covid-19 pandemic using twitter data. (2020)
4. Dousset, B., Mothe, J.: Getting insights from a large corpus of scientific papers on specialisted comprehensive topics-the case of covid-19. Procedia computer science 176, 2287–2296 (2020)
5. Hosseini, P., Hosseini, P., Broniatowski, D.A.: Content analysis of persian/farsi tweets during covid-19 pandemic in iran using nlp. arXiv preprint arXiv:2005.08400 (2020)
6. Martinez, I., Viles, E., Olaizola, I.G.: Data science methodologies: Current challenges and future approaches. Big Data Research 24, 100183 (2021)
7. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association 18(5), 544–551 (09 2011)

(a) Enero 2020 (b) Febrero 2020 (c) Marzo 2020

(d) Abril 2020 (e) Mayo 2020 (f) Junio 2020

(g) Julio 2020 (h) Agosto 2020 (i) Septiembre 2020

(j) Octubre 2020 (k) Noviembre 2020 (l) Diciembre 2020

Figura 5: Nubes de palabras año 2020.

8. Perdomo, L., Ordinez, L.: From global to local in the sneakers universe: A data science approach. In: XXVII Congreso Argentino de Ciencias de la Computación (CACIC)(Modalidad virtual, 4 al 8 de octubre de 2021) (2021)
9. Shafique, U., Qaiser, H.: A comparative study of data mining process models (kdd, crisp-dm and semma). International Journal of Innovation and Scientific Research 12(1), 217–222 (2014)
10. Tommasel, A., Diaz-Pace, A., Rodriguez, J.M., Godoy, D.: Capturing social media expressions during the covid-19 pandemic in argentina and forecasting mental health and emotions. arXiv preprint arXiv:2101.04540 (2021)
11. Wirth, R., Hipp, J.: Crisp-dm: Towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. vol. 1. Springer-Verlag London, UK (2000)
12. Zhang, Y., Chen, J., Liu, B., Yang, Y., Li, H., Zheng, X., Chen, X., Ren, T., Xiong, N.: Covid-19 public opinion and emotion monitoring system based on time series thermal new word mining. arXiv preprint arXiv:2005.11458 (2020)

(a) Enero 2021 (b) Febrero 2021 (c) Marzo 2021

(d) Abril 2021 (e) Mayo 2021 (f) Junio 2021

(g) Julio 2021 (h) Agosto 2021 (i) Septiembre 2021

(j) Octubre 2021 (k) Noviembre 2021 (l) Diciembre 2021

Figura 6: Nubes de palabras año 2021.

SIBDaCAr: Un Prototipo de Sistema de Cronotanatodiagnóstico para la República Argentina

Paola Azar[1,3], Darío Ruano[1,3] , Andrea Maldonado[1,3], Norma Herrera[1,3], Daniel Jaume[2,3] , Marcelo Martínez[4]

[1] Departamento de Informática, FCFMyN, Univ.Nacional de San Luis
[2] Departamento de Matemáticas, FCFMyN, Univ. Nacional de San Luis
[3] Laboratorio de Investigación y Desarrollo en Bases de Datos , Univ. Nacional de San Luis
[4] Jefe Interino del Cuerpo Médico Forense y Criminalístico de la Tercera Circunscripción Judicial de la Provincia de Mendoza

epazar18@unsl.edu.ar, dmruano@unsl.edu.ar, andreamaldonadoma@gmail.com, nherrera@unsl.edu.ar, djaume@unsl.edu.ar, drmarcelomartinez@hotmail.com

Abstract. El cronotanatodiagnóstico es el conjunto de observaciones y técnicas que permiten señalar el intervalo de tiempo donde con mayor probablidad se ha producido el proceso de muerte. La estimación de este intervalo, conocido como IPM (intervalo post mortem), es quizás una de las tareas más complicadas en medicina legal. El apoyo de herramientas informáticas con las que cuenta un médico forense de nuestro país para datar la muerte es escaso o nulo en algunos casos. En este artículo presentamos el desarrollo de un prototipo de un sistema integral de bases de datos (SIBDaCAr) que permita establecer el IPM y que pueda ser usado en el ámbito de nuestro país.
Palabras claves: Bases de Datos, Cronotanatodiagnóstico, IPM

1 Introducción

En la era actual, caracterizada por la evolución de las tecnologías de la información y las comunicaciones, las ciencias de la computación son transversales a la mayoría de nuestras actividades diarias, brindando las herramientas necesarias para abordar problemas complejos y contribuyendo en la búsqueda de soluciones eficientes a problemas de interés. La medicina legal y forense no escapa a esta realidad. Un tema de particular interés en este ámbito es la determinación de la data de muerte. Determinar la data de muerte es quizás uno de los problemas más complejos en medicina forense

El cronotanatodiagnóstico (o datación de la muerte) es el conjunto de observaciones y técnicas que permiten señalar el intervalo de tiempo donde con mayor probablidad se ha producido el proceso de muerte [1]. Este intervalo, conocido como IPM (intervalo post mortem), indica un período de tiempo circunscrito por el momento de muerte y el hallazgo del cadáver. Conseguir averiguar este intervalo de forma precisa es de suma importancia tanto en el ámbito penal como civil. La data de muerte puede, por ejemplo, permitir aceptar o rechazar coartadas durante la investigación de un crimen, y también puede tener consecuencias económicas relacionadas por ejemplo a herencias. La estimación del IPM es una de las tareas más complicadas en medicina legal, no existiendo

un método que sea completamente exacto para ello. Cuanto más prolongado es el IPM menos precisa es la estimación.

El apoyo de herramientas informáticas con las que cuenta un médico forense de nuestro país para datar la muerte es escaso o nulo en algunos casos. El software existente está basado en otras realidades: piden datos que en nuestro país no son viables y/o utilizan en sus cálculos fórmulas basadas en condiciones climáticas poco probables de ocurrir en nuestro país. Para complejizar aun mas la situación, no existe un núcleo común de datos que sea utilizado en todas las provincias para datar la muerte.

Por otro lado, los modelos matemáticos clásicos usados para la datación de la muerte ha demostrado ser limitados [1], [3] y [4]. Por esta razón se hace necesario también, adecuar y validar los mismos a las diferentes geografías de Argentina [5].

El objetivo general de este trabajo es el estudio y desarrollo de herramientas analíticas e informáticas de apoyo y soporte a las tareas de cronotanatación de cadáveres recientes para los sistemas forenses nacionales. La datación de muerte de cadáveres no recientes es extremadamente compleja y la abordaremos en una etapa posterior.

En este artículo presentamos el trabajo realizado que incluye desarrollo de un prototipo de un sistema de bases de datos que permita establecer el IPM y que pueda ser usado en el ámbito de nuestro país. Cabe señalar que este trabajo se ha realizado en el marco del *Proyecto D+i Estudio Analítico y Computacional de la Cronotanatología* [5].

Lo que resta del artículo está organizado de la siguiente manera; en la Sección 2 presentamos una breve reseña sobre cronotanatodiagnóstico donde introducimos las nociones básicas de la problemática para luego, en la Sección 3 analizar la realidad Argentina. En la Sección 4 presentamos nuestro aporte: el diseño e implementación de un prototipo de sistema de cronotanatodiagnóstico para la República Argentina. Finalizamos en la Sección 5 dado las conclusiones y el trabajo futuro.

2 Cronotanatodiagnóstico: un Problema Abierto

El cronotanatodiagnóstico se puede definir como un conjunto de observaciones, técnicas y métodos que permiten establecer un intervalo temporal (IPM) en el cual se ha producido con mayor probabilidad una muerte. Es una de las tres preguntas que se plantean en criminalística (lugar, data y causa de la muerte) y quizás sea uno de los problemas de mayor dificultad en medicina legal.

La precisión y la aplicabilidad de los procedimientos existentes dependen de las características y las circunstancias del fallecimiento y del tiempo transcurrido desde la muerte. Además, dependiendo del estado que tenga el cadáver, es el tipo de técnica que se utilizará. Se distinguen dos casos: cadáver reciente, aquel que no tiene signos evidentes de putrefacción, y cadáver no reciente, aquel con evidentes signos de putrefacción [2]. Como ya mencionamos, en este trabajo abordamos la datación de cadáveres recientes.

Pese a que la datación de la muerte nunca es una ciencia del todo exacta, hay un gran número de factores que ayudan a tener una idea más precisa del tiempo que ha transcurrido desde el fallecimiento. Estos factores denominados fenómenos cadavéricos, se

[5] Proyectos de Desarrollo e Innovación de la Facultad de Ciencias Físico Matemáticas y Naturales, Universidad Nacional de San Luis. Convocatoria 2021

pueden clasificar en inmediatos, mediatos o tardíos según el tiempo que tardan en aparecer. El caso de cadáveres recientes, los datos a tener en cuenta son:

- Signos de muerte molecular: se analizan cambios que se producen en el cadáver dependientes de circunstancias ambientales, que se pueden conocer y medir (fenómenos abióticos) y los de naturaleza físico química que tienen lugar en el cadáver tras la muerte (fenómenos bióticos). Dentro de los fenómenos abióticos encontramos: enfriamiento, deshidratación, livideces, hispotasis visceral. Dentro de los fenómenos bióticoo podemos mencionar rigor mortis, espamos, cambios físico químicos, tanatoquímica y microbiología.
- Signos paramédicos: elementos que rodean al cadáver y la situación en que se encuentra, características de la escena del crimen (por ejemplo si ha llovido), objetos de la víctima o de los alrededores (por ejemplo medios electrónicos con baterías que aun no se hayan agotado), etc.
- Signos de vida residual: reacción pupilar a luz, reacción pupilar a atropina y pilocarpina, contracción muscular, mortalidad de células espermáticas, entre otros.
- Signos derivados del cese de funciones vitales: aquí los datos provienen de observar el estado en que han quedado detenidas las funciones fisiológicas al interrumpirse tras la muerte.

Como puede observarse el volumen de datos que maneja un médico forense y la variabilidad de los mismos dependiendo de factores que son propios a cada hecho en sí, hace que la estimación de la data de muerte sea en extremo no-lineal. Los modelos matemáticos clásicos que dan los indicadores para establecer la data de muerte están basados fundamentalmente en modelos lineales, logarítmicos y sigmoidales (según los casos) de unas pocas variables, típicamente una o dos. Este enfoque ha demostrado ser limitado [1], [3], [4].

Para comprender un poco mas la complejidad del proceso de datación veamos un ejemplo. La temperatura corporal es uno de los datos más usado en la determinación del IPM. Una vez producida la muerte, la temperatura disminuye de forma gradual y progresiva hasta igualarse con el medio ambiente. Para calcular el intervalo post mortem mediante la temperatura, se sigue un modelo doble exponencial propuesto por Marshal y Hoare en el año 1962. Este modelo establece que es posible diferenciar una doble fase de enfriamiento: una meseta donde prácticamente no hay enfriamiento, y una fase final progresiva. Pero hay múltiples factores que afectan el proceso de enfriamiento de un cadáver: la temperatura ambiental, el peso de la persona dado que la grasa actúa como aislante térmico, las capas de ropas, las corrientes de aire, la humedad, la hipertermia y algunas enfermedades y situaciones previas al fallecimiento (intoxicaciones, sepsis, hemorragias) que alteran la curva de enfriamiento. Además, como los métodos de datación por temperatura dependen de la diferencia de temperatura corporal con relación a la ambiental, su efectividad se reduce notablemente en lugares con temperatura alta o con cambios bruscos de temperatura.

La complejidad y las dificultades que plantea la determinación del IPM son categóricas y es por que constituye uno de los problemas más complejos que afronta el médico forense.

3 Cronotanatodiagnóstico: Software y Realidad Argentina

Establecer en la actualidad el intervalo post mortem (IPM) de forma precisa sigue siendo un reto. Argentina no escapa a esta realidad. Como explicamos en la sección anterior, calcular el IMP es un problema no lineal, de múltiples variables cuyos valores se ven afectados por múltiples factores. Quizás esta sea una de las principales razones por las que el software de apoyo al cronotanatodiagnóstico sea escaso o de uso no masivo como ocurre en otras ciencias.

Veamos un ejemplo. El software AMAsoft (www.amasoft.deindex_e.html) utiliza dos métodos para realizar el proceso de datación: Nomograma de Henssge [3] y métodos no basados en la temperatura. En el caso del Nomograma de Henssge, no es adecuado para las condiciones climáticas habituales en la mayor parte nuestro país porque requiere variaciones térmicas suaves de menos de 5 grados centígrados a lo largo de día, con un rango de entre 16 a 24 grados centígrados. Esto implica que el Nomograma solo es utilizable en la región sur del país. Una instancia superadora del modelo de Henssge fue propuesta en [6] donde los autores proponen un modelo termodinámico de diferencias finitas, cuya inicialización (condiciones iniciales de frontera) es una lectura termométrica de la piel del cadáver. Lamentablemente, este enfoque no es aplicable a la realidad forense Argentina debido al tipo de equipos requeridos para tomar las lecturas y el entrenamiento que se requeriría del personal: formación en discretización del espacio y conocimientos elementales de la teoría de ecuaciones de diferencias finitas.

Con respecto a los métodos no basados en temperaturas usados en AMAsoft, éstos necesitan como datos de entrada: rigidez, livideces , excitabilidad mecánica de los músculos esqueléticos, excitación eléctrica de músculos de la mímica y excitabilidad química del iris (atropina, tropicamid/cyclopent, acetilcolina). Y es aquí cuando nos enfretamos con el segundo problema: no siempre existe la disponibilidad de esos datos. Mas aún, no existe un núcleo común de datos que sean usado por todos los forenses del país para realizar el proceso de datación. Si bien hay un consenso general sobre qué datos deberían usarse para datar la muerte, los datos que realmente se usan quedan determinados por la circunstacia de la muerte, por la disponibilidad de recursos y en algunos casos también por la experiencia del médico forense que interviene.

Todo lo expuesto justifica que el relevamiento hecho sobre software existente en nuestro país para cronotanatodiagnóstico (ver Sección 4.1) arrojó que las herramientas informáticas con las que cuenta un médico forense para datar la muerte son escasas o nulas en la mayoría de los casos. Resulta fundamental entonces proveer de herramientas tecnológicas que sirvan de apoyo al cronotanatodiagnóstico, las que deberán surgir de un enfoque multidisciplar del problema lo que asegurará la robustez de las mismas.

4 Nuestro Aporte: Un Prototipo para un Sistema de Datación

En este trabajo abordamos el problema de cronotanatodiagnóstico en Argentina proponiendo un cambio de paradigma: a partir del diseño de una base de datos específica, aplicar algoritmos de clustering para hallar clusters relevantes de casos, agrupados según múltiples factores extrínsecos e intrínsecos. En una segunda etapa, abordaremos el problema de desarrollar modelos matemáticos de cronotantadiagnóstico específicos para los diferentes clusters de casos hallados.

El objetivo final de este trabajo es diseñar un sistema de base de datos adecuado para la comunidad tanatológica argentina, que hemos llamado **SIBDaCAr** (sistema integral de base de datos cronotanatológica argentina). Lo que aquí presentamos es la primer etapa del trabajo que consistió en el desarrollo de un prototipo del sistema SIBDaCAr. Dada la complejidad del problema, decidimos usar la técnica de prototipado evolutivo. Describimos a continuación el trabajo realizado en el desarrollo de este prototipo.

4.1 Análisis de requisitos

Como todo proyecto de software, comenzamos con el análisis de requerimientos para detectar requerimientos funcionales y no funcionales del sistema usando como técnicas de elicitación entrevistas, observación y encuestas.

Las entrevistas inicialmente se realizaron con médicos forenses pertenecientes a las provincias de Mendoza, Córdoba, Tierra del Fuego y Corrientes. Luego de cada entrevista, se realizó un brainstorming con el equipo encargado de desarrollar el prototipo.

En lo que refiere a la observación, dado que por la temática involucrada es imposible realizar la observación directa de las actividades llevadas a cabo en el proceso de datación, se organizó un seminario de 4 encuentros donde médicos forenses expusieron casos de datación de distintas regiones del país y bajo distintas circunstancias.

Por lo explicado en la sección anterior, un punto importante era establecer un núcleo común de datos. Para ello se elaboró una encuesta que fue distribuida entre médicos forenses de todo el país. En esta encuesta se presentaban datos usados en cronotanatodiagnóstico organizados en dos categorías:

Datos tomados en el lugar del hecho: temperatura corporal, posición del cuerpo, temperatura del ambiente, reacción de la pupila a la luz, evaluación/descripción de las prendas en circunstancias de cadáver vestido, evaluación del contexto ambiental en el lugar del hallazgo.

Datos tomados en el Instituto Forense: temperatura hepática, temperatura rectal, temperatura auricular livideces, rigidez, opacidad de la córnea, concentración de potasio en humor vitreo, grado de putrefacción, excitabilidad mecánica de los músculos esqueléticos, excitación eléctrica de músculos de la mímica, evaluación del contenido gástrico, evaluación del contenido intestinal, estado de la vejiga, reacción pupilar por estimulación lumínica, reacción pupilar por estimulación química, movilidad del epitelio respiratorio, movilidad de los espermios, el mantenimiento del cadáver entre levantamiento y la autopsia.

Por cada dato se pidieron dos valoraciones:*el grado de importancia* que le asigna el encuestado a ese dato, en una escala de 0 a 5, donde 0 significa nada importante y 5 significa muy importante, y *la frecuencia de uso* que el encuestado hace de ese dato, usando la escala de valores: nunca, 20%, 40%, 60%, 80%, siempre.

Resultados obtenidos. Con respecto a los requisitos se documentaron los mismos especificando: objetivo del sistema, alcances y limitaciones, requisitos no funcionales y requisitos funcionales. En el caso de los requisitos funcionales se organizaron por funcionalidades y se describío de manera detallada cada uno de ello. La Figura 1 muestra

Identificador: R5	**Nombre: Detección de inconsistencias entre los datos ingresados**
Tipo (necesario/deseable): Deseable	**Funcionalidad:** Datación de la muerte
Crítico: NO	**Prioridad de desarrollo:** Media
Entrada: core de datos	**Salida:** Los datos ingresados no son consistentes entre si.
Descripción: Permitirá realizar la correlación de datos para la detección de posibles inconsistencias entre los mismos.	
Manejo de situaciones anormales: Emitir alerta	
Criterios de aceptación: El usuario deberá confirmar los datos ingresados	

Fig. 1. Documentación del requisito R5 perteneciente a la funcionalidad *Datación*.

la descipción de un requisito funcional. Por cuestiones de espacio no mostramos la documentación completa pero la misma se encuentra disponible para quien así requiera.

Con respecto a la encuesta, se recibieron el total 35 respuestas que se analizaron usando la plataforma data studio. Quizás lo más sorprendente fue ver que no hay una relación directa entre grado de importancia y frecuencia de uso. La figura 2 muestra a modo de ejemplo el resultado para el dato *reacción de la pupila a luz*. Como puede observarse es un dato considerado de importancia por un 72% de los encuestados aproximadamente (zonas en tonos de azul) pero que el 75% de los encuestados refiere que nunca lo usa. En la encuesta también se preguntaba sobre el software de apoyo utilizado, obteniendo como respuesta que, de los 35 encuestados, solo uno refiere usar el software SWISSWUFF (https://www.swisswuff.ch).

4.2 Elaboración de modelos

En la segunda etapa se elaboraron los siguientes modelos para SIBDaCAr:

- El modelo entidad relación y el modelo relacional de la base de datos
- El diagrama de componentes UML del sistema.
- El diagrama de casos de uso UML

Por cuestiones de espacio, solo mostramos el modelo relacional en la Figura 3, que fue generado usando MySQL Workbench. Los datos allí considerados son los que se establecieron como núcelo, pero sin descartar aquellos que se consideren importantes aunque tengan poca frecuencia de uso. El objetivo es que el proceso de datación utilice al menos los datos pertenecientes al núcleo, pero que se pueda mejorar la estimación en aquellos casos en los que se cuentan con otros datos adicionales al núcleo.

4.3 Implementación del prototipo

En la programación del prototipo se utilizó Python como lenguaje principal y Javascripts, CSS y HTML para la parte gráfica. La elección de Phyton se basó en que es software libre bajo licencia Python Software Foundation License, licencia muy parecida a la de GPL, con la facilidad de que se pueden distribuir los binarios del código sin tener que

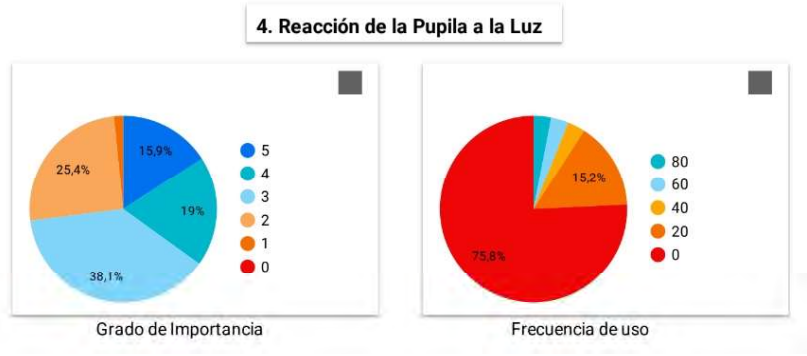


Fig. 2. Resultado de la encuesta para el dato *reacción de la pupila a la luz*

.

anexar las fuentes; es mulitparadigma y multiplataforma y ofrece frameworks de gran utilidad para el desarrollo de aplicaciones web.

La implementación del prototipo incluyó la generación de la bases datos. Esto fue realizado usando como motor de bases de datos MySQL dado que es software libre bajo la licencia GPL de código abierto, es personalizable porque la licencia GPL permite adecuarlo a necesidades específicas de la aplicación, es multiplataforma, permite varias capas de seguridad y soporta bases de datos de gran tamaño.

Por cuestiones de espacio solo mostramos las pantallas que consideramos más representativas. La Figura 4 muestra las pantallas correspondiente al menú principal y al ingreso de un nuevo expediente. El menú principal consta de 4 opciones: expedientes, cadáveres, autopsias y estimación (IPM). La Figura 5 muestra las pantallas correspondientes a los módulos de ingreso de datos del cadáver, ingreso de datos de la autopsia y el módulo datación. En el caso del módulo de datación se consideran tres posibilidades: nomograma de Henssge, datación usando el núcleo de datos y detección de casos similares usando algoritmos de clsutering. Para el caso del nomograma, el sistema recomendará o no su uso según las condiciones de temperatura ambiental, que se obtendrán automáticamente a partir de la ubicación ingresada al generar un expediente.

5 Conclusiones y Trabajo Futuro

En este artículo presentamos el desarrollo de un prototipo de un sistema integral de bases de datos (SIBDaCAr) que permita establecer el IPM y que pueda ser usado en el ámbito de nuestro país. El trabajo incluyó el análisis de requisitos, la elaboración de modelo y la programación del prototipo del sistema, usando la idea de prototipado evolutivo. El trabajo realizado hasta el momento nos permitió comprender la enorme complejidad del problema y visualizar las carencias de herramientas informáticas apropiadas a las condiciones de nuestro país para el cronotanatodiagnóstico. Como trabajo futuro nos proponemos estudiar el estudio de métodos de clustering . Ya hemos iniciado el desarrollo de un modelo de predicción basado en clustering, redes neuronales, y aprendizaje supervisado.

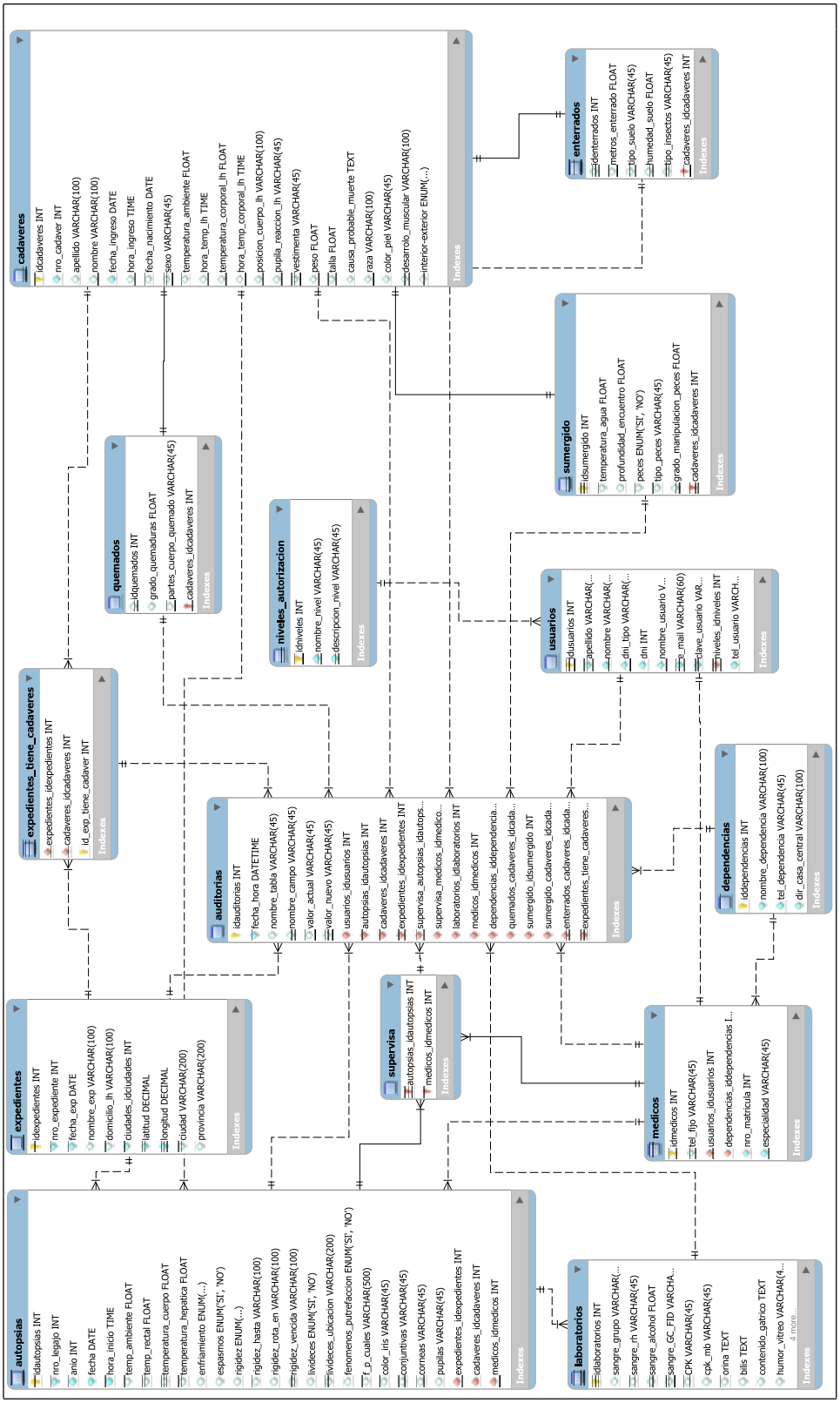


Fig. 3. Modelo Relacional de SIBDaCAr, realizado usando MySQL Workbench

Agradecimientos

Agradecemos a la Dra Inés Aparici (Médica Forense Subdirectora, Poder Judicial de Tierra del Fuego), al Dr. José Gálvez (Jefe de Gestión del Instituto Médico Forense del Poder Judicial de Corrientes) y al Dr. Moisés Dib (Jefe del Instituto de Medicina Forense de Córdoba) por la colaboración recibida para el desarrollo de este trabajo.

References

1. Maldonado, A. L. (2010) *La data de la muerte, un desafío no resuelto.*. Revista Española de Medicina Legal.
2. Trezza, F. C. (2006) *La data de la muerte. Las trasformaciones cadavéricas.* Ediciones Argentinas. Buenos Aires.
3. Henssge, C., & Madea, B. (2007). *Estimation of the time since death* Forensic science international, 165(2-3), 182-184.
4. Vidoli, G. M., Beasley, M. M., Jantz, L. M., Devlin, J. B., & Steadman, D. W. (2020). *The future of taphonomic research.* Estimation of the Time since Death: Current Research and Future Trends, 251?261.
5. Hayman, J., &Oxenham, M. *Estimation of the time since death in decomposed bodies found in Australian conditions*. Australian Journal of Forensic Sciences, 2017, 49(1), 31-44.
6. Wilk, L., Hoveling, R., Edelman, G., Hardy, H., Schouwen, S., Venrooij, H. & Aalders, M.C.G. (2020). *Reconstructing the time since death using noninvasive thermometry and numerical analysis.* Science Advances 6(22), 2020, Pages: eaba4243.

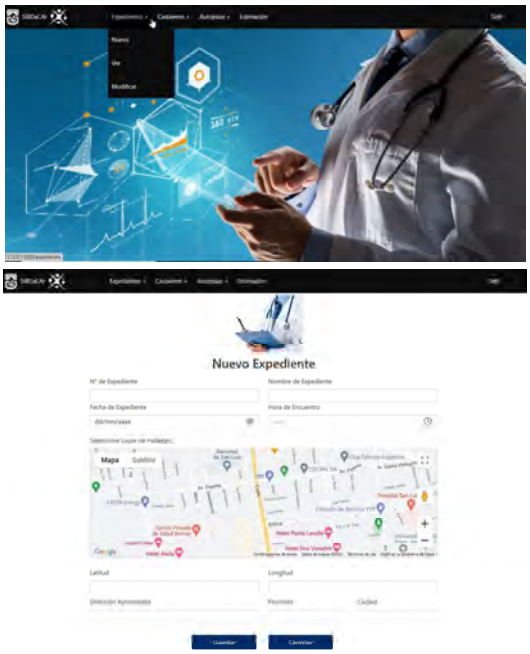


Fig. 4. Menú principal e ingreso de un nuevo expediente.

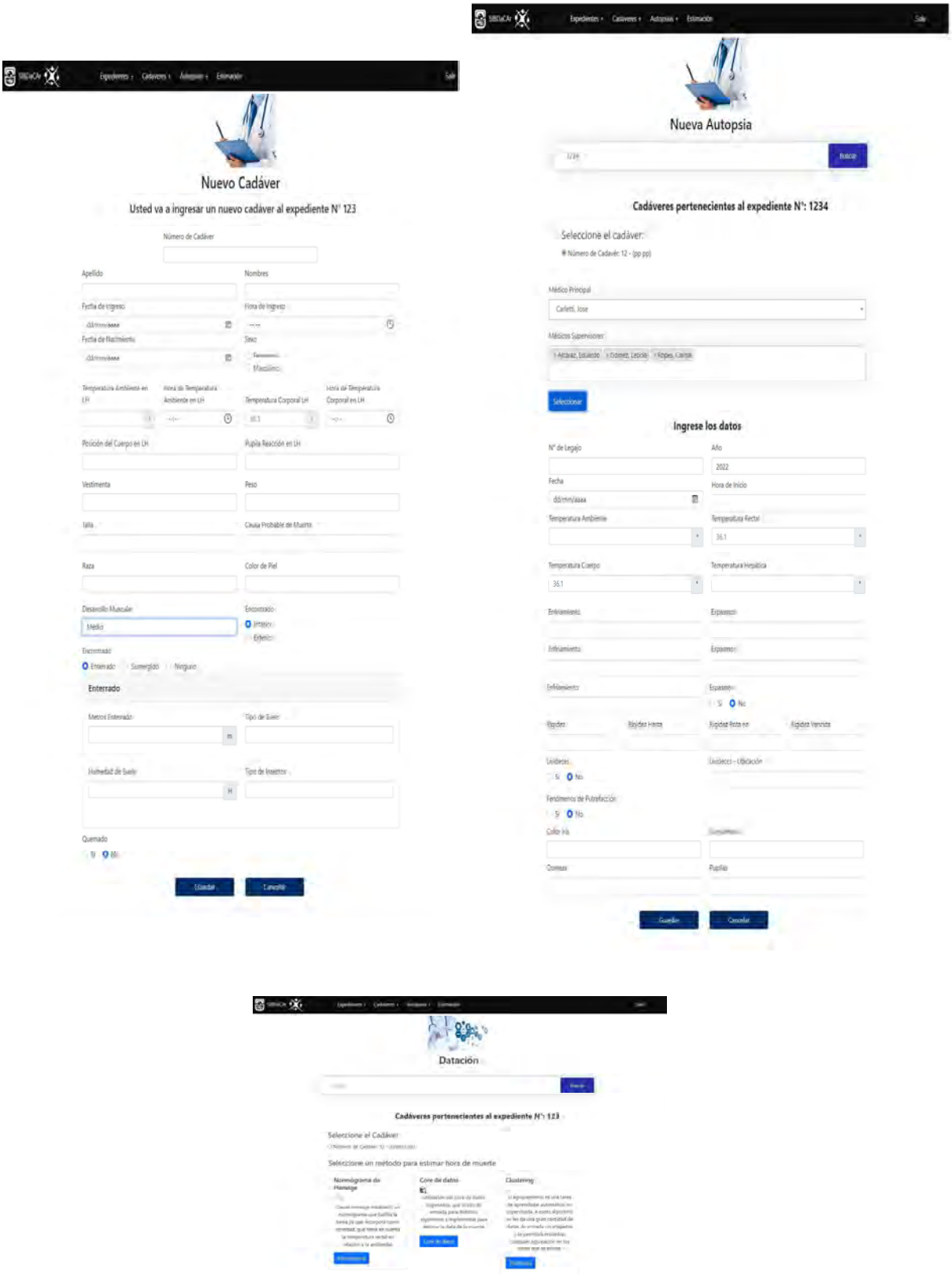


Fig. 5. Ingreso de datos de cadáver y autopsia, y módulo de datación.

Desarrollo de Feed Mashup con Línea de Productos de Software

Héctor Reinaga[1], Juan Enriquez[1] y Sandra Casas[1]

[1] GISP, Instituto de Tecnología Aplicada
Universidad Nacional de la Patagonia Austral (UARG)
Campus Universitario, Av. Gregores y Piloto Lero Rivera
Río Gallegos, Santa Cruz
{hreinaga, jenriquez, sicasas}@uarg.unpa.edu.ar

Abstract. Un mashup es una aplicación compuesta que integra dos o más tipos de componentes disponibles en la Web, creando un nuevo valor a partir de los componentes o artefactos que la componen. Las actuales herramientas y enfoques de desarrollo de estas aplicaciones carecen de algún modelo para integrar componentes similares. Las Líneas de Productos de Software es un enfoque de desarrollo de software cuyo principal objetivo es la reusabilidad, permitiendo crear una familia de productos donde cada producto posee características comunes, y difiere de otro en un conjunto de funcionalidades. En este trabajo se propone un enfoque para modelar, diseñar e implementar una aplicación Mashup desde una perspectiva de variabilidad, lo cual permitirá implementar una línea de productos de software para este dominio.

Keywords: Mashup, Línea de Producto de Software, Modelo de Características, Feature.

1 Problemas y Objetivos

Un mashup es una aplicación compuesta que integra dos o más tipos de componentes disponibles en la Web, creando un nuevo valor a partir de los componentes o artefactos que la compone, permitiendo su reuso y proporcionando una funcionalidad que no existía antes [1]. Un componente es cualquier segmento de datos, lógica de aplicación y/o interfaz de usuario que puede ser reutilizada y que es accesible ya sea local o remotamente [2]; y se basan en lenguajes estándar, tecnologías o protocolos de comunicación. La heterogeneidad de componentes y las diferentes tecnologías disponibles en la Web, llevan a otro desafío, el de seleccionar los modelos adecuados para la composición e integración de dichos componentes. Las actuales herramientas y enfoques de desarrollo de estas aplicaciones, carecen de algún modelo para integrar componentes similares.

Las Líneas de Productos de Software (LPS) son un enfoque de desarrollo de software cuyo principal objetivo es la reusabilidad, permitiendo crear una familia de productos donde cada producto posee características comunes, y difiere de otro en un conjunto de funcionalidades opcionales (variables) que implementa [3]. Esta diferencia funcional entre productos de una LPS se conoce como variabilidad [4].

Este trabajo propone un enfoque para modelar, diseñar e implementar aplicaciones Feed Mashup desde una perspectiva de variabilidad (características) y LPS. Así como resultado, se obtendrán artefactos para mejorar la reusabilidad, y la composición e integración de Feed en aplicaciones mashup.

Se establecen los siguientes objetivos específicos:

- Estudiar los enfoques y técnicas para el desarrollo de aplicaciones Feed Mashup, y especificar las mismas en términos de composición e integración.
- Aplicar técnicas del modelado de características para definir un enfoque que provea mecanismos para la reutilización, variabilidad y configuración necesarios para la integración y composición de aplicaciones Feed Mashup.
- Diseñar e Implementar una herramienta que dé soporte al enfoque propuesto.
- Evaluar y analizar la efectividad de las estrategias propuestas.

Asimismo, las hipótesis que se formulan son:

- Existen técnicas y/o herramientas que permitan mejorar la integración y composición de aplicaciones mashup, superando la limitación de los enfoques existentes.
- Un enfoque basado en variabilidad y LPS, permitirá el modelado, diseño e implementación de aplicaciones Feed Mashup, incrementando el nivel de automatización.

2 Antecedentes

Actualmente la Web ha dejado de ser un medio de comunicación en un solo sentido. Hoy existe un enorme ecosistema de aplicaciones en ejecución para diversos dispositivos, y algunas de ellas permiten la interacción con otras aplicaciones [2]. Con esto se ha marcado una nueva tendencia, la cual consiste en reutilizar componentes de diversas aplicaciones para formar nuevas aplicaciones que ofrezcan un valor agregado. Estos componentes o bloques, han hecho posible que aparezcan las aplicaciones denominadas Mashup. Técnicamente un Mashup se define como una aplicación compuesta que integra dos o más tipos de componentes disponibles en la Web [1]. Un componente es cualquier segmento de datos, lógica de aplicación y/o interfaz de usuario que puede ser reutilizada y que es accesible ya sea local o remotamente [2].

La Redifusión Web es un proceso por el cual un productor o distribuidor de contenidos produce información en formato digital a un suscriptor o una red de suscriptores [5]. Por lo tanto permite a un sitio informar a los interesados respecto a sus actualizaciones, logrando así personalizar los contenidos que ofrecen las publicaciones electrónicas, portales y sitios [6]. Aquí surge el concepto de Feed (fuente, canal), se define como documentos utilizados para transferir las actualizaciones de los contenidos digitales a los usuarios [7]. Los dos formatos más conocidos son RSS y Atom. RSS (Really Simple Syndication) es un dialecto del lenguaje XML.

Se considera las fuentes RSS como un componente Mashup entre otras tecnologías [8], y en consecuencia, no escapa a las dificultades que se encuentran en el proceso de desarrollo, debido al volumen y heterogeneidad de componentes mashup disponibles en la Web; ausencia de algún modelo para integrar componentes similares; y selección

de los modelos adecuados para la integración. Existen diversos enfoques orientados al desarrollo Mashup, en [9] se han llevado a cabo estudios para asistir en la selección de componentes proponiendo un framework basado en la calidad de los componentes; en [2] se aborda el desarrollo Mashup como una metodologia simple basada en componentes con la cual usuarios finales pueden crear sus propias aplicaciones.

A partir de un enfoque de variabilidad en el contexto de una LPS, que permita tener en cuenta el análisis sobre la heterogeneidad de los componentes en la Web, y los requerimientos de composición para integrar aplicaciones Mashup; proporcionaría nuevas opciones de modelado, diseño e implementación para componer e integrar este tipo aplicaciones.

Una LPS es una familia de sistemas (o productos) relacionados a un dominio en particular, cuyos artefactos de implementación son compartidos [10]. Los productos de una misma LPS poseen un conjunto de características en común, denominado núcleo, pero cada producto difiere de otro en un conjunto de funcionalidades opcionales (variables) que implementa [3]. Esta diferencia funcional entre productos de una LPS se conoce como variabilidad [4]. Para expresar características comunes se crean modelos de características y para manejar la variabilidad entre los productos de una LPS, modelos de variabilidad [11].

2.1 Modelo de Características

Las LPS se representan por medio de modelos. Una de las técnicas que existen para expresar estos modelos se denomina Modelos de Características (MC) [12].

Una "característica" (feature) es un rasgo o elemento distintivo que representa aspectos relevantes de un dominio de aplicación según el punto de vista del usuario o desarrollador. Las características se usan en las LPS para especificar y comunicar aspectos comunes y variables de la LPS [13]. Las características se relacionan entre sí con dependencias, entonces un modelo jerárquico puede ser creado, clasificando y estructurando las características, utilizando distintos tipos de relaciones. El método original [12] clasifica las características en obligatorias, opcionales o alternativas. El modelo se completa, con la especificación de las restricciones, que se conforma como un conjunto de reglas (expresiones lógicas formadas por características, conectivos lógicos y cuantificadores) [14].

Los MC se han aplicado en un amplio rango de dominios, y en particular el desarrollo de LPS en distintos dominios, como automotriz [14], telefonía móvil [15], robótica, geográficos [16], Gestor de Ventanas [12], TV Digital [17], entre otros, no encontrándose así para el dominio de aplicaciones Mashup.

La aplicación de MC en el dominio de aplicaciones Mashup, resulta así una oportunidad y a la vez un desafío que motiva el presente trabajo investigativo.

3 Metodología

El enfoque metodológico que se propone aplicar corresponde al DSR (Design Science Research). Este enfoque se dirige a la producción de artefactos, tales como

instanciaciones (muestran que modelos o métodos pueden ser implementados, como sistemas, prototipos e implementaciones).

En la revisión y análisis de la bibliografía se realizará una construcción del estado del arte en las áreas de Mashup, LPS, Desarrollo de Software Orientada a Feature, y MC aplicado a composición de Web Service. Los modelos se plantearán en forma gráfica y formal, empleando notaciones existentes en el campo del MC.

La herramienta se desarrollará a partir de lenguajes clásicos y apoyo de IDE/toolkit, (Feature IDE de Eclipse, FAMA, SPLOT, etc.). Para la evaluación del modelo se aplicarán métodos basados en el estándar de calidad SQuaRE [18], sobre diversos atributos (reusabilidad, mantenibilidad, flexibilidad, evolución y otros).

Los casos de estudio se aplicarán a la evaluación de la herramienta, para ello se desarrollarán con la misma, aplicaciones Feed Mashup con diferentes características y propósitos. En esta fase de la evaluación se usarán los enfoques [19] [20].

4 Conclusiones

En este trabajo, se propone un enfoque para modelar, diseñar e implementar aplicaciones Feed Mashup desde una perspectiva de variabilidad (características) y LPS. Así como resultado, se obtendrán artefactos para mejorar la reusabilidad, y la composición e integración de Feed en aplicaciones mashup. Para ello, y teniendo en cuenta una tecnología disponible y fundamental para desarrollar un Mashup, como son los Feeds, se procederá a estudiar y analizar distintos enfoques y técnicas para el desarrollo de aplicaciones sobre este dominio.

Posteriormente, se aplicarán técnicas del modelado de características para definir un enfoque que provea mecanismos para la reutilización, variabilidad y configuración necesarios para la integración y composición de aplicaciones Feed Mashup. A continuación, se realizará la implementación de la herramienta, de acuerdo al enfoque más conveniente; siendo la primera actividad, la definición de los requisitos y objetivos que debe cumplir, como así también sus limitaciones. En función de estas definiciones, se procederá a la especificación técnica y diseño, lo que refiere a las definiciones sintácticas y semánticas. La herramienta obtenida será sometida a una evaluación de calidad basada en el estándar de calidad SQuaRE [20]; y su validación mediante casos de estudios con diferentes características y propósitos, que permitan ajustar y corregir defectos, y además permitan garantizar el cumplimiento de los requisitos definidos.

Esta propuesta de trabajo es relevante dado que apunta a contribuir en dos campos: desarrollo web mashup, y desarrollo de software orientado a características, como así también, en dos aspectos esenciales de la Ingeniería de Software: la reutilización del software y el mantenimiento del mismo. Estos factores inciden directamente en los costos, esfuerzos y duración requeridos en el desarrollo del software.

Referencias

1. Daniel F., Muhammand I., Soi S., De Angeli A., Wikinson C., Casati F., y Marchese M.: Developing Mashup Tools for End.Users: On the Importance of the Application Domain. International Journal on Next-generation Computing, Vol 2. Nro. 2 (2012).
2. Daniel, F., y Matera M.: Mashups Concepts, Models and Architectures. Milano, Italy: ISBN: 978-3-642-55049-2. Springer (2014).
3. Garces. K., Parra, C., Arbolera, H., Yie, A. Y Casallas, R.: Administración de Variabilidad en una línea de producto basada en modelos. Proceedings of the Congreso Colombiano de Computación, Bogotá, Colombia (2007).
4. Pohl, K., Boeckle G., y Van Der Linden, F.: Software Product Line Engineering-Foundations, Principles, and Techniques, Springer Verlag, Berlin/Heidelberg (2005).
5. Holzner, S.: Secrets of RSS. : Peachpit Press (2006).
6. Powers, S.:What Are Syndication Feeds. : O'Reilly Media, Inc. (2005).
7. Yee, R.: Pro Web 2.0 Mashups: Remixing Data and Web Services. : Apress (2008).
8. Tinajero Diaz I.: Composición de sistemas con Mashups: El caso PhysicallTrello. Centro de Investigación de Matemática A.C . Zacatecas (2016).
9. Saeed A.A.: Quality-based Framework for Leveraging the Process of Mashup Component Selection. ISSN: 1651-4769. Department of Applied Information Technology. University of Gothenburg Sweden (2009).
10. Capilla, R., Bosch, J., Trinidad, P., Ruiz Cortés. A. y Hincheyd, M.: An overview of Dynamic Software Product Line architectures and techniques: Observations from research and industry. The Journal of Systems and Software 91, pp. 3-23 (2014).
11. Galindo J. A., Alférez M., Acher M., Baudry B. y Benavides D.: A variability-based testing approach for synthesizing video sequences. InProceedings of the 2014 International Symposium on Software Testing and Analysis, pp. 293-303 (2014).
12. Kang, K. , Cohen, S., Hess, J. , Novak, W., y Peterson, S.:Feature–Oriented Domain Analysis (FODA) Feasibility Study. Technical Report CMU/SEI90-TR-21, Software Engineering Institute, Carnegie Mellon University, November (1990).
13. Apel S., Batory D., Kästner C., y Saake G.: Feature-Oriented Software Product Lines. Berlin, Heidelberg: Springer Berlin Heidelberg (2013).
14. Schobbens P. Y., Heymans P., Trigaux J. C., y Bontemps Y.: Generic semantics of feature diagrams. Comput. Netw., vol. 51, nro. 2, pp. 456-479 (2007)
15. González A., Luna C., Zorzan F. y Szasz N.: Automatic Derivation of Behavior of Products in a Software Product Line. IEEE Latin America Transactions, Vol. 12, No 6 (2014).
16. Gherardi L., y Brugali D.: An eclipse-based Feature Models toolchain. An eclipse-based feature diagrams toolchain. In Eclipse-IT 2011. The Sixth Workshop of the Italian Eclipse Community, pp. 242-253 (2001).
17. Oyarzo F., Herrera F., Miranda M. y Casas S.: Línea de Producto de Software para aplicaciones de TVDi basado en patrones de diseño. ISSN 1852-4516. https://publicaciones.unpa.edu.ar/index.php/ICTUNPA/article/view/503 (2013).
18. ISO Standard 25000. System and Software Quality Requirements and Evaluation (SQuaRE). Disponible en: http://iso25000.com/.
19. Kitchenham B., Pickard L., y Pfleeger S. L.: Case Studies for Method and Tool Evaluation. IEEE Softw, Vol. 12, Nro. 4, pp. 52–62 (1995)
20. Pfleeger S. L.: Experimental Design and Analysis in Software Engineering: Part 2: How to Set Up and Experiment. SIGSOFT Softw Eng Notes, Vol. 20, Nro. 1, pp. 22–26 (1995).

Hacia un marco de desarrollo de sistemas de programación de la producción que permita la integración de chatbots

Daniel Díaz[1, 2], Sandra Oviedo[1, 2], Juan Manuel Cuneo[2], María del Carmen Becerra[2]

[1] Laboratorio de Informática Aplicada a la Innovación
Instituto de Informática, [2]Depto de Informática
FCEFN- UNSJ- San Juan, Argentina
ddiaz@iinfo.unsj.edu.ar

Resumen. La industria 4.0 es un conjunto de tecnologías, conceptos y procesos que ayuda a redefinir la cadena de valor de una empresa. La capacidad que tiene una empresa para absorber tecnologías de la industria 4.0 depende de varios factores, entre otros, su tamaño, sus recursos tecnológicos, y la actividad que desempeña. Dentro de una empresa la programación de la producción es el mecanismo que se utiliza para la planificación y control de la producción. En la literatura revisada se reportan muy pocos casos del uso de la tecnología de la industria 4.0 en relación a la programación de la producción en empresas con mediana capacidad de absorción. En este trabajo se describe una aproximación para incorporar la tecnología de chatbots a un sistema de programación de la producción en una industria con una mediana capacidad de absorción de tecnologías de la industria 4.0.

Palabras clave: programación de la producción, industria 4.0, chatbots

1. Introducción

La capacidad de una empresa para absorber tecnologías de la industria 4.0 se puede definir como la capacidad que tiene una empresa para apropiarse y desplegar tecnologías de la industria 4.0. La industria 4.0 presenta nuevos desafíos y oportunidades para los sistemas de programación de la producción [1]. La mayoría de los artículos relacionados con programación de la producción e industria 4.0 se enfocan en empresas con elevado poder de absorción [2] [3] , es decir, en el entorno sistemas ciberfísicos. Sin embargo, hay tecnologías de industria 4.0 que pueden aplicarse a los sistemas de programación de la producción en empresas con menor poder de absorción, tal es el caso de la tecnología de chatbots. Este trabajo presenta los posibles usos del chatbot en este dominio y el estado actual de un marco para desarrollar sistemas de programación de la producción que permite integrar dicha tecnología.

2. Sistemas de programación de la producción

Matemáticamente los problemas de programación de la producción son un tipo de problema de scheduling, más específicamente, son problemas de scheduling en el dominio industrial y son conocidos en la literatura inglesa como production scheduling o manufacturing scheduling. La definición más clásica de la palabra scheduling dice: "Scheduling es el problema de asignar recursos limitados a tareas en el tiempo con el objeto de optimizar uno o más objetivos" [4]. En la industria, la producción está a cargo del sistema de producción, el cual tiene como componente al sistema de planificación y control de la producción, una parte de este sistema es el sistema de programación de la producción o scheduling de producción. Normalmente el sistema de scheduling interactúa con otros sistemas de la empresa, tales como ERP (Enterprise Resource Planning) y MES (Manufacturing Execution System). Según Yen y Pinedo [5] un sistema de scheduling se compone de tres módulos: (1) módulos de base datos y base de conocimiento, (2) módulos del motor de scheduling y (3) módulo de interfaz de usuario. La parte esencial de todo sistema de scheduling es el motor de scheduling, es donde se vinculan las necesidades que tiene la empresa con los modelos y algoritmos de scheduling.

3. Chatbots

Conceptualmente un chatbot, es un programa informático que utiliza la inteligencia artificial y procesamiento de lenguaje natural para simular una conversación con un humano. En [6-8] se presenta arquitecturas generales de un chatbot, en las que se describen un módulo de procesamiento del lenguaje natural, el cual intenta analizar e interpretar las intenciones del usuario y las transforma en datos estructurados. El módulo de gestión del dialogo, que analiza la entrada transformada en datos estructurados comprensibles por el chabot, mantiene el contexto de la conversación. Por ultimo existe un módulo de generación de respuesta en lenguaje natural. En cuanto a los posibles usos del chatbot en la programación de la producción, la tabla 1 resume los casos que se han detectado.

Tabla 1. Posibles aplicaciones de chabot en un sistema de programación de la producción

Control de Stock	Un chatbot puede hacer de intermediario entre el responsable de control de stocks y el sistema de planificación de la producción para realizar consultas de los posibles planes de producción y el efecto que tiene en su stock, como así también, analizar como impactan diversos movimientos de SKU en un plan de producción.
Agilizar la gestión del trabajo de los operarios	Un chatbot puede informar ágilmente a un operario sobre cuál es su próxima tarea y su causa, lo que permite una coordinación fluida de los operarios y los hace más participes de las decisiones que toman los programadores de la producción. En ambientes de producción donde participan más de una planta industrial este tipo de prestaciones tecnológicas son muy útiles ya que agilizan el sistema de producción haciéndolo más flexible y adaptable para responder a los cambios del entorno.
Alerta de paradas no planificadas	Las paradas de máquinas no planificadas en una planta industrial suelen ser causales de grandes costos de producción, entre los que se puede mencionar el costo de retrabajo y el costo de descartar productos semielaborados, además de causar desconciertos en los agentes de la cadena de suministro. Un chatbot puede ayudar aquí informando a todos

	los agentes de la cadena de suministro el problema ocurrido y el nuevo programa de producción que pretende solucionar el problema ocasionado por la parada no planificada.
Información sobre el estado del pedido	Toda empresa debe tratar con las fechas de entregas de los productos. Un chatbot puede notificar e informar a todos, a los clientes y a los operarios sobre la actualización del estado del pedido en planta. Esto puede producir un ahorro de tiempo y costos de gestión tanto de los clientes como del equipo de atención al cliente.
Recomendación de programas de producción.	Una nueva forma de agregar valor para una empresa, puede ser aprovechar el histórico de los programas de producción para resolver nuevas situaciones que se presentan en una planta a la hora de decidir por alternativas de programas de producción. Un chatbot más los algoritmos de aprendizaje de máquina, pueden interactuar con el motor de scheduling para alcanzar juntos el mejor programa de producción que se adapte a una determinada situación en la planta.
Reducción de la complejidad de interfaces gráficas	La interacción humano computador comenzó con la línea de comando, luego surgieron las interfaces gráfica de usuario, las interfaces de voz, y estos últimos años, gracias al avance de la Inteligencia Artificial, los chatbot son el próximo paso en interfaces. Las interfaces de los sistemas de programación de la producción pueden reducir su complejidad al incorporar chatbot como asistentes de configuración y uso del sistema.

4. Marco para desarrollar sistemas de programación de la producción que permite integrar la tecnología de chatbots

En la figura 1 se presenta la arquitectura de un marco de trabajo para desarrollar sistemas de scheduling que incorporen chatbots. El marco está constituido por tres partes: El cliente, el servidor y los servicios en la Nube. En el cliente están los módulos de la interfaz de usuario, y el módulo cliente para la colaboración en tiempo real que permiten crear interfaces reactivas y la interfaz del chatbot. Del lado del servidor, el módulo de gestión de la colaboración contiene mecanismos de reactividad que permiten la colaboración en tiempo real y, en una escala más alta de tecnificación, permiten gestionar la programación de la producción en sistemas ciberfísicos [2]. El módulo Gestión de Solicitudes del chatbot es el encargado de realizar la integración entre el sistema y la plataforma en la nube de los servicios de IA para chatbot. El módulo de Gestión de la Seguridad autentifica al usuario y autoriza a los demás módulos a interactuar con solicitudes que provienen del cliente. El módulo motor de scheduling es un optimizador basado en modelos, su tarea es gestionar un modelo y los datos que representan un determinado problema de scheduling. Así, una vez que el usuario ha terminado de configurar el problema de scheduling en el cliente lo remite al servidor. En el servidor, el requerimiento pasa por los distintos módulos hasta alcanzar el motor de scheduling, este recibe el problema, y realiza un análisis del mismo llamando al módulo de Gestión del Almacén de datos, Soluciones y Configuración para determinar si el problema ha sido resuelto anteriormente, si es así, el modulo encargado de la gestión del almacén remite la solución correspondiente. Si el gestor del almacén responde que no existe solución almacenada para el problema, entonces el motor de scheduling llama al servicio de optimización en la nube para que resuelva el problema. El servicio de optimización en la nube remite la solución al motor de scheduling, este realiza los procesamientos necesarios y remite su salida al módulo de Visualización y Análisis, el cual formatea la salida y lo remite a la interfaz de usuario para que presente la solución.

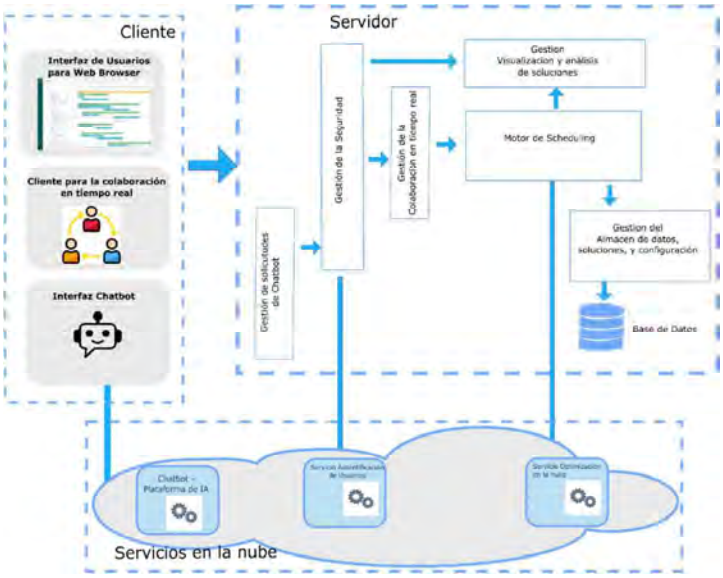


Figura 1. Arquitectura del marco para desarrollar sistemas de programación de la producción que permite integrar la tecnología de chatbots

4.1 Integración del chatbot

La integración del chatbot contiene dos grandes fases, la primera es la configuración del chatbot en el servicio seleccionado que será invocado desde el cliente, y la segunda fase atañe a la construcción del módulo gestión de solicitudes del chatbot, que se aloja en el servidor.

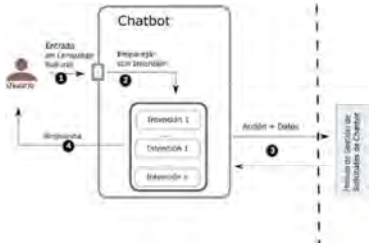


Figura 2. Esquema de trabajo del chatbot ante solicitudes al sistema de programación de la producción

En la figura **2** se describe el esquema de trabajo del chatbot. Ante una petición del usuario (1), el chatbot busca la intención que mejor responde a dicha petición. Luego (2), si la petición requiere hacer una solicitud al sistema, una acción y los datos capturados por la intención son enviados al módulo gestión de solicitudes del chatbot

para que el mismo resuelva y retorne la información al chatbot (3). El modulo gestión de solicitudes de chatbot, que está en el servidor transforma los requerimientos del chatbot en problemas de scheduling, es decir coloca una solicitud para resolver un problema y esta sigue el proceso descripto anteriormente, hasta obtener una solución al problema, es decir una respuesta, luego procesa esta respuesta y la remite al chatbot. Una vez recibida la información, el chabot contesta al usuario (4).

5. Conclusiones y trabajos futuros

En [1] se puede observar que en el entorno de los sistemas ciberfísicos de producción existe una gran penetración de las tecnologías de la industria 4.0., no ocurre lo mismo en empresas con una menor capacidad tecnológica. El marco presentado tiene por objeto aportar a reducir esta brecha y favorecer el acceso a una de las más importantes y prometedoras tecnologías de la industria 4.0.
Como trabajos futuros, se pueden mencionar que se está trabajando en la validación del marco propuesto, la que consiste en construir un pequeño sistema de scheduling que incorpore un chatbot, que en una primera etapa de soporte a los usos menos complejos del chatbot tales como dar información sobre el estado del pedido y permita una la reducción de la complejidad de interfaces gráficas.

Referencias

[1] M. Parente, G. Figueira, P. Amorim, and A. Marques, "Production scheduling in the context of Industry 4.0: review and trends," *International Journal of Production Research,* vol. 58, pp. 5401-5431, 2020/09/01 2020.

[2] D. A. Rossit, F. Tohmé, and M. Frutos, "Industry 4.0: Smart Scheduling," *International Journal of Production Research,* pp. 1-12, 2018.

[3] G. Guizzi, S. Vespoli, and S. Santini, "On The Architecture Scheduling Problem Of Industry 4.0," in *CIISE*, 2017, pp. 94-100.

[4] K. R. Barker, *Elements of sequencing and scheduling*. New York: John Wiley and Sons, 1974.

[5] B. P.-C. Yen and M. Pinedo, "On the design and development of scheduling systems," in *Fourth International Conference on Computer Integrated Manufacturing and Automation Technology*, 1994, pp. 197 - 204.

[6] M. McTear, Z. Callejas, and D. Griol, *The Conversational Interface: Talking to Smart Devices*: Springer International Publishing, 2016.

[7] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in *Artificial Intelligence Applications and Innovations*, Cham, 2020, pp. 373-383.

[8] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: A systematic review," *Expert Systems with Applications,* vol. 184, p. 115461, 2021/12/01/ 2021.

XIII Workshop Procesamiento de Señales y Sistemas de Tiempo Real **(WPSTR)**

Coordinadores

Horacio Villagarcia Wanza (UNLP)

Emanuel Frati (UNdeC)

Jorge Ierache (UM)

Diseño de un oxímetro de pulso. Prototipo de pruebas.

Lucas Barrera[1], Matías Rodríguez[1], Román Bond[1], Martin Morales[1,2], Diego Encinas[1,3]

[1]SimHPC-TICAPPS. Universidad Nacional Arturo Jauretche. Florencio Varela, 1888, Argentina.
[2]Centro CodApli. FRLP. Universidad Tecnológica Nacional. La Plata, 1900, Argentina.
[3]Instituto de Investigación en Informática (III-LIDI). Facultad de Informática, Universidad Nacional de La Plata - Centro Asociado CIC. La Plata, 1900, Argentina.

lucasxoom13@gmail.com, matydarkar@gmail.com,
{rbond,martin.morales,dencinas}@unaj.edu.ar

Resumen. Ante el inminente cambio de vida acontecido por la pandemia de COVID-19, se intensificaron los cuidados para aquellas personas más vulnerables a dicho virus, principalmente personas de edad avanzada o con problemas en el sistema respiratorio. Dadas las circunstancias, se hace imperioso el poder realizar controles a distancia sobre aquellos que puedan llegar a ser perjudicados en caso de contraer la enfermedad. Es por ello que se propone abordar el desarrollo de un oxímetro de pulso de bajo costo, integrando la placa MAX30102, una placa Arduino UNO y un visor OLED SS1306.

Palabras clave: Oxímetro, Arduino, Coronavirus, Covid, Max30102.

1 Introducción

Ante la situación de público conocimiento causada por el SARS-COV2 (COVID-19), y sumándose la apremiante disposición de los recursos económicos con los que cuentan los centros de salud. La medición y control de la saturación del oxígeno en sangre del paciente, es esencial para decidir cuándo aplicar los tratamientos necesarios de oxígeno a los pacientes comprometidos.

El Coronavirus, afecta principalmente a los pulmones, por ello la medida de la saturación de oxígeno es vital para saber cuándo es necesario utilizar los recursos de oxígenos intensivos [1].

Un oxímetro de pulso es un dispositivo médico que posibilita el cálculo de la saturación de oxígeno en sangre empleando un método no invasivo para el paciente, es decir, no es necesario obtener una muestra de sangre mediante una punción [2]. Este permite demostrar de manera confiable la saturación de oxígeno capilar periférica (SpO2) presentadas en el torrente sanguíneo [3] y las pulsaciones por minuto; variables que permiten identificar qué pacientes se encuentran en situación de riesgo y que necesitan por tanto ser hospitalizados como también recibir terapia de oxígeno.

Este tipo de estudios médicos, se realizan con dispositivos cuya tecnología es de sencilla implementación, y dada esta sencillez es posible adaptar su funcionamiento para facilitar el control a distancia para varios pacientes, con dispositivos de uso común

como son los dispositivos móviles. Esto también genera un impacto en los costos empleados para el desarrollo de dicho sistema de medición, por lo que se torna aún más accesible a una mayor cantidad de personas que requieren este control. La función del presente trabajo es explicar el accionar del procedimiento antes explicado y demostrar su validez como método preventivo.
Existen diferentes trabajos relacionados, de los cuales se destacan y se toman como referencia el de Juan Fabián Ramírez Hernández titulado "Análisis, diseño e implementación de un sistema modular de registro de variables cardíacas" [4] , el de Muhibul Haque Bhuyan titulado “Design, Simulation, and Implementation of a Digital Pulse Oxygen Saturation Measurement System Using the Arduino Microcontroller” [5] y por último el trabajo titulado “Low cost Pulse Oximeter using Arduino” de Amanda Aracely Castellanos Cárcamo [6],

3 Desarrollo

Como se hizo mención anteriormente, un Oxímetro de Pulso es un dispositivo médico que posibilita el cálculo de la saturación de oxígeno en sangre empleando un método no invasivo, es decir, que no es necesario obtener una muestra de sangre mediante alguna punción.
Se considera que una lectura de oxígeno normal oscila entre el 95 y el 100 por ciento de la muestra tomada [7].
El sistema realizará una medición en busca de hemoglobina, la cual es considerada una hemoproteína cuya función es la de transportar oxígeno mediante la sangre; esta absorbe diferentes cantidades y longitudes de onda de luz según el nivel de oxígeno que esté transportando. Para obtener la SpO2, es necesario: un par de luces LED (emisor) enfocados a un FotoDiodo (receptor), los LEDs deben estar colocados en una parte translúcida del cuerpo, como puede ser un dedo de la mano, se interpreta la lectura y realizan cálculos [2].

3.1 Fase de Captura

La Espectrofotometría es una de las técnicas experimentales más utilizadas para la detección específica de moléculas. Se caracteriza por su precisión, sensibilidad y su aplicabilidad a moléculas de distinta naturaleza (contaminantes, biomoléculas, etc) y estado de agregación (sólido, líquido, gas), se emplea un espectrofotómetro, en el que se puede seleccionar la longitud de onda de la luz que pasa por una solución y medir la cantidad de luz absorbida por la misma [8]. Basados en esta técnica, se puede observar en la Figura 1 y 2 el principio básico de la obtención de valores necesarios para el cálculo de saturación de oxígeno. La hemoglobina oxigenada (HbO2) permite pasar más luz roja-660 nm y absorbe más radiación infrarroja, por otro lado, la hemoglobina desoxigenada (Hb) absorbe más luz roja y permite pasar más radiación infrarroja-880nm.

El dispositivo MAX30102 (Figura 3) posee un LED rojo, un LED infrarrojo y un fotodetector en la parte superior del encapsulado. Como cada LED emite luz sobre el dedo del paciente, el fotodetector releva las variaciones lumínicas por cambios de volumen de oxígeno en la sangre. El sensor de pulso MAX30102 es un dispositivo que integra un pulsioxímetro y monitor de frecuencia cardiaca. Cuenta con infrarrojo, detectores fotoeléctricos, dispositivos ópticos y circuitos electrónicos de baja frecuencia con supresión de luz ambiental. Es compatible con el protocolo de comunicaciones I2C para facilitar la transmisión de información a Arduino, KL25Z u otros microcontroladores de pulso y oxigenación. El chip puede apagar el módulo o entrar en modo reposo. Para tener lectura del pulso y/o ritmo cardiaco se coloca el dispositivo en los dedos, lóbulo o muñeca. El MAX30102 es un dispositivo que integra un pulsioxímetro y un monitor de frecuencia cardiaca, es la evolución del sensor MAX30100 fabricado por Maxim Integrated.

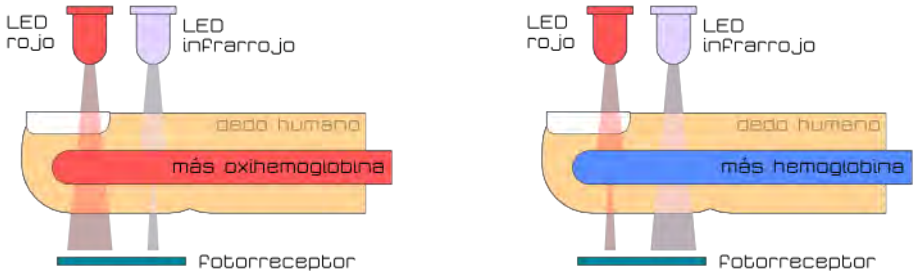


Fig. 1. Método de medición del dispositivo MAX 30102 [9]

Fig. 2. Espectro visible por el ojo humano [10]

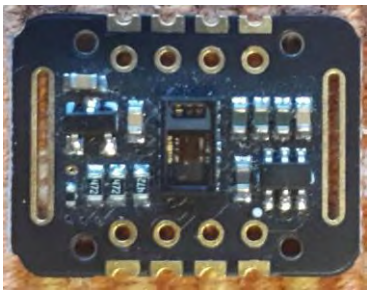


Fig. 3. Dispositivo MAX30102

3.2 Procesamiento de Datos

Arduino UNO (Figura 4) es una placa basada en el microcontrolador ATmega328P. Tiene 14 pines de entrada/salida digital (de los cuales 6 pueden ser usados con PWM), 6 entradas analógicas, un cristal de 16Mhz, conexión USB, conector jack de alimentación, terminales para conexión ICSP y un botón de reseteo. Posee toda la electrónica necesaria para que el microcontrolador opere, la energía requerida se obtiene por un puerto USB o con un transformador AC-DC.
Este dispositivo permite procesar todos los datos captados mediante el NODEMCU ESP8266, permitiendo generar códigos de manera eficaz y de rendimiento elevado al tener librerías específicas tanto para el procesamiento de datos como para la gestión a conexiones de Internet.
El display OLED SS1306 es una pantalla con una matriz de un color de 128 x 32 puntos. Debido a que la pantalla está basada en la tecnología LED, no necesita retroiluminación ya que tiene un alto contraste. El driver interno es un SSD1306 que se comunica por I2C, un protocolo rápido para este tipo de pantallas. Internamente todo el conjunto funciona a 3.3V, pero tanto la alimentación como los pines de entrada pueden trabajar a 5V, ideal para utilizarlo junto con el dispositivo MAX30102.

3.3 Uso de Datos

Para una mayor aproximación a la integración de los diferentes dispositivos, se especifica cómo se obtienen los datos de entrada para el cálculo de la oxigenación en sangre.
En cada pulsación de la sangre arterial se transmiten valores lumínicos. Considerando solo la sangre arterial, que denominaremos componente arterial (CA) y la cantidad de luz absorbida cambia de acuerdo a la cantidad de sangre, presencia de HbO2 o Hb. Llamaremos componente estático (CE) al formado por los tejidos, huesos, piel y la sangre venosa. La siguiente fórmula muestra como del cociente de la luz R(roja) e IR (infrarroja) se obtiene la SpO2:

$$\frac{(\text{CA luz R/CE luz R})}{(\text{CA luz IR/CE luz IR})} = \text{SpO}_2 \quad (1)$$

La medición de los cambios en la absorción de la luz permite estimar la saturación de oxígeno arterial y la frecuencia cardiaca. La SpO2 mostrada en la pantalla representa la media de la medición de los últimos segundos, los datos se actualizan cada 0.5 a 1 segundos.

Cálculo HR:
Calcula la media de las muestras IR (infrarrojo). Resta la media e invierte los valores. Calcula promedios de 4 posiciones. Calcula el límite promedio obtenido. Detecta los valles y picos como se visualiza en la figura 4. Si identifica dos o más picos, calcula la frecuencia cardiaca. Caso contrario, informa muestra como inválida. Obtiene un promedio de los picos. Dependiendo de la frecuencia de lectura, calcula la cantidad de pulsos por minuto.

Cálculo de SP02:
Primero, obtiene las muestras del led rojo (Y) y del led infrarrojo (X), de manera independiente. A continuación, obtiene el mínimo cerca de los valles y busca máximos entre 2 valles mientras elimina el ruido aportado por valores CE de cada tipo de lectura. Con el valor obtenido de Y calcula el numerador llamado n_nume, de manera similar, con el valor de X, obtiene el denominador n_denom. Luego, calcula la proporción utilizando la fórmula 2 . Como la señal puede variar entre latido y latido, se elige un valor medio de la proporción calculada. Finalmente, utilizando la media de la proporción obtiene una correspondencia en tabla de referencia llamada n_spo2_calc en el algoritmo de la librería del componente max30102.

$$\frac{n\_nume * 100}{denominador} \quad (2)$$

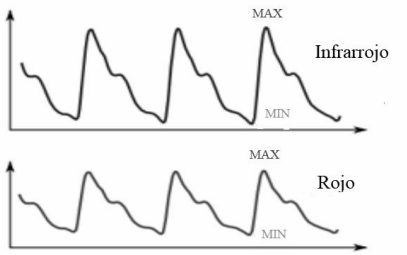


Fig. 4. Valles y Picos en la lectura de los leds rojo o infrarrojo [11]

2.4 Circuito y Pseudocódigo Implementado

A continuación, en la figura 5, se visualiza el diagrama del circuito implementado.

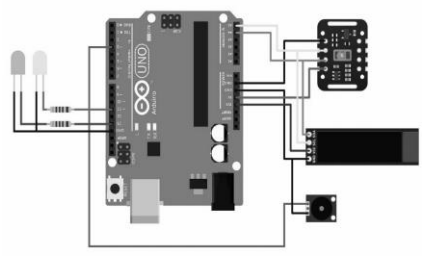


Fig. 5. Circuito del Oxímetro de Pulso

Pseudocódigo desarrollado.

1. Incluye librerías de display, max30102, cálculo de frecuencia cardiaca y SPO2.
2. Define estructuras de datos necesarias (buffers de lecturas de leds rojo e infrarrojo, SPO2,heartRate, flags de validación, asignación de puertos arduino).
3. Setup: imprime mensajes de inicio en display, configura sensor max30102, inicia leds de estado y buzzer.
4. Bucle de captura y procesamiento continuo (informando en el display.serial, led y buzzer los cambio de estado):
   a. Reconoce la presencia del dedo y resguarda hasta 4 lecturas por segundo. Descarta valores con el objetivo de normalizar lecturas.
   b. Si recaba suficientes lecturas válidas, informa los valores de SPO2 y HR.
   c. Mientras el sensor no reconozca la presencia, se mantiene en estado de espera hasta retomar el estado del punto a.

3.5 Análisis de Errores

Es conocido que existen distintos factores que pueden afectar la medición de oximetría, solo por nombrar alguna de ellas: anemia, movimiento durante la medición, sitio de colocación del sensor, esmalte de uñas, pigmentación de la piel, anestesia residual, entre otras.

La precisión y exactitud dependen de las diferentes marcas y estudios realizados que van desde más o menos 10% a menos de 2% en sujetos con saturaciones de oxígeno por encima de 70%. El funcionamiento de los oxímetros disminuye su precisión cuando las SpO2 están por debajo de 70% , lo cual llevaría a serias dudas de su interpretación en pacientes muy hipoxemicos.

Nos proponemos evaluar la concordancia entre dos métodos cuantitativos de medición. Al comparar un nuevo método de medición con un método estándar, una de las

propuestas en este trabajo es saber si la diferencia entre las mediciones de los dos métodos está relacionada con la magnitud de la medición.
Para nuestro análisis de errores se suma un oxímetro comercial aprobado por la ANMAT. La Administración Nacional de Medicamentos, Alimentos y Tecnología Médica es un organismo que se encuentra dentro del ámbito del Ministerio de Salud de la Nación [12]. Es autárquico, con jurisdicción en todo el territorio de la Nación. Entre sus funciones, rescatamos:
"*El control y fiscalización sobre la sanidad y calidad de las drogas, productos químicos, reactivos, formas farmacéuticas, medicamentos, elementos de diagnóstico, materiales y tecnología biomédicos y todo otro producto de uso y aplicación en la medicina humana.*"
En la figura 7 se muestra una porción del comunicado donde se aprueba el uso del dispositivo (oxímetro comercial) a utilizar para contrastar los resultados contra nuestro dispositivo:

"2020 - AÑO DEL GENERAL MANUEL BELGRANO"

Ministerio de Salud
Secretaria de Calidad en Salud
A.N.M.A.T.

ANEXO II

DECLARACIÓN DE CONFORMIDAD DE MODIFICACIÓN – PM CLASE I- II

2596-4

Nombre Descriptivo del producto:

OXIMETRO DE PULSO

Código de identificación y nombre técnico UMDNS:
Xuzhou Yongkang Electronic Science Technology / YONKER / FEMMTO

Modelos (en caso de clase II y equipos):

YK-80A, YK-80B, YK-80C, YK-81A, YK-81B, YK-81C, YK-82, K1, YK-820A, YK-820B, YK-820C

Fig. 7. Declaración ANMAT [13]

En la Figura 8, el dispositivo comercial que seleccionamos para el análisis de errores.

Fig. 8. Oxímetro de pulso Yonker YK-81A

A continuación, podemos ver una comparación que hicimos con un oxímetro comercial de mediciones de HR y SP02:

Tabla 1. Mediciones en paralelo.

Prototipo		Oxímetro Comercial		HR		SpO2	
HR	SpO2	HR	SpO2	Promedio	Desviación	Promedio	Desviación
93	100	59	95	76	34	97.5	5
187	98	58	94	122.5	129	96	4
68	99	62	96	65	6	97.5	3
125	99	86	93	105.5	39	96	6
187	96	62	96	124.5	133	96	0
187	96	59	*94*	*123*	130	95	2

Media de HR: 102,75
Desviación estándar HR: 78,5
Medias SP02: 96,3
Desviación estándar SP02: 3,3

4 Conclusiones

El presente proyecto permitió abordar distintas problemáticas tanto en lo que respecta al uso y configuración de hardware, como en software. El sistema al estar integrado por dispositivos que son de código abierto, posibilitó que muchas dudas y contrariedades fueran resueltas mediante la lectura e investigación en distintos foros; es por ello que se aconseja seguir empleando este tipo de dispositivos a futuro, ya que posibilitan al desarrollador sortear los obstáculos que se presentan a la hora de implementar mejoras de una forma más segura, y rápida.
Si bien el prototipo presentado realiza los cálculos correctos, no se puede garantizar que el rendimiento del sistema se mantenga a lo largo del tiempo. Para ello se recomienda establecer una simulación para poder probar el sistema en su completitud continuamente.
Por último, no se puede asegurar que las mediciones de HR sean confiables debido a la desviación estándar tan grande que hay entre el oxímetro comercial y el desarrollado. En cuanto a las mediciones de SP02 podemos garantizar que son confiables ya que la desviación resultó ser muy leve.

Referencias

1. Shi, Y., Wang, G., Cai, X. P., Deng, J. W., Zheng, L., Zhu, H. H., ... & Chen, Z. An overview of COVID-19. Journal of Zhejiang University-SCIENCE B, 21(5), 343-360. (2020).
2. Tremper, K. K. Pulse oximetry. Chest, 95(4), 713-715. (1989).
3. Basaranoglu, G., Bakan, M., Umutoglu, T., Zengin, S. U., Idin, K., & Salihoglu, Z. Comparison of SpO2 values from different fingers of the hands. Springerplus, 4(1), 1-3. (2015).
4. Juan Fabián Ramírez Hernández, Análisis, diseño e implementación de un sistema modular de registro de variables cardíacas. Universidad Nacional Autónoma de México.

5. Bhuyan, M. H., & Sarder, M. R.. Design, Simulation, and Implementation of a Digital Pulse Oxygen Saturation Measurement System Using the Arduino Microcontroller. International Journal of Biomedical and Biological Engineering, 15(2), 105-111. (2021).
6. Cárcamo, A. A. C., Reyes, M. G. M., & Urbina, S. M. S. Low cost Pulse Oximeter using Arduino. In 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON) (pp. 1-6). IEEE. (2019).
7. Franck, A. (2021). Covid-19: ¿Cómo saber si tengo una adecuada saturación de oxígeno en sangre?. 2022, Junio 22, de Noticias UNSL. Sitio web: http://www.noticias.unsl.edu.ar/05/05/2021/covid-19-como-saber-si-tengo-una-adecuada-saturacion-de-oxigeno-en-sangre/
8. Díaz, N. A., Ruiz, J. A. B., Reyes, E. F., Cejudo, A. G., Novo, J. J., Peinado, J. P., ... & Fiñana, I. T. Espectrofometría: Espectros de absorción y cuantificación colorimétrica de biomoléculas. Universidad de Córdoba, 1-8. (2010).
9. https://polaridad.es/monitorizacion-sensor-pulso-oximetro-frecuencia-cardiaca/
10. https://culturacientifica.com/2016/08/16/el-espectro-electromagnetico/
11. Mazón, A., Rojas, S., Sánchez, E., Ramírez, G., & Cabrera, A. Oxímetro de pulso para monitoreo no invasivo aplicado en el monitoreo atlético. In Memorias del VII Congreso Nacional de Tecnología Aplicada a Ciencias de la Salud. (2016)
12. Administración Nacional de Medicamentos, Alimentos y Tecnología Médica. https://www.argentina.gob.ar/anmat
13. Declaración de conformidad o PM 2596-4 ANMAT. https://helena.anmat.gob.ar/uploads/pdfs/dc_22559_30714943533_10782.pdf

Versión del Sistema Operativo XINU para la Arquitectura AVR con la Finalidad de ser Utilizado como RTOS Académico

Rafael Ignacio Zurita*, Candelaria Alvarez,
Miriam Lechner, and Alejandro Mora

Departamento de Ingeniería de Computadoras, Facultad de Informática,
Universidad Nacional del Comahue, Neuquen, Argentina
{rafa,candelaria.alvarez,mtl,alejandro.mora}@fi.uncoma.edu.ar
http://www.se.fi.uncoma.edu.ar/

Resumen Xinu es un sistema operativo académico desarrollado originalmente por Douglas Comer en la Universidad de Purdue, a fines de los 70. Desde entonces, ha sido portado a una gran variedad de plataformas de hardware y arquitecturas, y es utilizado, principalmente, como herramienta de investigación y educación. Xinu utiliza primitivas sencillas para proporcionar varios de los componentes y funcionalidades que existen en muchos sistemas operativos convencionales. Actualmente existen versiones de Xinu para x86 (PC), ARM, MIPS, y maquinas virtuales. Sin embargo, la mayoría de las versiones existentes se ejecutan en microprocesadores capaces de ejecutar sistemas operativos mas completos como Linux o Windows. En este artículo se presenta el trabajo realizado para lograr una implementación de XINU para un microcontrolador, de arquitectura Harvard, con sólo 32KB de memoria flash, y 2KB de RAM. Se evaluó el resultado portando un shell con varias utilidades de tipo UNIX, y también con un trabajo experimental de tesis de grado en donde se requirió el uso de un sistema operativo de tiempo real. Los resultados muestran que XINU tiene potencial para ser portado a otras familias de microcontroladores con pocos recursos, y ser utilizado también como RTOS académico en esas plataformas.

Keywords: Sistema Operativo de Tiempo Real, RTOS, Sistema Embebido, Sistema Operativo, Xinu, AVR, atmega328p, Arduino, microcontrolador

1. Introducción

Regionalmente, en Argentina, la enseñanza universitaria de programación de sistemas embebidos de tiempo real se realiza utilizando como herramienta de apoyo práctico algún sistema operativo de tiempo real (RTOS) [1]. Estos, se ejecutan principalmente en microprocesadores de baja complejidad, y mayormente, en microcontroladores.

*Agradecemos al profesor retirado Ing. Rodolfo del Castillo, quien nos presentó a Xinu como opción académica de diseño elegante.

Los sistemas operativos de tiempo real (RTOS) no son sistemas operativos completos. Están compuestos, básicamente, por un gestor de tareas apropiativo (preemptive), y de mecanismos para la sincronización y comunicación entre las mismas [2]. Su principal característica es que permite la ejecución concurrente de tareas o procesos de manera predecible, por lo que, usualmente, el planificador de CPU funciona en base a prioridades, y de manera round-robin en caso de que las prioridades sean las mismas.

Este reducido conjunto de características está directamente relacionado con el hardware donde se ejecuta. Los RTOS son utilizados mayormente en plataformas de cómputo basado en microcontroladores de bajos recursos para automatización y control. En los últimos años, con el crecimiento exponencial de desarrollo de dispositivos IOT, su uso ha crecido en ambientes de sensores conectados a Internet, también controlados generalmente por microcontroladores [3].

Desafortunadamente, no existe una gran variedad de RTOS académicos. Generalmente, en cursos especializados y en materias de grado relativas a sistemas embebidos de tiempo real, se emplea como herramienta práctica, algún RTOS de uso profesional, el cual suele estar diseñado y desarrollado para uso industrial. Algunos ejemplos de estas herramientas son FreeRTOS, ChibiOS/RT, Cesium RTOS, o VxWorks [4], [5], [6] y [7]. Los primeros dos son sistemas open source, los últimos son privativos. Todos estos RTOS proveen una API orientada al programador profesional de estos sistemas, y no al estudiante universitario de grado. Además, su estructura interna no suele ser fácil de comprender durante sólo un cuatrimestre (duración promedio de materias de grado universitaria). Esto limita el recorrido académico, y el RTOS seleccionado por el docente se acota a ser utilizado únicamente como herramienta de apoyo, sin tener la posibilidad de explorar otros caminos, como lo son el análisis de sus estructuras de datos y algoritmos internos, o la expansión y/o modificación de sus funcionalidades y servicios; requisitos necesarios usualmente en investigación. Por tal motivo se propone en este trabajo la modificación de un sistema operativo académico, para ser utilizado en ambientes de microcontroladores, y también como sistema operativo de tiempo real. Se seleccionó el sistema operativo Xinu, debido a que su estructura interna es elegante, su API es sencilla, y ya cuenta con un planificador de CPU apto para ser utilizado como RTOS. Como plataforma de hardware destino se seleccionó el microcontrolador AVR atmega328p, ya que es el microcontrolador original de las placas de desarrollo Arduino, muy disponible en el mercado regional.

2. Antecedentes

2.1. Sistemas embebidos de tiempo real y RTOS

Un sistema es de tiempo real si su correctitud está definida por la correctitud de los resultados computados, y también, por cumplir con los tiempos de respuesta definidos en sus especificaciones [8]. Si el sistema no es capaz de responder a un evento o tarea (definida en los requerimientos del sistema), en el

tiempo máximo permitido (tambien definido en las especificaciones del sistema) entonces se dice que el sistema falló, y no es de tiempo real.

Usualmente, al diseñar un sistema de tiempo real, se definen eventos y acciones a ser llevadas a cabo. Para cada uno de estos eventos se define también, con precisión, el tiempo de respuesta máximo permitido por parte del sistema. Luego, este diseño se divide usualmente en tareas independientes, que serán ejecutadas de manera concurrente, en tiempo de ejecución. Como software de apoyo, se utiliza un sistema operativo de tiempo real. Este software de apoyo es el encargado de ejecutar las tareas de manera concurrente, y también, da soporte para lograr los tiempos de respuesta para cada evento. Esto no es un proceso automático. Es decir, utilizar un sistema operativo de tiempo real no implica que el sistema logrado será de tiempo real. Para lograr el correcto funcionamiento del sistema, los desarrolladores deben especificar para cada tarea una prioridad, relacionar las tareas con los eventos, y definir como deben sincronizarse y comunicarse esas tareas. En tiempo de ejecución, el sistema operativo de tiempo real será el encargado de asegurar que siempre se está ejecutando la tarea de mayor prioridad. De esta manera, en tiempo de diseño y desarrollo, se podrá evaluar el modelo diseñado ante todas las situaciones posibles, para corroborar que siempre el tiempo de respuesta para un evento y/o tarea, es menor a la cota definida en los requerimientos. Si la cantidad de estados del sistema y situaciones posibles son inabarcables, los desarrolladores pueden realizar simulaciones y aproximaciones, para evaluar si el modelo diseñado cumplirá con los plazos requeridos.

Generalmente, el planificador de CPU (scheduller) de un sistema operativo de tiempo real es de prioridad fija (la prioridad de cada tarea se define en tiempo de compilación) y apropiativo (preemptive). Cuando el sistema está en ejecución, cada vez que el RTOS activa una tarea, esto es, coloca la tarea en estado de "listo para ejecutar", verifica su prioridad. Si esta prioridad es mas alta que la tarea que actualmente utiliza la CPU, entonces el RTOS realiza un cambio de contexto y coloca a la nueva tarea de mas alta prioridad a ejecutar. Si existen varias tareas listas con la misma prioridad que la tarea que está actualmente utilizando la CPU, entonces el RTOS realiza una expropiación de la CPU en intervalos regulares, empleando round-robin para asignarla a otra tarea.

Existen otros métodos y formas de desarrollar un sistema de tiempo real, pero el uso de un RTOS es la situación mas común. En síntesis, un RTOS debe contar con las siguientes caracterísitcas:

- Un RTOS (Real-Time Operating System) debe poder gestionar tareas diferentes (multiprogramación/multi hilos), y ser apropiativo (preemptive).
- Cada tarea debe contar con una prioridad, y el RTOS siempre ejecuta la tarea lista de mayor prioridad.
- El RTOS debe contar con mecanismos de sincronización y comunicación de tareas.
- Debe existir un sistema de herencia de prioridad.
- El RTOS debe permitir eventos asincrónicos, como pueden ser interrupciones de E/S.

Contando con estas características, los desarrolladores pueden diseñar un sistema cuyo funcionamiento es predecible, la cual es una condición necesaria para evaluar el diseño, y conocer si el sistema resultante cumple con los plazos definidos, y por lo tanto, ser de tiempo real.

2.2. Xinu

Xinu es un sistema operativo académico desarrollado originalmente por Douglas Comer en la Universidad de Purdue, a fines de los 70. Desde entonces, ha sido re-escrito y portado a una gran variedad de plataformas de hardware y arquitecturas, y es utilizado, principalmente, como herramienta de investigación y educación. Las versiones mas recientes (2015) de Xinu son para x86 (PC), ARM, MIPS, y maquinas virtuales [9]. A pesar de que Xinu comparte algunos conceptos y alguna terminología de componentes con UNIX, el diseño interno de Xinu difiere completamente de UNIX. Xinu es un sistema operativo con un diseño elegante, cuya estructura interna está conformada por una jerarquía multinivel de componentes esenciales, pero con las cuales puede luego desarrollarse otras más complejas. Tiene soporte para la creación dinámica de procesos, reserva dinámica de memoria, sistemas de archivos, soporte de red, y funciones de E/S independiente de los dispositivos. También ofrece servicios de comunicación y sincronización entre procesos. Xinu no implementa memoria virtual, ni tiene soporte para hardware de paginación. En tiempo de ejecución, la imagen de Xinu con un conjunto de aplicaciones se carga completamente en memoria RAM, utilizando un único espacio de direcciones para todos los procesos. Esto significa que los procesos comparten la memoria, aunque Xinu gestiona para cada proceso su propia pila.

A pesar de que las versiones de Xinu existentes son para procesadores de 32bits, el sistema fue igualmente seleccionado para este trabajo debido a que también es indicado en ambientes embebidos. El código fuente de Xinu es pequeño en comparación con otros sistemas operativos académicos (como por ejemplo MINIX). La versión de Xinu para x86 contiene aproximadamente 10 mil líneas de código fuente en lenguaje C. De estas, el 75 % pertenece a código fuente de drivers, soporte de red, y cabeceras .h; por lo que el kernel está desarrollado en sólo unas 2500 líneas de código en lenguaje C.

2.3. Trabajos Relacionados

En [10] y [11] se detallan el port del sistema operativo de tiempo real uC/OS-II a dos arquitecturas diferentes, ambas de 16 bits. Los trabajos describen ports que siguieron la metodología propuesta por la documentación oficial del RTOS en [13]. uC/OS-II es un RTOS open source, desarrollado por Micrium Inc. Ofrece una API orientada a la industria, y Micrium ofrece soporte comercial y licencias no libres para empresas que lo requieran. En [14] y [12] se detalla el proceso realizado para portar FreeRTOS y VxWorks a nuevas plataformas (para x86 en el caso del artículo de FreeRTOS, y MPC8313E para el caso de VxWorks). El eje de ambos trabajos fue exponer un ejemplo de como portar un sistema

operativo de tiempo real a nuevas arquitecturas, orientado a lectores que no tengan experiencia previa en este tipo de trabajo. En [15] la IEEE documenta un estándar que describe las características necesarias de un sistema operativo de tiempo real para sistemas embebidos compuestos por procesadores de 16 bits, sin MMU. El estándar reúne las características generales que pueden encontrarse de manera básica en casi todos los RTOS disponibles. En [16] se describe un modelo para el diseño de RTOS independientes del hardware. El foco está puesto en separar el código general del kernel RTOS de las características particulares del hardware. A partir de una descripción del hardware final se podrá, en tiempo de compilación, generar código específico faltante, por ejemplo, del cambio de contexto de tareas para un microcontrolador específico.

En nuestra búsqueda, no hemos encontrado trabajos recientes de cómo lograr un port de un sistema operativo académico convencional para ser ejecutado en ambientes embebidos de microcontroladores, y en caso de ser necesario, poder ser utilizado como RTOS. Sin embargo, los trabajos mencionados nos aportaron principios de diseño y metodologías, que pudimos emplear para lograr la versión de Xinu presentada en este artículo.

3. Metodologia y Desarrollo

El trabajo de transferencia de Xinu al microcontrolador AVR atmega328p se realizó utilizando, principalmente, la metodología descripta en [17].

Los pasos (etapas) de esta metodología se pueden observar en la Fig. 1 (b), y está directamente relacionada a la estructura interna de Xinu. En la Fig. 1 (a) se puede observar un diagrama de esta estructura, donde los componentes de software de Xinu están organizados en una jerarquía multinivel. Cuando un software está diseñado de esta manera, las interconexiones entre los componentes son claras, y el diseño interno es fácil de comprender. Ambas figuras fueron extraídas de [17].

No todas las etapas fueron necesarias en este trabajo, debido a que un RTOS no contiene todos los componentes de un sistema operativo tradicional. Además, del listado en la Fig. 1 (b) no fue necesario realizar los pasos 13 a 17, ya que el microcontrolador destino no cuenta con el hardware necesario para esos componentes.

Etapa 1. La primer tarea comprendió estudiar la arquitectura de hardware destino. Esta tarea tuvo una gran relevancia en este trabajo, debido a que la estructura de hardware del microcontrolador seleccionado difiere en gran medida de los procesadores donde Xinu se ejecuta. Los procesadores x86, ARM y MIPS soportados por Xinu presentan un modelo Von Neumann. Esto es, una única memoria para los programas y los datos (aún si algunos de estos contienen cachés de memoria interna en el procesador, diferentes para datos y código, ya que esta es transparente para el código binario del programa). En cambio, la arquitectura destino AVR posee una arquitectura Harvard. Esta arquitectura presenta al menos dos memorias accedidas por diferentes buses y direcciones:

(a)

Step	Description
1.	Learn about the new hardware
2.	Build cross-development tools
3.	Learn the compiler's calling conventions
4.	Build a bootstrap mechanism
5.	Devise a basic polled output function
6.	Load and run a sequential program
7.	Port and test the basic memory manager
8.	Port the context switch and process creation functions
9.	Port and test the remaining process manager functions
10.	Build an interrupt dispatcher
11.	Port and test the real-time clock functions
12.	Port and test a tty driver
13.	Port or create drivers for an Ethernet and other devices
14.	Port the network stack, including Internet protocols
15.	Port the remote disk and RAM disk modules
16.	Port the local and remote file system modules
17.	Port the Xinu shell and other applications

(b)

Figura 1. (a) Estructura interna de Xinu en jerarquía multinivel. (b) Serie de pasos (etapas) secuenciales necesarios para realizar un port de Xinu.

una memoria para los programas, y una memoria para los datos. La memoria para el código a ejecutar por la CPU, en el atmega328p, está compuesta de una memoria FLASH de 32KB. La memoria para los datos dinámicos es una pequeña memoria RAM de 2KB. Además, se cuenta con una tercera memoria de uso general para datos, no volatil, de 1KB. Esta tercera memoria presente en el AVR es de tipo EEPROM. AVR presenta además una mejora con respecto a una arquitectura Harvard teórica. Los procesadores AVR pueden almacenar en la memoria de programa (usualmente de mayor tamaño que la RAM) datos de solo lectura. Esto significa que las variables que sean constantes pueden estar almacenadas en la FLASH, y de esta manera, no ocupar espacio en RAM. Algo útil, siendo que las variables constantes no serán modificadas durante toda la vida del programa. Otra característica importante que se debió tener en cuenta es que la CPU del AVR es de 8 bits, y que las direcciones son de 16 bits. Un puntero de C es entonces de 16 bits, mientras que el dato natural con el que trabaja el procesador es de 8 bits.

Etapa 2. La segunda tarea consistió en contar con un cadena de herramientas de desarrollo para la arquitectura AVR. La computadora de desarrollo (host) seleccionada fue una PC con GNU/Linux, lo cual facilitó esta tarea en gran medida, ya que la distribución Linux utilizada trae empaquetado un compilador cruzado de C para AVR, los binutils (vinculador, ensamblador, etc) para AVR y la biblioteca de C avr-libc.

Etapas 3 a 6. La convención de llamadas utilizada fue la implementada por GCC. El mecanismo de bootstrap fue realizado utilizando el bootloader de Arduino, y la tarea de cargar y ejecutar un programa básico fue realizada compilando un programa en C, y utilizando los scripts ld del vinculador predeterminados para ese microcontrolador. De esta manera, el compilador avr-gcc junto con su ensamblador y vinculador, son capaces de generar un ejecutable que puede ser transferido a la flash del atmega328p y ser ejecutado por el bootloader de Arduino ya almacenado en la flash. La complejidad aquí radica en asignar, en tiempo de compilación, las direcciones ROM (flash) y RAM reales físicas a los distintos símbolos del programa binario (por ejemplo, el entry point del programa en C, o la dirección para el puntero de pila). En nuestro caso, esta asignación se realizó automaticamente via los scripts ld predeterminados del vinculador gcc para el microcontrolador atmega328p.

Etapa 7. La cuarta tarea consistió en adaptar las cuatro funciones básicas de gestión de memoria de Xinu, y verificarlas. Se adaptaron las funciones utilizando la asignación de direcciones RAM de la etapa anterior, la cual permite conocer los límites de la memoria física disponible. También se modificó el sistema de compilación, adaptando el archivo `compile/Makefile` para que utilice el compilador GCC para avr.

Etapas 8 y 9. Esta quinta tarea demandó la mayor cantidad de tiempo, y consistió en portar el cambio de contexto: rutina en ensamblador dependiente de la arquitectura, la cual resguarda el estado del procesador para un proceso en su pila, y carga un estado previo del proceso que continuará usando la CPU, desde la pila del proceso a reanudar. También se realizó, en esta tarea, el port de varias rutinas de creación y gestión de procesos. Una vez completada esta etapa fue posible contar con multiprogramación. Además, el mismo kernel de Xinu se convierte en un proceso, y los detalles de implementación, de la conversión de un código secuencial, en ejecución, a varios procesos, son tal vez los mas complejos de comprender. Encontramos aquí una dificultad extra de suma importancia para todo el port: la memoria RAM física disponible ya no fue suficiente para los componentes portados en esta etapa. Por tal motivo, se estudiaron alternativas, y utilizando diferentes técnicas en simultáneo, se lograron los objetivos de esta fase. A continuación se enumeran estas técnicas, las cuales pueden ser de interés para futuros ports, incluso de otros sistemas operativos a microcontroladores:

1. Utilizar opciones de optimización del compilador, para reducir el tamaño del ejecutable.
2. Reducir las estructuras de datos siendo portadas. Por ejemplo, el elemento PID de la estructura de datos PCB de la tabla de administración de procesos, era de 4 bytes de tamaño. Se utilizó un tipo de datos uint8, ya que para esta arquitectura es altamente probable que no haya nunca mas de 255 procesos activos (de hecho, no se cuenta con los recursos para que esto suceda). Con la misma metodología se estudiaron los demás elementos de cada estructura

de datos, con el fin de cambiar los tipos por tipos mas pequeños, y de esta manera, reducir sustancialmente el tamaño de las estructuras de datos en RAM, sin perder su semántica ni su funcionalidad.

3. Almacenar las constantes en la memoria de código flash (ROM). Para esta arquitectura y compilador, esta técnica es alcanzada anteponiendo las palabras reservadas `const __flash` al tipo de una variable o estructura siendo declarada tentativamente. Estas palabras reservadas le indican al compilador que la variable o estructura será de solo lectura, y que debe ser alojada en la memoria FLASH junto con el código ejecutable. Una estructura de ejemplo que fue alojada en FLASH de esta manera es la estructura `devtab[]`, la cual es un arreglo de estructuras, donde cada elemento del arreglo representa un dispositivo de E/S, y contiene además de información de administración del dispositivo, punteros a funciones que implementan las llamadas al sistema para el dispositivo en particular (open(), read(), write(), etc). Todo este arreglo de estructuras no cambia en tiempo de ejecución, por lo que pudo ser declarada como constante y ubicada en la FLASH del microcontrolador.

Aplicando estas técnicas conforme se fue portando las estructuras necesarias se logró reducir el tamaño del ejecutable a unos pocos KB. Cabe mencionar, que en un principio, Xinu requiere de varios MB de RAM cuando se utiliza en computadoras de propósito general. Finalmente, cuando se alcanzó la meta de alojar el ejecutable de Xinu en su etapa actual en FLASH, y con cierta disponibilidad de RAM, se pudo contar con un sistema Xinu multiprogramado capaz de trabajar con tareas cooperativas concurrentemente.

Etapas 10 a 12. En esta última fase de nuestro trabajo se escribieron dos drivers de dispositivos de E/S: uno para controla un temporizador (timer) de hardware del AVR, y el segundo para controlar el dispositivo UART, el cual será utilizado por Xinu como tty (CONSOLA de interfaz con el usuario). El driver del temporizador realizará interrupciones cada un milisegundo. Dependiendo del QUANTUM configurado en Xinu, cada una cierta cantidad de interrupciones el planificador de CPU de Xinu realizará un cambio de contexto, para poner en ejecución otra tarea en estado de LISTO para ejecutar.

4. Evaluación

Se realizaron tres pruebas de evaluación experimental, para verificar el funcionamiento de todos los componentes internos de esta nueva versión Xinu para AVR. Como hardware se utilizó una placa Arduino Nano, la cual presenta una interfaz USB para transferencia del firmware, y que Xinu finalmente utilizará como CONSOLA en tiempo de ejecución. Las tres pruebas fueron las siguientes:

En primer lugar se ejecutaron los programas presentados en el capítulo 2 de [17]. Estos programas muestran lo que es la ejecución concurrente de procesos, un ejemplo productor consumidor, y la sincronización entre procesos.

Luego, se realizó el port de todo el shell de Xinu, y de varias herramientas de tipo UNIX. Esto permitió utilizar la placa Arduino Nano como una antigua

estación UNIX, con un shell y varias utilidades. Esto verificó casi todos los componentes del sistema operativo portado.

Finalmente, se verificó este port de Xinu utilizándolo como RTOS:

- En dos ediciones de la materia *Programación de Sistemas Embebidos*, de la carrera Licenciatura en Ciencias de la Computación de nuestra universidad, donde se estudian conceptos de sistemas operativos de tiempo real, y de como usar un RTOS para desarrollar un sistema embebido; y
- En un trabajo experimental de tesis de grado, donde se requirió de un RTOS. El sistema controla varios sensores, entre ellos un magnetómetro y encoders ópticos, cuyos eventos asincrónicos requerían reportar el valor obtenido en tiempos límites, para que el cálculo de odometría de un robot experimental tuviese un error en los márgenes esperados.[19]

5. Conclusiones

En este artículo se presentó una descripción del trabajo realizado para lograr el port del sistema operativo académico Xinu, para ser utilizado en la arquitectura AVR (Arduino), como RTOS académico. El proceso demandó tres tareas principales. En primer lugar se estudiaron las características de los sistemas operativos de tiempo real, y la estructura interna de Xinu. Luego, se analizó la metodología empleada para realizar otros ports y se seleccionó una propuesta. Finalmente, el trabajo resultante fue evaluado a través de programas de prueba, utilizando un shell estilo UNIX, y en un trabajo experimental de tesis de grado.

Como trabajo futuro se espera evaluar su impacto académico, por ejemplo, observando el enlace entre materias en donde se pueda utilizar la misma herramienta. Una situación posible es utilizar Xinu en materias introductorias a la temática, como lo es el curso de *Sistemas Operativos*. Por lo que quedaría observar, si en materias superiores, como *Programación de Sistemas Embebidos*, de nuestras carreras, el uso de la misma herramienta tiene un mejor impacto de aceptación y agiliza recorridos académicos, utilizando esa ganancia para otros objetivos, por ejemplo, profundizando otros conceptos no vistos actualmente por falta de horas disponibles. También se propone evaluar un posible uso de Xinu en ambientes de microcontroladores, realizando nuevos ports, o promoviendo su uso en proyectos industriales.

Distribución. El código fuente del kernel Xinu elaborado en este trabajo, y aplicaciones de ejemplo, están disponible para descarga desde el sitio Web `http://se.fi.uncoma.edu.ar/xinu-avr/`

Referencias

1. Congreso Argentino de Sistemas Embebidos. Libro de artículos y reportes tecnológicos. (2020-2022). ISBN 978-987-46297-7-7, 978-987-46297-8-4, 978-987-46297-8-4.

2. Phillip A. Laplante; Seppo J. Ovaska. Fundamentals of Real-Time Systems. In Real-Time Systems Design and Analysis: Tools for the Practitioner , IEEE, 2012, pp.1-25, doi: 10.1002/9781118136607.ch1.
3. 2019 Embedded Markets Study Integrating IoT and Advanced Technology Designs, Application Development & Processing Environments. EE Times and Embedded https://www.embedded.com/wp-content/uploads/2019/11/EETimes_Embedded_2019_Embedded_Markets_Study.pdf
4. Warren Gay. Beginning STM32: Developing with FreeRTOS, libopencm3 and GCC 1st ed. ISBN 978-1484236239. Apress (2018).
5. Giovanni Di Sirio. ChibiOS/RT - The Ultimate Guide Book. 2020. https://www.chibios.org/dokuwiki/doku.php?id=chibios:documentation:books:rt:start
6. Cesium RTOS web page. https://weston-embedded.com/products/cesium
7. VxWorks web page. https://www.windriver.com/products/vxworks
8. Alan Burns, Andy Wellings. Real-Time Systems and Programming Languages: Ada, Real-Time Java and C/Real-Time POSIX (4th Edition). Addison Wesley. ISBN 978-0321417459 (2009).
9. The Xinu Web Page https://xinu.cs.purdue.edu/
10. Sobaihi, Khaled. (2007). Porting the µC-OS-II Real Time Operating System to the M16C Microcontrollers. doi:10.13140/2.1.2701.9846
11. Kolhare, Nilima. R. and Nitin I.Bhopale. Porting & Implementation of features of uC/OS II RTOS on Arm7 controller LPC 2148 with different IPC mechanisms. International journal of engineering research and technology 1 (2012).
12. Y. Zhang, F. Lu and X. Kong. VxWorks porting based on MPC8313E hardware platform. 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, 2010, pp. 246-249, doi: 10.1109/CMCE.2010.5610170.
13. Porting µC/OS-II https://micrium.atlassian.net/wiki/spaces/osiidoc/pages/163858/Porting+C+OS-II
14. H. Hsu and C. Hsueh. FreeRTOS Porting on x86 Platform. 2016 International Computer Symposium (ICS), 2016, pp. 120-123, doi: 10.1109/ICS.2016.0032.
15. IEEE Standard for a Real-Time Operating System (RTOS) for Small-Scale Embedded Systems. (n.d.). doi:10.1109/ieeestd.2018.84456
16. Gomes, R. M., & Baunach, M. (2018). A Model-Based Concept for RTOS Portability. 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA). doi:10.1109/aiccsa.2018.861286
17. D. Comer. Operating System Design - The Xinu Approach, Second Edition CRC Press, 2015. ISBN 9781498712439.
18. Rafael Zurita. Programación de Sistemas Embebidos. Materia del último año de la carrera Lic. en Ciencias de la Computación. Facultad de Informática, Universidad Nacional del Comahue. http://se.fi.uncoma.edu.ar/pse2020/
19. Candelaria Alvarez. Diseño e implementación de un sistema embebido de navegación por estima para la localización de robots móviles en ambientes de interiores, Abril 2022. Tesis de grado de la carrera Lic. en Ciencias de la Computación. Universidad Nacional del Comahue
20. Rafael Ignacio Zurita. Página web oficial del port de Xinu para arquitectura AVR. http://se.fi.uncoma.edu.ar/xinu-avr/

Diseño de un componente de comunicación para app móviles

Adriana Elizabeth Martín[1], Susana Beatriz Chavez [2], Sergio Rafael Flores[3]; A. Sara Zogbe [4],

Departamento e Instituto de Informática - F.C.E.F. y N. - U.N.S.J.
Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas",
Rivadavia, San Juan, CPA: J5402. San Juan, Tel 0264 4234129

[1]*arianamartinsj@gmail.com;* [2]*schavez@iinfo.unsj.edu.ar;* [3]*sergiorflores@gmail.com;*
[4]sarazogbe@yahoo.com.ar

Abstrac

El avance de las comunicaciones ha cambiado drásticamente la forma en que las personas y las máquinas interactúan entre sí, permitiendo el acceso instantáneo a información y servicios en tiempo real.

Se propuso un modelo de comunicación entre app móviles, que permitiera evaluar la disponibilidad de canales de comunicación, de manera de garantizar que un mensaje llegue a destino. Para lo cual se decidió trabajar sobre una plataforma con soporte a la programación reactiva. Esto obligó a analizar y entender qué propone el paradigma reactivo para que el desarrollo del software móvil sea una solución competitiva, haciendo uso de Apps reactivas.

Palabras claves: Paradigma Reactivo, Dispositivos móviles, IoT, Java, Android SO.

1. Introducción

Debido a que los procesadores multinúcleo se han convertido en un estándar, se han creado múltiples niveles de abstracción para simplificar la concurrencia y permitir un desarrollo más simple.

Para interactuar continuamente con su entorno, las apps reactivas deben adaptarse a la carga a la que se enfrentan, utilizando una mayor capacidad computacional cuando es necesario. Esto significa que debe hacer uso eficiente del hardware en un solo dispositivo (de uno o más núcleos), y pueda funcionar a través de varios nodos de cómputo a su disposición, dependiendo de la carga.

Se está desarrollando un componente de comunicación que se encarga de identificar todas las alternativas posibles de comunicación, que se encuentran disponibles en el hardware del dispositivo, con el objeto de garantizar la conectividad necesaria.

En un modelo de concurrencia, ya no hay memoria real compartida, los diferentes núcleos de un dispositivo se envían entre sí fragmentos de datos explícitamente, tal como lo hacen las computadoras conectadas en una red. Esto ocurre en lugar de ocultar el aspecto del paso del mensaje a través de variables marcadas como compartidas.

Un enfoque más disciplinado, consiste en mantener el estado local de una entidad concurrente y propagar datos o eventos a través de mensajes entre ellos.

Los protocolos de IoT, tanto para el sector Doméstico como para el Industrial, están resueltos siempre y cuando haya conexión real de internet (datos o wifi). Cuando estas conexiones no están disponibles, no es posible mantener la comunicación, solo queda esperar a que se restablezca la misma, con las consecuencias que esto pudiera acarrear. Por este motivo es que se propone el diseño de un componente (driver) que se encargue de guiar en forma automática la comunicación de acuerdo a un esquema de jerarquía y en forma transparente al usuario.

En la Figura 1 se muestra cómo funciona el sistema de comunicación en forma normal.

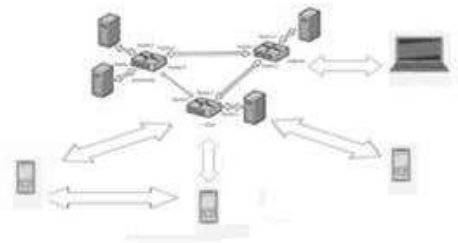


Figura 1. Comunicación normal

Cuando la comunicación se interrumpe debido a diversos factores, como por ejemplo fuera de cobertura, falta de crédito, o por cualquier otro motivo, existen otras formas de comunicación que siguen vigentes sin que hayan sido diseñadas para este propósito, SMS o llamadas de voz. Incluso, algunas compañías proveen de servicio gratuito al sistema de comunicación WhatsApp, que, si bien están destinados a un fin específico, se podrían utilizar para proveer de conectividad de manera temporal al móvil. De esa manera, un móvil puede hacer de anfitrión para proveer la comunicación, por ejemplo: como se muestra en la Figura 2 el móvil M1 tiene comunicación con un sistema que a su vez se comunica con el móvil M3. Si se rompe la comunicación de M1, el driver busca un anfitrión como por ejemplo el móvil M2 y accede a través de él al sistema

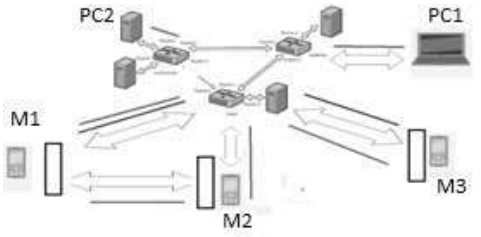


Figura 2. Comunicación a través de un anfitrión

Se está implementando un driver que se comunica directamente con el sistema operativo, tratando de discernir el mejor camino cuando se pierde la conexión natural de datos o wifi. La figura 3 muestra un esquema general de las funciones del driver.

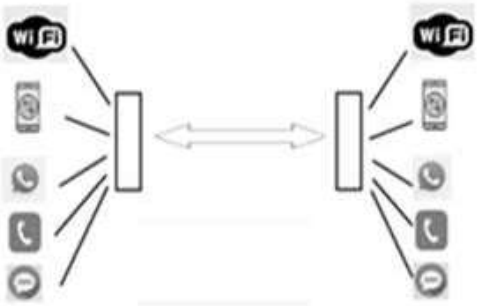


Figura 3. Funciones del Driver propuesto

2. MOTIVACIÓN

El objetivo de este trabajo es evaluar todas alternativas disponibles para manipular la comunicación entre aplicaciones que se originan en los distintos dispositivos, y construir un prototipo basado en una arquitectura reactiva donde la resiliencia y la comunicación asíncrona permitan que los sistemas estén exentos de errores, de modo que la comunicación siempre sea posible.

3. CONTEXTO

El presente trabajo se enmarca dentro del proyecto de investigación: Modelo de Sistema de Comunicación en Programación Reactiva, que ha sido aprobado por CICITCA , UNSJ. y está en desarrollo para el período 2020-2022.
Las unidades ejecutoras para dicho proyecto son el Departamento e Instituto de Informática de la FCEFyN de la UNSJ.

4. APORTES DE TRABAJO

Con esta propuesta de trabajo se espera contribuir a la profundización y consolidación del conocimiento en esta área temática por parte de cada uno de los integrantes de este proyecto

Desde el 2020 se realizaron diferentes publicaciones afines a la propuesta.

El prototipo propuesto se encuentra en un estado avanzado de desarrollo e implementación.

Se establecieron comunicaciones punto a punto entre de los distintos sistemas de comunicación sin hacer eco con el dispositivo, es decir, generando mensajes internos, enviándolos y recibiendolos en forma transparente al usuario. Evitando de esa forma la intervención del usuario en la comunicación. El entorno de desarrollo utilizado es Android Studio, con el lenguaje de programación nativo java.

Se está avanzando en la comunicación de las apps a través del canal de voz, por medio de una llamada, para lo cual se lograron las siguientes acciones:

a) Establecer la llamada con el móvil destino, es decir, se realiza la llamada y el móvil destino responde

b) Sincronizar la comunicación con el móvil destino, es decir, que sepa que no es una llamada de voz común, sino que se trata de una comunicación entre máquinas. Esto se puede realizar por

medio de una escucha de las notificaciones de llamadas, si es el número programado, se captura la llamada, y se envia un ok para sincronizar (en otra frecuencia).

c) El texto a enviar se debe modular con una frecuencia de 42KHz, luego tomar los datos, pasarlo a voz y modularlos.

d) Una vez sincronizado enviar la señal modulada.

e) En el dispositivo destino tomar la señal modulada con un formato de voz y aplicar el proceso de demodulación

f) El dato estaría disponible

La variedad de dispositivos con capacidades diferentes es enorme y está al alcance de la mano.

Resta configurar, adaptar, y poner a punto el prototipo propuesto: SiCo (Sistema de Comunicación).

De esta manera se contribuirá a la profundización y consolidación del conocimiento de esta área temática.

En cuanto a la movilidad de los dispositivos, los recursos pueden varían sin depender de los mismos, sino de los mecanismos de comunicación desde 4G pasando por GSM, solo WhatsApp, solo mensajería de texto o simplemente llamada de voz, es por ello que se necesita disponer de un mecanismo de selección automática del mejor sistema de comunicación disponible.

5. FORMACIÓN DE RECURSOS HUMANOS

El equipo de trabajo está compuesto por 4 docentes-investigadores de la línea de investigación presentada que figuran en este trabajo y Alumnos avanzados de las Carreras de Licenciatura en Sistemas de Información, Y Licenciatura en Ciencias de la Computación en estado de tesis; pertenecientes a la Universidad Nacional de San Juan.

Se está trabajando con dos tesinas en el ámbito de la programación para dispositivos móviles. Se espera iniciar otra tesina de grado en el área motivo de la presente propuesta de investigación.

6. INVESTIGACIONES FUTURAS

Por el momento la aplicación se está desarrollando bajo Android, pero se prevé a futuro implementar este componente en IOS.

7. BIBLIOGRAFÍA CONSULTADA

- Adam L. Davis: Reactive Streams in Java_ Concurrency with RxJava, Reactor, and Akka Streams-Apress (2018)
- How the Actor Model Meets the Needs of Modern, Distributed Systems: https://doc.akka.io/docs/akka/current/typed/guide/ac tors-intro.html
- José I. Rodriguez M. Tesis Doctoral "Metamodelo para la integración del Internet de las cosas y Redes sociales" Universidad Oviedo 2017
- González García Cristian, "MIDGAR: Plataforma para la generación dinámica de aplicaciones distribuidas basadas en la integración de redes de sensores y dispositivos electrónicos IoT," UNIVERSIDAD DE VIEDO, 2013.

- L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The Social Internet of Things (SIoT) – When social networks meet the Internet of Things: Concept, architecture and network characterization," Comput. Networks, vol. 56, no. 16, pp. 3594–3608, Nov. 2012.
- C. González García, B. C. Pelayo G-Bustelo, J. Pascual Espada, and G. Cueva-Fernandez, "Midgar: Generation of heterogeneous objects interconnecting applications. A Domain Specific Language proposal for Internet of Things scenarios," Comput. Networks, vol. 64, pp. 143– 158, May 2014.
- R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things," Comput. Networks, vol. 57, no. 10, pp. 2266–2279, Jul. 2013.
- J. Pascual Espada, O. Sanjuán Martínez, B. C. Pelayo GBustelo, and J. M. Cueva Lovelle, "Virtual Objects on the Internet of Things," Int. J. Interact. Multimed. Artif. Intell., vol. 1, no. 4, p. 23, 2011.
- B. Xu, L. Da Xu, H. Cai, C. Xie, J. Hu, and F. Bu, "Ubiquitous Data Accessing Method in IoT-based Information System for Emergency Medical Services," IEEE Trans. Ind. Informatics, vol. 3203, no. c, pp. 1–1, 2014. 100
- La Ciberseguridad en la Industria 4.0 https://www.incibe-cert.es/blog/ciberseguridad-industria-4-0

Colaboración ADS-B en la Predicción SSR

Oscar Bria✉[0000-0002-0001-4248] y Javier Giacomantone[0000-0001-9362-9323]

Instituto de Investigación en Informática (III-LIDI) - Facultad de Informática
Universidad Nacional de La Plata - Argentina

onb@info.unlp.edu.ar

Resumen En una estación de vigilancia del tránsito aéreo con SSR y ADS-B, la información única generada por cada sensor se puede utilizar para mejorar el rendimiento del otro. Cuando el SSR interroga Roll-Call en modo S, predice la posición de la aeronave para el siguiente giro de antena en base a estimaciones previas. Esta comunicación presenta un método de colaboración del rastreador del ADS-B con el SSR que mejora el desempeño del predictor que necesitan las interrogaciones Roll-Call.

Palabras Clave— ADS-B, SSR, Modo S, Colaboración entre Sensores

1. Introducción

El Radar de Vigilancia Secundario SSR (Secondary Surveillance Radar) y el Sistema de Vigilancia Dependiente Automática por Difusión ADS-B (Automatic Dependent Surveillance - Broadcast) son dos tecnologías cooperativas para la vigilancia del tráfico aéreo.

El SSR [1] determina en forma independiente la posición de la aeronave que, a su vez, es interrogada por el radar para recabar información complementaria. El SSR de Modo S (Selective) [2] es la segunda generación del SSR; permite interrogar en forma selectiva, interrogaciones Roll-Call, valiéndose de un código de identificación de 24 bits exlusivo de cada aeronave [3,4]. Los SSR suelen tener una tasa de actualización de la posición entre 4 y 10 segundos dependiendo de la velocidad de rotación de la antena.

El ADS-B [5] es dependiente de los Sistemas Globales de Navegación Satelital (por ejemplo, GPS) para la determinación de la posición que es emitida junto a otra información en forma expontánea por la aeronave y recibida en tierra por un sistema más simple que un radar. Los ADS-B suelen tener una tasa de actualización de la posición de alrededor de 1 segundo [6,8].

Todo sistema SSR o ADS-B suele incluir un rastreador (tracker) [7]. Un rastreador estima la posición de cada aeronave repitiendo una secuencia de dos pasos, un paso de predicción seguido de un paso de medición corregida. La medición corregida es en realidad una combinación de una medición original de un sensor con la predicción del último paso. La predicción se calcula en base a la historia cercana de mediciones corregidas. El paso de predicción es parte

fundamental en los SSR de Modo S para poder interrogar selectivamente a cada aeronave en el momento oportuno.

Hoy en día es común encontar un SSR Modo S y un ADS-B instalados en un mismo sitio, por lo tanto, con coberturas mayormente superpuestas. En lo que sigue se presenta un método de utilización del rastreador del ADS-B para mejorar la predicción que necesita el SSR de Modo S, cuando ambos sensores se ubican en un mismo sitio. Una de las motivaciones para proponer esta colaboración es que, como ya se mostró, la tasa de actualización de la posición del ADS-B es sensiblemente menor que la del SSR.

2. Método Colaborativo

El proceso de predicción de la ubicación de una aeronave bajo vigilancia para la visita de la próxima vuelta del SSR de modo S es fundamental para realizar interrogatorios selectivos. La próxima interrogación no sólo debe realizarse dentro del sector azimutal donde va a estar la aeronave, sino que también se debe predecir a qué distancia estará del radar para poder programar el intercalado de las interrogaciones del mismo período de manera eficiente.

El proceso básico de seguimiento del SSR tiene una cierta precisión como predictor que depende de la tasa de actualización de las mediciones del SSR, los errores de medición del SSR, la ubicación del objetivo en relación con el SSR, la complejidad del modelo de planta del rastreador y de la implementación numérica, y la complejidad de las maniobras que se admiten para las aeronaves.

La función de colaboración tiene como objetivo mejorar la precisión de la predicción básica del rastreador del SSR lo suficiente como para que la mejora resultante en el rendimiento del programador de interrogaciones Roll-Call sea significativa. La base lógica de la mejora es, por un lado, que dos medidas de una variable aleatoria combinadas dan una mejor estimación y, por otro lado, que la tasa de actualización de las medidas del ADS-B es generalmente mayor que la del SSR, con lo que se mejora la predicción aunque sólo se utilice ADS-B. Se debe destacar que la predicción podría mejorarse aún más si se incorporan otros datos de la aeronave.

Una mejora significativa podría ser, por ejemplo, poder predecir maniobras de aeronaves más abruptas con suficiente precisión. Maniobras que, de no ser predichas con suficiente precisión por el rastreador básico, podrían producir errores en la programación del Roll-Call con posible pérdida del rastro de estas aeronaves.

Si hay una mejora significativa, su cuantía puede depender de otros factores críticos, entre los que a priori se pueden contar: tasa de actualización relativa de los datos ADS-B con respecto a los datos SSR, calibración relativa de los datos ADS-B con respecto a los datos SSR, todas las características generales ya comentadas para el seguidor básico, pero en este caso relativas al tipo de rastreador compuesto que se implemente.

Caracterizar y ponderar todas las variables mencionadas es una tarea compleja que debe realizarse antes de cerrar el diseño preliminar de la función de colaboración. Además, la evaluación de la mejora con medidas en el radar puede

resultar muy difícil de realizar con vuelos reales *in situ*, o incluso con simuladores de carga. Por tanto, la definición de variables a almacenar para realizar estudios estadísticos y analíticos y/o simulaciones son de vital importancia no solo para el diseño sino también para la prueba de mejora. Las estadísticas deben incluir la contabilidad de fallas, que debe definirse en el diseño detallado y es una métrica de rendimiento importante cuando se busca visualizar mejoras.

A modo de ejemplo, supongamos que la composición se realiza tras las salidas de predicción de dos rastreadores en principio disjuntos: el rastreador básico del SSR [9] y un rastreador de los datos ADS-B. Se conocerán las predicciones de posición de ambos rastreadores para la próxima visita de cualquier objetivo. Estas predicciones se pueden combinar para obtener un mejor estimador que cualquier resultado por separado, si se cumplen ciertas condiciones (Figura 1). Una alternativa al esquema anterior es que los datos iniciales de ambos sensores alimenten un rastreador compuesto cuya salida es un mejor estimador que cualquier salida separada. Cualquier implementación deberá abordar los problemas del sistema de referencia común, calibración y sincronización ya mencionados. Además, la implementación debe contemplar aspectos como la ausencia momentánea o no de datos en alguno de los sensores. El diseño preliminar también debe determinar si los rastreadores se implementan para 2D o 3D.

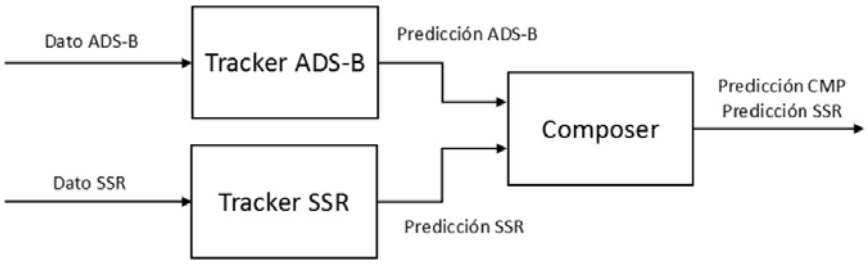


Figura 1: Diagrama de Bloques de la Colaboración.

La Figura 2 muestra las predicciones de un rastreador ADS-B (verde), un rastreaddor SSR (negro) y la composición de ambas estimaciones (azul) para una trayectoria (en rojo) con un cambio brusco de dirección. El móvil se dirige primero de Oeste a Este y luego de Sur a Norte, a velocidad constante en ambos tramos. La predicción del rastreador ADS-B es claramente mejor que la predicción del rastreador SSR durante e inmediatamente después del cambio brusco. Además, la predicción compuesta no es mejor que la predicción del rastreador ADS-B en la misma área para esta realización. Pero lo más importante es que no sabemos si todas, algunas o ninguna de las predicciones son suficientemente precisas para el programador de interrogaciones, no solo en la zona del cambio brusco sino en toda la trayectoria. Debe mencionarse que para este ejemplo se han hecho varias suposiciones y simplificaciones, entre ellas se ha asumido un

perfecto sincronismo entre los datos ADS-B y los datos SSR, lo cual no es real en la práctica; También se asumió que la frecuencia de actualización del SSR es periódica, lo cual tampoco es real. En ambos rastreadores se utilizó un filtro de Kalman [10] para modelar la trayectoria de velocidad constante con una componente de velocidad aleatoria con distribución normal con media cero [11]. En ambos casos, los valores de los parámetros del modelo son arbitrarios, elegidos solo con fines ilustrativos. Se observa que el error máximo del seguidor básico (negro) es aproximadamente igual a 1 paso a velocidad constante, para ambas coordenadas de este ejemplo 2D.

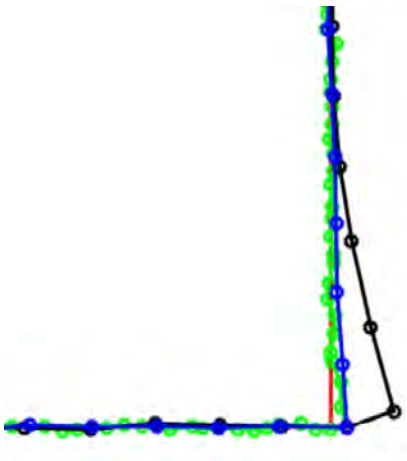


Figura 2: Ejemplo de Predicción con Colaboración.

El bloque denominado Composer en la Figura 1 calcula el promedio de los pronósticos ponderados por sus varianzas según el siguiente cálculo [12]:

$$P = \left[P_1^{-1} + P_2^{-1}\right]^{-1} \tag{1}$$

$$\hat{x} = P\left[P_1^{-1}\hat{x}_1 + P_2^{-1}\hat{x}_2\right] \tag{2}$$

Donde P_i son las covarianzas individuales de las predicciones de cada sensor y $\hat{x}_i$ son sus predicciones. Mientras que P y $\hat{x}$ son los valores respectivos de la composición.

Para un diseño preliminar, debe tenerse en cuenta que las respuestas espontáneas ADS-B y las respuestas de interrogación SSR no están sincronizadas. Además, las respuestas ADS-B ocurren en lo que puede considerarse intervalos regulares, pero las respuestas SSR no ocurren en intervalos regulares debido al movimiento relativo de la aeronave en relación con el SSR. Otro aspecto a considerar es la validación de las respuestas ADS-B.

3. Conclusiones y Trabajos Futuros

Se ha propuesto un método, se ha presentado un ejemplo conceptual simple y se han mencionado una serie de consideraciones prácticas para un futuro diseño preliminar de una funcionalidad de colaboracion ADS-B. Estas consideraciones sobre los compromisos de diseño permitirán mejorar la predicción que demanda el SSR de Modo S para las interrogaciones Roll-Call.

Deberá concretarse el diseño preliminar y evaluarse por simulación su desempeño en escenarios verosímiles que incluyan vuelos simulados en circuitos de espera (holding patterns) como los existentes en las cercanías de los aeropuertos [13]; y vuelos simulados dentro del denominado cono del silencio del SSR, donde la funcionalidad aquí presentada puede generar aportes significativos [14].

A posteriori, en la medida de lo posible, deberían implementarse pruebas sobre las instalaciones SSR y ADS-B de un sitio real, para relevar estadísticas que puedan verificar las mejoras efectivas aportadas por la funcionalidad.

Referencias

1. ICAO (International Civil Aviation Organization). Manual on the Secondary Surveillance Radar (SSR) Sytems. En: Doc 9684 AN/951 (2004)
2. U.S. Department of Transportation, Federal Aviation Adminitration Specification: Mode Select Beacon System (Mode S) Sensor. En: FAA-E-2716 (1983)
3. Stevens, M.: *Secondary Surveillance Radar.* Artech House (1988)
4. ICAO (International Civil Aviation Organization). *Annex 10, Third Edition of Volume IV.* (2014)
5. ICAO (International Civil Aviation Organization). *Technical Provisions for Mode S Services and Extended Squitter.* (2008)
6. RTCA: *Minimum Operational Performance Standards for* 1090 *MHz Extended Squitter Automatic Dependent Surveillane - Broadcast (ADS-B) and Traffic Information Services (TIS-B).* En: RTCA DO-260B (2011)
7. Sun, J.: *The 1090 Megahertz Riddle: A Guide to Decoding Mode S and ADS-B Signals. 2nd Edition.* TU Delft OPEN Publishing (2021)
8. EUROCAE (The Europena Organisation for Civil Aviation Equipment): *Technical Specification for a 1090 MHz Extended Squitter ADS-B Ground System.* En: EUROCAE ED-129B. (2016)
9. McDewitt, A.J.: A Tracker for Monopulse SSR. En: IEE Colloquium on State Estimation in Aerospace And Tracking Applications (1989)
10. Welch, G., Bishop, G.: An Introduction to the Kalman Filter. In: TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill (2006)
11. Broookner, E.: *Tracking and Kalman Filtering Made Easy*, Wiley-Interscience. (1998)
12. Bar-Shalom, Y., Rong Li, X., Kirubajaran, T.: *Estimation with Applications to Tracking and Navigation*, John Wiley & Sons, Inc. (2001)
13. FAA, Aeronautical Information Services: Aeronautical Chart Users Guide. En: aernav.faa.gov/userguide/20220714/cug-complete.pdf (2022)
14. Mariano, P., De Marco, P., Giacomini, C.: *Data Integrity Augmentation by ADS-B SSR Hybrid Techniques.* En: Integrated Communications Navigation and Surveillance Conference (2018)

Procesamiento de flujo de datos. Un caso de estudio: Análisis en tiempo real usando datos geolocalizados

Hugo Manuel Fajardo[1] and Waldo Hasperué [2]

[1] Universidad Nacional de Chilecito
[2] Instituto de Investigación en Informática (III-LIDI), Facultad de Informática, Universidad Nacional de La Plata
hfajardo@undec.edu.ar, whasperue@lidi.info.unlp.edu.ar

Abstract.
La sociedad hoy plantea crecientes demandas de soluciones informáticas, cuando estas soluciones requieren el procesamiento de grandes volúmenes de datos, las herramientas tradicionales de procesamiento muestran limitaciones e inconvenientes derivados de la cantidad de datos a procesar o del tiempo necesario para realizarlo. Surge así, la necesidad de herramientas específicas, llamadas herramientas de Big Data. Dentro de estas existe un grupo concreto para el procesamiento de flujos de datos (stream processing), entendiendo por flujo de datos la recepción y procesamiento continuo de datos ilimitados desde diferentes fuentes. Debido a su naturaleza sin límite, estos flujos no pueden descargarse de manera completa, y deben ser procesados en línea a cuando se reciben.
Dos de las principales herramientas para el procesamiento de flujos de datos son Apache Spark y Apache Flink, estas herramientas serán el objeto de estudio del presente trabajo. El caso de estudio a desarrollar tiene por finalidad comparar distintos aspectos de ambas herramientas. Como caso de estudio se propone obtener publicaciones que incluyan las expresiones *coronavirus* y/o *covid* (SARS-CoV-2), y agrupar las mismas de acuerdo a su geolocalización, ya que esto permitirá monitorear la evolución de la enfermedad de acuerdo a la localización de los usuarios y su participación en distintos lugares de la web (redes sociales, comentarios en publicaciones, etc.).

Keywords: Data Streaming, Stream Processing, Apache Spark, Apache Flink, Coronavirus, Covid19.

1 Introducción

1.1 Motivación

Existe una realidad en el mundo del procesamiento de datos y es que la cantidad de dispositivos que producen información aumenta exponencialmente, tanto en entornos personales y profesionales. Esta situación abre la puerta a una increíble explosión de creatividad e innovación en el dominio del análisis de datos en tiempo real, con la condición de que encontremos una manera de hacer que este análisis sea manejable.

1.2 Ventajas del procesamiento de flujos

El procesamiento de flujos resulta beneficioso en la mayoría de las situaciones en las que se generan datos nuevos y dinámicos de forma continua. Por lo general, las empresas comienzan desarrollando aplicaciones simples. Con el tiempo, estas aplicaciones evolucionan y se pasa al procesamiento más sofisticado en tiempo real, algunas incluyendo el uso de algoritmos de aprendizaje automático.

1.3 Herramientas de procesamiento de flujos

En el mundo del procesamiento de flujos Apache Spark [10] y Apache Flink [11] ocupan lugares centrales, tanto por funcionalidad que brindan como por preferencia de los usuarios.

Apache Spark es un motor de análisis para el procesamiento de datos a gran escala. Fue diseñado pensando en la rapidez y proporciona una API de alto nivel y un motor optimizado que admite grafos de ejecución general. Spark es una herramienta de propósito general desarrollada específicamente para el procesamiento por lotes que con el tiempo se adaptó y optimizó para el procesamiento de flujos.

Flink es un motor de procesamiento distribuido diseñado y construido específicamente para el trabajo con flujos de datos ilimitados y acotados. Proporciona un motor de transmisión de alto rendimiento y baja latencia permitiendo el procesamiento de eventos de tiempo y administración de estado.

1.4 Objetivo

El objetivo del presente trabajo es realizar una revisión de Apache Spark y Apache Flink como frameworks de procesamiento de flujos. Este trabajo forma parte de un trabajo final de la Especialización en Inteligencia de Datos orientada a Big Data de la Facultad de Informática de la UNLP.

La revisión entre los frameworks requiere el desarrollo de dos aplicaciones para el tratamiento del flujo de datos, ambas resolviendo el mismo problema. Una aplicación realizará el procesamiento del flujo de datos en Apache Spark, mientras que la otra realizará la misma tarea en Apache Flink. El problema puntual será la implementación de una técnica de minería de datos.

En 2019 la Organización Mundial de la Salud detectó un brote de enfermedad por coronavirus (SARS-CoV-2). El acelerado incremento de casos y alta propagación llevo a que se convierta en pandemia mundial, con impacto directo en la salud a nivel global. En este contexto, estados y empresas destinaron esfuerzos para desarrollar herramientas informáticas que contribuyan a controlar y hacer frente a esta tragedia humanitaria. Los desarrollos incluyen desde aplicaciones para detección de síntomas [15] [14], seguimiento de contagios, monitoreo de propagación [13] [15], control de aislamiento hasta monitoreo de circulación de personas [15].

Considerando la problemática vinculada a la pandemia por coronavirus (SARS-CoV-2), el presente trabajo se enfoca en un caso de uso concreto, actualmente en fase de desarrollo, y tiene la finalidad de filtrar publicaciones que incluyan las expresiones *coronavirus* y/o *covid*, y agrupar los mismos de acuerdo a su geolocalización, ya que

esto permitirá monitorear la evolución de la enfermedad de acuerdo a la localización de los usuarios.

2 Frameworks de Procesamiento de Flujos

2.1 Apache Spark

Apache Spark comenzó como un proyecto de investigación en la universidad de Berkeley en el año 2009 en el AMPlab. En 2013, el AMPlab contribuyó con Spark a *Apache Software Foundation*. Al año siguiente se constituyó como Top Level Project.

Apache Spark es un motor de computación unificado y un conjunto de bibliotecas para el procesamiento de datos en paralelo en clúster de computadoras basado en *Hadoop MapReduce.* Para poder llevar a cabo su propósito Spark está organizado en componentes. Spark Core es el componente principal, en él se encuentra la funcionalidad necesaria para la ejecución de trabajos y sirve de cimiento para los demás componentes. La abstracción central que proporciona el núcleo es conocida como RDD (Resilient Distributed Datasets o conjunto de datos resilientes). Spark SQL es un componente que proporciona funciones para manipular grandes conjuntos de datos distribuidos y estructurados utilizando un subconjunto de lenguaje SQL. Spark Streaming proporciona la funcionalidad necesaria para el procesamiento de flujos. GraphX es el componente para gráficos y computación paralela de gráficos en Spark.

2.2 Apache Flink

Flink se inició en Europa en 2010 como un proyecto colaborativo entre varias universidades y se convirtió en un proyecto de Apache Incubator en marzo de 2014. En diciembre de ese año fue aceptado como un Top Level Project de *Apache Software Foundation*.

Desde el punto de vista arquitectónico Flink se construye sobre un núcleo central conocido como: *Runtime Distributed Streaming Dataflow*. El núcleo reúne la funcionalidad básica, tanto para procesamiento de flujos como para procesamiento por lotes. En el núcleo se encuentran las capacidades para procesamiento distribuido, gestión de memoria y tolerancia a fallas. Sobre el núcleo se sitúan dos APIs, una para gestionar el procesamiento por lotes denominada *DataSetAPI* y otra para el procesamiento de flujos conocida como *DataStreamAPI*. Esta capa es muy importante ya que permite la interacción entre usuarios y el núcleo central. Sobre la capa de APIs se sitúan librerías específicas que permiten el trabajo con Machine Learning, procesamiento de gráficos Gelly y manejo de tablas SQL.

3 Implementación del caso de estudio

Las herramientas de procesamiento de flujos seleccionadas tienen diferentes orígenes, Apache Spark tiene su origen en el procesamiento por lotes y luego incluyó el procesamiento de flujos. Apache Flink por su parte tiene sus orígenes en el procesamiento de

flujos y posteriormente se hizo extensivo al procesamiento por lotes, esto motivó la selección de estos frameworks para la realización del presente trabajo.

La revisión de los frameworks de procesamiento de flujos no solo se enfoca en la performance de cada uno, sino que tiene por objetivo considerar otros aspectos: facilidad de instalación y de despliegue, fuentes de datos admitidas (tanto para la ingesta de datos como para la salida de los mismos), lenguajes de programación soportados, documentación disponible (tanto oficial como foros de ayuda de comunidades de usuarios).

La aplicación a desarrollar utilizará un lenguaje de programación soportado por ambas herramientas y accederá a algún medio en tiempo real para realizar la ingesta de datos. En una primera instancia se realizarán pruebas con la red social Twitter mediante su API, pero luego puede extenderse a la ingesta de otras fuentes de datos.

El procesamiento de datos tiene por finalidad filtrar las publicaciones que incluyan las expresiones *coronavirus* y/o *covid*, y agrupar los mismos de acuerdo a su geolocalización. La salida de la aplicación será un tablero donde se pueda visualizar la información agrupada por geolocalización. Este tablero busca reflejar en tiempo real el dinamismo y evolución la pandemia en diversas ciudades y/o comunidades de acuerdo a la goelocalización de las publicaciones de los usuarios.

4 Resultados Preliminares

La revisión de los frameworks considera varios aspectos, uno de ellos es facilidad de instalación y despliegue. En este apartado, ambos frameworks proporcionan la documentación necesaria para poder instalarlos en diversas plataformas.

Respecto a fuentes de datos, Spark concentra sus esfuerzos en admitir diversos formatos de archivos (orc, json, csv, text, avro), mientras que Flink se enfoca en proporcionar conectores de diversas fuentes de datos. Ambos frameworks soportan la conexión directa con Apache Kafka, un sistema de intermediación de mensajes de código abierto ampliamente utilizado. Spark permite el uso de sockets TCP, mientras que Flink no admite su uso.

En cuanto a lenguajes de programación, ambos frameworks soportan el uso de Java, Scala, SQL y Python. Spark, además soporta R. La API de lenguajes de Flink sigue aún en desarrollo, como es el caso de la API de Python, por lo que aún existe funcionalidad no soportada para Python.

En el apartado documentación disponible, ambos frameworks brindan casos de ejemplo básicos. Esto ayuda inicialmente, pero cuando la complejidad de los casos aumenta se debe recurrir a comunidades de usuarios. En estas situaciones, la comunidad de Spark proporciona más documentación, ejemplos y soporte que la comunidad de Flink.

Actualmente se está comenzando a realizar pruebas para evaluar: tiempo de procesamiento, consumo de CPU, latencia y rendimiento, de las cuales aún no hay resultados concluyentes.

References

1. Akidau T., Chernyak S., Lax R.: “Streaming Systems: The What, Where, When, and How of Large-Scale Data Processing”. O’Reilly ISBN: 978-1491983874 (2018).
2. Chambers B., Zaharia M.: “Spark. The Definitive Guide”. O’Reilly ISBN: 978-1-491-91221-8 (2018).
3. Hueske F., Kalavri V.: “Stream Processing with Apache Flink”. O’Reilly ISBN: 978-1-491-97429-2 (2019).
4. Maas G., Garillot F.: “Stream Processing with Apache Spark”. O’Reilly ISBN: 978-1-491-94424-0 (2019).
5. Miller H., Haller P., Müller N., Boullier J.: “Function Passing: A Model for Typed, Distributed Functional Programming”. ACM SIGPLAN Conference on Systems, Programming, Languages and Applications: Software for Humanity, Onward! November 2016 (2016).
6. Python Spark References https://spark.apache.org/docs/latest/api/python/index.html
7. Python Flink References https://nightlies.apache.org/flink/flink-docs-release-1.14/docs/dev/python/overview/
8. Saxena S., Gupta S.: “Practical Real-time Data Processing and Analytics: Distributed Computing and Event Processing using Apache Spark, Flink, Storm, and Kafka”. Packt Ltd ISBN: 978-1-78728-120-2 (2017).
9. Zaharia M., Chowdhury M., Franklin M.J., Shenker S., Stoica I.: “Spark: Cluster computing with working sets”. In 2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 10) (2010).
10. Apache Spark, https://spark.apache.org/ last accessed 2022/07/24.
11. Apache Flink, https://flink.apache.org/ last accessed 2022/07/28.
12. COVID-19 Ministerio de Salud. https://www.argentina.gob.ar/aplicaciones/coronavirus.
13. Coronavirus: así se usa la inteligencia artificial para rastrear la expansión de la enfermedad. https://www.iproup.com/economia-digital/11293-coronavirus-utilizan-inteligencia-artificial-para-rastrear-su-expansion.
14. Coronavirus: Datos, Salud Pública y Privacidad. https://www.udesa.edu.ar/sites/default/files/milab.pdf.
15. Crean herramientas para monitorear la propagación del coronavirus a través de celulares. https://www.argentina.gob.ar/noticias/crean-herramientas-para-monitorear-la-propagacion-del-coronavirus-traves-de-celulares.
16. En qué consiste el social listening, una herramienta poderosa para conectarse con los clientes. https://www.forbesargentina.com/innovacion/en-consiste-social-listening-una-herramienta-poderosa-conectarse-clientes-n7967.
17. Las redes sociales, la principal arma terrorista durante la pandemia de COVID-19. https://news.un.org/es/story/2020/11/1484342.
18. ¿Sirven las apps de rastreo para acorralar al coronavirus en América Latina?. https://chequeado.com/el-explicador/sirven-las-apps-de-rastreo-para-acorralar-al-coronavirus-en-america-latina/.
19. Tratamiento de datos personales ante el Coronavirus. https://www.argentina.gob.ar/noticias/tratamiento-de-datos-personales-ante-el-coronavirus.
20. What is streaming data?, https://aws.amazon.com/streaming-data/ last accessed 2022/06/15.
21. The Internals of Apache Spark, https://books.japila.pl/apache-spark-internals/overview/ last accessed 2022/06/15.

Interfaz cerebro computadora (BCI): Técnicas de Machine Learning aplicadas al análisis de actividad neurológica mediante un dispositivo de electroencefalografía (EEG)

Guillermo Eduardo De Jesus Jorge (gdejesusjorge@gmail.com),
Luis Miguel Luna (luismluna87@gmail.com), and
Director: Martin Bilbao (martinbilbao@ing.unp.edu.ar)

Universidad Nacional de la Patagonia San Juan Bosco

Abstract. En los últimos tiempos uno de los campos con mayor crecimiento a nivel mundial es el uso de Interfaces Cerebro Computadora (BCI), que permiten mediante la lectura de ondas cerebrales, tomar algún tipo de acción a partir de su análisis.
Por otro lado, existe un gran avance de las tecnologías de Inteligencia Artificial en la vida de las personas. Uno de sus campos de estudio de mayor auge es el de Machine Learning, que cuenta con grandes fuentes de información de acceso público.
El propósito de este proyecto es desarrollar un prototipo de herramienta no invasiva, mediante el uso de la tecnología EEG y BCI, que permita caracterizar la actividad eléctrica del cerebro, procesarla y convertirla en información que pueda ser interpretada por una máquina y tomar acción en base a la misma. Para lograr esto se utilizará un dispositivo de EEG comercial en conjunto con distintos modelos de Machine Learning. Con base en las métricas obtenidas, se seleccionará el modelo más apto.
Cumplir este objetivo ayuda a comprobar la viabilidad del uso de estas tecnologías para la mejora en la calidad de vida de las personas, y disponibilizar cualquier producto obtenido para el público en general.

Keywords: Electroencefalografía (EEG), Interfaces Cerebro Computadora (BCI), Inteligencia Artificial (IA), Machine Learning, Árboles de Decisión, JavaScript, Node.js, Open Source

1 Introducción

Las BCI en conjunto con la Inteligencia Artificial son una tecnología emergente, de rápido crecimiento, que proporciona un medio de comunicación entre el cerebro (que "habla" silenciosamente) y dispositivos de biomonitorización. Éstas interfaces pueden verse como una colaboración entre un dispositivo, que permite el paso de señales eléctricas desde las neuronas, con un sistema externo: una computadora, un brazo robótico, u otro tipo de actuador que ejecuta una acción específica. Es una tecnología muy poderosa que no depende de la participación de ningún músculo o vías neuronales para completar la comunicación, el comando y por tanto la acción que se desea producir.

Motivación Se pretende investigar la viabilidad del uso de la tecnología EEG y BCI para desarrollar herramientas que faciliten y mejoren la calidad de vida de las personas con capacidades motrices reducidas.

Objetivo Implementar un prototipo de herramienta no invasiva con la cual caracterizar la actividad eléctrica del cerebro, para luego procesarla y convertirla en un formato apto para visualización y análisis.

2 Solución propuesta

2.1 Arquitectura

La solución ideada para el desarrollo consta de 3 aplicaciones, la aplicación principal desarrollada en Node.js que realiza la adquisición de los datos, procesamiento y posterior predicción del movimiento en base a la información analizada. Y 2 aplicaciones soporte: un gráfico en vivo, y un laberinto.

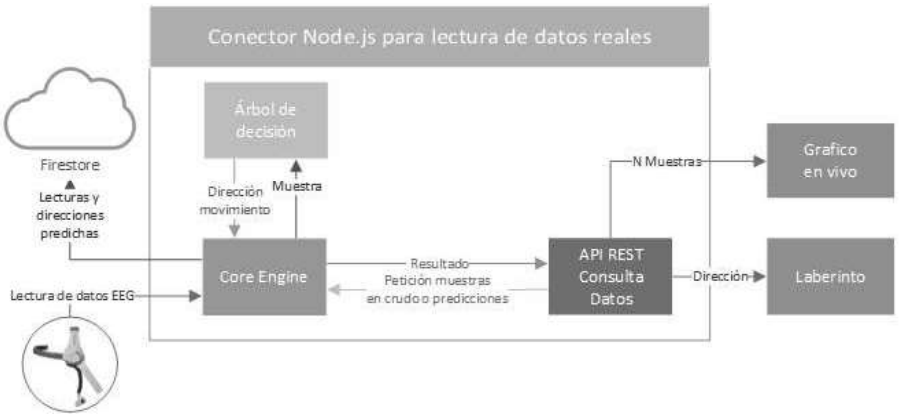


Fig. 1. Diagrama de interacción entre las distintas partes de la solución

Datos y procesamiento Se preparó un set de datos primarios mapeando los 4 movimientos básicos: arriba, abajo, derecha e izquierda. Éste conjunto de señales estocásticas contiene mucho ruido por lo que se requiere un análisis exhaustivo de cada una y las relaciones entre ellas. La imagen 2 representa el flujo de procesamiento de datos logrado.

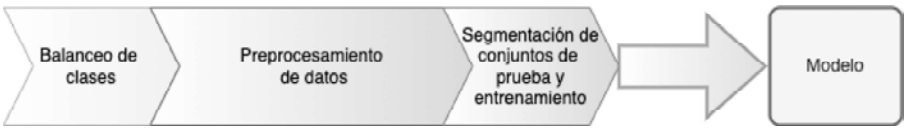


Fig. 2. Flujo de procesamiento de datos.

Análisis y predicciones En primer término esta etapa pretende servir como marco de comparación con respecto a los distintos modelos evaluados. Todas las pruebas se realizaron en iguales condiciones de uso del casco, con la misma cantidad de datos observados.

Proof. Filtrar muestras no deseadas por Q3 y desvío estándar es más eficiente que por el resto de los modelos estadísticos evaluados.

Proof. Se demostró una mejor predicción del modelo de Árbol de decisión frente a la Regresión lineal simple, y Naive Bayes.

Proof. Mejoramiento de predicción en base al aumento de datos y filtrado de valores.

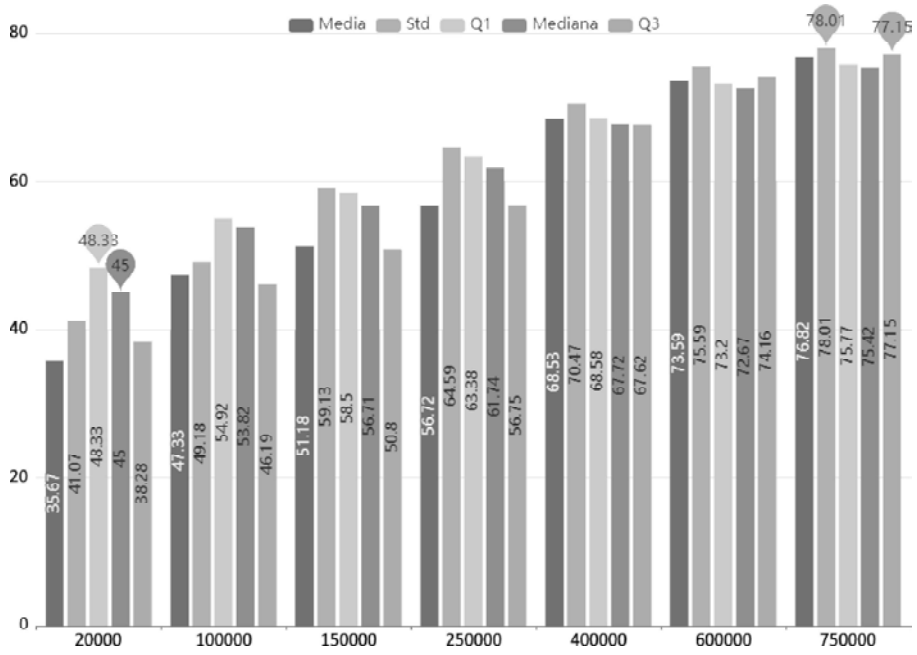


Fig. 3. Resultados filtrando y aumentando set de datos con un modelo de árbol de decisión, usando medidas estadísticas como criterio de remapeo de clases.

El codigo de este proyecto se encuentra alojado para acceso publico en un repositorio de Github.[5]
Una demostración del funcionamiento de la aplicación Laberinto desarrollada para este proyecto pueden encontrarse en YouTube. [6]

3 Conclusiones y trabajos futuros

3.1 Conclusiones

Este proyecto ha permitido implementar un prototipo que, utilizando un casco de lectura EEG comercial, logra obtener lecturas de actividad cerebral y convertirlas en un formato que permita portar la información a otras herramientas para visualización y proceso.

El uso del casco seleccionado permitió comprobar que la tecnología tiene mucho potencial para mejorar la calidad de vida humana, a pesar de que el modelo utilizado presentó inconvenientes técnicos. Se recomienda probar desarrollos similares con hardware más actualizado y mejores métodos de conexión.

Se pudo comprobar la versatilidad de JavaScript, dado que permitió desarrollar soluciones para diversos ámbitos de manera ágil. Sin embargo, en cuanto a modelos Machine Learning, el ecosistema aun no provee las facilidades que brindan otros lenguajes de programación.

Finalmente en cuanto a las herramientas desarrolladas concluimos que en las condiciones actuales no resultan utilizables en entornos reales pero el análisis y las pruebas realizadas demuestran que son una base firme para desarrollos futuros.

3.2 Trabajos futuros

En cuanto a trabajos futuros, se propone mejorar el pre-procesamiento de los set de datos, utilizar un modelo de casco más avanzado (Del mismo fabricante u otros), desarrollar una funcionalidad que permita re-entrenar el modelo de machine learning tomando muestras durante durante la utilización de la solución propuesta, aplicar la solución la solución propuesta a experimentos más avanzados (Por ejemplo complementando con robótica).

Referencias

1. Yongwook Chae, Sungho Jo, and Jaeseung Jeong. Brain-actuated humanoid robot navigation control using asynchronous brain-computer interface. In 2011 5th International IEEE/EMBS Conference on Neural Engineering, pages 519–524, 2011.
2. Luis Vergara, Yongrui Huang, Jianhao Yang, Pengkai Liao, and Jiahu Pan. Fusion of facial expressions and eeg for multimodal emotion recognition. Computational Intelligence and Neuroscience, 2017:1–8, Septiembre 2017. Article ID 2107451.
3. S.J. Russell and Norvig P.]. Artificial Intelligence. PEARSON-PRENTICE HALL, 2019.
4. George C. Canavos. Probabilidad y Estadística. McGraw-Hill/Interamericana de México, S.A. de C.V., 1988.
5. Repositorio público de proyecto en Github: https://github.com/luisluna-arg/TesinaLicInformaticaUNPJSB
6. Demostración de aplicación “Laberinto” en YouTube: https://youtu.be/joaJ0lnZQg0

XI Workshop Seguridad Informática **(WSI)**

Coordinadores

Javier Diaz (UNLP)

Hugo Ramón (UNNOBA)

Claudio Aciti (UNCPBA)

Atributos derivados para la clasificación de cadencias de tecleo en textos libres basados en el grado de desorden

Nahuel González
orcid.org/0000-0001-5570-6922

Laboratorio de Sistemas de Información Avanzados (LSIA),
Facultad de Ingeniería, Universidad de Buenos Aires,
Ciudad Autónoma de Buenos Aires, Argentina
ngonzalez@lsia.fi.uba.ar

Resumen Se proponen cuatro atributos derivados (*visibilidad*, *posición*, y *desorden*) para comparar muestras de cadencias de tecleo en textos libre, basados en el grado de desorden de los tiempos de retención y latencia, y se evalúa su rendimiento. Al verificar la identidad de los usuarios resultan en un FAR del 3,1 % y un FRR del 4,6 %, similares al de las métricas del estado del arte, y sin problemas de colinealidad. Al complementar a las anteriores, reducen el FAR y el FRR hasta el 2 %.

Keywords: seguridad informática, biometría comportamental, cadencias de tecleo, texto libre, aprendizaje automático, atributos derivados

1. Introducción

El análisis de los patrones característicos de escritura en un teclado convencional o un dispositivo móvil configura un dominio de la biometría comportamental denominado *cadencias de tecleo*. Dentro del ámbito de la seguridad informática, las cadencias de tecleo han sido ante todo empleadas como un segundo factor de autenticación (2/3FA) para la verificación de la identidad de los usuarios [1], aunque también han sido utilizadas, luego de ser extraídas en un ataque por canal lateral que revela la cadencia de tecleo pero no el texto ingresado, para reconstruir este último en base a los tiempos de escritura [2, 3] y para reducir la complejidad de un ataque por fuerza bruta [4].

Aunque el estado del arte en la verificación de usuarios por medio de cadencias de tecleo utiliza técnicas de aprendizaje profundo [5], recientemente se ha demostrado que puede mejorarse el rendimiento de clasificación en esta y otras modalidades de biometría comportamental, a la vez que se reduce el tiempo de entrenamiento y el tamaño del conjunto de datos necesario, si se complementa a las redes neuronales profundas con atributos derivados propios del dominio [6]. Más específicamente, estos han resultado efectivos para la detección de vida, con el objetivo de proteger a los sistemas de autenticación basados en cadencias de tecleo contra ataques de presentación con muestras sintetizadas [7].

En particular, la métrica R [8] ha demostrado tener un rendimiento excelente en diversos estudios comparativos [9, 10], aunque tiene la desventaja de requerir muestras grandes, del orden de 700 teclas, para alcanzar el rendimiento óptimo. La misma se basa en el grado de desorden global de los tiempos de escritura; ordena los tiempos entre teclas de las muestras a comparar y mide cuánto varían las posiciones relativas. Este idea no ha sido generalizada en la literatura del tema y, específicamente, no se ha estudiado cómo reducir la cantidad de teclas necesarias, o como aplicarla localmente a fragmentos de la muestra.

Contribuciones. El objeto de este estudio es proponer y evaluar el rendimiento de cuatro nuevos atributos basados en el grado de desorden, especializados para la verificación de cadencias de tecleo en textos libres. Las principales contribuciones ofrecidas son:

- Se proponen cuatro nuevos atributos basadas en el grado de desorden de los tiempos de escritura para comparar muestras de cadencias de tecleo.
- Se evaluó el rendimiento de los atributos propuestos para la verificación de identidad sobre un conjunto de datos realista, públicamente accesible, y que ha sido utilizado en estudios previos del tema [11].
- Se ofrece en forma abierta y gratuita el conjunto de datos de entrenamiento [11] para facilitar la verificación de los resultados.

Organización. El resto del artículo está organizado como se describe a continuación. La sección 2 reseña algunos estudios previos sobre el tema. La sección 3 describe los métodos propuestos. La sección 4 detalla la metodología del experimento. La sección 5 discute los resultados. La sección 6 resume las conclusiones.

2. Estudios previos

La métrica R de Bergadano, Gunetti, y Picardi [1] ha representado un hito en la disciplina, permitiendo por primera vez la verificación de cadencias de tecleo en el análisis de textos libres con bajas tasas de error. En el mismo artículo han propuesto también la métrica A, cuyo rendimiento no generaliza favorablemente al ser evaluada bajo otro conjunto de datos [12]. Killourhy y Maxion [9] comparan una decena de métricas y atributos, entre ellos $\mathcal{L}_1$ y $\mathcal{L}_2$, que se utilizan como parte del control en este experimento. El conteo de valores atípicos, denominado Z y también utilizado como control, fue propuesto por Haider y colaboradores [13]; aunque su rendimiento individual es pobre, suele ser útil para complementar a otros atributos en un esquema de clasificación debido a la baja correlación con ellos. Hasta donde alcanza nuestro conocimiento de la literatura del tema no se han propuesto, con la excepción de la métrica R, otros atributos basados en el grado de desorden global de los tiempos de escritura.

3. Métodos propuestos

3.1. Definiciones

Una muestra M de largo $l = l(M) = l(M')$ contiene una secuencia de teclas $k_1 \dots k_l$ y sus correspondientes tiempos $t_1 \dots t_l$, de retención (intervalo entre

eventos de presión y liberación) o latencia (intervalo entre presión de teclas sucesivas). Denominamos k_i a la i-ésima tecla de la muestra M; su *contexto de orden* m es la secuencia $k_{i-m} \dots k_i$ de teclas que la anteceden, junto con k_i. Al comparar dos muestras M y M' supondremos siempre que el largo y las secuencias de teclas de ambas son idénticas; cuando sea necesario, utilizaremos $t'_1 \dots t'_l$ para referirnos a los tiempos de la muestra M'.

3.2. V (visibilidad)

Diremos que la tecla k_i tiene *visibilidad descendente hasta n teclas atrás en la muestra M*, que denotamos con $v^-(M,i) = n$, si se cumplen simultáneamente las dos condiciones

$$(a)\ t_i \leq t_{i-n} \qquad\qquad (b)\ t_i > t_{i-j}\ \ \forall\ 0 < j < n$$

Idénticamente, diremos que la tecla k_i tiene *visibilidad ascendente hasta n teclas atrás en la muestra M*, que denotamos con $v^+(M,i) = n$, si se cumple simultáneamente que

$$(a)\ t_i \geq t_{i-n} \qquad\qquad (b)\ t_i < t_{i-j}\ \ \forall\ 0 < j < n$$

Dada una cota N, el *atributo V^- de visibilidad descendente* cuantifica la diferencia entre las muestras M y M', ambas de largo l, en la siguiente forma

$$V^-(M,M') = \frac{1}{N(l-N)} \sum_{i=N+1}^{l} |min\{N, v^-(M,i)\} - min\{N, v^-(M',i)\}| \quad (1)$$

que devuelve un valor entre 0 (cuando ambas muestras son idénticas) y 1.

Utilizando v^+ en la expresión de arriba en lugar de v^- obtenemos la expresión para V^+, el *atributo de visibilidad ascendente*. El objetivo de la cota N y los máximos dentro de la expresión es restringir la influencia de cada tecla al ámbito local y conservarla independiente del tamaño de la muestra; ya que de lo contrario, las teclas con mínimo y máximo tiempo en la muestra pueden tener una contribución exagerada si se encuentran cerca del final de la misma.

3.3. P (posición)

La *posición* de una tecla k_i en su contexto de orden m, que denotamos con $p(M,m,i)$, es el índice que le corresponde luego de ordenar las teclas del contexto en base a sus tiempos. Formalmente, el ordenar ambas filas de

$$\begin{array}{cccccc} k_{i-m} & k_{i-m+1} & \dots & k_{i-1} & k_i \\ t_{i-m} & t_{i-m+1} & \dots & t_{i-1} & t_i \end{array}$$

utilizando los valores temporales de $t_{i-m} \dots t_i$ resulta en una biyección $\sigma : [0,m] \to [i-m,i]$ tal que $t_{\sigma(0)} \leq t_{\sigma(1)} \leq \dots \leq t_{\sigma(m-1)} \leq t_{\sigma(m)}$ y

$$\begin{array}{cccccc} k_{\sigma(0)} & k_{\sigma(1)} & \dots & k_{\sigma(m-1)} & k_{\sigma(m)} \\ t_{\sigma(0)} & t_{\sigma(1)} & \dots & t_{\sigma(m-1)} & t_{\sigma(m)} \end{array}$$

Utilizando esta biyección, definimos

$$p(M, m, i) = \sigma^{-1}(i)$$

Por ejemplo, los tiempos de latencia al escribir HOLA, que es el contexto de orden 3 de la tecla A, pueden haber sido

H	O	L	A
153	278	176	190

Luego de ordenar por la segunda fila obtenemos

H	L	A	O
153	176	190	278

Y la posición de la tecla A en su contexto de orden 3, HOLA, es igual a 2 (contando desde cero) porque su valor temporal es mayor que el de la H y el de la L, pero menor que el de la O. Ahora, fijamos el orden del contexto a un determinado valor N y cuantificamos la diferencia entre las muestras M y M' utilizando el *atributo P de posición* en la siguiente forma

$$P(M, M') = \frac{1}{N(l-N)} \sum_{i=N+1}^{l} |p(M, N, i) - p(M', N, i)| \tag{2}$$

que devuelve un valor entre 0 (cuando ambas muestras son idénticas) y 1.

3.4. D (desorden)

Una vez más, consideremos la biyección $\sigma_i : [0, m] \to [i-m, i]$ que resulta de ordenar las teclas del contexto de orden m de la tecla k_i en base a los valores de tiempo $t_{i-m} \dots t_i$ en la muestra M.

$$\begin{array}{cccc} k_{\sigma(0)} & k_{\sigma(1)} & \dots & k_{\sigma(m-1)} & k_{\sigma(m)} \\ t_{\sigma(0)} & t_{\sigma(1)} & \dots & t_{\sigma(m-1)} & t_{\sigma(m)} \end{array}$$

El mismo proceso aplicado a la muestra M' resulta en otra biyección, llamémosla τ_i, que puede o no ser igual a σ_i ya que M y M' tienen igual secuencia de teclas pero no necesariamente de tiempos. Así, luego de ordenar, obtenemos

$$\begin{array}{cccc} k_{\tau(0)} & k_{\tau(1)} & \dots & k_{\tau(m-1)} & k_{\tau(m)} \\ t'_{\tau(0)} & t'_{\tau(1)} & \dots & t'_{\tau(m-1)} & t'_{\tau(m)} \end{array}$$

Ahora queremos medir el *grado de desorden local*, es decir cuánto difieren entre sí las biyecciones σ y τ para un contexto particular. El problema ha sido estudiado por Diaconis y Graham [14], que proponen diversas métricas de desorden. Entre otras, consideran

$$\begin{aligned} T(\sigma, \tau) = {} & \text{mínimo de transposiciones necesarias en la permutación} \\ & \text{que transforma } (\sigma(0), \dots, \sigma(m)) \text{ en } (\tau(0)), \dots, \tau(m)) \end{aligned}$$

Los mismos autores proveen un algoritmo de cálculo rápido basado en el número de ciclos en la permutación $\sigma^{-1}\tau$, denominado $C(\sigma^{-1}\tau)$, obteniendo

$$T(\sigma, \tau) = m + 1 - C(\sigma^{-1}\tau)$$

Fijamos el orden del contexto a un determinado valor N; el máximo de transposiciones posibles en una permutación de orden $N+1$ es N. Entonces, en base a la expresión anterior podemos definir el *atributo D de desorden local* para cuantificar la diferencia entre las muestras M y M', en la forma

$$D(M, M') = \left\lfloor \frac{N}{l} \right\rfloor \sum_{i=1}^{\lfloor l/N \rfloor} \frac{T(\sigma_{iN}, \tau_{iN})}{N} \tag{3}$$

donde $\lfloor \rfloor$ denota la función piso, σ_i es la biyección que corresponde a ordenar los tiempos del contexto de orden m de k_i en la muestra M, y τ_i en la muestra M'. Al igual que las anteriores, devuelve un valor entre 0 (para muestras que son idénticas) y 1.

4. Evaluación experimental

4.1. El conjunto de datos

La evaluación de los métodos propuestos se realizó con el conjunto de datos LSIA [15, 12], que contiene muestras de cadencias de tecleo en texto libre. Estas incluyen las secuencias de códigos de teclas presionadas, tiempos de retención (intervalo entre presión y liberación de tecla), y latencias (intervalo entre presión de teclas sucesivas). El conjunto de datos cuenta con 136 usuarios y 13.600 sesiones, de entre 150 y más de 1000 caracteres cada una. Se encuentra a disponibilidad del publico en forma abierta y gratuita [11].

4.2. Preprocesamiento y limpieza de los datos

Se utilizó una herramienta propietaria desarrollada por el LSIA para preprocesar las muestras del conjunto de datos, con el objetivo de eliminar valores inválidos o demasiado grandes que correspondan a pausas externas y no al ritmo natural de escritura. En particular, todos los valores mayores a 1500mseg. fueron marcados como inválidos, con el objetivo de que no sean incluídos en el cómputo de los atributos a evaluar.

4.3. Optimización de parámetros

Los atributos V, P, y D, descritos en la sección 3.4, requieren fijar un valor para el parámetro N, que determina el orden de los contextos considerados. No es evidente *a priori* qué valor resultaría óptimo, ameritando una búsqueda exhaustiva. Con el objetivo de optimizar el valor de N se llevó a cabo el proceso de evaluación del EER que se describirá en la siguiente sección, sobre la totalidad del conjunto de datos, pero con distintos valores de N. En particular, se evaluó $2 \leq N \leq 10$ para los atributos V^+ y V^-, $3 \leq N \leq 10$ para P, y $4 \leq N \leq 20$ para D.

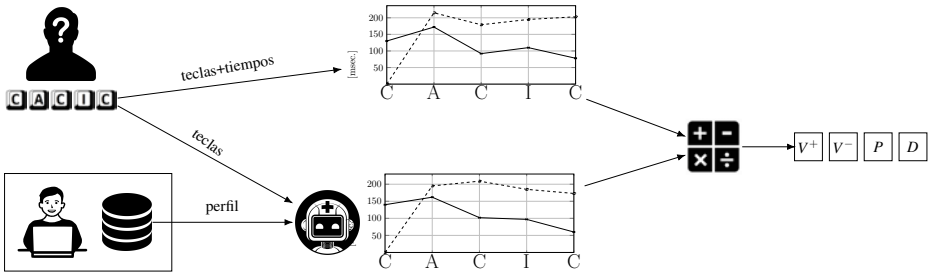


Figura 1: Esquema de evaluación de los atributos V^+, V^-, P, y D entre una muestra dada y su reconstrucción por contextos finitos

4.4. Evaluación del rendimiento individual de los atributos

En la sección 3.4 se ha indicado que los atributos V^+, V^-, P, y D sólo pueden comparar muestras con idéntica secuencia de teclas. Dado una muestra de un usuario cuya identidad se desea verificar, al tratar con texto libre es poco probable que el perfil del usuario cuente con una muestra similar. Por este motivo es necesario reconstruir cómo el usuario legítimo habría escrito el texto en cuestión. Para tal fin se utilizará el método de *modelado por contextos finitos* [15]. Este método, dada una secuencia de teclas, reconstruye la cadencia de tecleo del usuario legítimo; utiliza las muestras existentes en su perfil buscando para cada tecla observaciones anteriores de sus tiempos de retención y latencia, y prefiriendo aquellas observaciones cuyas teclas precendentes, o *contexto*, se corresponden con las de la muestra a reconstruir. Este método ha sido empleado previamente para la autenticación de usuarios por medio de su cadencia de tecleo [15], para la generación de muestras sintéticas y la detección de vida [7], y para la reconstrucción de un texto en base a los tiempos de escritura [3]. Para una descripción detallada se recomienda consultar las referencias.

La figura 1 muestra el esquema de evaluación de los atributos V^+, V^-, P, y D. Dada una muestra M de un usuario desconocido, se construye otra muestra M' con la misma secuencia de teclas, en base al perfil del usuario legítimo y utilizando contextos finitos. Como ambas muestras comparten la secuencia de teclas, pueden utilizarse las ecuaciones (1), (2), y (3) para calcular V, D, y P.

Cada usuario cuenta con 100 muestras en el conjunto de datos; de estas, 50 al azar se utilizaron como perfil para el modelado con contextos finitos. Las restantes 50 muestras de cada usuario legítimo, junto con 50 muestras adicionales de impostores tomadas al azar entre todas las muestras del resto de los usuarios, fueron evaluadas según el esquema anterior. Los 100 valores de V, D, y P resultantes fueron utilizados para trazar las curvas FAR/FRR y obtener el EER.

4.5. Evaluación del rendimiento conjunto

La utilización conjunta de diversos atributos derivados disminuye el error de clasificación si estos provéen suficiente ganancia de información y no se encuen-

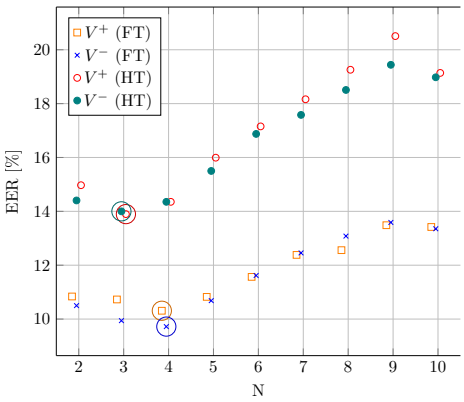


Figura 2: Rendimiento de V^+ y V^- con distintos valores de N.

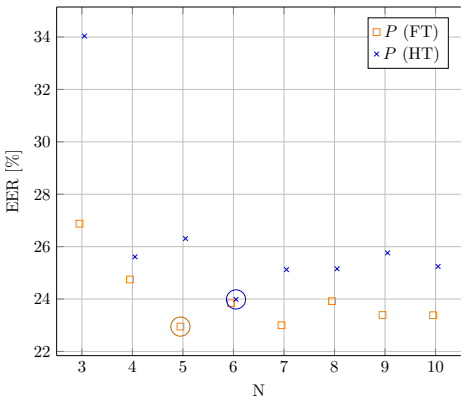


Figura 3: Rendimiento de P con distintos valores de N.

tran fuertemente correlacionados. Para evaluar el rendimiento conjunto de los atributos propuestos, se entrenó un clasificador SVM binario para verificar la identidad de los usuarios, con los parámetros que se indican en la Sección 4.6.

Se utilizaron los valores de los cuatro atributos propuestos (V^+, V^-, P, D), calculados tanto para tiempos de retención como para latencias, totalizando ocho atributos. El conjunto de entrenamiento se compuso de los valores correspondientes a las 50 muestras del perfil del usuario (ver sección anterior), marcadas como legítimas; además de 50 muestras adicionales tomadas al azar entre el resto de los usuarios, marcadas como impostores.

A modo de experimento comparativo, se evaluó idénticamente el rendimiento conjunto de cuatro métricas bien establecidas y de reconocido rendimiento: $\mathcal{L}_1$, $\mathcal{L}_2$, R, y Z, también para tiempos de retención y latencias. Finalmente, se evaluó el rendimiento conjunto de las métricas propuestas y de control, además de determinar la correlación entre todas ellas, para determinar si las primeras son útiles para reducir el error de clasificación.

4.6. Materiales y herramientas

Como se ha descrito en la Sección 4.2, se utilizó una herramienta propietaria desarrollada por el LSIA para preprocesar y limpiar las muestras del conjunto de datos. Para el modelado de cadencias de tecleo por medio de contextos se utilizó la implementación disponible en [16]. La clasificación se realizó con la implementación SVM de WEKA 3.8.4 [17], utilizando optimización secuencial mínima (SMO) [18] para el entrenamiento, núcleo polinomial, y calibrador logístico con parámetros por defecto.

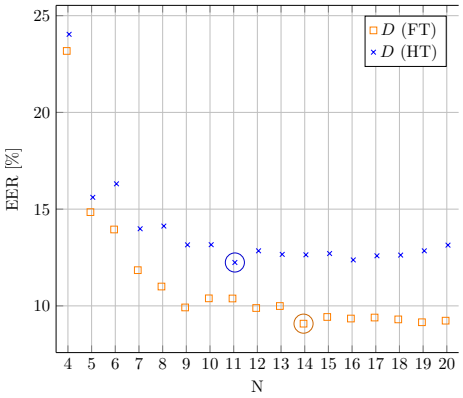


Figura 4: Rendimiento de D con distintos valores de N.

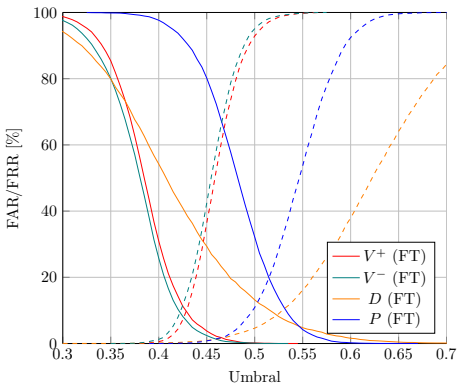


Figura 5: Curvas de FAR/FRR para latencias. Impostores en línea punteada.

5. Resultados y discusión

Los resultados de la optimización del parámetro N para los atributos V^+, V^-, P, y D se muestran en las figuras 2, 3, y 4, respectivamente. Los mínimos de EER para V^+ y V^- se obtienen con $N = 3$ para tiempos de retención y $N = 4$ para latencias. En el caso de P, los valores óptimos son $N = 6$ para tiempos de retención y $N = 5$ para latencias, mientras que para D son $N = 11$ y $N = 14$, respectivamente. Un detalle de las curvas de FAR/FRR para latencia con estos valores de N se muestran en la figura 5.

Conjunto	Atributos	FAR	FRR
Propuestas	V^+, V^-, P, D	3,1 % (± 0,7)	4,6 (± 1,5)
Control	$\mathcal{L}_1$, $\mathcal{L}_2$, R, Z	3,9 % (± 0.7)	3,9 % (± 0.4)
Todos	V^+, V^-, P, D, $\mathcal{L}_1$, $\mathcal{L}_2$, R, Z	**2,2 % (± 0.5)**	**1,9 % (± 0,6)**

Cuadro 1: Rendimiento conjunto de los atributos propuestos y las métricas de control

Con los valores óptimos de N, el rendimiento conjunto en la clasificación de los atributos propuestos se muestra en la tabla 1, junto con el rendimiento bajo el mismo esquema de las métricas de control y de ambos tipos simultáneamente, todas con intervalos de confianza del 95 %. Los valores de FAR y FRR obtenidos con los atributos propuestas muestran que las mismas son competitivas con las del estado del arte, ambas en el rango del 3–5 %; como los intervalos de confianza se solapan, la diferencia no tiene significación estadística. Lo que es más, la utilización conjunta reduce el error de clasificación hasta cerca del 2 %, aproxi-

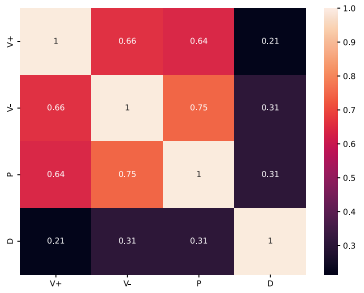


Figura 6: Matriz de correlación de V^+, V^-, P, y D para latencias.

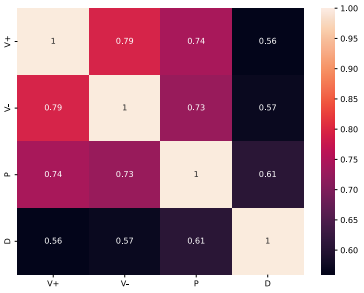


Figura 7: Matriz de correlación de V^+, V^-, P, y D para tiempos de retención.

madamente la mitad de los casos anteriores, y aquí los intervalos de confianza no se solapan por lo que sí hay significación estadística.

Las figuras 6 y 7 muestran las matrices de correlación de valores de los atributos propuestos para tiempos de retención y de latencia. En general, las correlaciones son mayores para tiempos de retención que para latencias. Como es esperable por su similitud, los atributos V^+ y V^- son las que muestran mayor correlación entre sí, pero ningún par de ellas supera el valor de $r = 0,79$. Un estudio reciente sobre los efectos negativos de la colinealidad en modelos de aprendizaje automático [19] establece un valor de corte aceptable de $r^2 = 0{,}6$ para evitar problemas, que corresponde a $r \approx 0,78$. Todos los pares de atributos, con la excepción de V^+ y V^- para tiempos de retención que se encuentra en el límite, presentan valores de correlación inferiores.

Limitaciones de este estudio. El conjunto de datos LSIA incluye sólo muestras en idioma castellano. Generalizar el rendimiento de los atributos propuestas a otros idiomas requerirá conjuntos de datos adicionales.

6. Conclusión

En el presente artículo se han propuesto cuatro atributos para comparar muestras de cadencias de tecleo en textos libre, basadas en el grado de desorden de los tiempos de retención y latencia, que exploran la idea subyacente en la métrica R [8] intentando aplicarla localmente y a muestras más reducidas. Individualmente, el atributo D para latencias con $N = 14$ alcanza el EER más bajo, de aproximadamente 10 %. Sobre el conjunto de datos utilizado para la evaluación, el empleo de los atributos propuestos para la verificación de la identidad de los usuarios resulta en un FAR del 3,1 % y un FRR del 4,6 %. En contraste, las métricas del estado del arte alcanzan un FAR y un FRR del 3,9 %, que luego de considerar intervalos de confianza resultan similares a los anteriores. Los atributos propuestos mostraron correlaciones mutuas suficientemente bajas

como para evitar los problemas acarreados por la colinealidad, y al complementar a las métricas del estado del arte redujeron el FAR y el FRR hasta casi el 2 %.

Futuras líneas de investigación. Los atributos propuestos prometen ser aplicables para la detección de vida y la discriminación de muestras sintetizadas. La exploración de estos usos queda relegado a futuras líneas de investigación.

Bibliografía

[1] Francesco Bergadano, Daniele Gunetti, and Claudia Picardi. User authentication through keystroke dynamics. *ACM Transactions on Information and System Security (TISSEC)*, 5(4): 367–397, 2002.

[2] John V Monaco. What are you searching for? a remote keylogging attack on search engine autocomplete. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 959–976, 2019.

[3] Nahuel González, Enrique P. Calot, Jorge S. Ierache, and Waldo Hasperux00E9;. The reverse problem of keystroke dynamics: Guessing typed text with keystroke timings only. In *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6, 2021. doi:10.1109/ICECET52533.2021.9698782.

[4] Dawn Xiaodong Song, David A Wagner, and Xuqing Tian. Timing analysis of keystrokes and timing attacks on ssh. In *USENIX Security Symposium*, volume 2001, 2001.

[5] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Julian Fierrez, and John V Monaco. Typenet: Scaling up keystroke biometrics. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2020.

[6] Sakorn Mekruksavanich and Anuchit Jitpattanakul. Biometric user identification based on human activity recognition using wearable sensors: An experiment using deep learning models. *Electronics*, 10(3):308, 2021.

[7] Nahuel González, Enrique P. Calot, Jorge S. Ierache, and Waldo Hasperué. Towards liveness detection in keystroke dynamics: Revealing synthetic forgeries. *Systems and Soft Computing*, 4:200037, 2022. ISSN 2772-9419. doi:https://doi.org/10.1016/j.sasc.2022.200037. URL `https://www.sciencedirect.com/science/article/pii/S2772941922000047`.

[8] Daniele Gunetti and Claudia Picardi. Keystroke analysis of free text. *ACM Transactions on Information and System Security (TISSEC)*, 8(3):312–347, 2005.

[9] Kevin S Killourhy and Roy A Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks*, pages 125–134. IEEE, 2009.

[10] Jiaju Huang, Daqing Hou, Stephanie Schuckers, Timothy Law, and Adam Sherwin. Benchmarking keystroke authentication algorithms. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017.

[11] Enrique P. Calot. Keystroke dynamics keypress latency dataset. Database, jan 2015. URL `http://lsia.fi.uba.ar/pub/papers/kd-dataset/`.

[12] Nahuel González, Enrique P Calot, and Jorge S Ierache. A replication of two free text keystroke dynamics experiments under harsher conditions. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2016.

[13] Ahmed Abbas y Abbas K. Zaidi Sajjad Haider. A multi-technique approach for user identification through keystroke dynamics. In *Systems, Man, and Cybernetics, 2000 IEEE International Conference on*, volume 2, pages 1336–1341. IEEE, 2000.

[14] Persi Diaconis and Ronald L Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):262–268, 1977.

[15] Nahuel González and Enrique P Calot. Finite context modeling of keystroke dynamics in free text. In *Biometrics Special Interest Group (BIOSIG), 2015 International Conference of the*, pages 1–5. IEEE, 2015.

[16] Nahuel González. Herramienta de análisis de cadencias de tecleo en texto libre, 7 2021. URL `https://github.com/lsia/herramientaGonzalez2021`.

[17] Ian H Witten, Eibe Frank, Mark A Hall, CJ Pal, and MINING DATA. Practical machine learning tools and techniques. In *DATA MINING*, volume 2, page 4, 2005.

[18] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[19] Ron Johnston, Kelvyn Jones, and David Manley. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of british voting behaviour. *Quality & quantity*, 52(4):1957–1976, 2018.

Syscall Top: Estrategias de monitoreo de llamadas al sistema en sistemas GNU/Linux

Fabián A. Gibellini, Sergio Quinteros, Germán N. Parisi, Milagros N. Zea Cárdenas, Leonardo Ciceri, Federico J. Bertola, Ileana M. Barrionuevo, Juliana Notreni, Analía L. Ruhl, Marcelo Auquer

Laboratorio de Sistemas / Dpto. de Ingeniería en Sistemas de Información/
Universidad Tecnológica Nacional / Facultad Regional Córdoba
Maestro M. Lopez esq. Cruz Roja Argentina S/N, Ciudad Universitaria (X5016ZAA) - Córdoba, Argentina
{fabiangibellini, ser.quinteros, germannparisi, milyzc, leonardorciceri, federicorbertola ilebarrionuevo, julinotreni, analialorenaruhl, marcelo.auquer}@gmail.com

Resumen. Los procesos que se ejecutan en un sistema GNU/Linux interactúan con el kernel por medio de llamadas al sistema, incluso los malwares. Existen distintas categorías de malware y en base a su comportamiento se puede inferir que existen patrones de llamadas al sistema, o syscalls, que permitirían descubrir qué tipo de malware se está ejecutando sobre un GNU/Linux. El presente trabajo pretende introducir formalmente la herramienta syscall top, la cual permite visualizar las llamadas al sistema interceptadas de todos los procesos y administrar reglas que permitan configurar acciones automáticas en contra de los procesos que no cumplan con dichas reglas. También se presentarán distintas estrategias reactivas frente a posibles ejecuciones de procesos sospechosos de ser ransomwares.

Palabras Claves: Seguridad. Syscalls. Kernel. Linux. Security. Malware. Ransomware.

1 Introducción

En los últimos años ha crecido la ciberdelincuencia y, con esta, la variedad de malwares o programas maliciosos. Según Raymond et al, el mayor desafío de crear un esquema completo de nombrado de malwares, se debe al número de muestras existentes de malware y a la frecuencia con la que nuevas muestras son descubiertas [1]. Si se considera la clasificación basada en comportamiento propuesta por C. Elisan [2], se puede distinguir entre ransomwares, keyloggers, spywares, gusanos, troyanos, etc. A su vez, un mismo malware puede comportarse como un virus cuando se propaga por un dispositivo de cómputo, como un gusano cuando se propaga a través de una red, mostrar comportamiento de botnet cuando se comunica con servidores de comando y control o cuando sincroniza con otras máquinas infectadas, y comportarse como un rootkit al ocultarse de un sistema de detección de intrusiones (IDS) [3].

Cada uno de estos malwares intentan generar algún daño y para lograrlo es normal que utilicen el kernel para acceder a los recursos que necesitan. Un ransomware, por ejemplo, es una forma de software malicioso utilizado en ataques, en los que no se

busca destruir irreversiblemente los datos, sino cifrar y cobrar por el servicio de recuperación de los datos cifrados [4] y para esto realiza operaciones de lectura y escritura sobre el disco, utilizando el kernel. Otro ejemplo es un keylogger, que es un software que se ubica entre el hardware y el sistema operativo e intercepta cada pulsación de tecla y la almacena, para lo cual también ejecuta estas operaciones por medio del kernel.

Una consultora de cibereconomía y ciberseguridad estima que los daños del cibercrimen tendrán un costo anual y global de seis mil millones de dólares en 2021 [5]. Estos costos incluyen daños y destrucción de datos, dinero robado, pérdida de productividad, robo de propiedad intelectual, robo de datos personales y financieros, malversación, fraude, interrupción posterior al ataque en el curso normal de los negocios, investigación forense, restauración y eliminación de datos perjudicados y sistemas, y daño a la reputación [6].
Es debido al impacto asociado a estos malwares que se hace necesaria una búsqueda permanente para encontrar nuevas técnicas de prevención, o actualizar continuamente las ya existentes, de forma tal que se minimice el impacto de estas amenazas. Entre estas técnicas podemos encontrar las técnicas de análisis de comportamiento o análisis dinámico.
Las técnicas de análisis de comportamiento utilizan las características del software en ejecución para identificar malware potencial [7]. Una de estas técnicas es el análisis de llamadas al sistema, en el que los comportamientos maliciosos se identifican mediante sus rastros de llamadas al sistema [8].

Las llamadas al sistema son muy utilizadas por los programas y los procesadores actuales optimizan su latencia de invocación. Por ejemplo, la llamada al sistema getpid() se completa en 61 nanosegundos [9].

Una llamada al sistema es solo una solicitud de espacio de usuario de un servicio del kernel. Cuando un programa quiere escribir o leer desde un archivo, comenzar a escuchar conexiones en un socket, eliminar o crear un directorio, o incluso terminar su trabajo, el programa usa llamadas al sistema (Figura 1). En otras palabras, una llamada al sistema es solo una función de espacio del núcleo que los programas de espacio del usuario llaman para manejar alguna solicitud. El kernel de Linux proporciona un conjunto de estas funciones y cada arquitectura proporciona su propio conjunto. Por ejemplo: el x86_64 proporciona 322 llamadas al sistema y el x86 proporciona 358 llamadas al sistema diferentes [10].

Es válido aclarar que la necesidad de acceder a los servicios que brinda el kernel a través de llamadas al sistema no es algo exclusivo de los malwares, sino que cualquier proceso lo realiza. La diferencia se encuentra en la manera de acceder, tanto hacia qué llamada, como la frecuencia y los parámetros con los que cada una es solicitada [11].

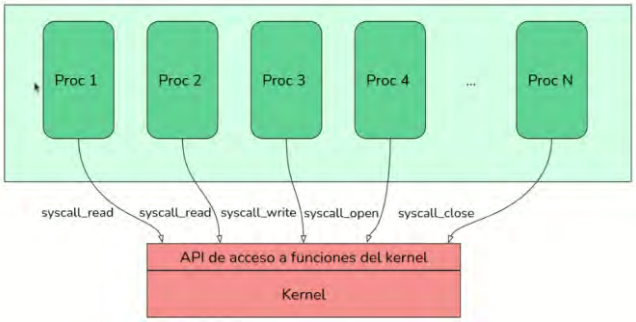


Figura 1. Representación gráfica de syscalls desde distintos procesos.

¿Por qué GNU/Linux? Linux es una opción popular para servidores [12] y sistemas integrados [13], por sus beneficios en rendimiento, fiabilidad y facilidad de desarrollo. Estos tipos de sistemas tienen un mayor riesgo de seguridad, ya que pueden ser una base de datos servidor, equipo de red o una unidad de control de un dispositivo crítico de seguridad.

El presente trabajo utiliza un módulo kernel ya presentado anteriormente como contador de llamadas al sistema en la edición anterior, el cual intercepta las syscalls y lleva un registro de cuántas llamadas al sistema se ejecutan por proceso permitiendo un monitoreo de las mismas [14]. A este contador de syscalls se lo denominó "Módulo kernel del Syscall Top" y para el cual se desarrolló una aplicación Syscall Top que es su Interfaz Gráfica de Usuario (GUI) que se ejecuta en el espacio de usuario. Además de permitir una mejor interpretación de la interacción de las llamadas al sistema de los programas con el kernel también permite configurar reglas con acciones definidas en el caso de que las syscalls superen ciertos umbrales definidos en dichas reglas.
Esta aplicación Syscall Top y sus posibles configuraciones como estrategias de monitoreo frente a procesos sospechosos de ser ransomwares son el eje central del presente trabajo.
Este trabajo está enmarcado dentro del proyecto de I+D "Sistema de detección de malware basado en patrones de llamadas al sistema en GNU/Linux.", código SIUTNCO0007850.

2 Desarrollo

Como se mencionó anteriormente la herramienta Syscall Top está compuesta por dos módulos: el módulo kernel que ya se presentó en la edición 2021 [14] y la aplicación Syscall Top que se presenta ahora.

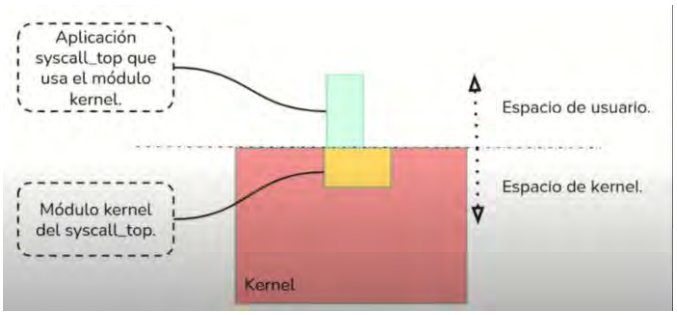


Figura 2. Arquitectura de Syscall Top. Módulo kernel y Aplicación.

El objetivo general de esta herramienta Syscall Top es monitorear ciertas llamadas al sistema revisando cada X tiempo, en base a su tasa de refresco, si dichas llamadas superan un umbral definido, y si lo superan, ejecutar una acción previamente definida. Cuando se inicia el Syscall Top se puede apreciar la siguiente salida (Figura 3) en la consola.

SyscallTop

PID	read	write	open	close	other	total	name
1	2	0	2	2	2	8	systemd(0)
1798	398	397	0	0	0	795	sshd(0)
391	326	0	176	176	204	882	systemd-journald(0)
1544	3	6	3	6	3	21	pickup(1)
1418	4	0	0	2	0	6	nscd(0)
1419	14	6	1	7	5	33	nscd(0)
1420	4	0	0	2	0	6	nscd(0)
1421	4	0	0	2	0	6	nscd(0)
1422	4	0	0	2	0	6	nscd(0)
845	1	0	0	1	0	2	systemd-logind(0)
848	0	0	81	27	378	486	dbus-daemon(4)
1593	23	0	0	0	0	23	slapd(0)
1624	19	12	2	10	10	53	nscd(0)
1625	13	12	0	2	0	27	slapd(0)
1626	12	11	0	3	0	26	slapd(0)
860	72	66	0	0	0	138	gmain(0)
1629	10	10	0	2	0	22	slapd(0)
1125	162	0	27	54	0	243	vmtnfo(0)
1009	3	0	0	0	0	3	in:imklog(0)
809	30	66	0	0	0	96	avahi-daemon(1)
1543	6	6	8	3	8	15	master(0)
1010	0	31	2	0	0	33	rs:main Q:Reg(1)
1800	266	209	62	149	189	875	bash(0)
818	0	0	0	0	21	21	cron(4)
1017	2	0	2	2	0	6	postgres(0)
1594	10	12	0	2	0	24	slapd(1)

Figura 3. Syscall Top en ejecución.

En la Figura 3 podemos visualizar la siguiente información:

- PID: Process Id del proceso en cuestión.
- read: Cantidad de llamadas al sistema de tipo read. Esta llamada al sistema lee bytes de un archivo referenciado por un file descriptor a un buffer.
- write: Cantidad de llamadas al sistema de tipo write. Esta llamada al sistema escribe bytes desde un buffer al archivo referenciado por el file descriptor.

- open: Cantidad de llamadas al sistema de tipo open. Esta llamada al sistema abre un archivo.
- close: Cantidad de llamadas al sistema de tipo close. Esta llamada al sistema cierra un file descriptor, por lo tanto, el archivo referenciado ya no puede ser accedido.
- other: Sumatoria de cantidad de llamadas al sistema de tipos stat, ptrace, fstat.
- stat y fstat: Estas llamadas al sistema obtienen información sobre un archivo, como por ejemplo tiempo.
- ptrace: La llamada al sistema permite observar y controlar la ejecución de un proceso.
- name: Nombre del proceso asociado al PID

Como se puede apreciar la interfaz es familiar a la herramienta Top [15] de un GNU/ Linux para lograr fácil apreciación de la información. Por otro lado, si se identificó un proceso y se quiere hacer énfasis en el mismo, puede hacérselo a través de la ejecución del siguiente comando (Tabla 1) y su salida será la de la Figura 4.

```
$ syscalltop ping
```

Tabla 1. Comando para ejecutar el syscall top monitoreando un solo proceso. Ejemplo con el proceso ping

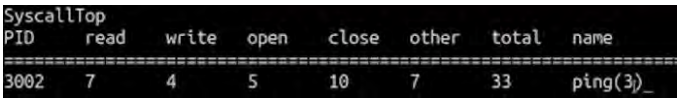


Figura 4. Salida en consola del syscall top top monitoreando un solo proceso. Ejemplo con el proceso ping.

Una vez que el Syscall Top está ejecutándose, además de brindar la información de llamadas al sistema en tiempo cuasi real, también está monitoreando si alguno de los procesos no cumple con reglas definidas en su archivo reglas.ini, el cual debe tener la siguiente estructura (Figura 5).

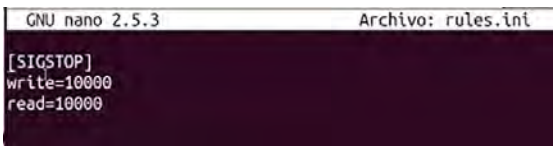


Figura 5. Estructura de archivo rules.ini. Archivo que define las reglas de acción del Syscall Top.

El ejemplo de la Figura 5 determina que el syscall top debe detener el/los procesos que superen al mismo tiempo las 10000 llamadas al sistema de write y read. Es decir, si un proceso X superó las 10000 llamadas de write, como también de read en los últimos Z segundos este proceso será detenido.
Generalizando esta estructura el archivo de reglas quedaría:

```
# Estructura de rules.ini
[<SEÑAL>]
syscallM=<umbralM>
…
syscallN=<umbralN>
```

Donde:
- <SEÑAL>: Señales (Signals) que interpreta el kernel Linux [16]. Representan las acciones a tomar cuando un proceso supere los umbrales definidos.
- <syscallM o N>: Llamadas al sistema que se definen para la regla.
- <umbralM o N>: Umbral que representa el máximo permitido de interacciones entre la llamada al sistema M o N.

Si un proceso cumple con todas esas reglas, a dicho proceso se le aplicará la SEÑAL definida.
En cuanto a las llamadas al sistema que pueden ser definidas como reglas, por el momento, solo se pueden definir las de *read* y *write*, debido a que por el momento, esta herramiento está siendo probada con ransomwares y el patrón detectado para este malware son este tipo de llamadas.
En lo referido a las señales, puede configurarse cualquier señal que interprete el kernel de linux, las que se pueden listar a través de un kill -l [17]. Entre las más conocidas podemos nombrar la SIGKILL y SIGSTOP.
Todo lo mencionado anteriormente permite pensar y diseñar distintas estrategias de reacción ante posibles ataques de ransomwares. A continuación se listan algunas estrategias, siempre teniendo en cuenta la tasa de refresco del Syscall top de T segundos.:
1. Detención de procesos: manteniendo el ejemplo anterior, una primera estrategia sería la de detener los procesos que superen las 10000 interacciones de write y read en los últimos T segundos para una revisión posterior manual por parte de una persona que decidirá si el proceso continúa o no (Tabla 2).

```
# Estructura de rules.ini
[SIGSTOP]
read=10000
write=10000
```

Tabla 2. Configuración de rules.ini para una estrategia de detención.

En esta estrategia, el usuario debe revisar manualmente los *jobs* detenidos para identificar los procesos en cuestión.

2. Matar procesos: se puede tener como premisa que todo proceso que supere las 10000 de *write* y *read* en los últimos T segundos es considerado malicioso y, por lo tanto, directamente se envía la señal de kill a ese proceso (Tabla 3).

```
# Estructura de rules.ini
[SIGKILL]
read=10000
write=10000
```

Tabla 3. Configuración de rules.ini para una estrategia de matar procesos.

Esta estrategia tiene el propósito de asegurar que el proceso sospechoso no se iniciará nuevamente. Sin embargo, al ser tan radical, tampoco permite dejar rastros de qué procesos fueron matados, por lo que los usuarios nunca se enterarán si hubo intentos de ejecución de código malicioso.

3. Detener y matar procesos: sirve contra malwares que han encontrado la manera de ejecutarse, en el caso que detecten que han sido detenidos. Es importante considerar en en este caso que los valores de *read* y *write* de la señal SIGKILL deben ser mayores a los de SIGSTOP (Tabla 4).

```
# Estructura de rules.ini
[SIGSTOP]
read=10000
write=10000

[SIGKILL]
read=11000
write=11000
```

Tabla 4. Configuración de rules.ini para una estrategia de detener y matar proceso.

Supongamos que en los últimos T segundos El proceso P se ha detenido cuando alcanzó las condiciones para el SIGSTOP, pero se volvió a iniciar y alcanzó las condiciones para el SIGKILL En este caso, al proceso P se le envía la señal de matar.

4. Notificar estado de procesos: asume que hay un proceso N en el espacio de usuario esperando esta notificación para tomar distintas acciones. También hay que aclarar que, para este caso, se debe adaptar el código del Syscall Top agregando el nombre de dicho proceso (Tabla 5).

```
# Estructura de rules.ini
[SIGUSR1]
read=10000
```

```
write=10000
```

Tabla 5. Configuración de rules.ini para una estrategia de notificar estados de procesos.

Suponiendo que en los últimos 5 segundos un proceso ha superado este umbral, entonces se enviará la señal SIGUSR1 al procesos N, para que éste desencadene sus acciones programadas. Estas acciones podrían ser enviar notificaciones a usuarios, ya sea a través de mail o algún otro medio, enviar datos del procesos a cierta aplicación que los procese en tiempo cuasi real, enviar datos a una persistencia de datos para su posterior análisis, por mencionar algunos ejemplos. Esto ya depende de la aplicación personalizada que se quiera añadir a este ecosistema del Syscall Top.

5. Estrategia combinada: se podrían combinar las anteriores, dando lugar a una estrategia más robusta (Tabla 6).

```
# Estructura de rules.ini
[SIGSTOP]
read=10000
write=10000

[SIGKILL]
read=11000
write=11000

[SIGUSR1]
read=10000
write=10000
```

Tabla 5. Configuración de rules.ini para una estrategia combinada.

Esta configuración permite detener los procesos que superen las 10000 interacciones de *read* y *write* en los últimos T segundos, como así también si este proceso se vuelve a iniciar, lo detiene cuando alcance las 11000 interacciones de *read* y *write*. Además, esta estrategía notifica a un proceso N, si los procesos sospechosos superan las 10000 interacciones también en los últimos T segundos.

3 Conclusiones

La herramienta presentada fue diseñada inicialmente para tener un mayor entendimiento del comportamiento de distintos malwares que se ejecutan sobre un sistema GNU/Linux.

En el proceso recorrido por el proyecto que enmarca este trabajo, se fue descubriendo y agregando funcionalidad para que Syscall Top no solo monitoree los procesos

sospechosos, sino que también podría tomar acciones reactivas frente a estos, tal como se comentan en las primeras tres estrategias descritas en la sección anterior.
El fuerte de Syscall Top es que es adaptable a cualquier entorno actualmente existente en el campo del análisis de llamadas al sistema como se explicó en las estrategias cuatro y cinco.
Esta herramienta pasa a formar parte de la última línea de defensa en el modelo de ciberseguridad de una infraestructura de red, ya que puede ser ejecutada en segundo plano en servidores Linux.
Syscall Top también puede ser utilizado en escritorios personales, no necesariamente servidores. El hecho de que sea de código abierto le permite ofrecer a quien necesite, la posibilidad de ser adaptado para cumplir con las necesidades de cada infraestructura.
El hecho de poder enviar señales a cualquier a cualquier aplicación de usuario, permite a la comunidad sumar módulos satélites con acciones definidas que interpretan estas señales, como por ejemplo un módulo de notificación por mail.
En cuanto a proyecciones futuras, esta herramienta da la posibilidad de explotar las diferentes señales que puede interpretar el kernel linux que sean consideradas que ayuden al análisis de malwares.
Si bien hasta el momento solo se han analizado y experimentado con procesos de ransomwares, el paso siguiente es estudiar procesos de otros malwares y experimentar y analizar posibles estrategias a ser configuradas en el Syscall Top para los mismos.

Referencias

1. Canzanese R. (2015). Detection and Classification of Malicious Processes Using System Call Analysis. Recuperado el 28 de Mayo de 2019 https://pdfs.semanticscholar.org/8060/eae74c98a66cfcc736f4fca61d46f4dbc1d4.pdf.
2. Elisan C. (2015) Advanced Malware Analysis. McGraw-Hill. Capítulo 2. ISBN: 9780071819756.
3. Ethan Rudd, Andras Rozsa, Manuel Gunther, and Terrance Boult. 2017. A survey of stealth malware: Attacks, mitigation measures, and steps toward autonomous open world solutions. IEEE Communications Surveys & Tutorials 19, 2 (2017), 1145–1172.
4. K. Savage, P. Coogan, and H. Lau, (2018). The Evolution of Ransomware. Secur. Response, p. 57, 2015.
5. Morgan S. (Mayo 2017). 2018 Cybersecurity Market Report. Recuperado el 28 de Mayo del 2019 de https://cybersecurityventures.com/cybersecurity-market-report/.
6. Morgan S. (Diciembre 2018). Cybercrime Damages $6 Trillion By 2021. Recuperado el 28 de Mayo del 2019 de https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021
7. M. Egele, T. Scholte, E. Kirda, and C. Kruegel, “A survey onautomated dynamic malware-analysis techniques and tools,” ACMComputing Surveys, 2008.
8. S. Forrest, S. Hofmeyr, A. Somayaji, and T. Longstaff, “A sense of selffor unix processes,” in IEEE Security and Privacy, 1996.
9. Haiquan Xiong, Zhiyong Liu, Weizhi Xu, Shuai Jiao LibVMI: a library for bridging the semantic gap between guest os and vmm 12th International Conference on Computer and Information Technology, IEEE (2012), pp. 549-556

10. System calls in the Linux kernel. Part 1. GitBooks. https://0xax.gitbooks.io/linux-insides/content/SysCall/linux-syscall-1.html
11. Searchable Linux Syscall Table for x86 and x86_64. https://filippo.io/linux-syscall-table/. Última visita 08/08/2021.
12. "Usage Statistics and Market Share of Operating Systems for Websites, August 2020", W3Techs, accessed 09.09.2020, https://w3techs.com/technologies/overview/operating_system.
13. "2019 Embedded Markets Study", AspenCore, accessed 09.09.202 0, https://www.embedded.com/wpcontent/uploads/2019/11/EETimes_Embedded_2019_Embedded_Markets_Study.pdf.
14. Gibellini F., Quinteros S., Parisi G., Zea Cárdenas M., Ciceri L., Bertola F., Barrionuevo I., Notreni J., Ruhl A. Monitoreo de Llamadas al Sistema como Método de Prevención de Malware. WSI - SEGURIDAD INFORMÁTICA, CACIC 2021, 27th Congreso Argentino de Ciencias de la Computación, CACIC 2021, Salta, Argentina.
15. Página oficial de Man. Ayuda de Top de Linux. https://man7.org/linux/man-pages/man1/top.1.html
16. Bovet D., Cesati M. Understanding the Linux Kernel. Capítulo 11. Third Edition. O'Reilly Media. ISBN 0-596-00565-2
17. Bovet D., Cesati M. Understanding the Linux Kernel. Capítulo 11. Third Edition. O'Reilly Media. ISBN 0-596-00565-2

Ataque por Sustitución de Algoritmo Criptográfico: Implementación de prueba para OpenSSL

Cipriano, Marcelo. García, Edith. Maiorano, Ariel.
Malvacio, Eduardo. Pazo Robles, María Eugenia

Facultad de Ingeniería del Ejército (FIE), Universidad de la Defensa Nacional (UNDEF)
{marcelocipriano; egarcia; maiorano;
emalvacio; mpasorobles}@fie.undef.edu.ar

Resumen. El presente artículo presenta la primera de una serie de implementaciones de diferentes esquemas kleptográficos. Reconociendo a la Kleptografía como un subcampo de la Criptovirología, que aborda los usos maliciosos de la criptografía y sus aplicacioines. Estas implementaciones se muestran a manera de prueba de concepto y aplicación experimental. De manera que permitan la validación y análisis de los paradigmas y herramientas criptológicas modernas para la creación de software malicioso. Asimismo se persigue la elaboración de técnicas de prevención, detección temprana y protección, para ser consideradas como un aspecto más de la Ciberdefensa. El código fuente de referencia ha sido publicado en Github.com.

Palabras clave: Keptografía, Criptovirología, Criptografía Maliciosa, Ataques por Sustitución de Algoritmos, Malware.

1 Introducción

Aunque comúnmente entendemos a la criptografía y a sus aplicaciones como herramientas de carácter defensivo, también pueden emplearse con usos ofensivos. Entre los usos maliciosos se encuentran los ataques basados en extorsión, software malicioso comúnmente denominado *Ransomware*, que causa la pérdida de acceso a la información. Por otra parte, la literatura también da cuenta de ataques en las etapas de diseño e implementación de en algoritmos criptográficos, comúnmente llamados *backdoors* o puertas traseras, que pueden vulnerar la privacidad, autenticidad, confidencialidad y seguridad en general de sus usuarios.

Se resume a continuación, de manera general, lo que podría considerarse el estado del arte y los alcances actuales de los conceptos de Criptovirología, Kleptografía y, como un ejemplo particular de esta última, de los llamados *Ataques por Sustitución de Algoritmos* –del inglés, *Algorithm-Substitution Attacks (ASAs)-,* haciendo foco principalmente en lo relativo al diseño e implementación de puertas traseras en algoritmos, protocolos y/o implementaciones de software criptográficos.

1.1 Contexto

En el año 1996, *Adam Young y Moti Yung* [1], introdujeron el concepto de "*Criptovirología"*, exponiendo que aunque comúnmente entendemos a la Criptografía y a sus aplicaciones como herramientas de carácter defensivo, que proporcionan privacidad, autenticidad, confidencialidad y seguridad a sus usuarios, también pueden emplearse con usos ofensivos. Entre otros, se ejemplifica el uso malicioso con el montaje de ataques basados en extorsión (para los cuales la Criptografía de Clave Pública ha sido muy utilizada) que causan la pérdida de acceso a la información (Integridad), pérdida de Confidencialidad y fuga de información.

Al margen de su utilización para la implementación de un virus, en palabras de los autores en aquel entonces, del tipo actualmente categorizado como *Ramsonware*, o *Software Malicioso Extorsivo*, se presenta luego el concepto de *"Kleptografía*", que abarca al diseño e implementación de *backdoors* o puertas traseras en algoritmos criptográficos, como fue expuesto sus trabajos posteriores [2-4].

En tales trabajos, se detallan diferentes metodologías de los llamados ataques *kleptográficos,* bajo el concepto de "*Secretly Embedded Trapdoor with Universal Protection*" *(SETUP).* Por citar un ejemplo particular, entre estos trabajos iniciales, se describe acabadamente un kleptograma para el algoritmo de intercambio de llaves *Diffie-Hellman*, pero también muestran cómo estos diseños podrían ser embebidos en otros sistemas, como los algoritmos de cifrado y de firma digital *ElGamal*, *DSA*, el algoritmo de firma *Schnorr,* y el *PKCS* de Menezes-Vanstone.

Luego aparecerían diseños de algoritmos *SETUP*, junto con otras técnicas o mecanismos kleptográficos, para el algoritmo *RSA* [5- 6, 8, 13].

Es importante destacar que estos diseños no se limitan a criptografía de llave pública. Se cuenta en la literatura con publicaciones que describen ejemplos aplicados a funciones de *hashing* donde se presentan colisiones para una versión de *SHA-1* con constantes modificadas [7], como también alternativas para protegerse de esta primitiva criptográfica (potencialmente comprometida) en algoritmos de nivel superior, o que dependen de las primeras, como *HMAC* y *HKDF* [14].

Por supuesto tampoco los *Generadores de Números Pseudo-Aleatorios* (*PRNG* por sus siglas en inglés) serían inmunes a este tipo de ataques, como se explica en [10, 12]. Estas publicaciones describen mecanismos que afectan las propiedades estadísticas de un generador haciéndolo muy sensible a la entropía de entrada. Por ejemplo, cuando las entradas tienen una distribución correcta, este mecanismo no tiene efecto, pero cuando la distribución estuviera sesgada, el generador malicioso empeora considerablemente. Se destaca que además de su uso malicioso más evidente, este mecanismo también se puede aplicar al testeo de generadores.

Por último podría resumirse, en línea con lo expresado en [14], que la seguridad de los esquemas criptográficos se mide tradicionalmente como la incapacidad de los adversarios con recursos limitados para violar un objetivo de seguridad deseado. El argumento de seguridad generalmente se basa en un diseño sólido de los componentes subyacentes. Podría decirse que uno de los fracasos más devastadores de este enfoque se puede observar al considerar adversarios con la capacidad de influir en el diseño, implementación y estandarización de primitivas criptográficas. El creer que esto no

era posible se ha comprobado *naive*. Es entonces que considerando el impacto y la relevancia actual [16-17] de las técnicas y mecanismos mencionados se justificaría esta línea de investigación.

1.2 Ataques por Sustitución de Algoritmos

Más recientemente se han publicado una serie de trabajos enfocados en la aplicación de técnicas *kleptográficas* en Algoritmos de Criptografía Simétrica, específicamente aquellos de Cifrado en Bloques [18-20].

Específicamente, *Bellare, Paterson y Rogaway* [18] abordan el concepto de *Ataques por Sustitución de Algoritmos* (*ASAs: Algorithm-Substitution Attacks*). En ese trabajo los autores, motivados por las revelaciones sobre la vigilancia masiva de las comunicaciones encriptadas, formalizan e investigan la resistencia de los esquemas de cifrado simétrico. El foco está puesto en los *ASA*, donde un algoritmo de cifrado subvertido reemplaza al real. Suponen que el objetivo del atacante –a quién llaman "*big brother"- es la subversión indetectable.* Ello significa que los textos cifrados producidos por el algoritmo de cifrado subvertido deben revelar los textos en claro al atacante y, sin embargo, ser indistinguibles de los producidos por el esquema de cifrado real para los usuarios.

Asimismo, formalizaron nociones de seguridad para capturar ese objetivo y luego ofrecen detalles acerca de ataques y defensa, en particular que pueden ser montados en una gran clase de esquemas de Cifrado Simétrico.

Como recuerdan los autores, los *ASA* se han tratado antes bajo varios nombres, abarcados en el concepto de *Kleptografía.* Mientras algunos criptógrafos parecen haber desestimado inicialmente a este subcampo, revelaciones recientes sugieren que esta actitud resultó ser ingenua. Estos ataques pueden estar sucediendo en la actualidad. Posiblemente en una escala masiva.

1.3 Ejemplo de Ataque por Sustitución de Algoritmo

Tal como se ha mencionado anteriormente, en este trabajo se presenta también un software experimental a modo de *Prueba de Concepto*, en el que se implementa una *puerta trasera criptográfica*. Tal esquema se muestra con fines de investigación, prueba de concepto y testeo para la implementación. De ninguna manera se lo considera destinado a otros usos y aplicaciones más allá de lo académico. El mismo fue desarrollado por el *Grupo de Investigación en Criptografía y Seguridad Informática,* perteneciente al *Laboratorio de Informática, Software Seguro y Criptografía (LISSyC),* de la *Facultad de Ingeniería del Ejército (FIE) de la Universidad de la Defensa Nacional (UNDEF).*

Este software libre y de código abierto, se alinea con otras publicaciones de GICSI, relacionados a la seguridad informática en general y a la criptografía en particular. El mismo se puede encontrar disponible en la plataforma de alojamiento de proyectos de software libre *Github.com* [22], bajo la *Licencia Pública General de GNU versión 3* (*GPLv3)*.

2 Sustitución de Algotimo AES en *engine* OpenSSL

De acuerdo con el primero de los dos tipos de ataque descriptos en [18], donde además se cita como ejemplo la aplicación al algoritmo *AES* con *128 bits* de llave en modo de operación *CBC* (*Cipher Block Chaining*), los autores muestran que los *Esquemas de Cifrado sin Estado* son típicamente subvertibles.

El tipo de ataque mencionado aplica a algoritmos de Cifrado Simétrico utilizando modos de operación que evidencian o exponen su vector de inicialización (IV: *Initialization Vector)*.

Siguiendo a *Young y Yung* en los conceptos mencionados en apartados anteriores, y luego a *Goh, Boneh, Pinkas y Golle* [21], los autores consideran el problema de agregar un mecanismo oculto, subrepticio, de recuperación de la llave *k* utilizada en esto algoritmos para diferentes protocolos.

Sugieren que cuando el servidor necesitara un valor aleatorio para ser utilizado como *nonce*, sería posible utilizar, en su lugar, un derivado de la llave *k* en cuestión, que permitiera al atacante recuperar la llave original. De esta forma, teniendo en cuenta a los esquemas de *Cifrado Simétrico en Bloques* que "exponen" el *Vector de Inicialización* en modo *CBC*, en lugar del *IV* explícito generado aleatoriamente, se utilizaría el resultado de cifrar la llave de cifrado, mediante otra llave k_s, que se denomina "*de subversión*", en poder del atacante.

Este trabajo describe la implementación de una versión kleptográfica del algoritmo *AES-CBC*, dentro, o en la forma, de un motor o *engine* de *OpenSSL*. Consideraremos el ataque más simple, en el que el *IV* simplemente se reemplaza por el resultado de una función *F* sobre la clave a utilizar.

La Figura 1 muestra el funcionamiento del mecanismo `cipher_aes_128_cbc` de la librería *OpenSSL*. Esto es, el cifrado mediante el algoritmo *AES* con llave o clave de *128* bits de longitud, en modo *CBC* y se lo compara con el funcionamiento de `cipher_aes_128_cbc_klepto`, implementado como *engine.* Este reemplazará la implementación original, para todos los llamados a estas funciones que se realicen desde programas que requieran a la librería esta funcionalidad.

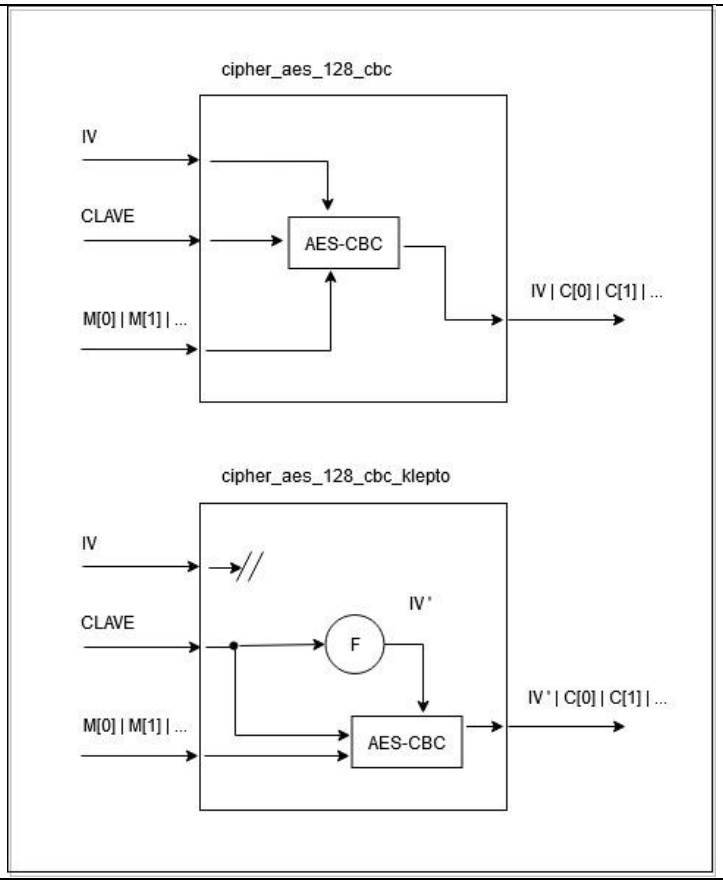


Figura 1. Ejemplo de implementación kleptográfica de `aes_128_cbc` en OpenSSL.

Exclusivamente y a modo de ejemplo, se presenta esta implementación de prueba. En un ataque verdadero requeriría de otro tipo de función, que pueda satisfacer determinadas propiedades que le permitan "*suplantar*" eficazmente al *engine* original y así engañar a su víctima.

En este caso, el *IV* producido originalmente, se descarta. Y se reemplaza por I*V′*. Este valor corresponde al resultado de aplicar la función *F* a la llave *k*, haciendo uso de la clave *k′*; siendo *k* la llave de cifrado original ("*CLAVE*" en el diagrama), y *k'* la llave de subversión, incluida o embebida en la función. El ataque se probará utilizando una conexión segura aplicando el protocolo *TLS* versión 1.2.

Una de las características más destacadas de la función *F* es que debería corresponder a algún esquema de cifrado que use la clave *k'*. Esta clave debe ser únicamente conocida por el atacante. De esta forma, solamente esta entidad maliciosa podría usufructuar los beneficios de esta vulnerabilidad del sistema. De forma adicional, se persigue también el ocultamiento del backdoor:

$$IV' = F_{k'}(k) \quad (1)$$

$$F_{k'}(k) = CIFRADO_{k'}(k) \quad (2)$$

2.1 Concepto de motor o "*engine*" en Librería OpenSSL

Desde su sitio web oficial [23], *OpenSSL* se define como un conjunto de herramientas robusto, de grado comercial y con todas las funciones para los protocolos de seguridad de la capa de transporte (*Transport Layer Security*, o *TLS)* y la capa de sockets seguros (*Secure Sockets Layer*, o *SSL*). A la vez, también es una librería de criptografía de propósito general.

En lo que respecta a los motores o *engines*[24-25], desde la versión 0.9.6 de *OpenSSL*, se agregó un nuevo componente para admitir implementaciones alternativas de funcionalidades criptográficas, más comúnmente utilizado para interactuar con dispositivos criptográficos externos, por ejemplo tarjetas aceleradoras.

Este componente, también entendido como objeto, se denominó motor o *engine*. Estos objetos actúan como contenedores para implementaciones de algoritmos criptográficos y admiten un mecanismo para permitir que se carguen dinámicamente en la aplicación en ejecución.

2.2 Detalles de la implementación

En el ejemplo presentado en este trabajo, se implementó, mediante un motor o *engine* OpenSSL, una versión kleptográfica del algoritmo AES 128 en modo CBC. Los archivos fuentes, junto con la configuración y script decodificador pueden descargase del repositorio en Github del GICSI [22].

En el caso del *engine OpenSSL* presentado en este trabajo, si bien "reemplaza" la implementación original de *AES128* en modo *CBC*, *OpenSSL* en su implementación de los protocolos *TLS* versiones *1.1* y *1.2,* siguiendo lo indicado en la sección 6.2.3.2 de la *RFC 4346* [26], particularmente lo descripto como alternativa (2)(b), generará el *IV* explícito cifrando un primer bloque a descartar por el destinatario. Por tal razón, esta implementación no realiza, estrictamente, un reemplazo de *IV* sino del primer bloque de texto cifrado generado.

Como fuera mencionado más arriba, este primer bloque será reemplazado por el resultado de F(k,k´), siendo k la llave de cifrado del algoritmo, y k' la llave de subversión, que se encuentra fijada en el código fuente del *engine* y en el *script* para realizar la decodificación.

En [18], para simplificar la presentación, los autores asumieron que la longitud del *IV* y la longitud de la llave *k* son las mismas. A la vez presentaron alternativas para cuando éste no fuera el caso. Para este ejemplo, con *AES-128-CBC*, esto no representa un problema.

Por otro lado, advertido también por los autores, el evitar *IVs* repetidos, evidenciados a través de textos cifrados, requeriría limitar la sustitución de *IV* a un texto cifrado. Esto requiere el uso de un *Esquema de Subversión con Estado*. En el presente

trabajo no se intentará evitar esta repetición ya que se considerará que las llaves no se repetirán en el escenario planteado. Dado que *OpenSSL* en su implementación de los protocolos *TLS* versiones 1.1 y 1.2, deriva una llave de *16 bytes* que no se podrá distinguir de *16 bytes* aleatorios. Como fuera indicado, para el caso de prueba para este trabajo, el resultado se obtendrá a partir de la *suma módulo 2* o *XOR* con el valor -fijo- de la llave de subversión. Téngase en cuenta además que el reemplazo se realiza sólo inicialmente, es decir, sólo una vez.

3 Resultados experimentales

3.1 Captura de conexión segura

Para la captura de paquetes de red de una conexión segura se realizó utilizando el utilitario *s_client* de *OpenSSL*. Se trata básicamente de un comando permite la realización de pruebas y depuración o *debug* sobre conexiones *TLS*.

Al margen de lo anterior, téngase en cuenta que todo software del sistema que se encuentre enlazado dinámicamente con la librería *OpenSSL* utilizaría este *engine* kleptográfico.

El comando ejecutado fue el siguiente:

```
$ echo "GET /" | openssl s_client -ign_eof -tls1_2 -cipher \
AES128-SHA -connect www.google.com:443 –servername
www.google.com
```

Listado 1. Ejemplo de ejecución de `s_client` para generación de conexión segura.

Habiendo configurado previamente el sistema para la utilización del *engine* para todo caso en que se requiera a *OpenSSL* utilizar *AES,*-128-*CBC*. AsÍ se realiza una petición al sitio web de Google para obtener su página de inicio.

La captura se registra en un archivo *.pcap,* que luego alimentará al script decodificador, como se muestra en el apartado siguiente.

3.2 Decodificación de captura

Como fuera adelantado, el otro componente necesario para probar el ataque, es el script decodificador. Este script tomara por entrada la captura generada de acuerdo al **Listado 1** y recuperará el *IV* explícito enviado.

Luego aplicará la operación *XOR* sobre éste y la llave de subversión fija *k'* para obtener así la llave *k*, que fue utilizada para cifrar.

Finalizado este procedimiento, se identificarán los mensajes *TLS* de tipo "*Application Data"* para descifrarlos y presentarlos como salida.

```
$ python decodificar.py
```

```
IV explícito:
b'\xdf\xbeQ\xfd\xcd\xc41\xae9L\xf0\xc7\rk\x15\t'
llave: b' A\xae\x022;\xceQ\xc6\xb3\x0f8\xf2\x94\xea\xf6'
data: b"\xbc)Q\xe7\\\xab\x17`...
decifrado: b'GET /...\x05\x05\x05\x05\x05'
```

Listado 2. Salida de script decodificar.py con ejemplo de datos descifrados.

Tal como puede apreciarse en el **Listado 2,** se encuentra descifrada la petición *HTTP* con el método *"GET"* que fuera enviada en el ejemplo. También puede identificarse el *relleno o padding,* de *5 bytes*, que corresponden al final del bloque en texto en claro.

4 Trabajo futuro

Se planea la continuación del desarrollo de diferentes esquemas *kleptográficos* para la experimentación y evaluación de su factibilidad y eficacia.

En particular, se intentará abordar el estudio de otras funcionalidades criptográficas como por ejemplo: generación de números aleatorios, funciones de *hashing*, y algoritmos de cifrado asimétrico.

5 Conclusiones

Se han presentado las generalidades, a través de un ejemplo de implementación, de una técnica *kleptográfica*. Aunque específico, este primer ejemplo representa un caso posible de aplicación a un algoritmo en particular, presente en todas las *ciphersuites* aceptadas para *TLS 1.2* . Incluso, es la única definida como obligatoria.

Aunque el escenario aquí presentado, requiere e implica el compromiso inicial del sistema, el ataque es una *puerta trasera* o *backdoor* no detectable *"over the wire",* ni siquiera monitoreando la red. Además persiste, aún frente a actualizaciones del software del Sistema Operativo, incluyendo aún, a aquellas que actualicen la librería OpenSSL,

Es por estas y otras capacidades, que se puede apreciar la potencia y alcance de este tipo de *malware*.

Referencias

1. Young, Adam L. and Moti Yung. "Cryptovirology: extortion-based security threats and countermeasures." Proceedings 1996 IEEE Symposium on Security and Privacy (1996): 129-140.
2. Young, Adam L. and Moti Yung. "The Prevalence of Kleptographic Attacks on Discrete-Log Based Cryptosystems." CRYPTO (1997).
3. Young, Adam L. and Moti Yung. "Kleptography: Using Cryptography Against Cryptography." EUROCRYPT (1997).
4. Young, Adam L. and Moti Yung. "Malicious cryptography - exposing cryptovirology." (2004).
5. Young, Adam L. and Moti Yung. "A Space Efficient Backdoor in RSA and Its Applications." Selected Areas in Cryptography (2005).
6. Young, Adam L. and Moti Yung. "An Elliptic Curve Backdoor Algorithm for RSASSA. " Information Hiding (2006).
7. Albertini, Ange, Jean-Philippe Aumasson, Maria Eichlseder, Florian Mendel and Martin Schläffer. "Malicious Hashing: Eve's Variant of SHA-1." Selected Areas in Cryptography (2014).
8. Young, Adam L. and Moti Yung. "Cryptography as an Attack Technology: Proving the RSA/Factoring Kleptographic Attack. " The New Codebreakers (2015).
9. Russell, Alexander, Qiang Tang, Moti Yung and Hong-Sheng Zhou. "Cliptography: Clipping the Power of Kleptographic Attacks." ASIACRYPT (2015).
10. Indarjani, Santi. Sugeng, Kiki. Widjaja, Belawati. "Modification Attack Effects on PRNGs: Empirical Studies and Theoretical Proofs." (2015).
11. Young, Adam L. and Moti Yung. "Cryptovirology: the birth, neglect, and explosion of ransomware." Commun. ACM 60 (2017): 24-26.
12. Teseleanu, George. "Random Number Generators Can Be Fooled to Behave Badly." IACR Cryptology ePrint Archive (2018).
13. Markelova, A. V. "Vulnerability of RSA Algorithm." (2018).
14. Fischlin, Marc. Janson, Christian. Mazaheri, Sogol. "Backdoored Hash Functions: Immunizing HMAC and HKDF." (2018): 105-118.
15. Xiao, Dianyan and Yang Yu. "Klepto for Ring-LWE Encryption." Comput. J. 61 (2018): 1228-1239.
16. Yogi, Manas. Aparna, S.. "Novel insights into Cryptovirology A Comprehensive Study." International Journal of Computer Sciences and Engineering. 6. (2018): 1252-1255.
17. Zimba, Aaron. Chishimba, Mumbi. "On the Economic Impact of Crypto-ransomware Attacks: The State of the Art on Enterprise Systems." European Journal for Security Research. (2019).
18. Bellare M., Paterson K.G., Rogaway P. "Security of Symmetric Encryption against Mass Surveillance." Advances in Cryptology. CRYPTO 2014. Lecture Notes in Computer Science, vol 8616. Springer, Berlin, Heidelberg (2014).
19. Mihir Bellare, Joseph Jaeger, and Daniel Kane. "Mass-surveillance without the State: Strongly Undetectable Algorithm-Substitution Attacks." In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS '15). Association for Computing Machinery, New York, NY, USA, 1431–1440. (2015).
20. Dunkelman, Orr and Léo Perrin. "Adapting Rigidity to Symmetric Cryptography: Towards 'Unswerving' Designs." Proceedings of the 5th ACM Workshop on Security Standardisation Research Workshop. (2019).

21. E.-J. Goh, D. Boneh, B. Pinkas, and P. Golle. "The design and implementation of protocol-based hidden key recovery". In C. Boyd and W. Mao, editors, ISC 2003, volume 2851 of LNCS, pages 165-179. Springer. (2003).
22. Proyecto keplto-openssl-engine. GICSI. Repositorio en Github. En línea: https://github.com/gicsi/keplto-openssl-engine..
23. Home page. OpenSSL. Sitio oficial. En línea: https://www.openssl.org/.
24. ENGINE cryptographic module support. OpenSSL. Sitio oficial. En línea: https://www.openssl.org/docs/man1.0.2/man3/engine.html.
25. README.ENGINE. OpenSSL. Repositorio en Github. En línea: https://github.com/openssl/openssl/blob/master/README.ENGINE.
26. RFC 4346. The Transport Layer Security (TLS) Protocol Version 1.1. IETF. En línea: https://tools.ietf.org/html/rfc4346.
27. RFC 5246. The Transport Layer Security (TLS) Protocol Version 1.2. IETF. En línea: https://tools.ietf.org/html/rfc5246.

Aspectos de seguridad en un sistema de IOT para controlar la calidad del aire

Paula Venosa[1], Sofía Martin[1], Patricio Bolino[1], Paula Durán[1], Lautaro Canales[1]
[1] LINTI - Facultad de Informática - Universidad Nacional de La Plata
pvenosa@info.unlp.edu.ar, smartin@linti.unlp.edu.ar, patriciobolino@gmail.com, paumduran@hotmail.com, lautarocanales@gmail.com

Abstract. La Internet de las cosas (IoT) ha crecido y ampliado su uso a lo largo de los últimos años, aplicándose en ámbitos de la industria, ahorro de energía, mejoras de servicios gubernamentales, entre otros. Por lo que su seguridad ha generado mayor interés debido a los datos que se adquieren, transmiten y procesan. En este artículo se evalúan aspectos de seguridad a tener en cuenta concentrándose en un caso particular de implementación y se proponen posibles mitigaciones para mejorar las debilidades de seguridad existentes.

Keywords: Seguridad IoT; Automatización CO2; Seguridad;.

1 Introducción

Actualmente el uso de tecnologías que permiten automatizar funciones avanza cada día más, permitiendo transformar las ciudades en "inteligentes" con el fin de mejorar sus servicios y la calidad de vida de sus ciudadanos [1], [2]. Esta transformación liderada por los dispositivos IoT (Internet of Things o Industrial Internet of Things) está ayudando tanto a las grandes metrópolis, como a las pequeñas ciudades a mejorar los servicios que prestan y a permitir la comunicación de la información. Su expansión ha permitido pensar en aspectos que permitan prácticas de ahorro de energía inteligentes, infraestructuras con servicios automatizados para simplificar la vida en los hogares, y el paradigma de una industria 4.0 [3].

Algunos de los aspectos a tener en cuenta cuando se trabaja con dispositivos IoT, es profundizar no solamente en las necesidades técnicas que se espera cubrir con los mismos sino también poner especial atención al nivel de protección ante ataques que puedan afectar a la privacidad de los datos, de la organización y de los usuarios, o impedir el correcto funcionamiento de su actividad [4].

Las organizaciones que conectan a sus redes dispositivos IoT deben tener en consideración la importancia de garantizar la seguridad, a través del fortalecimiento de la red donde se conectan los dispositivos [5], del análisis y configuración de los dispositivos, del monitoreo del tráfico y actividad de los mismos así como de contar con un plan para gestionar los incidentes en dicho marco [6].

El análisis de vulnerabilidades es un paso importante a la hora de construir arquitecturas que transmiten datos por red y en particular aquellas que utilizan dispositivos IoT ya que permite identificar debilidades y mitigar así posibles

amenazas, para mejorar el nivel de seguridad de los sistemas. El paso más importante para evaluar las mejoras para securizar un sistema es en la etapa de diseño, gestionando luego un proceso contínuo de mejoras cuando el sistema ya ha sido implementado.

Es por ello que desde el LINTI, en el marco de la línea de investigación de seguridad en IOT, se trabaja desde el año 2017 en el análisis y mitigación de vulnerabilidades en diferentes proyectos: “Análisis y mejoras de seguridad a una aplicación prototipo en IoT” [7], “Hackeamos para construir robots seguros” [8] e “Identificación de vulnerabilidades en ambientes IOT” [9]. En el marco de este último se realizó el trabajo que se presenta en este artículo.

2 Ciberseguridad y aplicaciones IoT

En la actualidad la seguridad en sistemas informáticos ha cobrado cada vez mayor importancia debido al crecimiento en su uso que hace que las organizaciones basen sus principales funciones en dichos sistemas. Una de las áreas involucradas son aquellos sistemas que incluyen dispositivos IoT, la cual ha cobrado mucha relevancia debido a la masividad de su uso [10] y a la importancia de proteger los datos confidenciales que estos procesan. A su vez, deben implementarse controles en la seguridad física ya que este tipo de dispositivos se encuentran dispuestos por fuera de los centros de datos y suelen quedar excluidos del “radar” de los equipos de seguridad informática, comunicaciones e infraestructuras.

La creciente disponibilidad de los mismos ha permitido su implementación con diversos fines, dado que en general son placas que cuentan con varios puertos de conexión, con capacidad de gestionar funcionalidades simples en función de los valores recibidos. Estos puertos permiten conectar diferentes tipos de sensores, actuadores, dispositivos RFID, entre otros y en general el procesamiento local es mínimo, debido a la capacidad del procesador [2]. Esta característica propia de los dispositivos IoT genera que en muchos casos la información se envíe a servidores remotos que centralizan los datos de uno o más dispositivos.

Una de las aplicaciones de los dispositivos IoT, como mencionamos anteriormente, es permitir sensar información del ambiente para, no solo poder reportar los datos, sino también generar alguna acción concreta que dé cuenta de un cambio ocurrido. Este tipo de desarrollos permite automatizar el control de temperaturas, movimiento en un ambiente delimitado, entre otras cosas.

La información enviada, en muchos casos, es a través de la tecnología Wi-Fi, es por ello que la seguridad de los datos se torna crítica y en función de la capacidad del procesador entran en juego las limitaciones para implementar protocolos de encriptación u otras medidas para incrementar la seguridad.

Diferentes organizaciones referentes en controles de ciberseguridad para tecnologías IT comenzaron a desarrollar estándares y frameworks dentro de sus programas de buenas prácticas enfocados a los dispositivos y las aplicaciones IOT. Entre ellos podemos nombrar a la iniciativa OWAP [11] con su Top Ten de vulnerabilidades comunes en IOT, NIST (National Institute of Standards and

Technology) [12] con su programa de ciberseguridad para IOT que incluye estándares, guías y herramientas y colabora con fabricantes, gobiernos, consumidores y universidades, CIS (Center for Internet Security) [13] con el desarrollo de una guía de controles que recorre diversos aspectos tales como la protección de los datos, la gestión en el control de los accesos, las configuraciones de seguridad, las protecciones para el correo electrónico y los navegadores web, el monitoreo de las redes, las defensas ante malware, y la IoTSF que se autodefine como un "super Blue Team[1]" de usuarios, profesionales de seguridad, proveedores de productos de hardware y software de IoT, operadores de red, y otros actores del mundo del IoT, que tienen como misión colaborar en mejorar la seguridad de las aplicaciones de IoT, a través de grupos de trabajo, documentación y distintas herramientas que proveen en el marco de su comunidad [14], entre otros.

Existen diferentes propuestas de metodologías para implementar planes de control de amenazas que pueden afectar a un escenario, para lo cual es necesario evaluar procesos, el hardware del dispositivo, el software, las interfaces cableadas e inalámbricas, la autenticación y la seguridad en la capa de aplicación [6]. Cabe destacar que, si bien la mayoría de los escenarios comparten similitudes, la implementación de cada uno debe evaluarse en función del servicio que presta.

Las organizaciones mencionadas anteriormente dan cuenta de que hoy día ya existen herramientas para evaluar la seguridad de los dispositivos IoT, por lo que es necesario al momento del desarrollo de software y hardware IoT minimizar las amenazas y vulnerabilidades más comunes. Estas precauciones permiten evitar incidentes de seguridad que se traduzcan en la pérdida de confianza hacia estas nuevas tecnologías.

En este contexto se propone un análisis de un caso de uso aplicable a diferentes ámbitos de trabajo, que implementa esta tecnología y los diferentes aspectos evaluados para definir mejoras tanto a nivel del software como a nivel físico.

3 Ciberseguridad y aplicaciones IoT

Debido a la importancia que ha cobrado estos últimos años el control de la calidad del aire en ambientes cerrados o pocos ventilados, han surgido desarrollos que automatizan el control de estos valores permitiendo establecer un ambiente seguro de trabajo. Este servicio fue pensado construyendo un sistema con un dispositivo IoT junto con un sensor de dióxido de carbono en cuanto al hardware y un desarrollo de software que acompañe la visualización y registro histórico de los datos.

El escenario que detallaremos se compone de un dispositivo IoT que realiza mediciones del entorno, una red que tiene conexión a Internet y diferentes dispositivos (computadoras y equipos móviles) que interactúan para permitir la visualización de los datos del dispositivo a través de un portal web.

[1] Blue Team: grupo de especialistas en seguridad que rastrean ciber incidentes y realizan análisis de los sistemas para garantizar la seguridad, identificar posibles fallos y verificar la efectividad de cada medida.

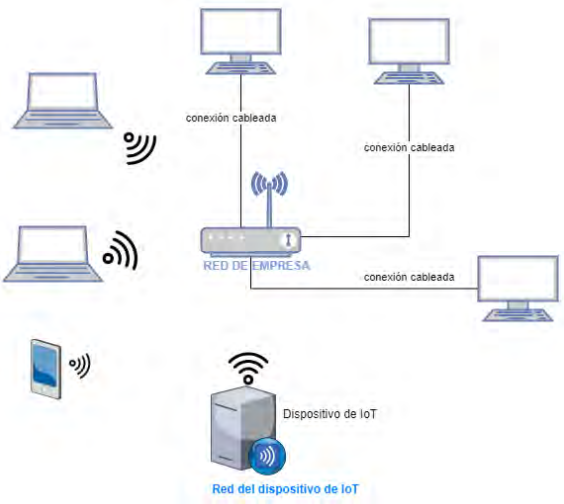


Fig 1. Arquitectura del escenario

3.1 Arquitectura del prototipo

El prototipo que permite sensar datos del ambiente e informarlos, como así también generar una alarma en función de la configuración inicial está compuesto por:

- Una placa NodeMCU[15].
- Un sensor CO2.
- Un servidor web.

En forma general, el rol de la placa es sensar los datos de ambiente y, en función de un programa interno, generar una alarma cuando los valores no son los normales. Además todos los datos censados son enviados por medio de una conexión wireless a un servidor web..

La placa NodeMCU es muy utilizada en dispositivos IoT por su tamaño, facilidades de uso y bajo costo. Una de las ventajas de esta placa es que tiene incorporado el microchip ESP8266 System-on-Chip (SoC) que simplifica la conexión y el procesamiento local de la comunicación. El microchip ESP8266 es un microcontrolador capaz de ejecutar código de forma local en varios lenguajes de programación y habilita la conexión a redes inalámbricas desde el programa que se grabe. Este es un módulo Wi-Fi pequeño, utilizado para establecer una conexión entre un microcontrolador o procesador y una red inalámbrica. El mismo puede trabajar tanto como un sistema independiente como conectado a otras redes inalámbricas disponibles. Se entiende un sistema independiente a la capacidad de disponer de una red inalámbrica propia a la cual se pueden conectar dispositivos.

En el caso particular de este proyecto se utilizan ambos modos disponibles para diferentes propósitos:

- Levantar una red propia (con su propio SSID) que permite generar las configuraciones iniciales necesarias que detallaremos más adelante.

- Para conectarse a una red externa que permite acceder a Internet y mandar los datos al registro externo o servidor web.

Al momento de evaluar las placas disponibles para el sistema, el hecho de que el firmware y las herramientas de desarrollo sean open source es importante para el desarrollo y adaptación a los proyectos propios. Otras de las características relevantes es la interfaz USB a UART, la cual permite comunicarse mediante USB con el SoC. Posee a su vez un circuito capaz de regular tensiones de entrada desde 4.8V hasta 12V, a un valor de operación de 3.3V. Por último, se puede mencionar que cuenta con pulsadores que permiten reiniciar o borrar la memoria del módulo en casos de fallas del sistema.

Otra de las partes del prototipo es el sensor de CO2 que permite detectar los valores del ambiente, también tiene la capacidad de ser calibrado para establecer los valores normales de referencia. Y, por último, el servidor web cuenta con un portal que permite acceder a los datos provistos por el sensor, como así también tiene la posibilidad de guardar un histórico de los niveles captados en un registro externo. El acceso a los datos es a través de un usuario y clave para seguridad del acceso.

3.2 Funcionamiento general del sistema

Los dispositivos que se utilizan para generar alguna alarma en función de los valores medidos del entorno requieren que se indique cuáles son los valores que son normales al momento de sensar. Por lo cual una de las primeras acciones necesarias es realizar una calibración antes de su utilización ya que, a través de luces externas se indican los posibles estados según los valores tomados en las mediciones. Dicho dispostivo posee tres luces para indicar el estado del ambiente, verde (aún está ventilado), amarillo (es una alerta de que pasó los niveles bajos) y rojo (el ambiente no está correctamente ventilado, puede ser el aire estancado o viciado que aumenta el riesgo de contagios en caso de COVID).

En el caso que los valores censados por el dispositivo sobrepasen el umbral establecido, es decir que los niveles de toxicidad del entorno pueden llegar a generar complicaciones para la salud de las personas, se encenderá una alarma sonora. Esta alarma tiene la capacidad de poder desactivarse en forma manual para evitar ruido constante por medio de un botón. En el caso que se mantenga apretado una cantidad de segundos determinada, habilita la configuración para calibrar el sensor de CO2.

El puerto USB de la placa es utilizado para la programación del mismo, a través del cual se graba el software que realizará el control de los valores del ambiente y enviará la información al servidor web. Además de permitir la carga del programa provee la posibilidad de cargar el dispositivo eléctricamente.

Hay dos formas de realizar la conexión con el sistema de medición de CO2 para obtener los datos de las mediciones. En la primera de ellas, el dispositivo IoT puede conectarse a la red como un host más para habilitar el acceso a la información recolectada a través de un servicio web. Si la red tiene Internet, es posible configurar un servidor para que el dispositivo le envíe la información de las mediciones. En la segunda opción, el dispositivo IoT puede abrir su propia red Wi-Fi para publicar los

datos de las mediciones sin necesidad de que el usuario ni el dispositivo tengan que estar conectados a una red externa. De este modo, la nueva red Wi-Fi con nombre RedA aloja el sitio web a través de la IP privada del dispositivo para que el usuario pueda visualizar los datos de las mediciones.

Para poder configurar el dispositivo, se puede abrir una nueva red Wi-Fi, denominada Config_RedA que permite al usuario modificar la red Wi-Fi a la que se conecta el dispositivo para disponibilizar la información, configurar la red Wi-Fi propia o reiniciar el dispositivo con alguna configuración anterior.

4 Análisis

Como se detalló anteriormente los entornos con dispositivos IoT que intercambian información crítica para el entorno pueden ser vulnerables a modificaciones, para lo cual se realizó un análisis de posibles vulnerabilidades tanto del hardware como del software.

En una primera etapa se analizaron las posibles vulnerabilidades de hardware, en este sentido se consideraron los aspectos físicos del dispositivo. La evaluación incluyó las consecuencias que provocarían las modificaciones y el comportamiento del dispositivo. De dicha evaluación se tuvieron en cuenta aspectos tales como el acceso a la recalibración, la modificación del programa que permite sensar, acceso físico al botón de apagado de la alarma.

Se encontraron cuatro vulnerabilidades de hardware, dos de ellas relacionadas con el modo en el que se calibra el dispositivo lo que permite que los valores límites sean erróneos y por consecuencia se encienda la luz indicadora incorrecta. Las otras dos vulnerabilidades tienen que ver con puertos y botones específicos del dispositivo que pueden ser manipulados si un atacante tiene acceso físico al mismo.

Estos casos de vulnerabilidades provocarían las siguientes consecuencias:

- En caso que se recalibre el dispositivo para que tome niveles concentrados de CO2 como niveles normales podría provocar que no detecte saturación del aire cuando debería hacerlo. Este caso se conoce como un falso negativo dado que no genera una alerta porque considera como normales valores que no corresponden.
- En caso que se recalibre el dispositivo para que tome niveles demasiado bajos de CO2 como niveles normales podría provocar que la alarma suene, aunque los niveles de CO2 no sean altos. Este caso se considera un falso positivo, dado que estaría alertando que los niveles de dióxido son peligrosos cuando no lo son.

En cuanto al acceso físico a diferentes periféricos del dispositivo las posibles consecuencias son:

- El acceso al puerto USB podría provocar la reescritura del programa dado, que al estar a simple vista, y no solicitar contraseña ni validación para hacerlo, facilita su uso malicioso para re-configurar o cambiar el código fuente. En caso que esta acción se realice podría provocar que el dispositivo

no realice las funcionalidades para lo cual fue programado, o en su defecto con un funcionamiento erróneo o malintencionado.

- En el caso del botón de silenciado que apaga la alarma al instante, una persona podría apagarla de forma casi instantánea y de esta forma las personas presentes en el ambiente no advertirían del peligro de la situación.

En una segunda etapa se realizó el análisis de las vulnerabilidades de software, para lo cual se evaluaron los protocolos y servicios utilizados por el dispositivo mediante capturas de tráfico de red a través de técnicas de sniffing. Durante el análisis, se evaluaron diferentes aspectos, tales como los protocolos de la capa de aplicación utilizados, la seguridad de las contraseñas y el método de autenticación para habilitar la red de configuración.

Cabe mencionar que la cantidad de tráfico presente en los entornos de uso de los dispositivos y la frecuencia en que se configura los mismos pueden dificultar el acceso al tráfico de un dispositivo en particular.

En esta etapa al momento de investigar el modo de autenticación, se pudo determinar a través de las capturas que se utilizaba el protocolo WPA2 para las redes propias de Wi-Fi (RedA y Config_RedA) y el protocolo WWW-Authenticate [16], a través de un popup desde el sitio web, para poder ingresar a la red de configuración del dispositivo IoT para hacer modificaciones. En caso de que las credenciales ingresadas por el usuario sean las correctas, el dispositivo se desconectará de la red actual y abrirá la red de configuraciones.

Se encontraron cuatro vulnerabilidades de software, dos de ellas tienen que ver con el uso de credenciales por defecto. Las dos vulnerabilidades restantes se relacionan con el uso de protocolos inseguros.

El uso de credenciales por defecto se da tanto en el acceso a las redes propias como en el acceso a la configuración a través del protocolo WWW-Authenticate. Dichas contraseñas se encuentran establecidas y no es posible modificarlas. En el caso de las credenciales utilizadas en el protocolo WWW-Authenticate, se configuraron un usuario y contraseña comunes y no permiten su modificación desde el sistema, lo cual genera que sean fácilmente descubiertas en un ataque por fuerza bruta o diccionario. En el caso de las redes inalámbricas, la contraseña y nombres de las redes configuradas en el prototipo si bien no son comunes, serán las mismas utilizadas en todos los medidores de CO2. Esto trae como consecuencia que quien conoce o descubre la clave de uno de ellos puede acceder a todos los medidores, aunque no esté autorizado a ello.

Como se describió anteriormente estas configuraciones seleccionadas generan que sean más predecibles las contraseñas y permite su acceso a través de ataques por diccionario o fuerza bruta.

Respecto al uso de protocolos inseguros, el análisis se centró en el tipo de autenticación utilizado, el protocolo WWW-Authenticate base, que se implementa a través del intercambio de los datos dentro de las cabeceras del paquete, lo cual permitiría descifrar la contraseña por medio de un ataque de diccionario o fuerza bruta sobre dichas cabeceras. Por otra parte, al utilizar HTTP y WebSocket [17], protocolos que no encriptan la información, esto facilita el acceso a la información transmitida y por lo tanto a las credenciales de configuración.

5 Conclusiones

Este artículo presenta un caso particular de la implementación de dispositivos IoT, las características del prototipo, un análisis realizado con posibles vulnerabilidades y propuestas de mitigación. Además, se detallan la forma de comunicación entre las diferentes partes del escenario y las consecuencias que podrían suceder en los diferentes casos de vulnerabilidad. En cuanto a las vulnerabilidades de hardware se detallan los cuidados a nivel del dispositivo, su alcance físico y los periféricos que podrían generar intervenciones externas. Por lo tanto, se recomienda deshabilitar los puertos mediante un bloqueo lógico o físico, modificar el comportamiento del dispositivo cuando se interactúa con los botones y posicionar el dispositivo en un lugar que no sea de fácil acceso para el público, pero en el cual pueda seguir cumpliendo su funcionalidad.

Con respecto a las vulnerabilidades de software se recomienda utilizar protocolos seguros para prevenir los ataques más comunes, establecer una política de contraseñas fuertes que a través del mismo sistema fuercen al administrador el cambio de la misma luego de un determinado tiempo. Por último, el desarrollo del programa que se ejecuta localmente en la placa y lleva a cabo la configuración, reforzar el modo de autenticación utilizando protocolos que eviten el fácil acceso a la información transmitida.

A partir de esta investigación se puede concluir que existen similitudes y particularidades en el análisis y mitigación de vulnerabilidades de arquitecturas de IoT respecto a arquitecturas tradicionales. En cuanto a las metodologías y herramientas para detectar vulnerabilidades, es posible utilizar las mismas en ambos casos.

Si hablamos de detección de vulnerabilidades en sistemas de IoT, tenemos que tener presentes los mecanismos de seguridad implementados en los protocolos de transmisión de información, al igual que en otras arquitecturas. También resulta similar el análisis de la seguridad de los servicios, como ser el mecanismo de autenticación y la gestión de claves, así como la seguridad de su configuración por defecto. Mientras que, si pensamos en vulnerabilidades de los dispositivos y problemas de seguridad física, hay que tener en cuenta características inherentes a los sistemas de IoT, donde los dispositivos suelen residir en lugares públicos y abiertos, y además tener limitaciones de hardware que dificultan su protección y su configuración segura. Por otro lado, al momento de evaluar las medidas para lograr la mayor seguridad en los dispositivos, se debe tener en cuenta las capacidades de los dispositivos para la implementación de protocolos más seguros y cómo esto podría aumentar el retardo y el consumo de energía que provocan.

Referencias

[1]Smart cities: Background paper. GOV.UK. Recuperado 19 de julio de 2022, de https://www.gov.uk/government/publications/smart-cities-background-paper

[2]Forecast: The Internet of Things, Worldwide, 2013. (s. f.). Gartner. Recuperado 21 de julio de 2022, de https://www.gartner.com/en/documents/2625419

[3]Blanco-Novoa et al., (2017). An Open-Source IoT Power Outlet System for Scheduling Appliance Operation Intervals Based on Real-Time Electricity Cost. 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.141

[4]Díaz, F. J. et al., (2017, abril). Estrategias de IOT para lograr ciudades digitales seguras, más inclusivas y sustentables. XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires). http://sedici.unlp.edu.ar/handle/10915/62410

[5]Díaz, F. J. et al., (2021). Investigación en ciberseguridad en un año de pandemia. XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja). http://sedici.unlp.edu.ar/handle/10915/120528

[6]Monzón, G. et al., (2019). Modelo de seguridad IoT. XXV Congreso Argentino de Ciencias de la Computación (CACIC) (Universidad Nacional de Río Cuarto, Córdoba, 14 al 18 de octubre de 2019). http://sedici.unlp.edu.ar/handle/10915/91363

[7]Pertini, B. (2017). Análisis y mejoras de seguridad a una aplicación prototipo en IoT [Tesis, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/72059

[8] Presentación del Proyecto "Hackeamos para construir robot seguros" https://www.youtube.com/watch?v=XHICnjsjcVs. Último acceso 27 de julio.

[9] Proyecto del LINTI subsidiado por la OEA, año 2021-2022. https://www.youtube.com/watch?v=AuhPA45LYkQ&t=7s

[10] Antonio Liñán Colina, Alvaro Vives, Marco Zennaro, Antoine Bagula, Ermanno Pietrosemoli. Internet of Things IN 5 DAYS. https://archive.org/details/IoT5days. Último acceso 25 de Julio de 2022.

[11] OWASP (Web Open Application Security Project) www.owasp.org. Último acceso: 21 de Julio de 2022.

[12] NIST Cybersecurity for IoT Program. https://www.nist.gov/itl/applied-cybersecurity/nist-cybersecurity-iot-program. Último acceso: 21 de Julio de 2022.

[13]CIS Controls v8 Internet of Things Companion Guide White paper. Recuperado 21 de julio de 2022, de https://www.cisecurity.org/white-papers/cis-controls-v8-internet-of-things-companion-guide/

[14]IoT Security Foundation. https://www.iotsecurityfoundation.org/. Último acceso 27 de julio de 2022.

[15]NodeMCU Official Website. Disponible online: http://www.nodemcu.com. Último acceso: 21 de Julio de 2022.

[16]WWW-Authenticate. https://developer.mozilla.org/es/docs/Web/HTTP/Headers/WWW-Authenticate. Último acceso: 1 de Agosto de 2022

[17]Request for Comments: 6455. The WebSocket Protocol. https://datatracker.ietf.org/doc/html/rfc6455. Último acceso: 1 de Agosto de 2022.

Generador Binario Pseudoaleatorio Basado en la Combinación de Registros de Desplazamiento con Retroalimentación Lineal, mediante Funciones por Mayoría

Andrés Francisco Farías – Andrés Alejandro Farías

Departamento Académico de Ciencias Físicas, Matemáticas y Naturales
Universidad Nacional de La Rioja, La Rioja. Argentina
(afarias665@yahoo.com.ar, andres_af86@hotmail.com)

Abstract
El presente documento expone el procedimiento de construcción de un Generador binario pseudoaleatorio basado en la combinación de registros de desplazamiento con retroalimentación lineal (Linear Feedback Shift Register, LFSR). El proceso incluye la descripción del modelo, la estructura de cada generador, selección de las funciones booleanas que cuenten con las mejores propiedades criptográficas, la definición de la combinación final. Por último, para verificar la aleatoriedad de las secuencias obtenidas, se aplican a las mismas un conjunto de pruebas estadísticas de aleatoriedad.

Keywords: LFSR, cipher, key, Boolean function, non-linearity

1 Introducción

Un requisito fundamental de ese tipo de generadores se relaciona con la calidad de la secuencia generada. Entre otras características se exige imprevisibilidad y facilidad de implementación, pero, fundamentalmente un período con una longitud significativa.

Es en esos términos que se propone un modelo que responda a tales exigencias. La modalidad elegida se basa en la combinación no lineal de secuencias producidas por cuatro LFSR [1], [2].

El procedimiento de construcción de un generador pseudoaleatorio de ese estilo requiere de varias etapas:

- Definición esquemática del modelo.
- Función por mayoría.
- Elección de los distintos LFSR.
- Selección de funciones booleanas de cuatro variables en base a sus propiedades criptográficas.
- Conformación del generador con los componentes ya seleccionados.
- Clave y el procedimiento para generar los estados iniciales de los LFSR.

- Elección de las pruebas estadísticas a utilizar y los criterios de análisis de los resultados.
- Puesta en funcionamiento y realización de las pruebas de aleatoriedad necesarias sobre las secuencias obtenidas.

2 Definición esquemática del modelo

El generador propuesto en este trabajo, está conformado por cuatro LFSR, que tienen cada uno, dos funciones booleanas de filtrado no lineal, que producen secuencias binarias, las que luego son combinadas mediante funciones mayoría. Las secuencias obtenidas alimentan dos funciones booleanas de cuatro variables, cuyos resultados son sometidos a una operación XOR, según la figura 1:

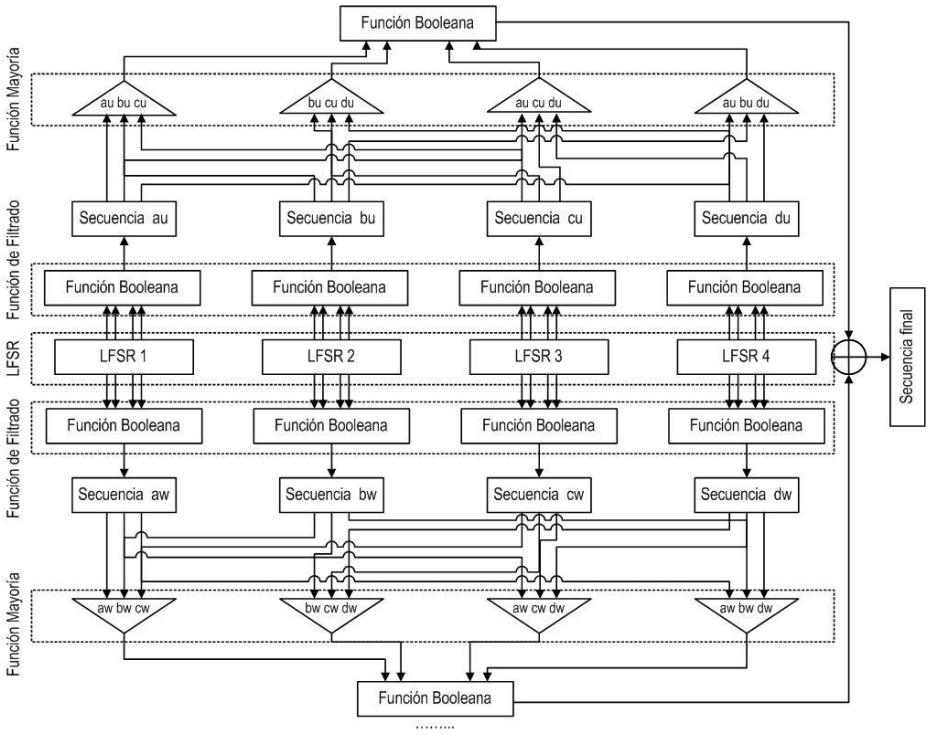


Fig. 1. Esquema generador binario pseudoaleatorio

3 Función por mayoría

En el esquema se indica la combinación de secuencias mediante función por mayoría, el número de secuencias debe ser impar y el valor binario es el que más se repite:

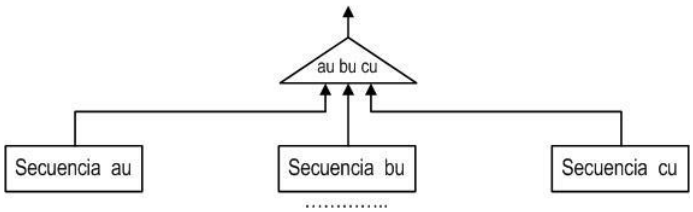


Fig. 2. Función por mayoría

4 Elección de los distintos LFSR

Las longitudes y polinomios primitivos de cada LFSR, que componen el generador, son las siguientes [3], [4], [5].

Tabla 1. LFSR, longitudes y polinomios primitivos del Generador

LFSR	Longitud	Polinomios primitivos
1	47	$P(x) = x^{47} + x^{32} + x^{24} + x^{11} + 1$
2	61	$P(x) = x^{61} + x^{57} + x^{26} + x^{3} + 1$
3	59	$P(x) = x^{59} + x^{54} + x^{46} + x^{26} + 1$
4	53	$P(x) = x^{53} + x^{50} + x^{41} + x^{20} + 1$

5 Selección de las funciones booleanas

5.1 Propiedades criptográficas deseables .

A continuación se indican algunas de las propiedades criptográficamente más significativas, adoptadas para este trabajo [6], [7], [8].

- **Función Balanceada**: Una función booleana de n-variables $\boldsymbol{f}$ es balanceada si $\boldsymbol{w(f) = 2n - 1}$. Esta propiedad es deseable para evitar ataques criptodiferenciales. La función es balanceada cuando el primer coeficiente del espectro de Walsh-Hadamard, es igual a cero: $\boldsymbol{F(0) = 0}$.
- **No Linealidad**: Valores altos de esta propiedad reducen el efecto de los ataques por criptoanálisis lineal. La No Linealidad de una función booleana puede ser calculada con la transformada de Walsh-Hadamard, $NL_f = \frac{1}{2} \cdot (2^n - |WH_{max}(f)|)$
- **Grado Algebraico**: El grado algebraico de una función, es el número de entradas más grande que aparece en cualquier producto de la Forma Normal Algebraica. Es deseable que sean valores altos.
- **SAC**: El Criterio de Avalancha Estricto requiere los efectos avalancha de todos los bits de entrada. Una función booleana se dice que satisface SAC sí y solo sí, la Ecuación 3, es balanceada para toda u con w(u)=1, $f(x) \oplus f(x \oplus u)$

Siguiendo los criterios arriba indicados las funciones booleanas aceptadas, son:

Tabla 2. Funciones de cuatro variables adoptadas

f_{NAF}
$f_{5775} = a \oplus b \oplus a \cdot b \oplus a \cdot c \oplus a \cdot d$
$f_{4722} = a \oplus b \oplus a \cdot c \oplus b \cdot c \oplus c \cdot d$
$f_{4529} = a \oplus c \oplus a \cdot c \oplus b \cdot c \oplus c \cdot d$
$f_{4393} = a \oplus c \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$
$f_{2402} = b \oplus a \cdot c \oplus b \cdot c \oplus d \oplus c \cdot d$
$f_{3981} = a \oplus a \cdot c \oplus b \cdot c \oplus d \oplus c \cdot d$
$f_{3338} = a \oplus a \cdot b \oplus c \oplus b \cdot c \oplus b \cdot d$
$f_{3672} = a \oplus a \cdot b \oplus c \oplus a \cdot c \oplus a \cdot d$
$f_{5911} = a \oplus b \oplus a \cdot b \oplus b \cdot c \oplus b \cdot d$
$f_{5056} = a \oplus b \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$

6 Conformación del generador combinacional

El generador combinacional queda de la siguiente manera:.

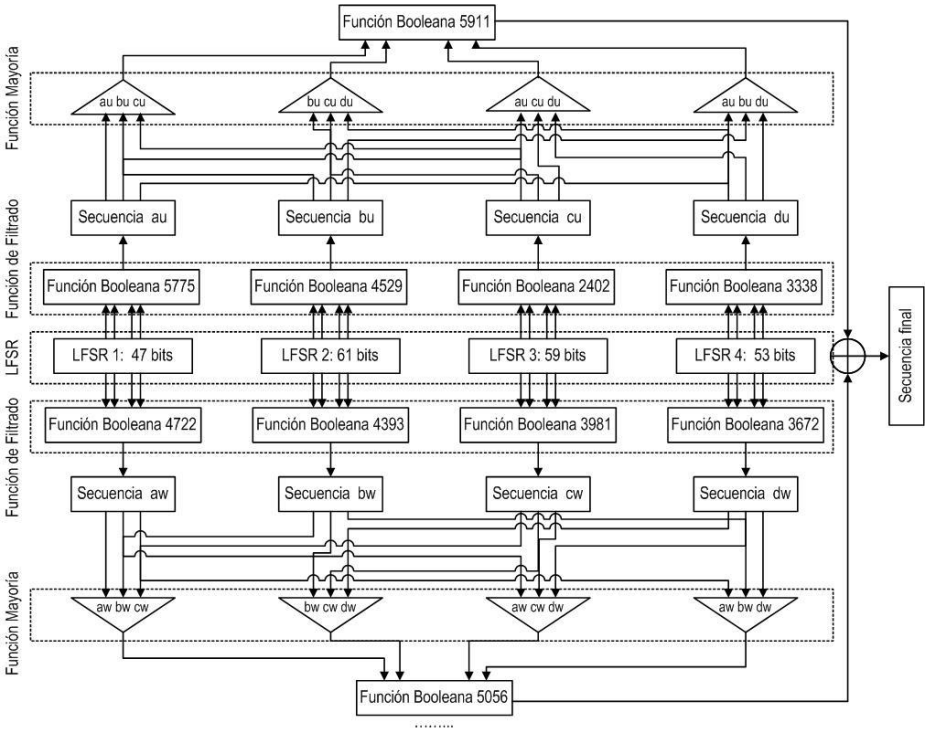


Fig. 3. Generador Combinacional

7 Clave

Para originar los estados iniciales de los distintos LFSR se realiza un proceso que utiliza una clave de 32 caracteres, que expresada en código ASCII (American Standard Code for Information Interchange), tiene longitud de 256 bits.

Se aceptan solamente las letras del alfabeto inglés (minúsculas y mayúsculas) y los números del sistema de numeración decimal, es decir un total de 62 caracteres.

La clave es sometida a un proceso criptográfico, que se indica en la Figura 5.

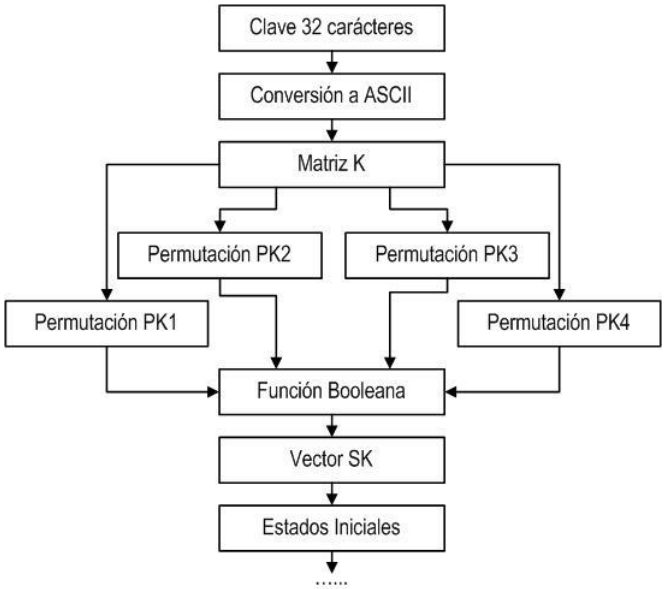


Fig. 4.. Clave para el generador

8 Permutaciones

8.1 Generador congruencial multiplicativo

El generador tiene la siguiente expresión: [9]

$$x_{i+1} = (a_x \cdot x_i)mod\ m_x \qquad (1)$$

Donde: $a_x = multiplicador, m_x = módulo, x_0 = semilla$

Tabla 3. Vectores, módulos, multiplicadores y semillas

Vector	módulo	multiplicador	semilla
PK1	1048576	1747	3249
PK2	1048576	1753	3271
PK3	1048576	1759	3301
PK4	1048576	1777	3347

8.2 Generación de los estados iniciales

La función booleana que procesa los cuatro vectores $PK1, PK2, PK3$ y PK4 es la siguiente: $MK = PK2 \oplus (PK1 \cdot PK3) \oplus (PK2 \cdot PK3) \oplus (PK1 \cdot K4) \oplus (PK2 \cdot PK4)$

De la operación resulta un vector SK[j] de 256 bits, que es el que proveerá los estados iniciales de los LFSR, en forma secuencial.

9 Elección de las pruebas estadísticas

Fueron seleccionadas algunas pruebas de la Norma NIST Special Publication 800-22, del trabajo de Rukhin (et al.) [10].

9.1 Prueba de frecuencia

El propósito de esta prueba es determinar si el número de unos y ceros en una secuencia es aproximadamente el mismo que se espera de una secuencia verdaderamente aleatoria. La prueba evalúa la cercanía de la fracción de unos a ½, que es decir, el número de unos y ceros en una secuencia debe ser aproximadamente el mismo. Todas las pruebas posteriores dependen de la aprobación de esta prueba.

9.2 Prueba de frecuencia en un bloque

La meta de esta prueba es determinar si la frecuencia de unos en un bloque de M bits es aproximadamente $M / 2$, como se esperaría bajo un supuesto de aleatoriedad.

9.3 Prueba de rachas

Una racha de longitud k consta de exactamente k bits idénticos y está acotada antes y después con un poco del valor opuesto. El propósito de la prueba de rachas es determinar si el número de rachas unos y ceros de varias longitudes es lo esperado para una secuencia aleatoria.

9.4 Prueba de rachas de unos en un bloque

El fin de esta prueba es determinar si la longitud de la ejecución más larga de las dentro de la secuencia probada es consistente con la longitud de la serie más larga de las que cabría esperar en una secuencia aleatoria. Tenga en cuenta que una irregularidad en la longitud esperada de la serie más larga implica que también hay una irregularidad en la longitud de la serie más larga de ceros.

9.5 Prueba de sumas acumuladas

Determina si la suma acumulativa de las secuencias parciales que ocurren en la secuencia probada es demasiado grande o demasiado pequeña en relación con el comportamiento esperado de esa suma acumulada para secuencias aleatorias.

9.6 Prueba de entropía aproximada

El enfoque de esta prueba es la frecuencia de todas las posibles superposiciones patrones de m bits en toda la secuencia. El propósito de la prueba es comparar la frecuencia de bloques superpuestos de dos longitudes consecutivas / adyacentes (m, m + 1) contra el resultado esperado para un secuencia aleatoria.

10 Pruebas sobre el generador

Se analizaron cien secuencias binarias, obtenidas del generador a partir de cien claves distintas.

El nivel de significancia adoptado para las pruebas estadísticas es de $\alpha = 0{,}01$. La hipótesis nula es:

$$H_0 \rightarrow p\_valor > 0{,}01$$

Debido al gran volumen de procesamiento requerido, se desarrolló un programa escrito en lenguaje C++, con los algoritmos correspondientes al generador y a las pruebas estadísticas. Es decir que el software calculó las secuencias binarias y simultáneamente realizó las pruebas sobre las mismas.

11 Interpretación de los resultados

Teniendo los resultados se realizan dos procesos para la interpretación de los mismos:

- Proporción de muestras que pasan las pruebas.
- Prueba de Uniformidad de los p-valor
    - Tabla de frecuencia e histograma
    - Prueba de Bondad de Ajuste

Se aplica la prueba de Bondad de Ajuste χ^2 aplicando la siguiente expresión:

$$\chi^2 = \sum_{i=1}^{10} \frac{\left(F_i - \frac{s}{10}\right)^2}{\frac{s}{10}} \qquad (2)$$

Donde: $F_i = \text{Frecuencia de la clase i}$ $s = \text{Cantidad de muestras}$

El primer procedimiento se realiza considerando los resultados de todas las pruebas y el segundo se realiza en forma individual. En todos los casos se deben superar todas las pruebas para aceptar los resultados.

11.1 Proporción de muestras que pasan las pruebas

Para el análisis de los resultados, se determina la proporción de muestras que superan las pruebas, y con esos datos se construye un gráfico de puntos, luego se verifica si los mismos caen dentro de los límites superior e inferior, donde k es el número de muestras.

$$LS, LI = (1-\alpha) \pm 3 \cdot \sqrt{\alpha \cdot (1-\alpha)/k} \qquad (3)$$

En nuestro caso k = 100 y el nivel de significancia elegido es: α = 0.01, los límites quedan:. $LS = 1{,}02$ y $LI = 0{,}96$

Se consideran todas pruebas, los resultados se indican en la tabla

Tabla 4. Pruebas

Pruebas	Proporción	Superior	Inferior
Frecuencias	0,98	1,02	0,96
Frecuencias en un Bloque	1,00	1,02	0,96
Rachas	1,00	1,02	0,96
Rachas de Unos en un Bloque	0,97	1,02	0,96
Sumas Acumuladas Adelante	0,98	1,02	0,96
Sumas Acumuladas Atrás	0,99	1,02	0,96
Entropía Aproximada	0,98	1,02	0,96

En el gráfico se aprecia el resultado, en definitiva la secuencia que entrega el generador supera las pruebas de aleatoriedad.

:

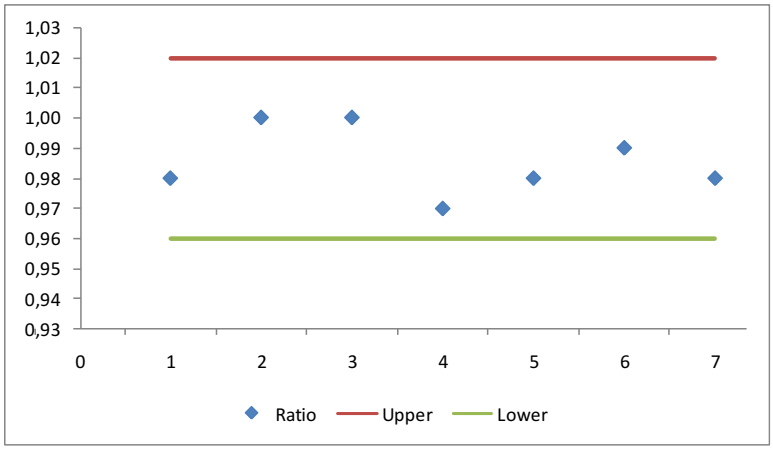


Fig. 5. Gráfico de puntos

11.2 Prueba de bondad de ajuste

Este control se ejecuta para cada prueba sobre las cien muestras, con los resultados de las frecuencias de p-valor obtenidos.

Tabla 5. Pruebas χ^2

Pruebas	χ^2	$\chi^2 ref$	Pasa
Frecuencias	0,658	0,0001	Sí
Frecuencias en un Bloque	0,172	0,0001	Sí

Rachas	0,679	0,0001	Sí
Rachas de Unos en un Bloque	0,817	0,0001	Sí
Sumas Acumuladas Adelante	0,043	0,0001	Sí
Sumas Acumuladas Atrás	0,720	0,0001	Sí
Entropía Aproximada	0,834	0,0001	Sí

11.3 Análisis final

En base a los resultados de las pruebas se realiza una tabla resumen.

Tabla 6. Análisis final

Análisis	Pruebas	Resultados
Proporción de secuencias que pasan las pruebas	Todas	Supera
Distribución uniforme de p-valor	Frecuencias	Supera
	Frecuencias dentro de un bloque	Supera
	Rachas	Supera
	La más larga racha de unos en un bloque	Supera
	Sumas acumuladas adelante	Supera
	Sumas acumuladas atrás	Supera
	Entropía estimada	Supera

En definitiva las secuencias que entrega el generador son pseudoaleatorias.

12 Conclusiones

La generación de bits aleatorios de alta calidad criptográfica resulta de alto interés, en consecuencia, se desarrolló un generador de secuencias binarias pseudoaleatorias de elevado período y complejidad lineal. Para ello se implementó un dispositivo que combina mediante función por mayoría, secuencias producidas por LFSR que sufren un filtrado no lineal con el auxilio de funciones booleanas

Los LFSR que componen cada generador tienen polinomios de conexión primitivos, lo que asegura un elevado período en la secuencia resultante.

La función booleana es la responsable del proceso no lineal, asegura las mejores prestaciones criptográficas, partiendo de criterios tales como ser balanceadas y tener alta no linealidad.

Realizado el proceso de selección, las funciones, las mismas fueron incorporadas al generador, que luego se puso en funcionamiento para generar las secuencias respectivas y con distintos valores de claves.

Los resultados fueron sometidos a un conjunto de pruebas de aleatoriedad, que mostraron valores positivos, por lo que el modelo propuesto se considera válido para la generación de secuencias pseudoaleatorias de buena calidad criptográfica.

13 Referencias

[1] Massodi, F., Alam, S. and Bokhari, M., "A Analysis of Linear Feedback Shift Registers in Stream Ciphers", International Journal of Computer Application, 16 (17), pp. 0975 – 887, 2012.

[2] Menezes, A.,Van Oorschot, P. and Vanstone, S., "Handbook of Applied Cryptography", Massachusetts Institute of Technology, 1996.

[3] Parr, C. and Pelzl, L., Understanding Cryptography, Springer, 2010.

[4] Stahnke, W., "Primitive Binary Polynomials", Mathematics of Computation, 27. 124, pp. 977-980, 1973.

[5] Seroussi, G., "Table of Low-Weight Binary Irreducible Polynomials", Computer Systems Laboratory, 1998.

[6] Clark, J., Jacob, J., Maitra, S., Stanica, P.: Almost Boolean Functions: The Design of Boolean Functions by Spectral Inversion. Computational intelligence. 20. (3), 450—462 (2004)

[7] Braeken, A.: Cryptographic Properties of Boolean Functions and S-Boxes. Faculteit Ingenieurswetenschappen. Katholieke Universiteit Leuven (2003)

[8] Elhosary, A., Hamdy, N., Farag, I., Rohiem, I.: State of the Art in Boolean Functions Cryptographic Assessment. International Journal of Computer Networks and Communications Security.1. (3), 88--94 (2013)

[9] Fishman, G.: Multiplicative Congruential Random Number Generators with Modulus 2ß : An Exhaustive Analysis for ß = 32 and a Partial Analysis for ß = 48. Mathematics of Computation. 54. (189), 33--344 (1990)

[10] Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S., "A Statistical Prueba Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", National Institute of Standards and Technology, (2000).

Generador Binario Pseudoaleatorio Basado en la Combinación de Registros de Desplazamiento con Retroalimentación Lineal, mediante Suma Real con Acarreo

Andrés Francisco Farías – Andrés Alejandro Farías

Departamento Académico de Ciencias Físicas, Matemáticas y Naturales
Universidad Nacional de La Rioja, La Rioja. Argentina
(afarias665@yahoo.com.ar, andres_af86@hotmail.com)

Abstract
El trabajo consiste en el desarrollo de un generador binario pseudoaleatorio basado en la combinación de registros de desplazamiento con retroalimentación lineal (Linear Feedback Shift Register, LFSR), mediante la suma real con acarreo de las secuencias de salida de las funciones de filtrado no lineal, que se alimentan de los registros de los LFSR. Se incluye la descripción del modelo, la estructura de cada generador, selección de las funciones booleanas que cuenten con las mejores propiedades criptográficas, la definición de la combinación final. Por último, para verificar la aleatoriedad de las secuencias obtenidas, se aplican a las mismas un conjunto de pruebas estadísticas de aleatoriedad.

Keywords: LFSR, cipher, key, Boolean function, non-linearity

1 Introducción

El generador se basa en LFSR [1], [2],de distintas longitudes, que tienen, cada uno, dos funciones de filtrado no lineal que se abastecen de las secuencias producidas por los mismos LFSR, después en grupo de a cuatro, esos resultados se combinan mediante un proceso de suma real con acarreo. De esto se obtienen dos secuencias que se someten a una operación de XOR, para obtener un resultado final, que es sometido luego a pruebas de aleatoriedad.

El desarrollo de un generador pseudoaleatorio de estas características requiere de varias etapas:

- Presentación del modelo.
- Selección de los distintos LFSR.
- Búsqueda de funciones booleanas de cuatro variables en base a sus propiedades criptográficas.
- Composición del generador con los componentes ya seleccionados.
- Clave y el procedimiento para generar los estados iniciales de los LFSR.

- Elección de las pruebas estadísticas a utilizar y los criterios de análisis de los resultados.
- Puesta en funcionamiento y realización de las pruebas de aleatoriedad necesarias sobre las secuencias obtenidas.

2 Definición esquemática del modelo

El generador propuesto en este trabajo, está conformado por cuatro LFSR, que tienen cada uno, dos funciones booleanas de filtrado no lineal, que producen secuencias binarias, las que luego son combinadas mediante un procedimiento de suma real con acarreo.. Las secuencias obtenidas son sometidas a una operación XOR, según la figura:

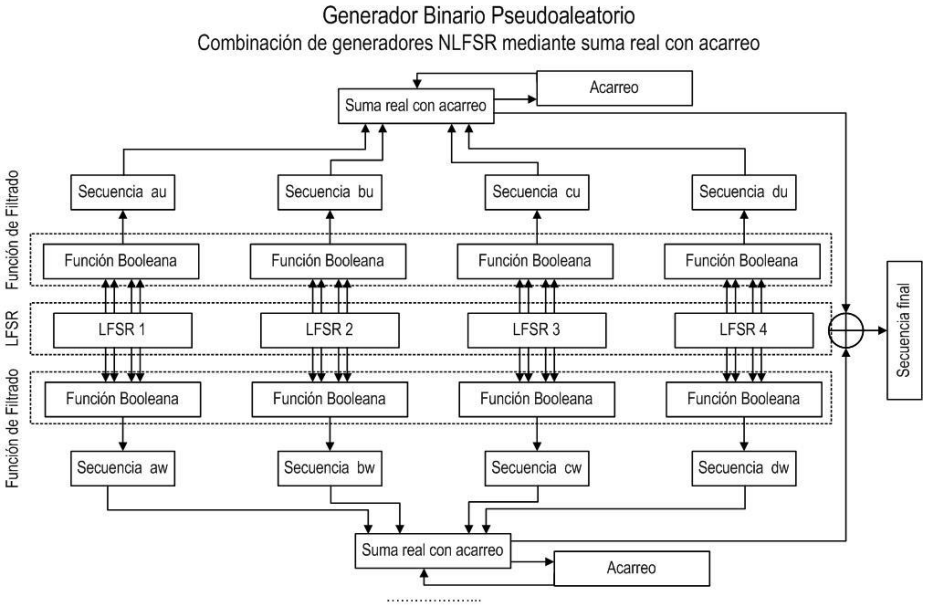


Fig. 1. Esquema generador binario pseudoaleatorio

3 Elección de los distintos LFSR

Las longitudes y polinomios primitivos de cada LFSR, que componen el generador, son las siguientes [3], [4], [5].

Tabla 1. LFSR, longitudes y polinomios primitivos del Generador

LFSR	Longitud	Polinomios primitivos
1	31	$P(x) = x^{31} + x^{25} + x^{23} + x^{8} + 1$
2	37	$P(x) = x^{37} + x^{22} + x^{14} + x^{2} + 1$

3	41	$P(x) = x^{41} + x^{32} + x^{31} + x^{27} + 1$
4	43	$P(x) = x^{43} + x^{27} + x^{22} + x^{5} + 1$

El generador propuesto en este trabajo, está conformado según la figura:

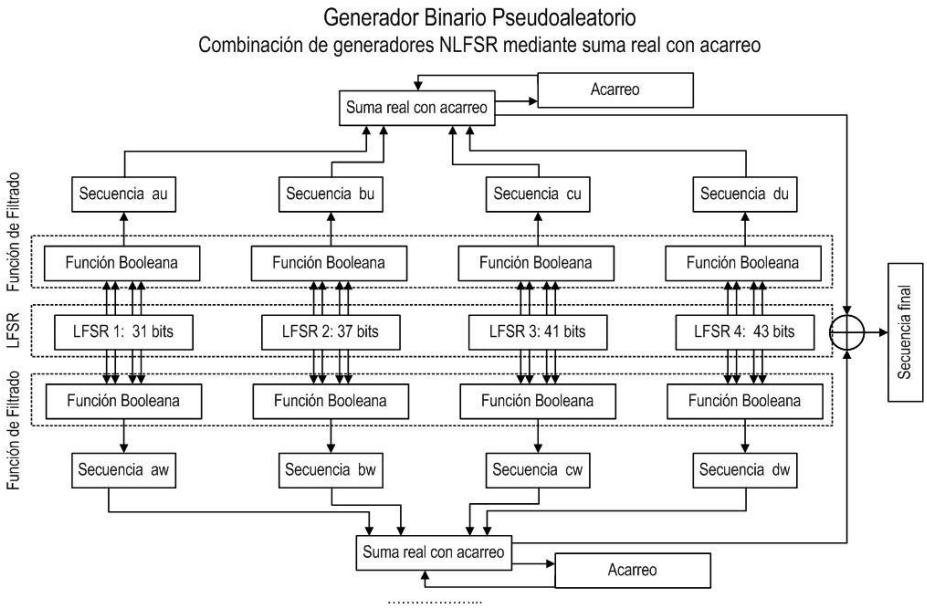


Fig. 2. Generador binario pseudoaleatorio

4 Selección de las funciones booleanas

4.1 Propiedades criptográficas deseables .

A continuación se indican algunas de las propiedades criptográficamente más significativas, adoptadas para este trabajo [6], [7], [8].

- **Función Balanceada**: Una función booleana de n-variables $\boldsymbol{f}$ es balanceada si $\boldsymbol{w(f) = 2n - 1}$. Esta propiedad es deseable para evitar ataques criptodiferenciales. La función es balanceada cuando el primer coeficiente del espectro de Walsh-Hadamard, es igual a cero: $\boldsymbol{F(0) = 0}$.
- **No Linealidad**: Valores altos de esta propiedad reducen el efecto de los ataques por criptoanálisis lineal. La No Linealidad de una función booleana puede ser calculada directamente de la transformada de Walsh-Hadamard, (Ecuación 2):

$$NL_f = \frac{1}{2} \cdot (2^n - |WH_{max}(f)|) \qquad (1)$$

- **Grado Algebraico**: El grado algebraico de una función, es el número de entradas más grande que aparece en cualquier producto de la Forma Normal Algebraica. Es deseable que sean valores altos.

- **SAC**: El Criterio de Avalancha Estricto requiere los efectos avalancha de todos los bits de entrada. Una función booleana se dice que satisface SAC sí y solo sí, la Ecuación 3, es balanceada para toda u con w(u)=1. $f(x) \oplus f(x \oplus u)$

4.2 Tabla de resultados.

Siguiendo los criterios arriba indicados las funciones booleanas aceptadas, son:

Tabla 2. Funciones de cuatro variables adoptadas

f_{NAF}
$f_{84} = a \cdot c \oplus b \cdot c \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$
$f_{89} = a \cdot c \oplus b \cdot c \oplus d \oplus a \cdot d \oplus b \cdot d$
$f_{100} = a \cdot c \oplus b \cdot c \oplus d \oplus a \cdot b \cdot d \oplus c \cdot d$
$f_{176} = c \oplus a \cdot c \oplus b \cdot c \oplus a \cdot d \oplus b \cdot d$
$f_{199} = c \oplus a \cdot c \oplus b \cdot c \oplus d \oplus c \cdot d$
$f_{381} = c \oplus a \cdot b \cdot c \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$
$f_{468} = c \oplus d \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$
$f_{536} = a \cdot b \oplus b \cdot c \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$

5 Conformación del generador combinacional

El generador combinacional queda de la siguiente manera:.

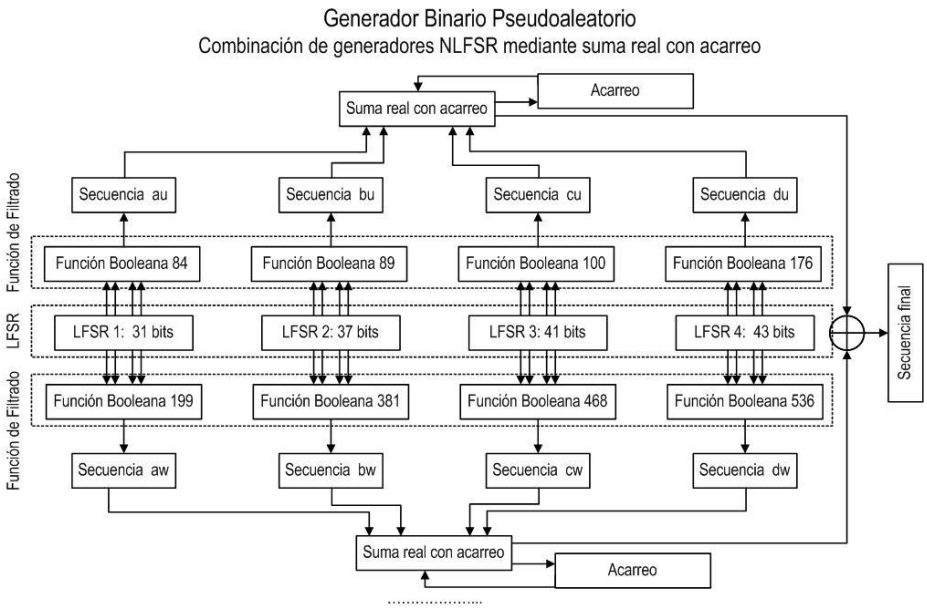


Fig. 3. Generador Combinacional

6 Clave

Para obetener los estados iniciales de los distintos LFSR se realiza un proceso que utiliza una clave de una longitud de 32 caracteres, que expresada en código ASCII (American Standard Code for Information Interchange), tiene longitud de 256 bits.

Para simplificar la introducción de la clave, se aceptan solamente las letras del alfabeto inglés (minúsculas y mayúsculas) y los números del sistema de numeración decimal, es decir un total de 62 caracteres.

La clave es sometida a un proceso criptográfico, que se indica en la Figura 4.

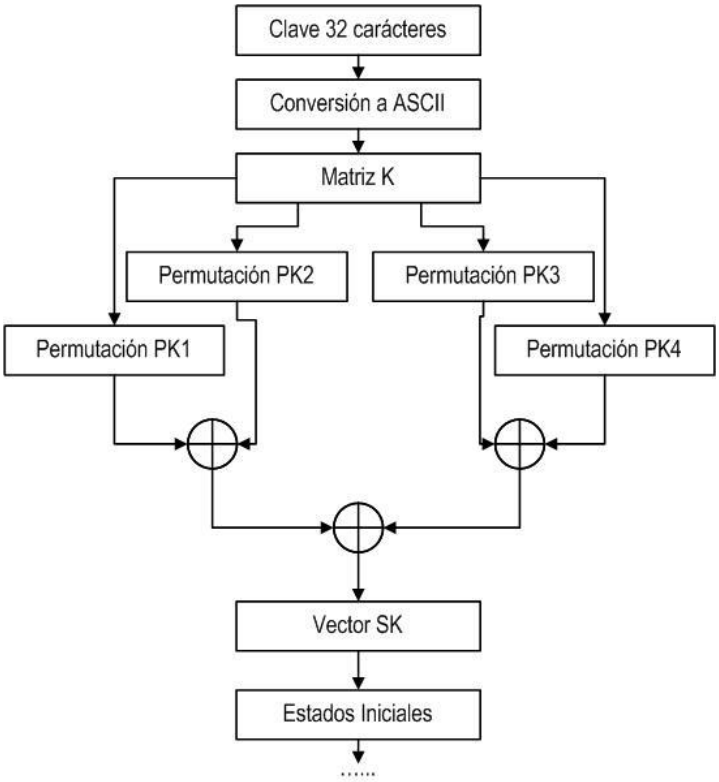


Fig. 4.. Generador para estados iniciales

7 Permutaciones

7.1 Generador congruencial multiplicativo

El generador tiene la siguiente expresión: [9]

$$x_{i+1} = (a_x \cdot x_i)mod\ m_x \qquad (2)$$

Donde: $a_x = multiplicador, m_x = módulo, x_0 = semilla$

Tabla 3. Vectores, módulos, multiplicadores y semillas

Vector	módulo	multiplicador	semilla
PK1	1048576	2741	3249
PK2	1048576	2749	3271
PK3	1048576	2753	3301
PK4	1048576	2767	3347

7.2 Generación de los estados iniciales

De la operación resulta un vector SK[j] de 256 bits, que es el que proveerá los estados iniciales de los LFSR, en forma secuencial.

8 Elección de las pruebas estadísticas

Fueron seleccionadas algunas pruebas de la Norma NIST Special Publication 800-22, del trabajo de Rukhin (et al.) [9].

8.1 Prueba de frecuencia

El propósito de esta prueba es determinar si el número de unos y ceros en una secuencia es aproximadamente el mismo que se espera de una secuencia verdaderamente aleatoria. La prueba evalúa la cercanía de la fracción de unos a ½, que es decir, el número de unos y ceros en una secuencia debe ser aproximadamente el mismo. Todas las pruebas posteriores dependen de la aprobación de esta prueba.

8.2 Prueba de frecuencia en un bloque

La meta de esta prueba es determinar si la frecuencia de unos en un bloque de M bits es aproximadamente $M / 2$, como se esperaría bajo un supuesto de aleatoriedad.

8.3 Prueba de rachas

Una racha de longitud k consta de exactamente k bits idénticos y está acotada antes y después con un poco del valor opuesto. El propósito de la prueba de rachas es determinar si el número de rachas unos y ceros de varias longitudes es lo esperado para una secuencia aleatoria.

8.4 Prueba de rachas de unos en un bloque

El fin de esta prueba es determinar si la longitud de la ejecución más larga de las dentro de la secuencia probada es consistente con la longitud de la serie más larga de las que cabría esperar en una secuencia aleatoria. Tenga en cuenta que una irregularidad en la longitud esperada de la serie más larga implica que también hay una irregularidad en la longitud de la serie más larga de ceros.

8.5 Prueba de sumas acumuladas

Determina si la suma acumulativa de las secuencias parciales que ocurren en la secuencia probada es demasiado grande o demasiado pequeña en relación con el comportamiento esperado de esa suma acumulada para secuencias aleatorias.

8.6 Prueba de entropía aproximada

El enfoque de esta prueba es la frecuencia de todas las posibles superposiciones patrones de m bits en toda la secuencia. El propósito de la prueba es comparar la frecuencia de bloques superpuestos de dos longitudes consecutivas / adyacentes (m, m + 1) contra el resultado esperado para un secuencia aleatoria.

9 Pruebas sobre el generador

Se analizaron cien secuencias binarias, obtenidas del generador a partir de cien claves distintas.

El nivel de significancia adoptado para las pruebas estadísticas es de $\alpha = 0{,}01$. La hipótesis nula es:

$$H_0 \rightarrow \text{p\_valor} > 0{,}01$$

Debido al gran volumen de procesamiento requerido, se desarrolló un programa escrito en lenguaje C++, con los algoritmos correspondientes al generador y a las pruebas estadísticas. Es decir que el software calculó las secuencias binarias y simultáneamente realizó las pruebas sobre las mismas.

10 Interpretación de los resultados

Teniendo los resultados se realizan dos procesos para la interpretación de los mismos:

- Proporción de muestras que pasan las pruebas.
- Prueba de Uniformidad de los p-valor
    - Tabla de frecuencia e histograma
    - Prueba de Bondad de Ajuste

Se aplica la prueba de Bondad de Ajuste χ^2aplicando la siguiente expresión:

$$\chi^2 = \sum_{i=1}^{10} \frac{\left(F_i - \frac{s}{10}\right)^2}{\frac{s}{10}} \qquad (3)$$

Donde: $F_i = \text{Frecuencia de la clase i} \quad s = \text{Cantidad de muestras}$

El primer procedimiento se realiza considerando los resultados de todas las pruebas y el segundo se realiza en forma individual. En todos los casos se deben superar todas las pruebas para aceptar los resultados.

10.1 Proporción de muestras que pasan las pruebas

Para el análisis de los resultados, se determina la proporción de muestras que superan las pruebas, y con esos datos se construye un gráfico de puntos, luego se verifica si los mismos caen dentro de los límites superior e inferior, donde k es el número de muestras.

$$LS, LI = (1 - \alpha) \pm 3 \cdot \sqrt{\alpha \cdot (1 - \alpha)/k} \qquad (4)$$

En nuestro caso $k = 100$ y el nivel de significancia elegido es: $\alpha = 0.01$, los límites quedan:. $LS = 1{,}02$ y $LI = 0{,}96$

Se consideran todas pruebas, los resultados se indican en la tabla

Tabla 4. Pruebas

Pruebas	Proporción	Superior	Inferior
Frecuencias	1,00	1,02	0,96
Frecuencias en un Bloque	0,97	1,02	0,96
Rachas	1,00	1,02	0,96
Rachas de Unos en un Bloque	0,97	1,02	0,96
Sumas Acumuladas Adelante	0,99	1,02	0,96
Sumas Acumuladas Atrás	1,00	1,02	0,96
Entropía Aproximada	1,00	1,02	0,96

En el gráfico se aprecia el resultado, en definitiva la secuencia que entrega el generador supera las pruebas de aleatoriedad.

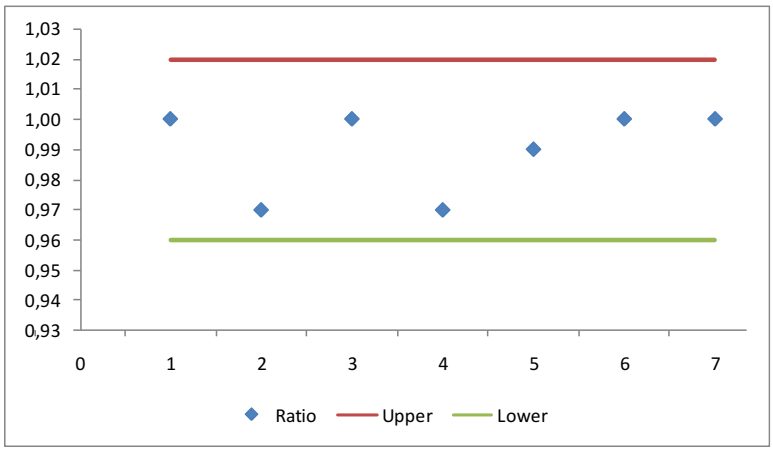


Fig. 5. Gráfico de puntos

10.2 Prueba de bondad de ajuste

Este control se ejecuta para cada prueba sobre las cien muestras, con los resultados de las frecuencias de p-valor obtenidos.

Tabla 5. Pruebas χ^2

Pruebas	χ^2	$\chi^2 ref$	Pasa
Frecuencias	0,936	0,0001	Sí
Frecuencias en un Bloque	0,319	0,0001	Sí
Rachas	0,924	0,0001	Sí
Rachas de Unos en un Bloque	0,384	0,0001	Sí
Sumas Acumuladas Adelante	0,401	0,0001	Sí
Sumas Acumuladas Atrás	0,456	0,0001	Sí
Entropía Aproximada	0,163	0,0001	Sí

10.3 Análisis final

En base a los resultados de la pruebas se realiza una tabla resumen.

Table 6. Análisis final

Análisis	Pruebas	Resultados
Proporción de secuencias que pasan las pruebas	Todas	Supera
Distribución uniforme de p-valor	Frecuencias	Supera
	Frecuencias dentro de un bloque	Supera
	Rachas	Supera
	La más larga racha de unos en un bloque	Supera
	Sumas acumuladas adelante	Supera
	Sumas acumuladas atrás	Supera
	Entropía estimada	Supera

En definitiva las secuencias que entrega el generador son pseudoaleatorias.

11 Conclusiones

El generador obtenido entrega secuencias binarias pseudoaleatorias de elevado período y complejidad lineal. Para ello se diseñó un dispositivo que combina en forma no lineal las secuencias producidas por cuatro LFSR, que cuentan con ocho funciones de filtrado no lineal, que luego se combinan mediante dos sumas reaesl con acarreo.

Los LFSR tienen polinomios de conexión primitivos, lo que asegura un elevado período en la secuencia resultante.

Las funciones booleanas de filtrado no lineal, son las responsables del proceso de entregar secuencias no lineales y aseguran las mejores prestaciones criptográficas, si cumplen determinados criterios. Realizado el proceso de selección, las funciones fueron incorporadas al generador y luego puestas a funcionar para generar las secuencias respectivas con distintos valores de claves y ser sometidas a las pruebas de aleatoriedad respectivas.

Las secuencias obtenidas, superaron todas las pruebas lo que demuestra que el generador funciona de acuerdo a lo previsto.

12 Referencias

[1] Massodi, F., Alam, S. and Bokhari, M., "A Analysis of Linear Feedback Shift Registers in Stream Ciphers", International Journal of Computer Application, 16 (17), pp. 0975 – 887, 2012.

[2] Menezes, A.,Van Oorschot, P. and Vanstone, S., "Handbook of Applied Cryptography", Massachusetts Institute of Technology, 1996.

[3] Parr, C. and Pelzl, L., Understanding Cryptography, Springer, 2010.

[4] Stahnke, W., "Primitive Binary Polynomials", Mathematics of Computation, 27. 124, pp. 977-980, 1973.

[5] Seroussi, G., "Table of Low-Weight Binary Irreducible Polynomials", Computer Systems Laboratory, 1998.

[6] Clark, J., Jacob, J., Maitra, S., Stanica, P.: Almost Boolean Functions: The Design of Boolean Functions by Spectral Inversion. Computational intelligence. 20. (3), 450—462 (2004)

[7] Braeken, A.: Cryptographic Properties of Boolean Functions and S-Boxes. Faculteit Ingenieurswetenschappen. Katholieke Universiteit Leuven (2003)

[8] Elhosary, A., Hamdy, N., Farag, I., Rohiem, I.: State of the Art in Boolean Functions Cryptographic Assessment. International Journal of Computer Networks and Communications Security.1. (3), 88--94 (2013)

[9] Fishman, G.: Multiplicative Congruential Random Number Generators with Modulus 2ß : An Exhaustive Analysis for ß = 32 and a Partial Analysis for ß = 48. Mathematics of Computation. 54. (189), 33--344 (1990)

[10] Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S., "A Statistical Prueba Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", National Institute of Standards and Technology, (2000).

Cifrador de Flujo Basado en un Generador Binario Pseudoaleatorio, con Clave de 256 Bits

Andrés Francisco Farías – Andrés Alejandro Farías

Departamento Académico de Ciencias Físicas, Matemáticas y Naturales
Universidad Nacional de La Rioja, La Rioja. Argentina
(afarias665@yahoo.com.ar, andres_af86@hotmail.com)

Abstract
El presente documento expone el procedimiento de construcción de un cifrador de flujo basado en un generador binario pseudoaleatorio conformado por la combinación no lineal de registros de desplazamiento con retroalimentación lineal con funciones no lineales de filtrado. Para verificar la aleatoriedad de las secuencias obtenidas, se aplican a las mismas un conjunto de pruebas estadísticas de aleatoriedad.

Keywords: NLFSR, LSFR, Clave, Período, Polinomios Primitivos, Pruebas de Aleatoriedad, XOR, Bits aleatorios

1 Introducción

Se trata de un dispositivo conformado por un generador de números binarios pseudoaleatorios, con una clave de 256 bits, basado en el uso de distintos registros de desplazamiento de reatroalimentación lineal (LFSR, sigla en inglés), combinados mediante funciones booleanas balanceadas y de alta no linealidad.

La secuencia cifrante binaria pseudoaleatoria entregada por este generador es sometida a una operación XOR, con la secuencia binaria de los carácteres del texto plano a cifrar en código ASCII binario, de esto se obtiene una nueva secuencia binaria, que es el texto cifrado en código ASCII binario [1], [2].

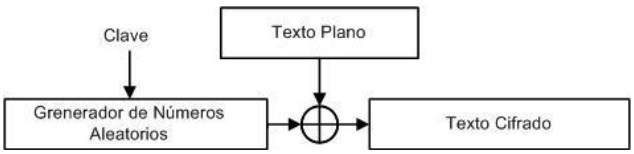


Fig. 1. Cifrador de Flujo

Para desencriptar se realizar una operación XOR entre el texto cifrado en código ASCII binario y la misma secuencia pseudoaleatoria binaria producida por el generador de números binarios pseudoaleatorios, con la que se realizó el cifrado. Que

da como resultado el texto plano en código ASCII binario.

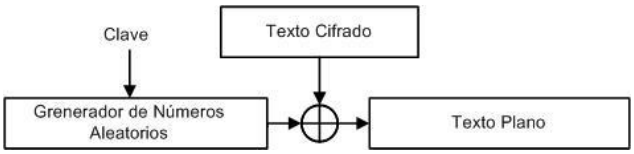


Fig. 2. Descifrador de Flujo

2 Definición del modelo para el generador pseudoaleatorio

Los componentes principales, son LFSR y las funciones de filtrado no lineal que son funciones booleanas de cuatro variables. Para el generador en estudio se dispone de tres de estos LFSR con dos funciones de filtrado no lineal, que entrega dos secuencias, se tiene que el conjunto produce seis secuencias aleatorias. La combinación se producen según el criterio parada y arranque; en la figura 3.

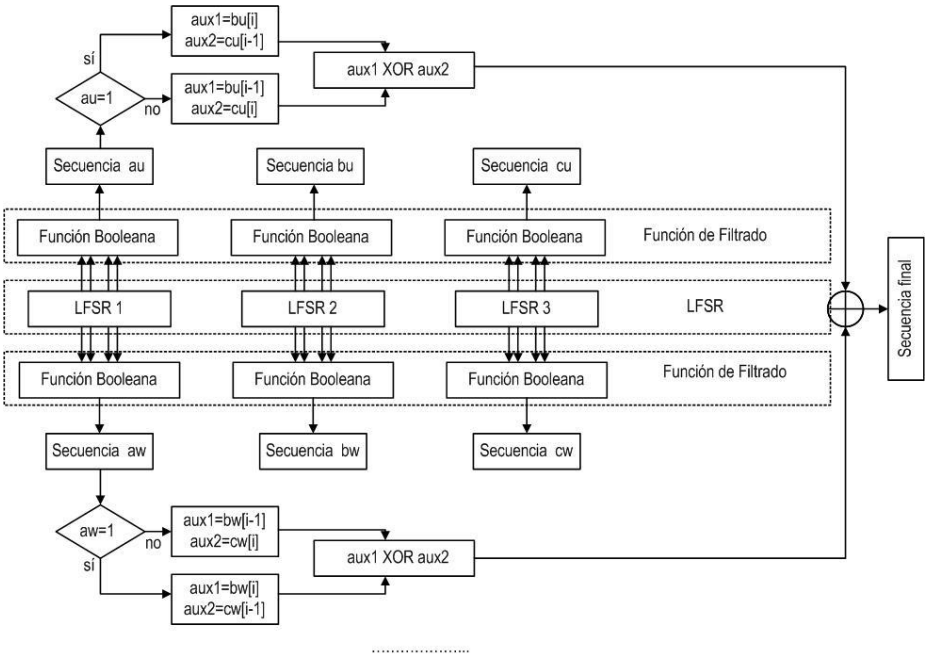


Fig. 3. Esquema generador aleatorio

3 Elección de los LFSR

Las longitudes y polinomios primitivos de cada LFSR, que componen el generador, son las siguientes [3], [4], [5].

Tabla 1. LFSR, longitudes y polinomios primitivos del Generador

LFSR	Longitud	Polinomios primitivos
1	47	$P(x) = x^{47} + x^{32} + x^{24} + x^{11} + 1$
2	61	$P(x) = x^{61} + x^{57} + x^{26} + x^{3} + 1$
3	59	$P(x) = x^{59} + x^{54} + x^{46} + x^{26} + 1$

El generador propuesto en este trabajo, está conformado según la figura:

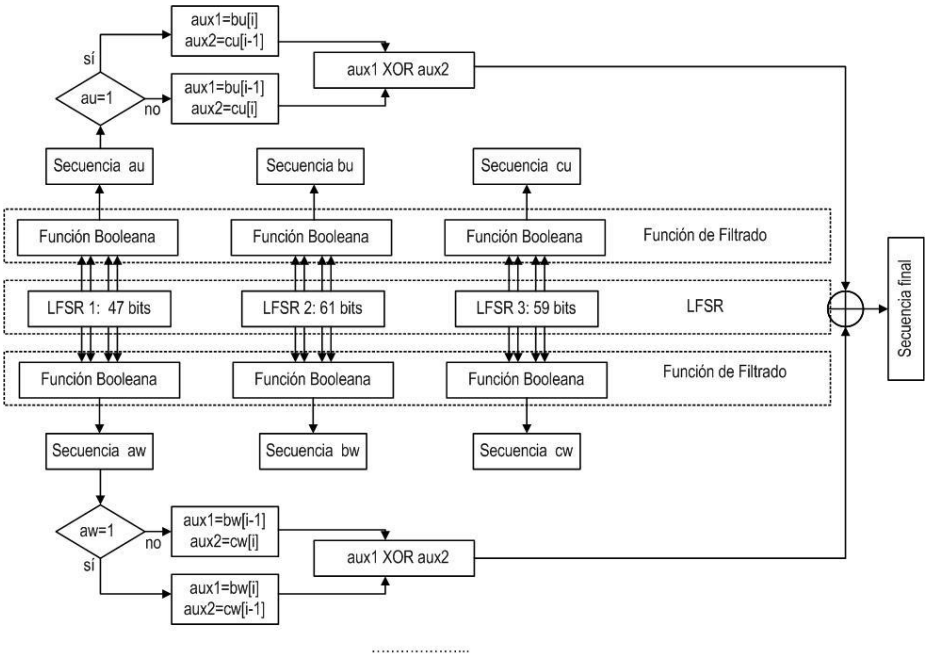


Fig. 4. Esquema generador aleatorio

4 Selección de las funciones booleanas

4.2 Propiedades criptográficas deseables

A continuación se indican algunas de las propiedades criptográficamente más significativas, adoptadas para este trabajo [6], [7], [8].

- **Función balanceada:** Esta propiedad es deseable para evitar ataques criptodiferenciales. La función es balanceada cuando el primer coeficiente del espectro de Walsh-Hadamard, es igual a cero: $F(0) = 0$
- **No linealidad:** Valores altos de esta propiedad reducen el efecto de los ataques por criptoanálisis lineal. La No Linealidad de una función booleana puede ser calculada directamente de la transformada de Walsh-Hadamard:

$$NL_f = 1/2 \cdot (2^n - |WH_{max}(f)|) \qquad (1)$$

- **Grado algebraico:** El grado algebraico de una función, es el número de entradas más grande que aparece en cualquier producto de la Forma Normal Algebraica. Es deseable que sean valores altos.
- **SAC:** El Criterio de Avalancha Estricto requiere los efectos avalancha de todos los bits de entrada. Una función booleana se dice que satisface SAC sí y solo sí, es balanceada para toda u con w(u)=1, para: $f(x \oplus u)$

Siguiendo los criterios arriba indicados las funciones booleanas aceptadas, son:

Tabla 2**.** Funciones de cuatro variables adoptadas

f_{NAF}
$f_{5775} = a \oplus b \oplus a \cdot b \oplus a \cdot c \oplus a \cdot d$
$f_{4722} = a \oplus b \oplus a \cdot c \oplus b \cdot c \oplus c \cdot d$
$f_{4529} = a \oplus c \oplus a \cdot c \oplus b \cdot c \oplus c \cdot d$
$f_{4393} = a \oplus c \oplus a \cdot d \oplus b \cdot d \oplus c \cdot d$
$f_{2402} = b \oplus a \cdot c \oplus b \cdot c \oplus d \oplus c \cdot d$
$f_{3981} = a \oplus a \cdot c \oplus b \cdot c \oplus d \oplus c \cdot d$

5 Conformación del generador combinacional

El generador combinacional queda de la siguiente manera:.

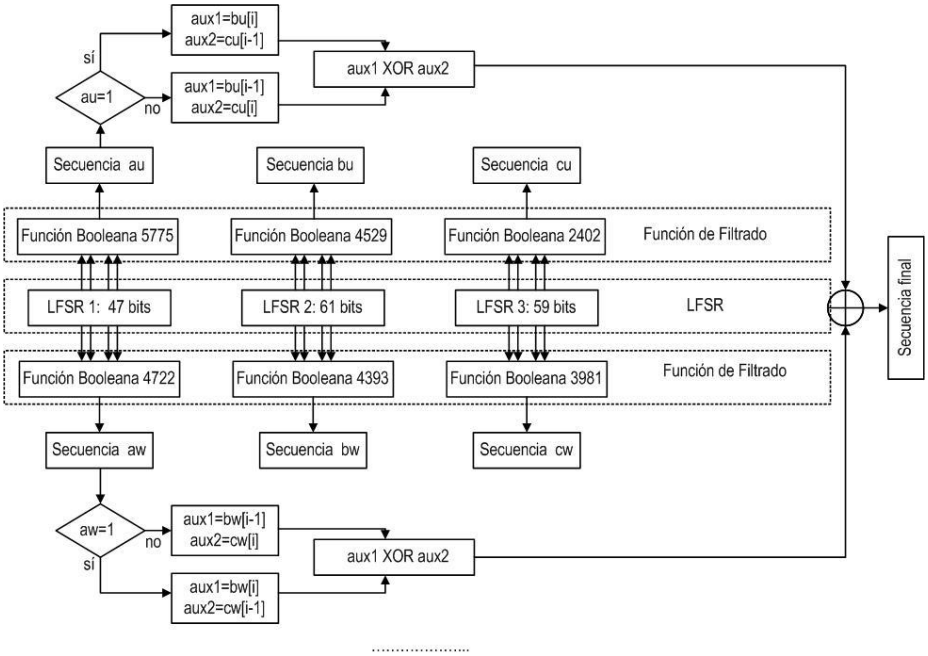


Fig. 5. Generador Combinacional

6 Clave

Para originar los estados iniciales de los distintos LFSR se realiza un proceso que utiliza una clave de una longitud de 32 caracteres, que expresada en código ASCII (American Standard Code for Information Interchange), tiene longitud de 256 bits.

Para simplificar el procedimiento de introducción de la clave, se aceptan solamente las letras del alfabeto inglés (minúsculas y mayúsculas) y los números del sistema de numeración decimal, es decir un total de 62 caracteres.

La clave es sometida a un proceso criptográfico, que se indica en la Figura 6.

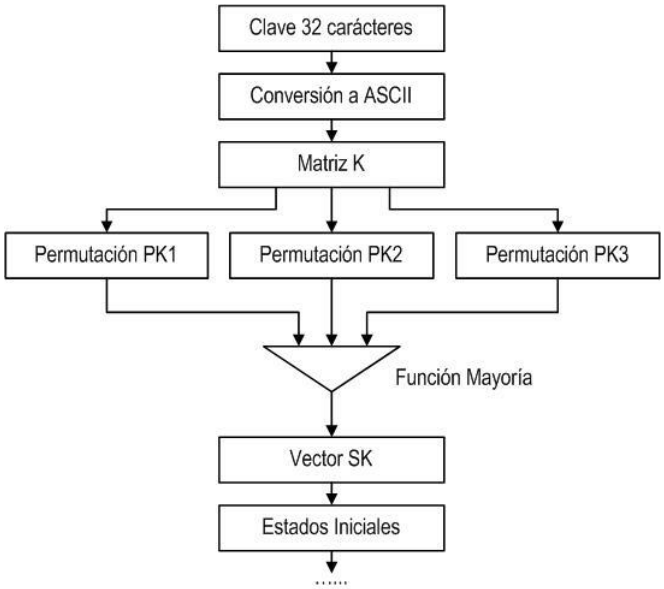


Fig. 6.. Clave del generador

7 Elección de las pruebas estadísticas

Fueron seleccionadas algunas pruebas de la Norma NIST Special Publication 800-22, del trabajo de Rukhin (et al.) [9].

7.1 Prueba de frecuencia

El propósito de esta prueba es determinar si el número de unos y ceros en una secuencia es aproximadamente el mismo que se espera de una secuencia verdaderamente aleatoria. La prueba evalúa la cercanía de la fracción de unos a ½, que es decir, el número de unos y ceros en una secuencia debe ser aproximadamente el mismo. Todas las pruebas posteriores dependen de la aprobación de esta prueba.

7.2 Prueba de frecuencia en un bloque

La meta de esta prueba es determinar si la frecuencia de unos en un bloque de M bits es aproximadamente $M / 2$, como se esperaría bajo un supuesto de aleatoriedad.

7.3 Prueba de rachas

Una racha de longitud k consta de exactamente k bits idénticos y está acotada antes y después con un poco del valor opuesto. El propósito de la prueba de rachas es determinar si el número de rachas unos y ceros de varias longitudes es lo esperado para una secuencia aleatoria.

7.4 Prueba de rachas de unos en un bloque

El fin de esta prueba es determinar si la longitud de la ejecución más larga de las dentro de la secuencia probada es consistente con la longitud de la serie más larga de las que cabría esperar en una secuencia aleatoria. Tenga en cuenta que una irregularidad en la longitud esperada de la serie más larga implica que también hay una irregularidad en la longitud de la serie más larga de ceros.

7.5 Prueba de sumas acumuladas

Determina si la suma acumulativa de las secuencias parciales que ocurren en la secuencia probada es demasiado grande o demasiado pequeña en relación con el comportamiento esperado de esa suma acumulada para secuencias aleatorias.

7.6 Prueba de entropía aproximada

El enfoque de esta prueba es la frecuencia de todas las posibles superposiciones patrones de m bits en toda la secuencia. El propósito de la prueba es comparar la frecuencia de bloques superpuestos de dos longitudes consecutivas / adyacentes (m, m + 1) contra el resultado esperado para un secuencia aleatoria.

8 Pruebas sobre el generador

Se analizaron cien secuencias binarias, obtenidas del generador a partir de cien claves distintas.

El nivel de significancia adoptado para las pruebas estadísticas es de $\alpha = 0{,}01$. La hipótesis nula es:

$$H_0 \rightarrow \text{p\_valor} > 0{,}01$$

Debido al gran volumen de procesamiento requerido, se desarrolló un programa escrito en lenguaje C++, con los algoritmos correspondientes al generador y a las pruebas estadísticas. Es decir que el software calculó las secuencias binarias y simultáneamente realizó las pruebas sobre las mismas.

9 Interpretación de los resultados

Teniendo los resultados se realizan dos procesos para la interpretación de los mismos:

- Proporción de muestras que pasan las pruebas.
- Prueba de Uniformidad de los p-valor
    - Tabla de frecuencia e histograma
    - Prueba de Bondad de Ajuste

Se aplica la prueba de Bondad de Ajuste χ^2aplicando la siguiente expresión:

$$\chi^2 = \sum_{i=1}^{10} \frac{\left(F_i - \frac{s}{10}\right)^2}{\frac{s}{10}} \qquad (2)$$

Donde: $F_i = \text{Frecuencia de la clase i}$ $s = \text{Cantidad de muestras}$

El primer procedimiento se realiza considerando los resultados de todas las pruebas y el segundo se realiza en forma individual. En todos los casos se deben superar todas las pruebas para aceptar los resultados.

9.1 Proporción de muestras que pasan las pruebas

Para el análisis de los resultados, se determina la proporción de muestras que superan las pruebas, y con esos datos se construye un gráfico de puntos, luego se verifica si los mismos caen dentro de los límites superior e inferior, donde k es el número de muestras.

$$LS, LI = (1 - \alpha) \pm 3 \cdot \sqrt{\alpha \cdot (1 - \alpha)/k} \qquad (3)$$

En nuestro caso $k = 100$ y el nivel de significancia elegido es: $\alpha = 0.01$, los límites quedan:. $LS = 1{,}02$ y $LI = 0{,}96$

Se consideran todas pruebas, los resultados se indican en la tabla

Tabla 3. Pruebas

Pruebas	Proporción	Superior	Inferior
Frecuencias	0,99	1,02	0,96
Frecuencias en un Bloque	1,00	1,02	0,96
Rachas	1,00	1,02	0,96
Rachas de Unos en un Bloque	1,00	1,02	0,96
Sumas Acumuladas Adelante	0,99	1,02	0,96
Sumas Acumuladas Atrás	0,99	1,02	0,96
Entropía Aproximada	0,99	1,02	0,96

Las secuencias que produce el cifrador superan las pruebas de aleatoriedad.

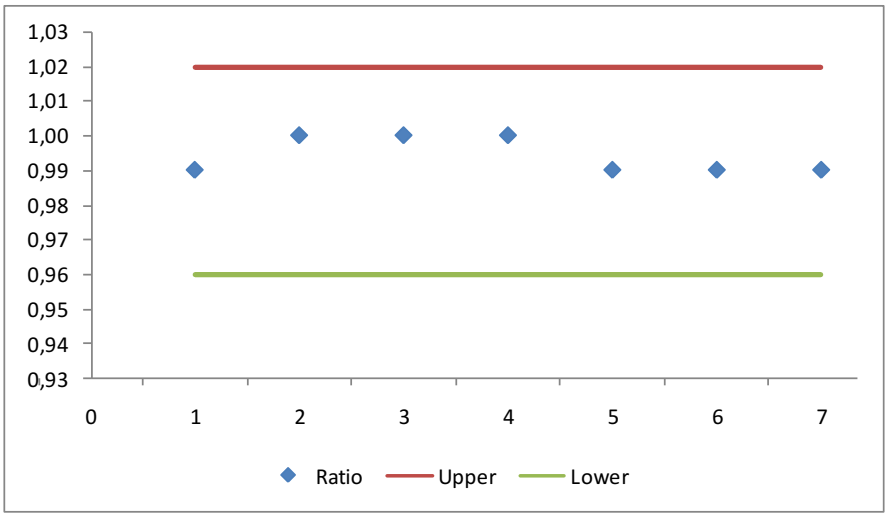


Fig. 7. Gráfico de puntos

9.2 Prueba de bondad de ajuste

Este control se ejecuta para cada prueba sobre las cien muestras, con los resultados de las frecuencias de p-valor obtenidos.

Tabla 4. Pruebas χ^2

Pruebas	χ^2	$\chi^2 ref$	Pasa
Frecuencias	0,225	0,0001	Sí
Frecuencias en un Bloque	0,936	0,0001	Sí
Rachas	0,720	0,0001	Sí
Rachas de Unos en un Bloque	0,335	0,0001	Sí
Sumas Acumuladas Adelante	0,262	0,0001	Sí
Sumas Acumuladas Atrás	0,798	0,0001	Sí
Entropía Aproximada	0,658	0,0001	Sí

9.3 Análisis final

En base a los resultados de la pruebas se realiza una tabla resumen.

Tabla 5. Análisis final

Análisis	Pruebas	Resultados
Proporción de secuencias que pasan las pruebas	Todas	Supera

Distribución uniforme de p-valor	Frecuencias	Supera
	Frecuencias dentro de un bloque	Supera
	Rachas	Supera
	La más larga racha de unos en un bloque	Supera
	Sumas acumuladas adelante	Supera
	Sumas acumuladas atrás	Supera
	Entropía estimada	Supera

En definitiva las secuencias que entrega el generador son pseudoaleatorias.

10 Comparación

10.1 Comparación de frecuencias

Diferencias entre texto plano y texto cifrado.

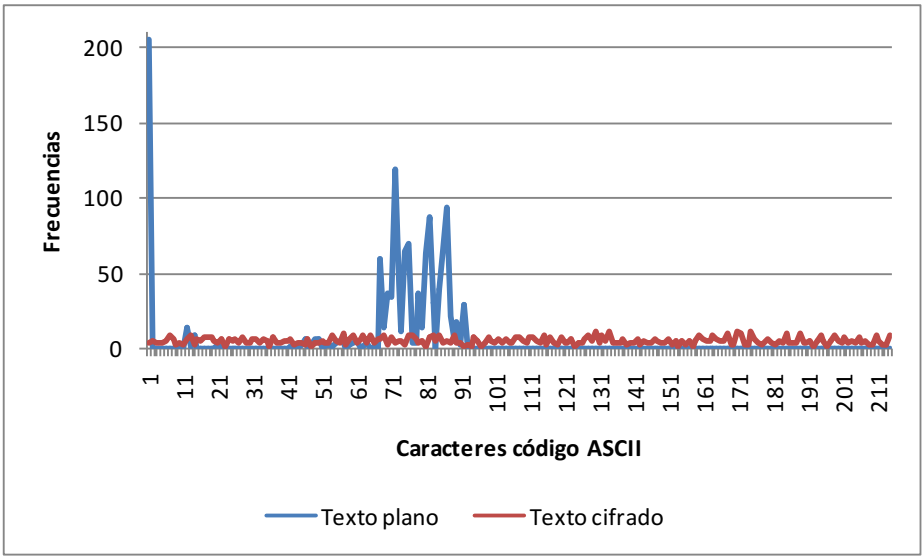


Fig. 8. Frecuencias de caracteres del texto plano y cifrado

11 Conclusiones

Se ha presentado un cifrador de flujo con algunas características interesantes tales como clave de mayor longitud y la incorporación de un generador basado en la combinación de LFSR.

Sobre este diseño se pueden implementar otras variantes para lograr futuras versiones que contemplen entre otras cosas: claves más largas y nuevos generadores binaries pseudoaleatorios.

La respuesta de esta versión fue buena y entregó un texto cifrado con una frecuencia de carácteres con cierta uniformidad, lo que hace difícil un criptoanálisis basado en la estadística de aparición de carácteres.

Finalmente se realizaron pruebas estadísticas de aleatoriedad sobre cien secuencias obtenidas del mismo texto plano con cien claves distintas, las que dieron resultados positivos.

12 Referencias

[1] Massodi, F., Alam, S. and Bokhari, M., "An Analysis of Linear Feedback Shift Registers in Stream Ciphers", International Journal of Computer Application, 16 (17), pp. 0975 – 887, 2012.

[2] Canteaut, A. and Filio, E., "Ciphertext only reconstruction of stream ciphers based on combination generators. Fast Software Encryption 2000", Lecture Notes in Computer Science, 1978, pp. 165–180, 2001.

[3] Menezes, A.,Van Oorschot, P. and Vanstone, S., "Handbook of Applied Cryptography", Massachusetts Institute of Technology, 1996.

[4] Parr, C. and Pelzl, L., Understanding Cryptography, Springer, 2010.

[5] Constantinescu, N., "Combining Linear Feedback Shift Registers", in Annals of University of Craiova, Math. Comp. Sci. Ser, 2009, 36 (2), pp. 42–46.

[6] Braeken, A.: Cryptographic Properties of Boolean Functions and S-Boxes. Faculteit Ingenieurswetenschappen. Katholieke Universiteit Leuven (2003)

[7] Elhosary, A., Hamdy, N., Farag, I., Rohiem, I.: State of the Art in Boolean Functions Cryptographic Assessment. International Journal of Computer Networks and Communications Security.1. (3), 88--94 (2013)

[8] Fishman, G.: Multiplicative Congruential Random Number Generators with Modulus 2ß : An Exhaustive Analysis for ß = 32 and a Partial Analysis for ß = 48. Mathematics of Computation. 54. (189), 33--344 (1990)

[9] Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S., "A Statistical Prueba Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", National Institute of Standards and Technology, (2000).

Cifrador de Bloque con Doble Red de Feistel y Funciones Booleanas de Alta No Linealidad

Andrés Francisco Farías – Andrés Alejandro Farías

Departamento Académico de Ciencias Físicas, Matemáticas y Naturales
Universidad Nacional de La Rioja, La Rioja. Argentina
(afarias665@yahoo.com.ar, andres_af86@hotmail.com)

Abstract. Cifrador de bloque, basado en la doble Red Feistel de 48 rondas cada una, con bloques de 256 bits de longitud y clave de 128 bits. Donde las Cajas S, de sustitución son reemplazadas por funciones booleanas que siguen los siguientes según criterios de buenas propiedades criptográficas: balance, cumplimiento del critero de avalancha estricta (SAC en inglés) y alta no linealidad.
Key Words: NLFSR, Cipher, key, boolean function, non-linearity.

1 Introducción

El presente documento expone el desarrollo de un cifrador de bloque, basado en una doble Red de Feistel que permite el cifrado y descifrado utilizando la misma estructura, donde para el caso del descifrado se utilizan las subclaves cambiando el orden de las mismas [1], [2]. La clave adoptada es de 16 caracteres, es decir 128 bits. Se utiliza como Función de Feistel, en lugar de las clásicas Cajas S (S-Box), funciones booleanas de cuatro variables, balanceadas y de alta no linealidad [3].

2 Esquema del cifrador

El cifrado de bloque se denomina así por realizar el proceso de cifrado trabajando sobre cadenas de texto de igual longitud. En este caso se utilizaron bloques de 256 bits. Luego esos bloques son ensamblados siguiendo el modo de encadenamiento de bloques de cifrado de propagación (PCBC, Propagating Cipher Block Chaining) [4]. Básicamente la estructura del cifrador está conformada por dos Redes de Feistel en que consta de: Las propia Redes de Feistel para: Cifrado y descifrado, subclaves.y función de las redes de Feistel

2.1 Redes de Feistel para el cifrado

En la figura 1, se indica la disposición del cifrador de bloque.

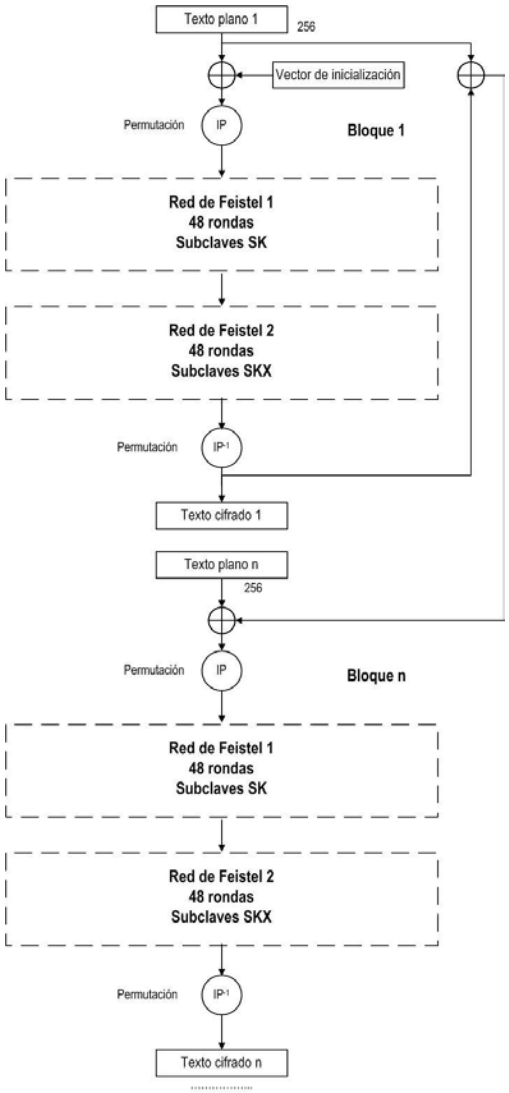


Fig. 1. Redes de Feistel para cifrado en modo PCBC

Cada bloque al ingresar a la red sufre una permutación dada por una matriz PI, luego de ello se divide al bloque en dos bloques, uno izquierdo y otro derecho, de 128 bits cada uno, a partir de ese momento esos bloques entran en las redes de Feistel. Finalmente los bloques resultantes del final de las rondas se concatenan para un formar un bloque de 256 bits, que es sometido a una nueva permutación IPI, que da como resultado el texto cifrado.

2.2 Redes de Feistel para el descifrado

Se indica en la figura 2:

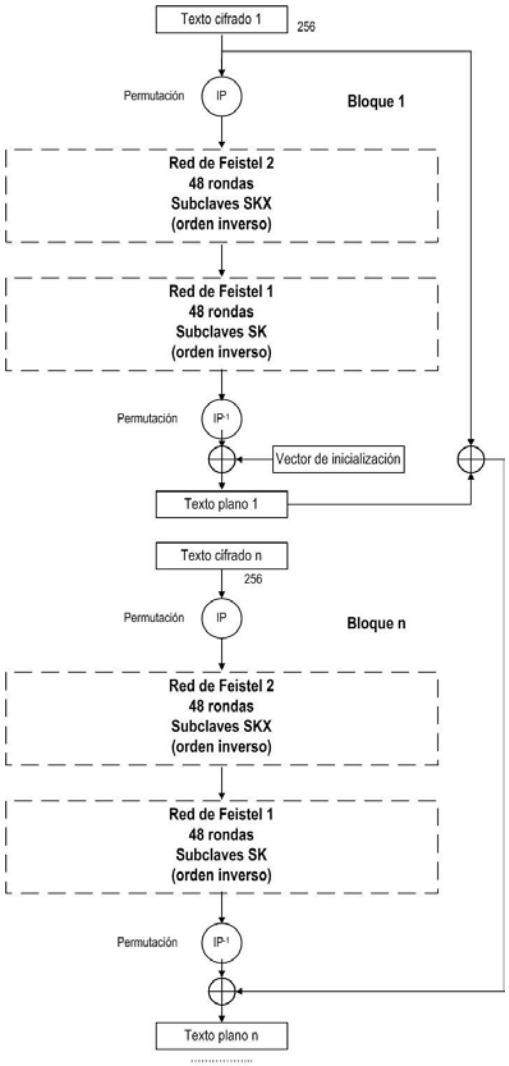


Fig. 2. Redes de Feistel para descifrado en modo PCBC

Las redes de Feistel para descrifrado son similares a la anterior, pero en este caso se toma el texto cifrado y se lo divide en bloques de 256 bits. Las permutaciones PI e IPI son las mismas que se utilizaron para el cifrado:

2.3 Clave y subclaves

Como se dijo previamente, la clave está conformada con 16 caracteres (128 bits), de las que se generan dos conjuntos de 48 subclaves de 128 bits, para cada una de las redes. Esos pares son ensamblados y luego sometidos a la permutación PC2, para obtener las subclaves finales.

2.4 Función de Feistel

La función de Feistel tiene la configuración que se indica en la figura 3:

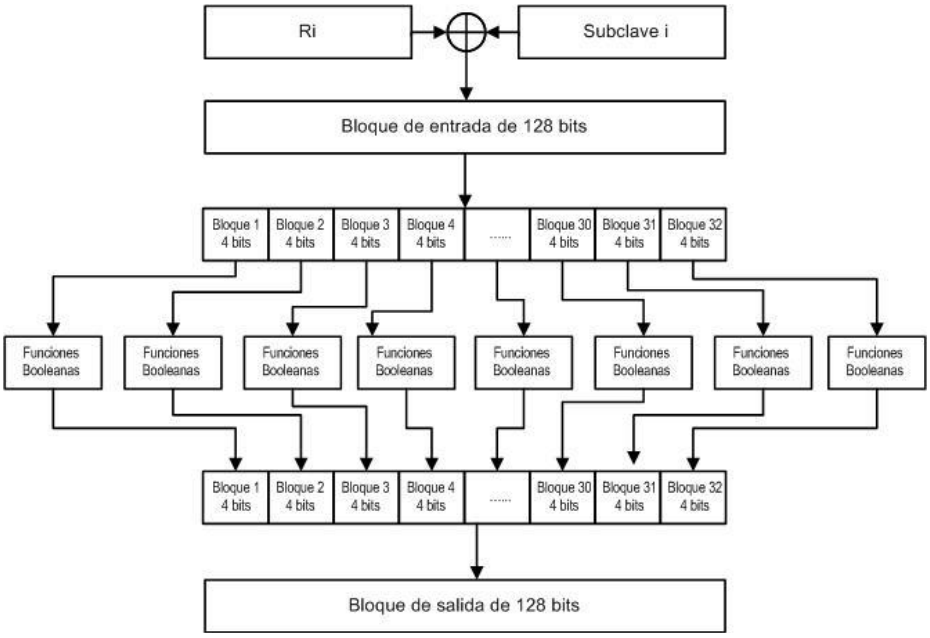


Fig. 3. Función de Feistel

La mitad del bloque de texto, la parte derecha de 128 bits, es sometida a una operación XOR con la subclave de 128 bits. Luego se divide en bloques de cuatro bits que alimentan a funciones booleanas balanceadas y de alta no linealidad, de esto resulta una salida de 128 bits.

2.5 Bloques con funciones booleanas

A continuación en la figura 4, se indica la estructura de cuatro bits.

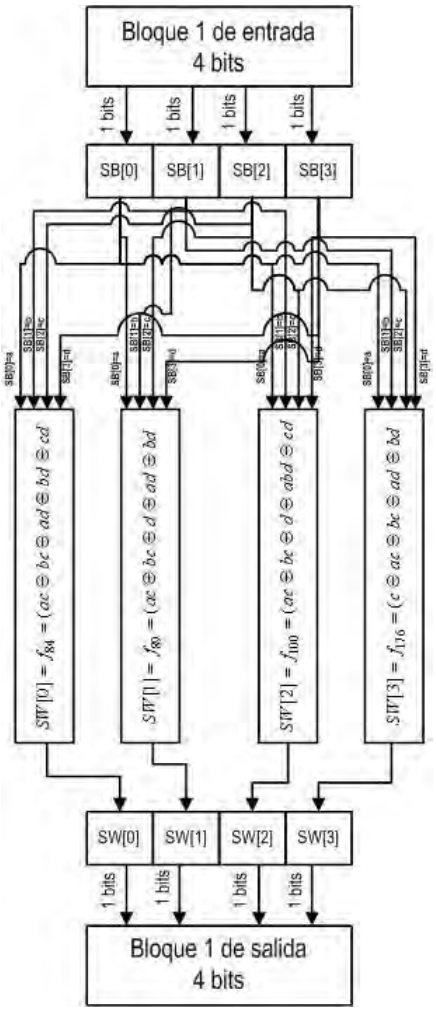


Fig. 4 Bloque de 4 bits con funciones booleanas de alta no linealidad

3 Propiedades criptográficas deseables adoptadas

A continuación se indican algunas de las propiedades criptográficamente más significativas, adoptadas para este trabajo [5], [6] y [7].

- **Función Balanceada:** Una función booleana de n-variables f es balanceada si $w(f) = 2n - 1$. Esta propiedad es deseable para evitar ataques criptodiferenciales. La función es balanceada cuando el primer coeficiente del espectro de Walsh-Hadamard, es igual a cero: $F(0) = 0$.
- **No Linealidad:** Valores altos de esta propiedad reducen el efecto de los ataques por criptoanálisis lineal. La No Linealidad de una función booleana puede ser

calculada de la transformada de Walsh-Hadamard.: $NL_f = \frac{1}{2} \cdot (2^n - |WH_{max}(f)|)$

- **Grado Algebraico:** El grado algebraico de una función, es el número de entradas más grande que aparece en cualquier producto de la Forma Normal Algebraica. Es deseable que sean valores altos.
- **SAC:** El Criterio de Avalancha Estricto requiere los efectos avalancha de todos los bits de entrada. Una función booleana se dice que satisface SAC sí y solo sí, $f(x) \oplus f(x \oplus u)$, es balanceada para toda u con w(u)=1.

4 Permutación

Se recurre a una matriz con una distribución aleatoria de las posiciones, para obtenerla se utiliza un generador de números aleatorios, en esta ocasión se adopta un generador congruencial multiplicative [8].

4.1 Generador congruencial multiplicativo:

El generador tiene la siguiente expresión:

$$x_{i+1} = (a_x \cdot x_i) mod\ m_x$$

Donde: $a_x = multiplicador, m_x = módul, x_0 = semilla$

Tabla.1. Matrices

Matriz	módulo	multiplicador	semilla
IP	1048576	1279	1153
PC1	1048576	1597	1531
PC2	1048576	1933	1759

5 Elección de las pruebas estadísticas

Algunas pruebas de la Norma NIST 800-22, del trabajo de Rukhin (et al.) [9].

Prueba de frecuencia: El propósito de esta prueba es determinar si el número de unos y ceros en una secuencia es aproximadamente el mismo que se espera de una secuencia verdaderamente aleatoria. La prueba evalúa la cercanía de la fracción de unos a ½, que es decir, el número de unos y ceros en una secuencia debe ser aproximadamente el mismo. Todas las pruebas posteriores dependen de la aprobación de esta prueba.

Prueba de frecuencia en un bloque: La meta de esta prueba es determinar si la frecuencia de unos en un bloque de M bits es aproximadamente M / 2, como se esperaría bajo un supuesto de aleatoriedad.

Prueba de rachas: Una racha de longitud k consta de exactamente k bits idénticos y está acotada antes y después con un poco del valor opuesto. El propósito de la prueba de rachas es determinar si el número de rachas unos y ceros de varias longitudes es lo esperado para una secuencia aleatoria.

Prueba de rachas de unos en un bloque: El fin de esta prueba es determinar si la longitud de la ejecución más larga de las dentro de la secuencia probada es consistente con la longitud de la serie más larga de las que cabría esperar en una secuencia aleatoria. Tenga en cuenta que una irregularidad en la longitud esperada de la serie más larga implica que también hay una irregularidad en la longitud de la serie más larga de ceros.

Prueba de sumas acumuladas: Determina si la suma acumulativa de las secuencias parciales que ocurren en la secuencia probada es demasiado grande o demasiado pequeña en relación con el comportamiento esperado de esa suma acumulada para secuencias aleatorias.

Prueba de entropía aproximada: El enfoque de esta prueba es la frecuencia de todas las posibles superposiciones patrones de m bits en toda la secuencia. El propósito de la prueba es comparar la frecuencia de bloques superpuestos de dos longitudes consecutivas / adyacentes (m, m + 1) contra el resultado esperado para un secuencia aleatoria.

6 Pruebas sobre el generador

Se analizaron cien secuencias binarias, obtenidas del cifrador a partir de cien claves distintas. El nivel de significancia adoptado para las pruebas estadísticas es de $\alpha = 0{,}01$. La hipótesis nula es: $H_0 \rightarrow p\_valor > 0{,}01$

7 Interpretación de los resultados

Teniendo los resultados se realizan dos procesos para la interpretación de los mismos:

- Proporción de muestras que pasan las pruebas.
- Prueba de Uniformidad de los p-valor
    - Tabla de frecuencia e histograma
    - Prueba de Bondad de Ajuste

El primer procedimiento se realiza considerando los resultados de todas las pruebas y el segundo se realiza en forma individual. En todos los casos se deben superar todas las pruebas para aceptar los resultados.

7.1 Proporción de muestras que pasan las pruebas

Para el análisis de los resultados, se determina la proporción de muestras que superan las pruebas, y con esos datos se construye un gráfico de puntos, luego se verifica si los mismos caen dentro de los límites superior e inferior, donde k es el número de muestras. $LS, LI = (1-\alpha) \pm 3 \cdot \sqrt{\alpha \cdot (1-\alpha)/k}$

En nuestro caso $k = 100$ y el nivel de significancia elegido es: $\alpha = 0.01$, los límites quedan:. $LS = 1{,}02$ y $LI = 0{,}96$. Los resultados se indican en la tabla

Tabla 2. Pruebas

Pruebas	Proporción	Superior	Inferior
Frecuencias	0,98	1,02	0,96
Frecuencias en un Bloque	0,97	1,02	0,96
Rachas	0,99	1,02	0,96
Rachas de Unos en un Bloque	0,99	1,02	0,96
Sumas Acumuladas Adelante	0,97	1,02	0,96
Sumas Acumuladas Atrás	0,99	1,02	0,96
Entropía Aproximada	1,00	1,02	0,96

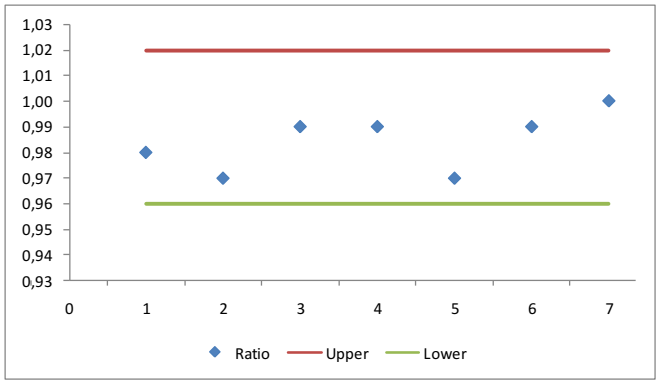


Fig. 5. Gráfico de puntos

7.2 Prueba de bondad de ajuste

Este control se ejecuta para cada prueba sobre las cien muestras, con los resultados de las frecuencias de p-valor obtenidos.

Tabla 3. Pruebas χ^2

Pruebas	χ^2	$\chi^2 ref$	Pasa
Frecuencias	0,130	0,0001	Sí
Frecuencias en un Bloque	0,946	0,0001	Sí
Rachas	0,883	0,0001	Sí
Rachas de Unos en un Bloque	0,154	0,0001	Sí

Sumas Acumuladas Adelante	0,019	0,0001	Sí
Sumas Acumuladas Atrás	0,067	0,0001	Sí
Entropía Aproximada	0,699	0,0001	Sí

7.3 Análisis final

En base a los resultados de la pruebas se realiza una tabla resumen.

Tabla 4. Análisis final

Análisis	Pruebas	Resultados
Proporción de secuencias que pasan las pruebas	Todas	Supera
Distribución uniforme de p-valor	Frecuencias	Supera
	Frecuencias dentro de un bloque	Supera
	Rachas	Supera
	La más larga racha de unos en un bloque	Supera
	Sumas acumuladas adelante	Supera
	Sumas acumuladas atrás	Supera
	Entropía estimada	Supera

En definitiva las secuencias que entrega el generador son pseudoaleatorias.

8 Comparación de frecuencias

Diferencias entre texto plano y texto cifrado.

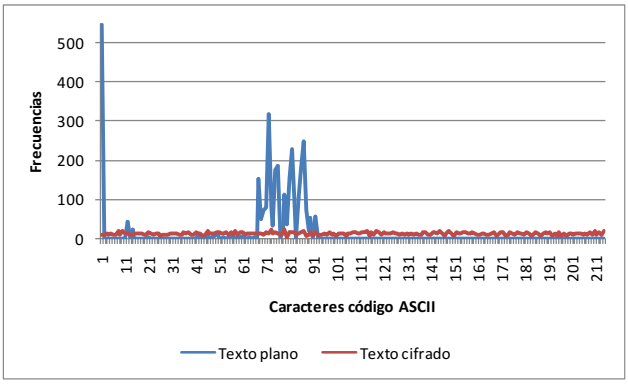


Fig. 6. Frecuencias de caracteres del texto plano y cifrado

9 Conclusiones

Se ha presentado un cifrador de bloque con algunas características interesantes tales como clave de mayor longitud y la incorporación de funciones booleanas de alta no linealidad.

Sobre este diseño se pueden implementar otras variantes para lograr futuras versiones que contemplen entre otras cosas: claves más largas y mayor cantidad de funciones booleanas y otros métodos de concatenación de bloques.

La respuesta de esta versión fue buena y entregó un texto cifrado con una frecuencia de caracteres con cierta uniformidad, lo que hace difícil un criptoanálisis basado en la estadística de aparición de caracteres.

Los cifrados son herramientas útiles cuando se necesita dar seguridad a información de tipo confidencial.

10 Referencias

[1] Massodi, F., Alam, S. and Bokhari, M., "An Analysis of Linear Feedback Shift Registers in Stream Ciphers", International Journal of Computer Application, 16 (17), pp. 0975 – 887, 2012.

[2] Canteaut, A. and Filio, E., "Ciphertext only reconstruction of stream ciphers based on combination generators. Fast Software Encryption 2000", Lecture Notes in Computer Science, 1978, pp. 165–180, 2001.

[3] Menezes, A.,Van Oorschot, P. and Vanstone, S., "Handbook of Applied Cryptography", Massachusetts Institute of Technology, 1996.

[4] Parr, C. and Pelzl, L., Understanding Cryptography, Springer, 2010.

[5] Constantinescu, N., "Combining Linear Feedback Shift Registers", in Annals of University of Craiova, Math. Comp. Sci. Ser, 2009, 36 (2), pp. 42–46.

[6] Braeken, A.: Cryptographic Properties of Boolean Functions and S-Boxes. Faculteit Ingenieurswetenschappen. Katholieke Universiteit Leuven (2003)

[7] Elhosary, A., Hamdy, N., Farag, I., Rohiem, I.: State of the Art in Boolean Functions Cryptographic Assessment. International Journal of Computer Networks and Communications Security.1. (3), 88--94 (2013)

[8] Fishman, G.: Multiplicative Congruential Random Number Generators with Modulus 2ß : An Exhaustive Analysis for ß = 32 and a Partial Analysis for ß = 48. Mathematics of Computation. 54. (189), 33--344 (1990)

[9] Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S., "A Statistical Prueba Suite for Random and Pseudorandom Number Generators for Cryptographic Applications", National Institute of Standards and Technology, (2000).

Teoría de Grafos para la Identificación de Nodos Maliciosos en la Red

Tatiana S. Parlanti[1], Carlos A. Catania[1], and Luis G. Moyano[2]

[1] Universidad Nacional de Cuyo, Facultad de Ingenieria, Laboratorio de Sistemas Inteligentes (LABSIN), Mendoza, Argentina
[2] División de Física Estadística e Interdisciplinaria, Centro Atómico Bariloche, Bariloche, Argentina
{tatiana.parlanti,harpo}@ingenieria.uncuyo.edu.ar

Resumen Se explora el reconocimiento de las botnets en una red a partir de su representación como grafo, extrayendo características a sus nodos y poniendo a prueba algoritmos de agrupamiento. Se logra la separación del 88 % de las botnets junto al ∼0,14 % de los nodos benignos.

Keywords: Redes complejas, seguridad de redes, aprendizaje máquinas

1. Introducción

Una estrategia posible para la detección de nodos maliciosos en la red consiste en la construcción de modelos estadísticos utilizando información obtenida mediante técnicas aplicadas en las áreas de las redes complejas. Se trata de algoritmos tomados de la teoría de grafos para extraer características de las redes para luego ser utilizadas como información de entrada en modelos de aprendizaje estadístico. Estas características permiten capturar la estructura topológica del grafo y permite exponer aspectos adicionales de los nodos maliciosos [2,3].

En este trabajo se explora la viabilidad de reconocer el comportamiento de las botnets en una red de computadoras a partir de su representación como grafo, extrayendo para tal fin distintas características que dan información de sus nodos en cuanto al grado y centralidad de cada uno. Para ello, se pondrán a prueba diferentes algoritmos de *clustering* o agrupamiento, con el objetivo de diferenciar los nodos, no las conexiones, que estén infectados de los que no. A diferencia de otros trabajos anteriores que han considerado el número de paquetes transferidos entre los distintos nodos, en este trabajo se analizan el número de bytes transferidos, ya que se considera que contar con el dato del tamaño de los paquetes puede ofrecer mayor información para el reconocimiento de los canales de comando y control (C&C) de una botnet.

El trabajo se encuentra enmarcado en el proyecto de tesis doctoral de la Lic. Parlanti, realizado en la Universidad Nacional de Centro de la Pcia. de Bs.As. bajo la dirección de los Dres. Moyano y Catania, financiado por CONICET.

2. Materiales y Métodos

Se utilizó la base de datos CTU-13, el cual es un conjunto de 13 capturas con datos de tráfico de botnets que fue capturado en la Universidad CTU, República

Checa, en 2011 [4]. Particularmente se filtraron aquellas observaciones cuyo protocolo fuera del tipo TCP o UDP, ya que se consideró que los otros protocolos observados no ofrecían información, y se almacenó la dirección IP del nodo de origen y destino, así como la cantidad de bytes transmitidos entre ambos nodos.

Con esta información se construyó un grafo ponderado y dirigido por cada captura, donde los pesos están dados por la suma de bytes de las distintas conexiones que comparten los mismos nodos de origen y destino, diferenciando entre cantidad de bytes transmitidos del primero al segundo (`SrcBytes`), y su correspondiente conexión inversa (`DstBytes`), eliminando previamente aquellas conexiones que no hubieran logrado transmitir byte alguno. Luego, se extrajeron características de cada nodo, descriptas en la Tabla 1, cuyas definiciones comprendidas en la Teoría de Grafos se puede consultar en [5].

Tabla 1. Descripción de las características extraídas.

	Característica	**Descripción**
ID	Grado Entrante	Número de aristas que entran a un vértice
OD	Grado Saliente	Número de aristas que salen de un vértice
IDW	Grado Entrante Ponderado	Suma de los pesos de las aristas que entran a un vértice
ODW	Grado Saliente Ponderado	Suma de los pesos de las aristas que salen de un vértice
BC	Centralidad de Intermediación	Número de caminos más cortos que pasan por un vértice y minimizan la suma de los pesos de las aristas
LCC	Coeficiente de Agrupamiento Local	Cuantifica qué tan interconectado está un vértice con sus vecinos, a partir de las conexiones de sus vecinos entre sí. Se consideró el grafo como no dirigido, y no se tuvieron en cuenta los pesos de las aristas
AC	Centralidad Alfa	Variante de la centralidad de autovector, donde el vértice está sujeto a distinta importancia dependiendo de factores externos. Se tomaron los valores $\alpha = 0{,}01,\ e = 1$

Finalmente, cada uno de los nodos fue etiquetado a partir de la información provista por el equipo de StratosphereIPS [1], bajo el supuesto de que todo nodo que no tuviera la identificación de "botnet" sería considerado "normal".

Una vez calculadas cada una de las características antes mencionadas, se procedió a utilizarlas como conjunto de entrenamiento de los distintos modelos de agrupamiento, a excepción de las características de la novena captura, la cual fue conservada para testear en una etapa posterior a la abarcada en este trabajo.

Entre los algoritmos de *clustering* conocidos [7], se compararon los siguientes: **k−means** usando tanto el algoritmo Hartigan-Wong como el Lloyd-Forgy, tomando por k los cuadrados de los números naturales entre 2 y 15, ambos inclusive; así como **CLARA** usando la distancia euclideana y la Manhattan, con los mismos valores de k que para $k-$means, tomando 300 muestras sobre las que se aplica el algoritmo PAM.

Para llevar a cabo los experimentos aquí planteados, así como la extracción de características, se utilizó un procesador AMD Ryzen 7 1700 Eight-Core de 64 GB de memoria, con sistema operativo Ubuntu 20.04.4 LTS. Se trabajó con las versiones `3.8.10` de `Python` y `4.2.1` de `R`. El análisis de los distintos grafos para la extracción de características se llevó a cabo usando el paquete `igraph` (versión

1.3.1) de R, así como la biblioteca homónima (versión 0.9.9) de Python. Así se calcularon LCC y AC en el primero, mientras que el resto en el segundo.

3. Resultados y Discusión

3.1. Sobre las Características

En primer lugar se realizó una comparación sobre la información que brindan los paquetes y bytes transmitidos entre nodos. Si bien pareciera existir cierta correlación entre ambas, como se observa en la Figura 1, al hacer foco en una muestra particular de los datos, en este caso aquellas observaciones con 16 paquetes o menos y a lo sumo 1000 bytes, se aprecia que para una misma cantidad de paquetes la cantidad de bytes puede ser diferente (gráfica derecha), lo que representa un dato de interés. Es por ello que, a diferencia de trabajos como [2,3], en lugar de tomar la cantidad de paquetes como peso, se decidió utilizar la cantidad de bytes.

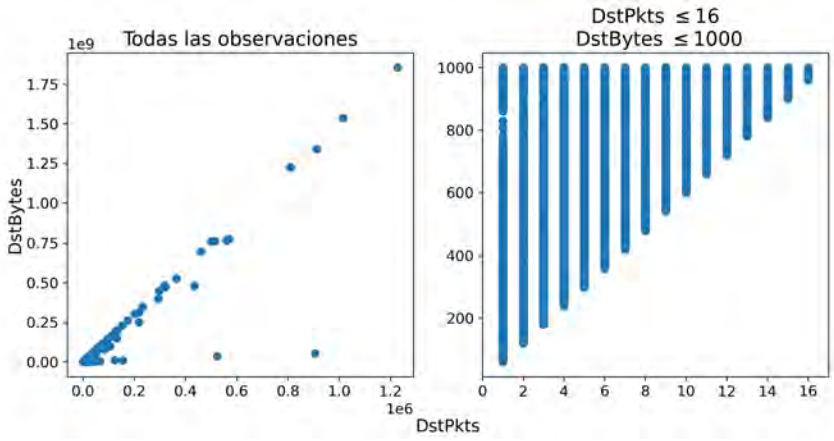


Figura 1. La gráfica de la izquierda compara la relación entre paquetes y bytes transmitidos desde nodos destino para cada una de las observaciones; la de la derecha se concentran en los casos en que se transmitieron a lo sumo 16 paquetes y 1000 bytes como máximo. Las conexiones desde nodo origen presentan un comportamiento similar.

Respecto al tiempo que demandó la extracción de características de cada grafo, en la Tabla 2 se observa el promedio de éste así como la desviación estándar en segundos. Se destaca que BC es la que requiere mayor tiempo, tomando unas ~ 21 horas promedio para su cálculo. Así también, es la que tiene mayor variación, junto con AC, en relación a la cantidad de vértices de cada grafo.

3.2. Agrupamiento

Como se explicó en la Sección 2, una vez calculadas las características de interés sobre un total de 2 874 213 nodos, de los cuales 25 están infectados, se entrenaron distintos modelos de agrupamiento, con el objetivo de diferenciar entre nodos malignos y benignos. Dada la desproporción entre ambos, es de esperar que durante el agrupamiento se encuentre un cluster principal, que contenga a la

Tabla 2. Tiempo promedio en capturas (seg.) de la extracción de características.

Característica	Tiempo Promedio	Desviación Estándar
BC	77355,149458	84083,199324
AC	507,160385	908,155491
LCC	0,087248	0,068996
ODW	0,036261	0,026509
IDW	0,036241	0,026511
OD	0,003525	0,002684
ID	0,003377	0,00256

Tabla 3. Agrupamiento usando $k-$means y CLARA. Se especifican: (i) porcentaje de nodos no infectados que quedaron fuera del cluster principal; (ii) porcentaje de nodos infectados que quedaron fuera del cluster principal; (iii) tiempo empleado (seg.).

	k-means						**CLARA**					
k	**Alg. Hartigan-Wong**			**Alg. Lloyd-Forgy**			**Dist. Euclidea**			**Dist. Manhattan**		
	(i)	**(ii)**	**(iii)**	**(i)**	**(ii)**	**(iii)**	**(i)**	**(ii)**	**(iii)**	**(i)**	**(ii)**	**(iii)**
4	0,0005	0	28,057	0,0005	0	30,651	0,0206	36	49,593	0,0203	36	29,625
9	0,003	8	48,293	0,0034	8	64,265	**0,1243**	**84**	**107,559**	**0,1376**	**88**	**68,215**
16	0,0289	36	66,089	0,0379	68	103,720	0,2964	88	185,435	0,2599	88	121,003
25	0,0875	80	95,352	0,0872	80	208,62	0,3179	88	277,426	0,3439	88	179,842
36	**0,1861**	**84**	**171,429**	**0,1876**	**84**	**480,524**	0,4437	88	381,026	0,4164	88	257,241
49	0,1991	84	326,304	0,3693	88	811,729	0,6354	88	498,378	0,5725	88	345,199
64	0,4067	88	425,199	0,4903	88	1846,021	0,9636	88	646,739	0,9193	88	448,270
81	0,5692	88	515,123	0,7593	88	2724,904	1,134	88	823,091	1,1839	92	579,934
100	0,9523	88	572,051	0,8478	92	4589,811	1,1173	92	1023,322	1,1234	92	716,886
121	0,8453	92	569,522	1,3271	92	5973,175	1,2073	92	1331,300	1,2207	92	898,435
144	1,2516	100	828,449	1,3512	92	11048,307	2,3659	100	1735,531	1,4612	92	1204,218
169	2,1877	100	791,413	1,9773	100	13722,263	1,3678	100	2159,850	1,2173	92	1552,681
196	79,4276	100	594,627	2,3576	100	17771,940	1,9427	92	2755,081	9,7819	100	2048,137
225	79,2468	100	700,582	2,3729	100	22760,47	2,632	92	3475,695	11,2066	100	2652,382

mayoría de los nodos benignos, por tener características similares. En el Tabla 3 se observan los resultados obtenidos usando $k-$means y CLARA, para diferentes valores de k, algoritmos y distancia, especificando las siguientes métricas utilizadas para su evaluación:

- **i)** Porcentaje de nodos no infectados que quedaron fuera del cluster principal.
- **ii)** Porcentaje de nodos infectados que quedaron fuera del cluster principal.
- **iii)** Tiempo empleado (segundos).

Tanto para $k-$means como para CLARA, en la mayoría de los casos se logró un cluster principal que, a juzgar por (i), contiene la mayor cantidad de nodos benignos. Sin embargo, dicho cluster no es completamente homogéneo, ya que por (ii) se observa que éste también contiene botnets. Entonces, se busca maximizar el cluster principal, minimizando la cantidad de nodos infectados que pueda contener. Teniendo en cuenta esto, además del tiempo empleado, se concluye que utilizar el algoritmo CLARA con la distancia Manhattan es la mejor opción,

ya que con $k = 9$ se logra un cluster principal que sólamente excluye el 0,1376 % de los nodos no infectados y el 88 % de los infectados, es decir un total de 3956 nodos no infectados quedan excluidos, pero incluye únicamente a 3 nodos infectados, y sólo demanda $\sim$ 1,14 minutos. Por otro lado, CLARA usando la distancia euclideana y el mismo valor de k, deja menos nodos no infectados por fuera del cluster principal (3574), pero incluye un nodo infectado más y toma $\sim$2 minutos. En contraposición, para obtener resultados similares implementando $k-$means, ya sea con el algoritmo de Hartigan-Wong o el de Lloyd-Forgy, son necesarios 36 clusters. Sin embargo, en ambos casos el cluster principal contiene 4 nodos infectados (el 84 % restante queda fuera) y toman $\sim$ 3 y 8 minutos, respectivamente. Finalmente, vale aclarar que si bien hay casos donde el 100 % de los nodos infectados quedan por fuera del cluster principal, no fueron tenidos en cuenta ya que de igual manera aumenta el número de nodos benignos que no pertenecen a dicho cluster, siendo que además toman más tiempo.

4. Conclusiones y Trabajo Futuro

De los resultados preliminares mediante técnicas de cluster se desprende que la aplicación de teoría de grafos para la extracción de características en redes con millones de nodos facilita la discriminación de comportamiento. La utilización del número de bytes transferido entre los nodos demostró ser adecuada. En general todos los algoritmos de clustering analizados fueron capaces de agrupar a la mayoría de los nodos benignos excluyendo a los nodos con comportamiento malicioso. Sin embargo, aquellas características que se focalizan en medir la centralidad de un nodo demandan un tiempo considerable, lo que dificulta su aplicación en escenarios de tiempo real. Es por esto último que como trabajo futuro se propone analizar otras técnicas para la extracción de características de los grafos como las redes convolucionales para grafos [5,6].

Referencias

1. Stratosphere research laboratory. https://www.stratosphereips.org/, última vez visitado en Agosto 2022
2. Abou Daya, A., Salahuddin, M.A., Limam, N., Boutaba, R.: Botchase: Graph-based bot detection using machine learning. IEEE Transactions on Network and Service Management 17(1), 15–29 (2020)
3. Chowdhury, S., Khanzadeh, M., Akula, R., Zhang, F., Zhang, S., Medal, H., Marufuzzaman, M., Bian, L.: Botnet detection using graph-based feature clustering. Journal of Big Data 4(1), 1–23 (2017)
4. Garcia, S., Grill, M., Stiborek, J., Zunino, A.: An empirical comparison of botnet detection methods. computers & security 45, 100–123 (2014)
5. Hamilton, W.L.: Graph representation learning. Synthesis Lectures on Artifical Intelligence and Machine Learning 14(3), 1–159 (2020)
6. Welling, M., Kipf, T.N.: Semi-supervised classification with graph convolutional networks. In: J. International Conference on Learning Representations (ICLR 2017) (2016)
7. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Transactions on neural networks 16(3), 645–678 (2005)

XI Workshop Innovación en Educación en Informática **(WIEI)**

Coordinadores

Cecilia Sanz (UNLP)

Beatriz Depetris (UNDTF)

Marcelo De Vincenzi (UAI)

TEARA: Educational Treatment of Children with ASD, mediated through augmented reality.

Mónica. R. Romero[1], Ivana Harari[1], Javier Diaz[1], Estela Macas[2]

[1]National University of La Plata, Faculty of Informatics, Research Laboratory in New Information Technologies (LINTI).
Calle 50 y 120, 1900 La Plata. Buenos Aires.

[2] International Ibero-American University - UNINI MX. Mexico

[1]monica.romerop@info.unlp.edu.ar, [1]iharari@info.unlp.edu.ar, [1]jdiaz@info.unlp.edu.ar, [2]estela.macas@doctorado.unini.edu.mxc

Orcid ID:https://orcid.org/0000-0002-6099-7039
Orcid ID: https://orcid.org/0000-0001-6350-7739
Orcid ID:https://orcid.org/0000-0002-4225-3829
Orcid ID:https://orcid.org/0000-0002-1237-1154

Abstract. The treatments that used since the 1960 as educational proposals for children with autism spectrum disorder (ASD) are becoming obsolete over time. This research proposes an educational treatment for children with autism mediated through augmented reality called (TEARA), as a response to the challenges and constant change of a globalized world, which requires the establishment of new methods, strategies and treatments that allow improve the quality of life of these children with autism. The methodology was approached through a mixed, exploratory, descriptive, and purposeful study where a multidisciplinary team participated, we developed a training system called Hope, which reinforces and promotes teaching-learning processes, finally after several cycles of intervention, deep observation and the compilation of results, it was established that TEARA can be used by professionals, parents and people who accompany children with ASD.

Keywords: ***ASD, Treatment, Education, Hoope, Augmented Reality, TEARA***.

1 Introduction

Autism spectrum disorder onwards ASD is a neurological disorder that is complex[1], has no cure, in addition to being considered one of the enigmas that even medicine does not fully understand[two]. Autism spectrum disorder becomes one of the most complex to treat[3]–[5], although over time specialists have managed to diagnose this disorder at a younger age, which allows an early intervention plan to be drawn up[5],[6].

Regarding ASD and strategies in the educational field, various investigations have been planted where information and communication technologies have been used for

two decades to reinforce certain areas in children.[8], [9]and they manifest themselves in different ways such as: software and hardware. Experience tells us that the projects that were raised individually and methodically, although it is true, strengthen certain areas and provide assertive results.[10], [11] . It is no less true that those projects that are worked on from multidisciplinary groups and with a prolonged duration achieve greater benefits.[12]–[14].

In recent years we have seen how emerging technologies make their way into our society, virtual, augmented and mixed reality are present as an innovative element in education[15], [16], and even more in children with special educational needs (NEE)[17],[18]; That is why when we started this research in 2017 we planned to use augmented reality (AR) in teaching-learning processes.[18],[19], however in the development and conceptualization we rely on different strategies such as: Design thinking general conceptualization of the product[twenty-one], User Experience (UX), we use a user-centered design (UDC)[22]that allowed us to develop the training system called Hope.

Once these processes have completed, we seek to maximize their results, we elaborate a pedagogical technological intervention plan where we design the TEARA program. This study structured as follows: Section 2 explains the materials and methods that used, for data collection several techniques such as surveys, interviews and deep observation were conducted. Section 3 presents the results of the study, finally Section 4 presents a critical reflection on the proposal and future challenges of the TEARA program.

2 Material and method

This research presents a mixed approach since it uses a qualitative and quantitative method.[23]. The scope of this work is exploratory because programs that include the integration of innovative technologies, specifically augmented reality, to support children with ASD have not yet defined. This research is descriptive because it seeks to know in detail about the benefits of including TEARA in the treatment of children with ASD and purposeful since this research defines an Educational Treatment of Children with ASD, mediated through TEARA augmented reality.

This research focuses on two modalities such as documentary and experimental, this is because the experimentation was carried out through the Hope application, it is closely related to the deductive method because it bases its development on evaluating TEARA on children with ASD. , allowing to identify if this program is proactive for interventions where it is intended to use emerging technologies, especially augmented reality.

The field work of this research conducted in the city of Quito at the Ludic Place Therapeutic Center, the population made up of the group of children with ASD who used the Hope training system (5 children), and professionals who accompanied this experience, medical, academic (5) and information and communication technology (5) personnel.

The children who participated are 5 children, three of the male gender and 2 of the female gender, keeping the privacy of the data defined from now on by a capital letter: Eidan (E), Matias (M), Santiago (S), Valeria (V) and Ana (A). The children have a confirmed diagnosis of moderate and severe ASD respectively, the children regularly attend the Ludic Place Therapy Center 2-4 times per week. Their legal representatives previously signed an informed consent agreement.

2.1 Educational Treatment of Children with ASD, mediated through augmented reality TEARA

TEARA designed to strengthen teaching and learning processes for children with ASD, mild, moderate, and severe, it can used by children from 4 years old to 12. TEARA uses a training system called Hope that teaches through dance, that is, configured so that the child learns to dance and progressively seeks to include new learning, showing greater complexity as time goes by, it can used in academic centers, therapy centers, or at home.

TEARA proposes the use of Hope software, which developed through a friendly interface, in a playful space, allowing children and their caregivers to interact through emerging technology, specifically augmented reality. This system has activities and teaches dance steps. This system allows the child with ASD mobility of options to configure the environment by adding or removing options for use.

2.2 What do I need to apply TEARA

4 main aspects have defined that will help the use of TEARA to be successful.

- People: The people involved in TEARA are children, parents, or legal guardians, academic or medical staff of a child with ASD.
- Place and adaptations: To work with TEARA it is essential to define a place where the activities will take place, this can be a meeting room, a classroom, the living room of a house, the important thing is that you have a place of 4 square meters so that the proposed activities can be conducted.
- Training system: The necessary instruments for the application of the treatment are of distinct types: Hope software, free version; hardware: television or laptop, Kinect. The following figure 1 shows the training system.

Fig.1. Training system, on the left software and hardware for operation, on the right Hope main menu.

- Intervention plan: To apply TEARA it is necessary to start from a correct planning followed by an intervention plan that in turn defines phases: socialization, diagnosis, intervention, monitoring, evaluation.

2.2 TEARA phases.

Planning Phase: This phase allows sessions to devoted to the proper planning of the intervention. In this phase, important actions defined, such as the Therapy Center where TEARA will used, communication with managers, sessions to explain this process to participating teachers or medical personnel., which is the place where we will conduct the sessions, the necessary materials. This phase made up of 9 sessions that indicated below:

Sessions with academic staff, doctors, parents.

- Socialization sessions (1): it is important to explain what TEARA for what it is for, the work methodology and what aspects of the children will be reinforced through its use. Parents participate in the socialization meetings and decide whether to give your informed consent for the use of TEARA in your clients.
- Sessions to define work team (1): in this session what sought is to integrate a preferably multidisciplinary work team that will be responsible for the execution of TEARA.
- Adaptation sessions (1): In these sessions the work team will look for the right place to use the training system, define the place and install the Hope software, check that the Kinect works correctly, in addition to checking the connection of the laptop or television necessary for the execution of the sessions.
- Test and training sessions (2): The people who are part of the work team have to start the tests of the training system, when the tests have a favorable result, the training can begin, the intention is that the teacher who will participate with the child with ASD must be proficient in the use of the Hope system because he will be the one to guide the child in the first sessions.
- Participant Identification Sessions (1): This session identifies child participants with ASDs who will use TEARA.

- Sessions to define curricular intervention plan (2): The curricular plan must be thought and worked for each of the participants according to their needs, in the curricular plan objectives of different types are generally identified: cognitive, communications or procedural the intention is to define a path to follow according to the strengths or weaknesses that each child with ASD presents.
- Sessions to define the form of evaluation (1): once the intervention with TEARA has conducted, the progress of the children must evaluated in accordance with the curricular plan drawn up to know to what extent the use of TEARA was favorable or not, to reinforce certain teaching-learning processes or skills in children with ASD who have participated

Execution Phase: This phase allows TEARA to run the training system through the Hoope software. This phase is up of 25 sessions that indicated below: Sessions with children with ASD:

- Initial diagnosis sessions: it is essential that prior to using the system an initial diagnosis of the participant conducted, this evaluation done to determine the current state of processes such as imitation, perception, gross motor skills and fine motor skills, in this section the professionals They must make a record of the data obtained.
- Hope training system sessions: these sessions serve to reinforce teaching-learning processes, imitation, fine and gross motor skills, perception, visual-motor coordination. These sessions last 22 minutes and held 3 times a week. The number of TEARA sessions defined in the intervention plan; however, it recommended that the number of sessions be 25 sessions to have more assertive results.
  The first sessions the child accompanied by the teacher or therapist in charge so that in a coordinated process they can gradually use the TEARA system individually.

Review Phase: This phase allows evaluating the TEARA process, that is, the use of the training system through the Hope software. This phase consists of 2 sessions that indicated below:

- Sessions to conduct the evaluation (1): This session allows the work team to conduct an evaluation of the progress obtained after the application of TEARA, the same form of evaluation that used in the diagnostic session used and it verified if any of the teaching-learning processes improved after the intervention.
- Feedback sessions (1): this session allows evaluating whether the objectives of the intervention plan met and if opportunities for improvement are evident, it conducted in a meeting where it discussed with those interested in the process. The following Figure 2 below shows the phases of TEARA as a summary.

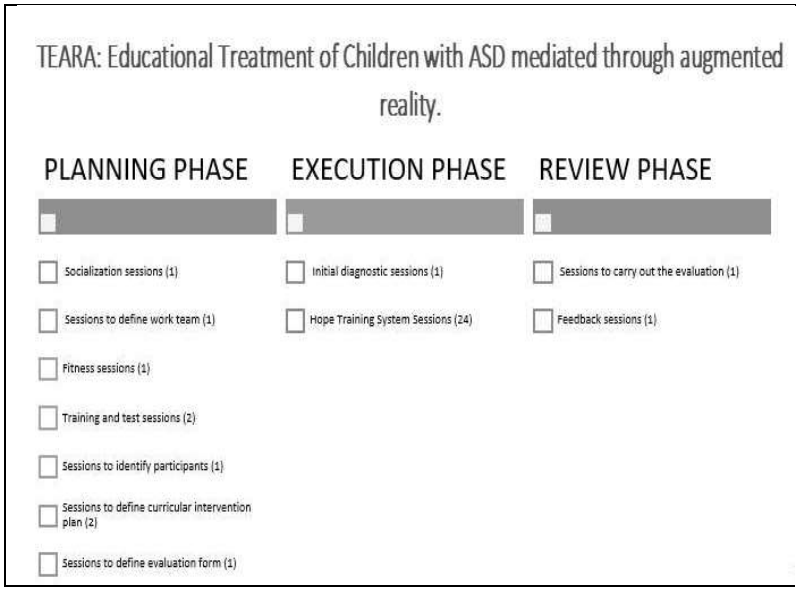


Fig.2. Phases of the Educational Treatment of Children with ASD mediated through augmented reality TEARA.

2 Results

TEARA used in a group of children with ASD and with the support of a multidisciplinary team, in this process each of the phases defined in its methodology considered, TEARA used for three months, through 9 sessions of the phase planning, 25 sessions of the execution phase and finally 2 sessions of the review or evaluation phase. For each session, actions conducted so that children with ASD use the system and through it learn dance steps, using their body, each option of the system allows children to reinforce the processes defined in advance in the pedagogical intervention plan, TEARA allowed strengthen imitation, perception, gross and fine motor skills, visual coordination, motor movement.

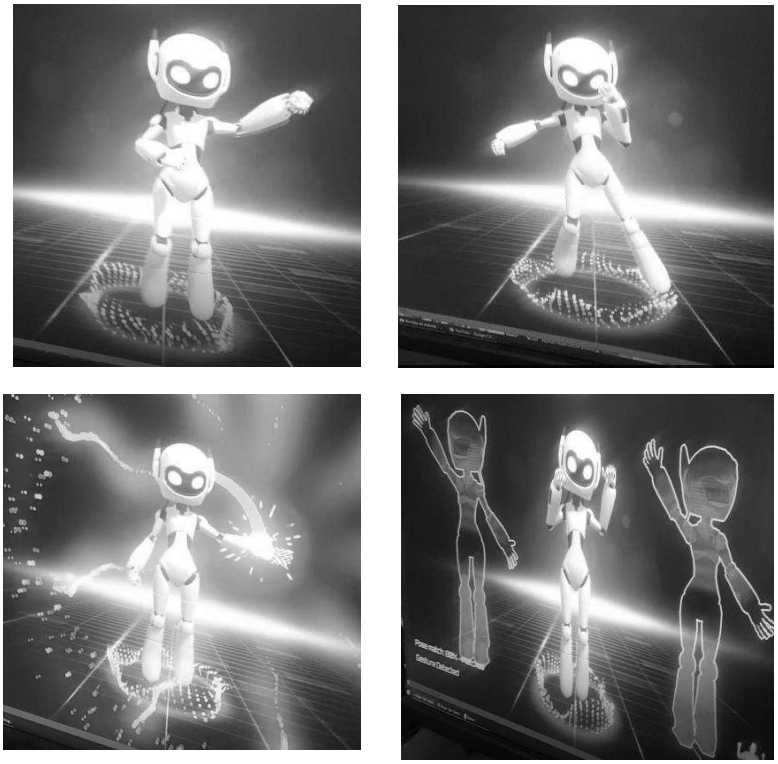


Fig3. Images of the Hope software training system shown.

The multidisciplinary team assigned for this project conducted the diagnostic sessions, keeping a record of the evaluation obtained from the processes before the use of TEARA, a scale of 1-5 used, where 1 means that the process is incomplete and 5 means that the process is incomplete dominated and compared with the evaluation sessions at the end of the intervention, the following results were obtained after the application of TEARA in the participants, Figure 4 presents the results grouped by process in the graphs each participant is shown evaluating in the diagnostic session before TEARA and after use.

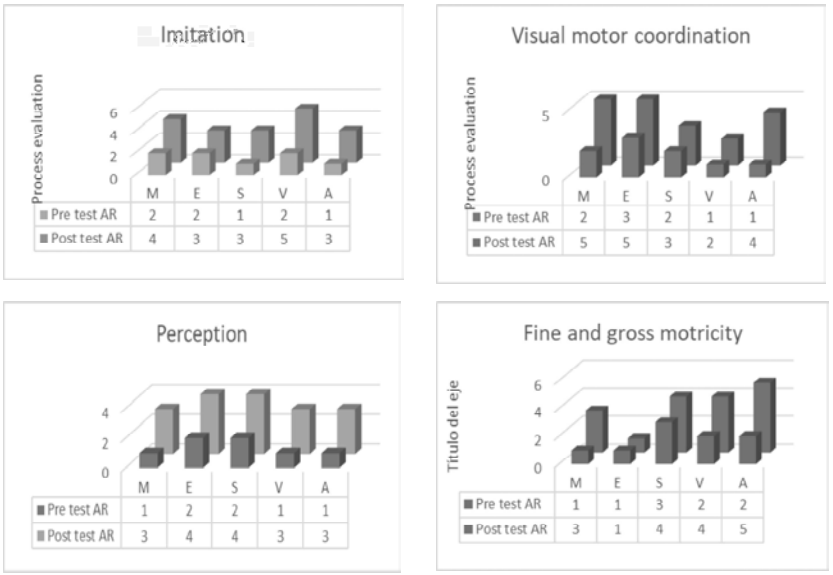


Fig.4. Results obtained in the diagnostic session of children with ASD contrasted with the results obtained after the use of TEARA.

3 Discussion and conclusion

The advancement of science and technology is increasing, the teaching-learning processes are being renewed and strengthened over the years, there is also evidence that Information and Communication Technologies (ICT) are assertive in treatments in children with ASD, because they cause a special motivation for their use, they have been used for about a decade[24].

However, from a review of the literature it was possible to show that the interventions of children with ASD are maintained over the years, these treatments that have been widely disseminated have been maintained over the years, without a change, updates or innovation for decades, we analyze its limitations and after reviewing that new technologies have strengthened various motor, cognitive and communication areas, we conducted this research defining a new Educational Treatment mediated through augmented reality that we call TEARA as a disruptive way and innovative way of strengthening teaching-learning processes in children with ASD.

TEARA was developed as a doctoral research, carried out in the LINTI New Computer Technologies Research Laboratory of the National University of La Plata, Argentina, the time allocated to its analysis, design and implementation was

approximately five years, during this time work was carried out taking into account the needs of children with ASD and with the opinion of experts who are part of their immediate context, we refer to the team of medical professionals, academics and parents who share their day-to-day lives, which have allowed through their knowledge and experience to develop a training system called Software Hope that, from a playful proposal, supports cognitive and communicational areas of ASD children.

In this investigation, the phases on which TEARA is based are indicated, which are well defined, being 3: planning, execution and review, the treatment is based on the execution of 36 sessions where 09 of them are destined for a planning stage, 25 of them have the intention of intervening directly with the child with ASD and the final 2 sessions try to carry out an evaluation and feedback on the usefulness of this treatment.

TEARA can used in both therapy centers and primary settings, or in the homes of children with ADS. It designed to reinforce certain teaching-learning processes such as imitation, visual coordination, perception, motor skills, however, it has shown that in addition to these processes it helps children with ASD by promoting verbal and non-verbal communication.

The training system that is a fundamental part of this proposal has been continuously improved by adding better options, the software is very easy to use and very intuitive, it is designed so that in a first stage the therapists or teachers accompany the child with ASD and that in subsequent sessions as the child gets used to its use, the child can use it alone.

The intention behind TEARA is the definition of a methodology for the inclusion of new technologies as a teaching-learning strategy, which broadly encompasses the considerations that must be taken into account before, during and after their use; the intention is not only the use of a software created for children with ASD, but also a comprehensive intervention proposal that is structured in a complete way, that is, in the planning phase several actions are carried out and it is essential to allocate time to the elaboration of an intervention plan that defines objectives in a particular way for each child, the diagnostic sessions and subsequent evaluation allowed us to verify that TEARA was effective in strengthening the teaching-learning processes.

TEARA had a very positive impact on the children who used it, the processes for the most part improve remarkably, when reviewing the records of the diagnostic sessions and the evaluation sessions an increase is evident, the children showed particular interest in the use of the Hope system , and according to the clinical observation, the children finished the therapy with a better mood predisposition, their behavior favored the activities that were carried out after the intervention. Figure 5 shows a child using TEARA.

Fig5.TEARA envisioned a child with ASD participating through the training system using Hope Software.

The limitations that we find when using TEARA can be defined in two central points; On the one hand, there is a certain fear on the part of the teachers and therapists of the use of new strategies, at the beginning the professionals presented resistance to a change in the way of imparting the therapy, the same one that is always carried out in the same way, on the other On the other hand, the limitation is that to use TEARA, the center must have a laptop or a television in addition to acquiring a Kinect device whose market value is not high but can be considered a budgetary limitation. As future work, we encourage the use of TEARA in children with mild ASD as the participants diagnosed with moderate and severe ASD.

References

[1] I. Málaga, R. B. Lago, A. Hedrera-Fernández, N. Álvarez-álvarez, V. A. Oreña-Ansonera, and M. Baeza-Velasco, "Prevalence of autism spectrum disorders in USA, Europe and Spain: Coincidences and discrepancies.," *Medicina (B. Aires).*, vol. 79, no. 1, pp. 4–9, 2019.

[2] I. Journal and N. E. Issn, "El Trastorno del Espectro Autista (TEA) y el uso de las Tecnologías de la información y comunicación (TIC)," *Int. J. New Educ.*, 2019, doi: 10.24310/ijne2.2.2019.7447.

[3] A. Hervás Zúñiga, N. Balmaña, and M. Salgado, "Los trastornos del espectro autista : aportes convergentes," *Pediatr. Aten. Primaria*, vol. XXI, no. 2, pp. 92–108, 2017,

[Online]. Available: https://www.pediatriaintegral.es/wp-content/uploads/2017/xxi02/03/n2-092-108_AmaiaHervas.pdf.

[4] L. E. Contini, F. Astorino, and D. C. Manni, "Estimación de la prevalencia temprana de Trastornos del Espectro Autista. Santa Fe-Argentina," *Boletín Técnico*, vol. 13, pp. 12–13, 2017.

[5] E. Bleuler, E. Minkowski, and S. Manual, "El trastorno del espectro autista : aspectos etiológicos , diagnósticos y terapéuticos," vol. 55, no. 55, 2017.

[6] K. Chawarska, A. Klin, R. Paul, S. Macari, and F. Volkmar, "A prospective study of toddlers with ASD: Short-term diagnostic and cognitive outcomes," *J. Child Psychol. Psychiatry Allied Discip.*, vol. 50, no. 10, pp. 1235–1245, 2009, doi: 10.1111/j.1469-7610.2009.02101.x.

[7] P. M. Ruiz-Lázaro, M. Posada de la Paz, and F. Hijano Bandera, "Trastornos del espectro autista: Detección precoz, herramientas de cribado," *Pediatría Atención Primaria*, vol. 11, pp. 381–397, 2009, doi: 10.4321/s1139-76322009000700009.

[8] M. T. Sánchez Rodríguez, S. Collado Vázquez, P. Martín Casas, and R. Cano de la Cuerda, "Neurorehabilitation and apps: A systematic review of mobile applications," *Neurologia*, vol. 33, no. 5. Spanish Society of Neurology, pp. 313–326, Jun. 01, 2018, doi: 10.1016/j.nrl.2015.10.005.

[9] S. Suparjoh, "The Potential of Augmented Reality to Support the Interest-based Learning of Children with Autism Spectrum Disorder (ASD)," *Adv. Soc. Sci. Educ. Humanit. Res.*, vol. 388, no. Icse, pp. 50–56, 2019.

[10] I. J. Lee, C. H. Chen, C. P. Wang, and C. H. Chung, "Augmented Reality Plus Concept Map Technique to Teach Children with ASD to Use Social Cues When Meeting and Greeting," *Asia-Pacific Educ. Res.*, vol. 27, no. 3, pp. 227–243, 2018, doi: 10.1007/s40299-018-0382-5.

[11] C. Lasheras Díaz, "La realidad aumentada como recurso educativo en la enseñanza de Español como lengua extranjera. Propuesta de intervención a partir de un manual," p. 63, 2018, [Online]. Available: https://reunir.unir.net/bitstream/handle/123456789/7039/LASHERAS DÍAZ%2C CARLOS.pdf?sequence=1&isAllowed=y%0Ahttps://reunir.unir.net/handle/123456789/7039.

[12] O. Gali-Perez, B. Sayis, and N. Pares, "Effectiveness of a Mixed Reality system in terms of social interaction behaviors in children with and without Autism Spectrum Condition," *ACM Int. Conf. Proceeding Ser.*, 2021, doi: 10.1145/3471391.3471419.

[13] P. M. Kellidou, M. Kotzageorgiou, I. Voulgari, and E. Nteropoulou Nterou, "A Review of Digital Games for Children with Autism Spectrum Disorder," *ACM Int. Conf. Proceeding Ser.*, pp. 227–234, 2020, doi: 10.1145/3439231.3439270.

[14] C. Pamparău and R. D. Vatavu, "A Research Agenda Is Needed for Designing for the User Experience of Augmented and Mixed Reality: A Position Paper," *ACM Int. Conf. Proceeding Ser.*, pp. 323–325, 2020, doi: 10.1145/3428361.3432088.

[15] K. Khowaja *et al.*, "Augmented reality for learning of children and adolescents with autism spectrum disorder (ASD): A systematic review," *IEEE Access*, vol. 8, pp. 78779–78807, 2020, doi: 10.1109/ACCESS.2020.2986608.

[16] J. Rodríguez Medina, "Mediacion entre iguales, competencia social y percepcion interpersonal de los ninos con TEA en el entorno escolar," 2019, doi: 10.35376/10324/39475.

[17] M. Romero, E. Macas, I. Harari, and J. Diaz, "Is It Possible to Improve the Learning of Children with ASD Through Augmented Reality Mobile Applications?," *Commun. Comput. Inf. Sci.*, vol. 1194 CCIS, pp. 560–571, 2020, doi: 10.1007/978-3-030-42520-3_44.

[18] M. Romero, E. Macas, I. Harari, and J. Díaz, "Eje integrador educativo de las TICS: Caso de EstudioNiños con trastorno del espectro autista.," *SAEI - Simp. Argentino Educ.*

en Informática, pp. 171–188, 2019.

[19] M. Romero, J. Díaz, and I. Harari, "Impact of information and communication technologies on teaching-learning processes in children with special needs autism spectrum disorder," *XXIII Congr. Argentino Ciencias la Comput.*, pp. 342–353, 2017, [Online]. Available: https://www.researchgate.net/publication/341282542.

[20] M. Romero and I. Harari, "Uso de nuevas tecnologías TICS -realidad aumentada para tratamiento de niños TEA un diagnóstico inicial," *CienciAmérica Rev. Divulg. científica la Univ. Tecnológica Indoamérica*, vol. 6, no. 1, pp. 131–137, 2017, [Online]. Available: https://dialnet.unirioja.es/descarga/articulo/6163694.pdf.

[21] M. Romero, I. Harari, J. Diaz, and E. Macas, "Proyecto Esperanza: Desarrollo de software con realidad aumentada para enseñanza danza a niños con transtorno del espectro autista.," *Rev. Investig. Talent.*, vol. 9, no. 1, pp. 99–115, 2022.

[22] M. Romero, I. Harari, J. Diaz, and E. Macas, "Hoope Project: User-centered design techniques applied in the implementation of augmented reality for children with ASD.," *Int. Conf. Human-Computer Interact. (pp. 277-290).*, no. Springer, Cham., pp. 277–290, 2022.

[23] M. B. L. Roberto Hernández Sampieri, *Metodología de la Investigación*. 2010.

[24] M. Romero, I. Harari, J. Diaz, and J. Ramon, "Augmented reality for children with Autism Spectrum Disorder. A systematic review.," *Int. Conf. Intell. Syst. Comput. Vision, ISCV 2020*, vol. 5, 2020, doi: 10.1109/ISCV49265.2020.9204125.

SETIC: un Software Educativo sobre el Funcionamiento de las Partes de un Computador

Maximiliano Gauthier[1], Paola D. Budán[2],

[1] Colegio San Agustín, Mar del Plata, maximilianogauthier@gmail.com
[2] Universidad Nacional de Santiago del Estero, Dpto. de Informática, pbudan@unse.edu.ar

Abstract. In this paper, we present the design and development of a prototype called *SETIC*, which is software for teaching information and communication technology concepts. *SETIC's* prototype includes several multimedia resources and is destinated for secondary school students who must learn different abstract concepts. *SETIC* was designed following the MeiSE methodology that combines a traditional cycle of life system development, with agile methods, also considering a pedagogical point of view as aggregate value for the software.

Keywords: prototype – MeiSE methodology – multimedia resources – educational software - information and communication technology concepts.

1 Introducción

Cuando se trabaja con alumnos del nivel secundario en asignaturas de enseñanza de la tecnología, como por ejemplo con adolescentes de los últimos años que cursan NTICs, es posible observar dificultades para que los mismos entiendan y puedan explicar los procesos computacionales que lleva a cabo el computador en el manejo de la entrada de datos, su procesamiento, almacenamiento hasta llegar a la salida de la información en algún tipo de formato. Estos conocimientos son básicos para quienes desean articular, posteriormente, con el cursado de alguna carrera Informática en niveles educativos superiores y no desertar rápidamente. Atendiendo específicamente a la problemática particular que se presenta en el *4º año del Colegio San Agustín de la ciudad de Mar del Plata*, se propone diseñar y prototipar un Software Educativo (SE) cuyo principal objetivo es *fomentar el conocimiento de los procesos informáticos que se lleva a cabo dentro del computador*. Este software se denomina *SETIC (Software Educativo de las Tecnologías de Información y Comunicación)* y se enmarca en la asignatura *NTICx* que por el plan educativo de la provincia de Buenos Aires es curricular. Es necesario aclarar que en 4to año los alumnos toman contacto por primera vez con este tipo de contenidos, después de haber transitado todo el nivel primario y más de la mitad del nivel secundario. Anteriormente, no tienen ningún espacio con abordaje de saberes en alfabetización digitales. El dictado de esta asignatura es dependiente de cada institución, se puede aplicar de forma extracurricular o no reglada, y los contenidos pueden ir desde un acercamiento teórico por falta de dispositivos, hasta el uso de programación con robótica e impresión 3d en

el caso de las instituciones con más recursos económicos. Desde el *punto de vista técnico-informático*, para el desarrollo de *SETIC* se siguen los pasos detallados en la Metodología MeiSE [1], que plantea un ciclo de vida de desarrollo de software incremental guiado por prototipos, y en la que se deja libertad en cuanto a cómo alcanzar esos prototipos, por ejemplo, usando algún método ágil. Para el desarrollo de SETIC, el SE se considera compuesto de mini-proyectos: *(a) Tipos de computadores, (b) Partes del computador, (c) Periféricos (entrada, salida, e/s, almacenamiento), (d) Tipos de datos, (e) Digitalización de la información*, y *(f) Representación binaria de los datos digitales.* Cada uno de éstos se aboca a un tema en particular relacionado con la currícula de las NTICx. En este trabajo, la etapa de definición y desarrollo se detalla para *SETIC* en general, mientras que lo que corresponde a la etapa de despliegue e implementación se contextualiza en el mini-proyecto *Periféricos (entrada, salida, e/s, almacenamiento)*.

El trabajo se estructura de la siguiente manera: en primer lugar, se describe el contexto de aplicación del software y sus objetivos generales; luego se resume la metodología elegida para el desarrollo en sus etapas generales. Posteriormente, se describen las actividades que conforman cada una de estas etapas, y por último se esbozan algunas conclusiones y trabajo a futuro.

2 *SETIC* y su contexto de aplicación

El presente trabajo se enmarca en el cursado de la asignatura *Desarrollo de Software Educativo*, que forma parte de las obligaciones curriculares de la *Maestría en Informática Educativa* que ofrece la Universidad Nacional de Santiago del Estero[1]. Desde esta obligación curricular se identifica una situación problemática plausible de ser resuelta mediante el desarrollo de un SE. Así, *SETIC* se diseña y comienza a prototipar, teniendo en cuenta la población de los alumnos de *4to año de secundaria que asisten al Colegio San Agustín de la ciudad de Mar del Plata*. Sintéticamente, cuando se abordan contenidos relacionados al procesamiento de la información, estos alumnos requieren un alto grado de abstracción para comprenderlos, pues todos los procesos que se llevan a cabo son electrónicos y no pueden ser tangibles ni visibles en tiempo real, al mismo tiempo que se llevan procesan a una altísima velocidad. Esto genera la falta de comprensión y análisis de dichos procesos, que terminan siendo memorizados, en su gran mayoría, por repetición conceptual teórica pero no empírica. Así, el alumno queda en una instancia de incertidumbre ante el contenido, sin terminar de entender bien el proceso que se está llevando a cabo, y eso puede ser un factor fundamental de decisiones futuras ante qué carreras elegir posteriormente. Al no realizar una alfabetización digital intermedia entre lo teórico y lo práctico, los alumnos no logran comprender cómo realmente funciona un computador, sin lograr aprender ni aprehender su funcionamiento real. *SETIC* busca situarse entre la teoría y la práctica, agilizar los procesos cognitivos de comprensión alejándose de la abstracción y demostrarlo visualmente, a fin de intentar conseguir un aprendizaje significativo que logre estimular vocaciones relacionadas a la informática.

[1] Resolución Ministerial 2270/2014

Así, los objetivos generales de *SETIC* son: *(i) construir una herramienta que permita identificar las partes del computador y el funcionamiento individual de cada una de ellas; (ii) ilustrar la diferencia entre los periféricos de ingreso o egreso de datos; (iii) identificar los componentes principales y las interacciones/relaciones entre ellos en el proceso; (iv) mostrar las funciones de cada uno de los componentes; (v) simular el flujo de datos del proceso computacional desde su ingreso hasta su salida pasando por el procesamiento y/o almacenamiento*. Cabe aclarar que este tipo de iniciativas cuentan con el apoyo de los directivos del Colegio San Agustín. Se decide generar un Software y no reutilizar uno existente dado que se quiere obtener un producto a medida para el Colegio.

3 MeiSE: Metodología para el Desarrollo del Software Educativo

Para el diseño y prototipado inicial de *SETIC,* se selecciona la metodología MeiSE [1], porque presenta las etapas necesarias para balancear la calidad técnica de un software con los aspectos requeridos tanto para su uso pedagógico como constituirse en una herramienta didáctica. Esto es así dado que MeiSE incluye criterios en su desarrollo para que el contenido pueda ser comprendido por el alumno a la vez que procura atender las características didácticas.

Brevemente, la Metodología consta principalmente de dos grandes *etapas*: *definición* y *desarrollo*. La primera se divide en tres *fases*: la *fase conceptual*, durante la cual se identifican los requerimientos del sistema y se define el plan de desarrollo; la *fase de análisis y diseño inicial*, en la que se propone la arquitectura de base para la solución y se establecen las características pedagógicas y de comunicación que regirán el desarrollo; finalmente la *fase de plan de iteraciones*, en la cual se divide el proyecto en partes funcionales que permitan mejor control en su desarrollo. En la *etapa de desarrollo* se tienen: la *fase de diseño computacional*, en la que se realizará un diseño computacional detallado de un incremento específico del software; la *fase de desarrollo*, durante la cual se implementa la arquitectura en forma incremental; y la *fase de despliegue*, donde se realiza la transición del producto ejecutable al usuario final. Estas tres últimas etapas se repiten iterativamente para cada incremento del software. Esta metodología ha sido empleada en trabajos tales como [9,8,5] inicial de *SETIC*.

4. Desarrollo de SETIC

A continuación, se detallan las actividades llevadas a cabo para las fases correspondientes a la etapa de *definición y desarrollo* y a la *etapa de implementación y despliegue.*

4.1 Fase Conceptual

En esta fase se analizan las necesidades educativas, las alternativas de solución (si

existieran), se identifica la funcionalidad del software, se conforma el equipo de trabajo y el plan inicial de desarrollo, y se establecen los criterios de calidad. Para ello, se espera obtener como artefactos el *modelo instruccional*, una *lista de requerimientos*, el *estudio de alternativas* de solución, un *plan inicial* y *un modelo de aceptación* [1].

El *modelo instruccional* elegido para *SETIC* es el de ASSURE de Heinich et al. [2], basado en el enfoque constructivista conjugado con la importancia de los conocimientos previos y el contexto del estudiante. Cada tema tendrá contenidos que podrán ser abordados de forma secuencial, con navegabilidad por los contenidos, con un desplazamiento y recursividad de estos. Además, el alumno dispondrá de la posibilidad de regresar a temáticas de interés desde cualquier punto del SE. Para revertir la situación de que los procesos abstractos son impartidos en su gran medida de forma teórica, este SE propone darle una "visibilidad" mediante la representación de dichos procesos con simulaciones interactivas donde el alumno puede aprender mediante un conjunto de estímulos audiovisuales más allá de los textuales. Se prevén autoevaluaciones de diferentes abordajes (múltiple choice, verdadero o falso, puzzle), de carácter lúdico, que vaya otorgando al estudiante "puntos" por intentos y reintentos de evaluaciones.

Con respecto a los *requerimientos funcionales iniciales* se detecta que el SE debe: (i) identificar a los usuarios para su seguimiento en el proceso de aprendizaje; (ii) permitir a los usuarios seleccionar la temática; (iii) registrar los intentos de cada usuario con las actividades que desarrolla; (iv) generar una auto evaluación de conocimientos previos para comparar luego de transitar el SE; (v) mostrar la evolución a cada estudiante con la barra de avance; (vi) orientar al usuario en qué etapa se encuentra , y sugerir la siguiente; (vii) guardar los procesos que se llevan en la interfaz humano-máquina , intentos realizados, caminos tomados por cada usuario; (viii) mostrar cómo se produce el ingreso y egreso de los datos mediante periféricos de entrada y salida. Como *requerimiento docente* se destaca el generar informes de aprendizaje para el docente en el que se consignen los resultados obtenidos por cada alumno según el camino tomado, la velocidad de aprendizaje según los contenidos previos.

En cuanto al *estudio de alternativas* de solución, se contempla la *alternativa 1*: uso de videotutoriales separados por temáticas, en los que el alumno elige la temática a abordar, visualiza el contenido multimedial en tramos, y que al finalizar el mismo se despliegan las opciones de autoevaluación; y la *alternativa 2* bajo el formato de podcast educativos cortos que expliquen los contenidos, y luego se ejecute la ventana de elecciones múltiples, donde el alumno deba elegir la respuesta según lo recién oído. En caso de error puede elegir volver a escuchar o avanzar. Cada elección dotará de un puntaje según el perfil del alumno en el sistema. Se elige la alternativa 1, formato de video tutoriales, porque se acerca más al objetivo de mostrar visualmente mediante la simulación, presentación o explicación multimedia los procesos informáticos que se llevan a cabo dentro del computador.

El *plan inicial de desarrollo* se consigna en la Tabla I, mientras que en la Fig. 1 se le asigna a cada tarea los responsables de llevarla a cabo.

			1S JUN	2S JUN	3S JUN	4S JUN	1S JUL	2S JUL	3S JUL	4S JUL	1S AGO	2S AGO	3S AGO	4S AGO
Actividad	1	Resolución de modelo educativo												
Actividad	2	Propuesta de interfaz de usuarios												
Actividad	3	Esquema de navegación por módulos												
Actividad	4	Decisión de contenidos educativos												
Actividad	5	Decisión relación contenidos y módulos												
Actividad	6	Proponer autoevaluaciones												
Actividad	7	Prototipo inicial												
Actividad	8	Creación de interfaz usuario												
Actividad	9	Carga de contenido multimedia												
Actividad	10	Realización de formato de autoevaluaciones												
Actividad	11	Asignación de puntajes												
Actividad	12	Formatear reportes												
Actividad	13	Puesta en marcha incial												

Tabla I. Plan inicial de desarrollo para *SETIC*

Para generar el modelo de aceptación se permitirá intervenir a un grupo real de alumnos de la institución. Se realizará un muestreo de los resultados en 60 días de trabajo.

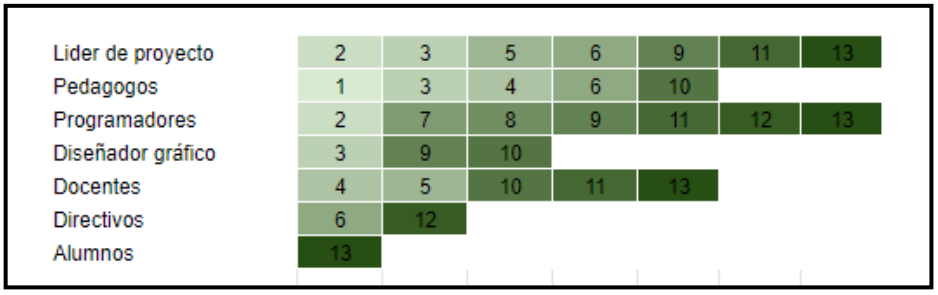


Fig. 1. Tareas y actores responsables de llevarlas a cabo

Se prevé realizar la medición de la calidad de SETIC en 3 niveles: (1) *Docentes*: facilidad para incorporar contenidos, posibilidades de monitoreo de los datos recopilados, resultados pedagógicos sobre propuestas didácticas. Devoluciones de ventajas y desventajas encontradas; (2) *Directivos*: se involucrará al directivo y su visión sobre la planificación del SE; y (3) *Alumnos*: es el momento donde se realizará un muestro del impacto académico, para medir impacto sobre el aprendizaje, facilidad de uso devoluciones sobre el sistema de puntajes y opiniones sobre el formato multimedial e interactivo.

4.2 Análisis y Diseño Conceptual

En esta fase se detalla una primera arquitectura del SE, se elabora el modelo educativo y se elabora el modelo comunicacional. Los artefactos que se esperan de esta fase están relacionados a estos modelos, y especialmente se destacan el modelo de interfaz, modelo de navegación y prototipo de la interfaz de usuario [1].

La arquitectura general de *SETIC* se describe en la Fig. 2, donde los módulos representan cajas negras cuyas funcionalidades están autocontenidas en el nombre del módulo.

El primer prototipo de *SETIC* se desarrollará en APP INVENTOR [3-6], por ser una plataforma para desarrollo rápido de los módulos para el manejo de los usuarios y perfiles de los actores involucrados. Además, permite que los recursos multimedia necesarios se ejecuten de forma local o a links externos para lograr más dinamismo en

el sistema, y generar una plataforma más liviana.

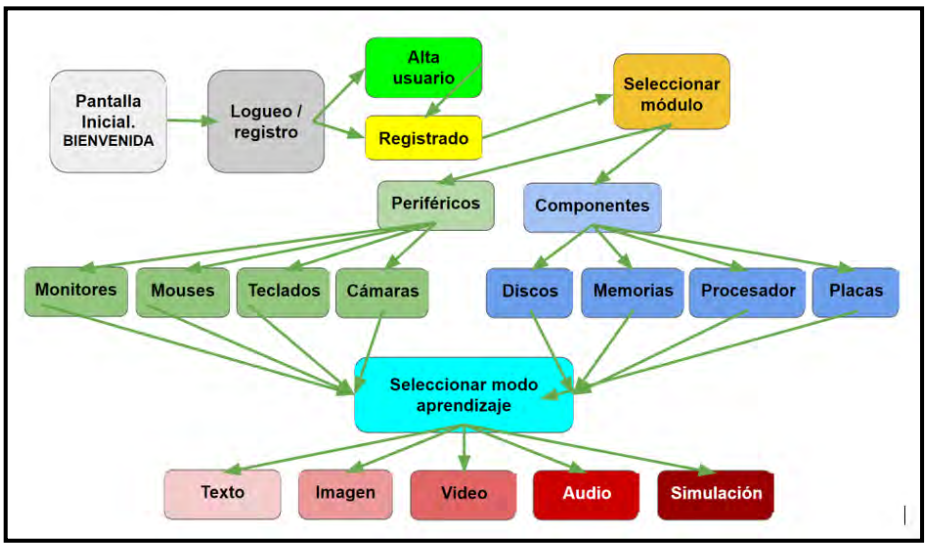


Fig. 2. Arquitectura modular para *SETIC*

La ejecución de los distintos tipos de evaluaciones podrá será implementada de forma rápida mediante el uso de botones, check y armado de puzles, pues la misma soporta el manejo de objetos con Canvas, interacciones por pantallas y sensores.

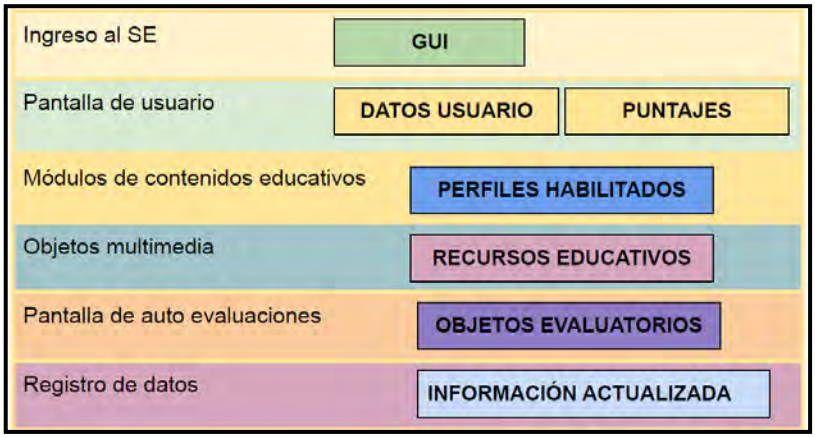


Fig. 3. Arquitectura de las interfaces para *SETIC*

Para el uso en PC, en esta primera etapa se instalará el emulador BLUESTACK[2] para poder ejecutar la aplicación sin tener que crearla para Windows. La Fig. 3 ilustra una primera aproximación de arquitectura de interfaces. Por otra parte, la Fig. 4

[2] https://www.bluestacks.com BlueStacks App Player está diseñado para permitir que aplicaciones de Android puedan ejecutarse en computadoras de Windows y Macintosh

muestra los modelos de interfaz con el usuario sea para actividades de aprendizaje como para aquellas de evaluación.

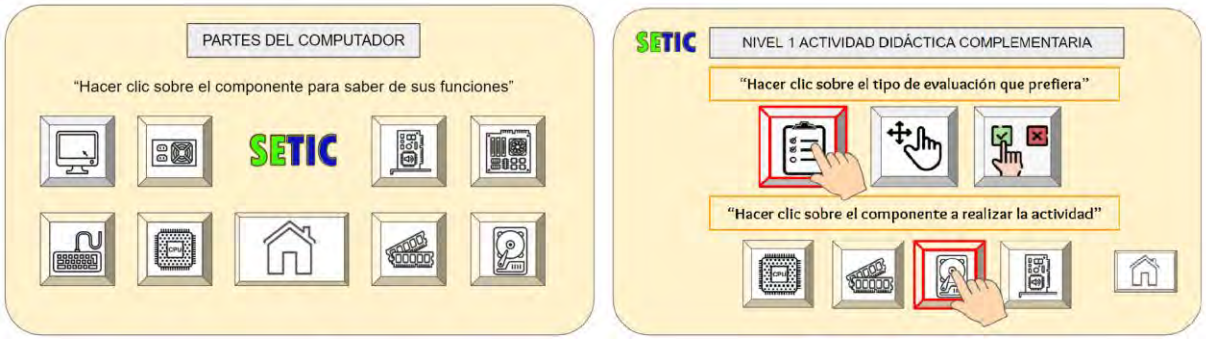


Fig. 4. Modelo de interfaz de la pantalla "Partes de un Computador" y de las Actividades de Autoevaluación.

Y, por último, en la Fig.5 se pueden apreciar los modelos de pantallas para un usuario que ya se encuentra haciendo uso de la aplicación.

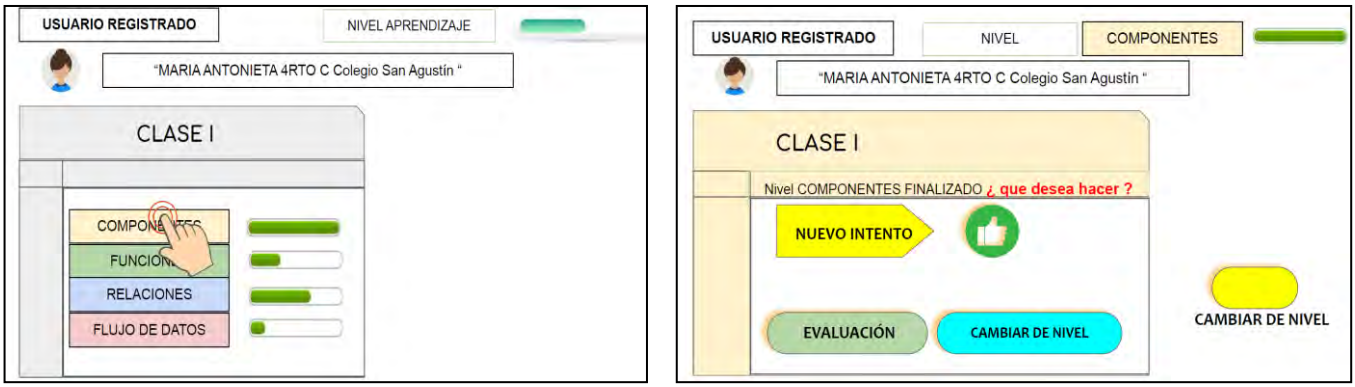


Fig. 5. Modelo de interfaz para un usuario que se encuentra navegando por *SETIC*

4.3. Plan de Iteraciones

El desarrollo iterativo e incremental del proyecto se planifica en bloques a realizar en tiempos calendarios al que llamamos iteraciones. Las iteraciones serán en formato de mini-proyectos, donde cada uno proporcionará un resultado concreto para los usuarios sobre el producto final, de manera que los mismos puedan obtener los beneficios del proyecto de forma incremental. Para cada uno de esos mini-proyectos se prevé el desarrollo de etapas comunes: *Etapa 1*-Determinar del formato educativo y resultados esperables; *Etapa 2*-Establecer cuáles serán los contenidos curriculares en formato multimedia; *Etapa 3*-Determinar cuáles serán las pantallas interactivas entre los usuarios, considerando la exposición de contenidos y las autoevaluaciones; *Etapa 4*-

Configurar reportes. Los mini-proyectos que forman para obtener el *SETIC* están pensados para que el docente puede proponerlos de manera gradual, según cómo quiera que los contenidos sean aprendidos por los usuarios (alumnos) y con cierta coherencia entre los mismos. Todos ellos tienen como propósito presentar una escalabilidad para el aprendizaje, para que cada docente puede decidir utilizar secuencialmente o en forma individual cada mini-proyecto. Estos son: *(a) Tipos de computadores, (b) Partes del computador, (c) Periféricos (entrada, salida, e/s, almacenamiento), (d) Tipos de datos, (e) Digitalización de la información*, y *(f) Representación binaria de los datos digitales.*
Hasta aquí se han detallado las actividades para las fases incluidas en la etapa de *definición y desarrollo.* En lo que sigue, nos enfocaremos en los productos que se han desarrollado en la etapa de *implementación*.

4.4 Productos Obtenidos en las Primeras Iteraciones

Para el desarrollo de *SETIC* se prevé la construcción de recursos multimediales de variada naturaleza, según el mini-proyecto que se aborde. A modo ilustrativo, si se toma como referencia el mini-proyecto *Periféricos (entrada, salida, e/s, almacenamiento)*, el recurso multimedial que forma parte de este proyecto es un objeto de aprendizaje (OA)[3] [4,7] cuyas generalidades se pueden apreciar en la Fig. 6 y que se puede profundizar en el enlace https://youtu.be/V_z2ZoW6Zv0.

Fig.6. Recurso multimedial creado para el mini-proyecto *Periféricos.*

A su vez, dentro de este recurso, el alumno cuenta con las opciones de un juego interactivo. Se trata de un juego, en el cual el alumno debe hacer clic sobre los componentes indicados (tipos de periféricos), y el juego le presentará opciones a elegir, otorgándole luego puntos por aciertos o desaciertos en un tiempo determinado.

[3] Es importante notar que el OA es solo un recurso inserto en *SETIC*, y que simplemente se lo concibe como un módulo en el mini-proyecto.

Al finalizar dará un "*score*" tiempo/aciertos/desaciertos. Si bien esta actividad es de carácter lúdico, lleva implícita una autoevaluación para el alumno y una evaluación de la evolución del alumno de la interacción con la actividad y el contenido. Es por ello que el tipo de imágenes que se incorporan al juego deben ser representativas y claras en cuanto al componente detallado. Parte de esta actividad puede apreciarse en el recurso https://view.genial.ly/629b9bf17dcdb50019b9f470/interactive-content-fichas-de-refuerzo-repaso-ampliacion. A los fines de dejar ilustrado el modelo de navegación para cuando el usuario ingresa, por ejemplo, a la opción *tipo de pantallas* del OA referenciado, es de utilidad la fig. 7.

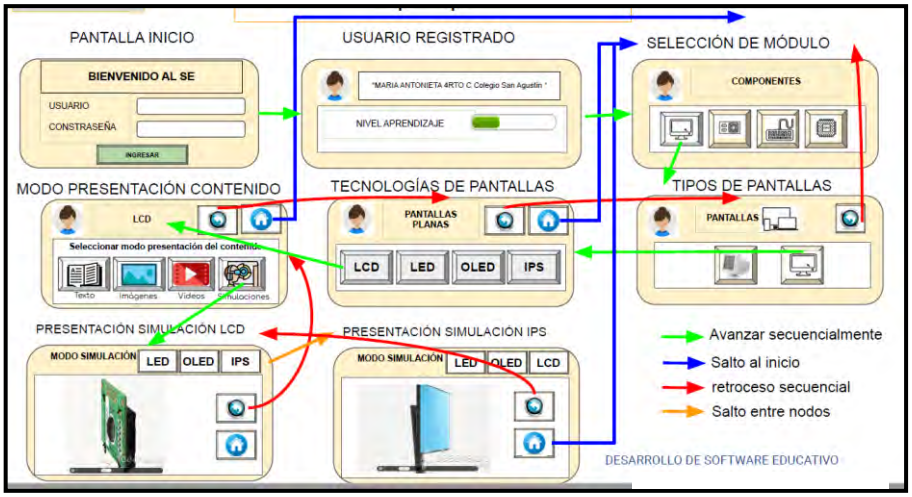


Fig.7. Modelo de navegación mini-proyecto *Periféricos – Tipos de pantallas*

Siguiendo líneas de diseño técnico y pedagógico como las ejemplificadas, las iteraciones sucesivas para lograr *SETIC* parten de la hipótesis de que el aprendizaje de los temas propuestos está ligado a la dificultad inherente al nivel de abstracción que deben manejar los alumnos para entender el proceso del "viaje" de la información desde los dispositivos que ingresan datos al sistema (mini-proyecto periféricos de entrada), su procesamiento en el dispositivo (mini-proyectos partes del computadora y representación binaria de los datos), sus opciones de guardado (mini-proyectos dispositivo de almacenamiento, internos, externos o en la nube) hasta su posible salida mediante los dispositivos que muestren el dato o información sea textual, sonora, gráfico (mini-proyecto periféricos de salida).

5 Conclusiones y Trabajo Futuro

En este trabajo se ha detallado el proceso de diseño de *SETIC,* compuesto de seis mini-proyectos y el desarrollo de las iteraciones del mini-proyecto *Periféricos*. *SETIC* viene a cubrir una necesidad real y, si bien se enmarca en una asignatura puntual, puede ser utilizado como un medio pedagógico y didáctico para facilitar el proceso de

enseñanza de la asignatura, como así también pare fortalecer el autoaprendizaje, el repaso o incluso ser utilizado en diferentes experiencias de nivelación/ articulación para materias correlativas del mismo nivel educativo u otros niveles educativos superiores (terciarios, universitarios, cursos, capacitaciones). *SETIC* actualmente está atravesando la fase de prueba y el link de la APK se encuentra en https://drive.google.com/file/d/1rXGqdhFi5odTFZjW6gkxDClJu4DU1eoP/view?usp=sharing
El trabajo plasma las bondades que un SE brinda para que el alumno se acerque a un contenido tan abstracto como el proceso computacional mediante el uso de distintos formatos de presentación del contenido, sobre todo el animado y/o simulado. Desde el punto de vista técnico, la metodología MeiSE seguida permite obtener el producto sin exigentes requerimientos de programación, permitiendo que los esfuerzos se destinen a las utilidades pedagógica y didáctica del software. Pedagógica porque considera el modelo de aprendizaje, y didáctica debido a que permite obtener una herramienta para mejorar el proceso de enseñanza. Sin embargo, sería deseable que *esta separación de tareas mencionadas como mini-proyectos esté explícita en la metodología, es decir, que el abordaje de MeiSE contemplara la existencia de mini-´proyectos para aquellos autores que desean desarrollar los recursos y no tienen diversas habilidades en la programación*. Como línea de trabajo futura queda el desafío del prototipado de los restantes mini-proyectos y la profundización de la evaluación de la herramienta *in-situ*. Los resultados que se pudieran obtener de esta experiencia nos motivarían a contemplar un dominio más amplio relacionando con más contenidos de interpretación abstracta.

Referencias

1. Abud, A. "MeiSE: Metodología de Ingeniería de Software Educativo, Vol. 2, N 1." Revista Internacional de Educación en Ingeniería, Instituto Tecnológico de Orizaba, Drizaba, Veracruz, México (2009).
2. Hernández-Alcántara, Mariana, Genaro Aguirre-Aguilar, and Jorge Arturo Balderrama-Trápaga. "Revisión del modelo tecnoeducativo de Heinich y colaboradores (ASSURE)." Los Modelos Tecno-Educativos, revolucionando el aprendizaje del siglo XXI (2014): 61-72.
3. Inventor, MIT App, and M. I. T. Explore. "App inventor." línea]. Disponible en: http://appinventor. mit. edu/explore/.[Accedido: 26-may-2015] (2017).
4. Massa, Stella Maris. Objetos de aprendizaje: Metodología de desarrollo y Evaluación de la calidad. Diss. Universidad Nacional de La Plata, 2013.
5. Ocsa, Alexander, et al. "Propuesta Para El Diseño Y Desarrollo De Aplicaciones M-Learning: Caso, Apps De Historia Del Perú Como Objetos De Aprendizaje Móviles." Nuevas Ideas En Informática Educativa TISE (2014): 873-878.
6. Patton, Evan W., Michael Tissenbaum, and Farzeen Harunani. "MIT app inventor: Objectives, design, and development." Computational thinking education. Springer, Singapore, 2019. 31-49.
7. Sanz, C., F. Barranquero, and L. Moralejo. "Metodología para la creación de Objetos de Aprendizaje CROA. Consultado el 15 de abril del 2017." (2015).
8. Torres-Carrion, P., C. González González, and Jaime-Edwin Basurto-Ortiz. "Diseño de un juego serio para la mejora de la conciencia fonológica de los niños con dislexia." IEEE 11 Congreso Colombiano de Computación. 2016.
9. Valdivia, Ricardo Castro, María Reina Zarate Nava, And Jesús Leonardo López Hernández. "Aplicación de TI en el proceso de enseñanza-aprendizaje para la adquisición de conocimientos de matemáticas 1er. grado de educación secundaria “Imat”." (2017).

Utilización de Encuestas para el seguimiento y diagnóstico continuo

Paola Caymes Scutari[1,2] y Germán Bianchini[1],

[1] Laboratorio de Investigación en Cómputo Paralelo Distribuido (LICPaD, UTN-FRM)
Rodríguez 273 CP 5500 Mendoza, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

{pcaymesscutari, gbianchini}@frm.utn.edu.ar

Abstract. El proceso o sistema de evaluación, entendido como un proceso formativo de mejora continua, ha de contemplar distintos tipos de evaluación (autoevaluación, coevaluación y heteroevaluación), tanto formativa como sumativa. A su vez, la implementación de metodologías mixtas que combinan actividades síncronas y asíncronas, y el uso de modelos de aula invertida, requieren que el docente tenga un seguimiento permanente del grado de avance de los alumnos, a fin de asistir, orientar, y/o revisar los saberes relacionados con cada actividad o tema propuesto. En este artículo se presenta una experiencia implementada en el marco de la carrera de Ingeniería en Sistemas de Información, mediante el uso de encuestas de Moodle como una herramienta para el seguimiento en cuanto al nivel de comprensión y maduración de los diferentes temas.

Keywords: encuestas, estadística, alumnos, aprendizaje, evaluación formativa.

1 Introducción

Desde el punto de vista formativo, la reflexión, la constancia, y el seguimiento continuo, son factores primordiales para que la evaluación se encuentre integrada y sea útil al proceso de aprendizaje, tanto si es realizada por los propios alumnos, como por el docente. En este sentido, no son los instrumentos, las técnicas, o las posibles calificaciones en sí mismos los que constituyen la evaluación, sino que se trata de un proceso reflexivo continuo e integrado, que además es importante naturalizar e incorporar en la formación de los profesionales, ya que la actividad de evaluación es y será parte intrínseca en el ejercicio de la profesión, siempre que se analicen, propongan, y/o evalúen proyectos, planes o cronogramas de trabajo, impactos, riesgos, uso de diferentes recursos, instalaciones, presupuestos, etc. [7].

El proceso o sistema de evaluación, entendido como un proceso formativo de mejora continua, que a partir de métodos cuantitativos y cualitativos permita identificar y recolectar datos para diagnosticar y determinar el nivel de comprensión de los temas, de dominio de saberes, y del logro de los resultados del aprendizaje, sin dudas debe ser continuo, y debe contemplar distintos tipos de evaluación, por lo que

es deseable que se aborde desde diferentes puntos de vista [3]. Tanto puede contemplarse quién evalúa, cuándo evalúa, y para qué evalúa. Considerando *quién*, pueden distinguirse la autoevaluación, la coevaluación, y la heteroevaluación [7], [12]. Considerando *cuándo*, se identifican tres momentos en relación al proceso de producción personal y aprendizaje determinados por el antes, el durante, y el después [7], [4], [11]. Finalmente, teniendo en cuenta *para qué*, se considera que la evaluación puede ser formativa o sumativa [7], [10].

Esta clasificación no es única ni cerrada, pero nos permite caracterizar la experiencia abordada en este artículo, con el claro objetivo de implementar de una forma sencilla y colaborativa el seguimiento de los alumnos, en cuanto al grado de avance y comprensión de los sucesivos temas, especialmente en un esquema de aprendizaje autorregulado [4], [13]. Se trata de una experiencia que combina autoevaluación con héteroevaluación, pues cada alumno responde un breve cuestionario (implementado mediante una encuesta, como se presenta en la sección 2) de acuerdo a sus saberes; y una vez finalizado el cuestionario de seguimiento, el docente en conjunto con los alumnos analiza las respuestas y focaliza las inquietudes o falencias a revisar. A su vez, se trata de una actividad propuesta para su realización antes de comenzar con la puesta en práctica o aplicación de saberes, a fin de aclarar las posibles dudas o inseguridades que hayan surgido durante el proceso de aprendizaje autorregulado. Y finalmente, se trata de una actividad formativa, pues no busca calificar sino diagnosticar y subsanar continuamente los aspectos que, por el asincronismo del aprendizaje, la simple falta de participación, por inseguridad, o por timidez, no se manifiestan por sí solos entre las inquietudes o preguntas de los estudiantes. La actividad se implementa a través de encuestas de Moodle [8], [9] por ser la plataforma que la institución brinda para su Campus Virtual, pero no es restrictivo.

Consideramos que el seguimiento es importante como una construcción que permite que tanto los docentes como los propios alumnos detecten y tomen conciencia de las debilidades o falencias que presentan, y tomen acciones para superar la situación previo a avanzar hacia nuevas temáticas y a la realización de las evaluaciones sumativas. La experiencia presentada en este artículo fue llevada a cabo en dos asignaturas correspondientes a la carrera de Ingeniería en Sistemas de Información de la Facultad Regional Mendoza (Universidad Tecnológica Nacional). Una de las asignaturas es “Matemática Discreta”, asignatura obligatoria dictada en el primer semestre, correspondiente al primer nivel de la carrera, con cursos de entre 60 y 100 alumnos (en otras universidades los niveles se denominan “años”). La otra asignatura se denomina “Computación Paralela”, y se trata de una asignatura electiva dictada en el segundo semestre, destinada a alumnos de tercero, cuarto y quinto nivel, en un curso de entre 30 y 40 alumnos.

El contexto de ambas asignaturas es bastante diverso, dado que en el caso de Computación Paralela los alumnos ya cuentan con ciertos conocimientos previos y de base, y además de que ya se encuentran habituados a la vida universitaria, son ellos mismos quienes eligen inscribirse en la asignatura. Por su parte, los alumnos de Matemática Discreta son en su mayoría alumnos ingresantes, provenientes de diferentes contextos, y suelen tener dificultades para adaptarse al volumen y ritmo de estudio universitario, motivo por el que se produce un mayor desgranamiento. En la sección 2 se describe más detalladamente la experiencia, el recuso utilizado en

Moodle (subsección 2.1) y la metodología propuesta para su utilización (subsección 2.2). En la sección 3 se presentan los resultados obtenidos, y en la sección 4 se expresan las principales conclusiones.

2 Descripción de la Experiencia

Si bien resulta bastante normal que los docentes obtengan información de diagnóstico y seguimiento al involucrar a los alumnos en el repaso y relación de temas nuevos con los conocimientos previos, o de la práctica con la teoría, las nuevas tendencias educativas requieren ir un poco más allá. Un ejemplo lo constituyen los modelos de aprendizaje invertido [5], en los cuales el docente brinda el material y recursos (bibliografía, material de estudio, resúmenes, organizadores gráficos, videos para su visualización, audios o podcasts, herramientas, etc.) necesarios para que el alumno se prepare, revise, tome nota de dudas, y estudie, antes de asistir o participar en la clase presencial/síncrona. Y durante la clase se proponen actividades seleccionadas o diseñadas para permitir la aplicación y/o puesta en práctica de los saberes correspondientes para su resolución, con la guía y supervisión de los docentes. La necesidad de implementar este tipo de modelos se vio acelerada por la pandemia de COVID-19 que impuso una modalidad de trabajo y aprendizaje virtual/remota [6], en la que los docentes tuvimos que adaptar y reinventar nuestras prácticas a las posibilidades que brindaba la tecnología hogareña. Así, las restricciones en la calidad de conexión para el dictado de clases 100% sincrónicas, puso de manifiesto la necesidad de comenzar a implementar un modelo de aprendizaje invertido, en el que el estudiante pudiera regular su aprendizaje, y los encuentros síncronos tuvieran un objetivo más productivo que instructivo. Sin lugar a dudas, es un esquema de trabajo que requiere gran compromiso por parte del alumno, del futuro profesional, que pasa a ser protagonista de su propio aprendizaje, en lugar de ser un espectador. Así es que el docente, antes de proponer la actividad específica, ha de diagnosticar el estado colectivo de los estudiantes, en cuanto al nivel general de comprensión alcanzado, y los temas o conceptos que manifiestan mayor dificultad, a fin de acompañar el proceso de comprensión y aprendizaje, previo a la puesta en práctica.

Como se mencionó anteriormente, la experiencia que se presenta a continuación, surgió en el ciclo 2020 para la asignatura Computación Paralela durante el aislamiento por la pandemia de COVID-19, donde todas las actividades académicas tuvieron lugar de forma virtual/remota [1], y para la asignatura Matemática Discreta se comenzó a utilizar en el ciclo 2021 en vista de la utilidad que brindaban [2]. La metodología de trabajo tomó características del modelo de aprendizaje invertido, en el sentido de que tanto se brindaban las clases síncronas como el material de estudio para el autoaprendizaje. Pero si en una clase presencial es muchas veces difícil que todo el alumnado participe, plantee sus dudas e inquietudes, y se involucre, las videoconferencias llevaron esa dificultad a otro nivel. A partir de la necesidad de traspasar las pantallas y de alguna manera lograr una retroalimentación por parte de los alumnos, surgió la idea de utilizar las encuestas de Moodle [9] como una herramienta para poder realizar un diagnóstico continuo, e identificar los puntos a

fortalecer y/o reformular, generar reflexión e identificar dudas. También podrían haberse utilizado cuestionarios u otros tipos de recursos para implementar este tipo de actividades de seguimiento, pero a efectos prácticos, las encuestas fueron los recursos que mejor se adecuaban a los objetivos de la actividad.

2.1 Encuestas de retroalimentación de Moodle

El modulo o recurso ***encuesta de retroalimentación*** (también denominado *feedback*, *encuesta*, o *retroalimentación*), permite crear y realizar encuestas personalizadas [9], (ver Fig. 1). Constituye un recurso que permite aplicar encuestas con el propósito de conocer el estado de conocimiento de los estudiantes, y no de calificarlo. Esta subsección no pretende ser exhaustiva, sino que intenta condensar las principales características y posibilidades del recurso, contempladas en la vista general, la edición de preguntas, y el análisis de resultados.

Fig. 1: Encuesta de retroalimentación para la Unidad 3 de Computación Paralela 2020 (recurso de Moodle, utilizada a través del aula virtual).

Las preguntas/respuestas en las encuestas de retroalimentación no tienen calificación, y, como se aprecia en la Fig. 2, se puede configurar tanto como una encuesta anónima como con identificación, dependiendo del objetivo específico del docente y de la cátedra (por ejemplo, en cátedras que además implementen tutorías personalizadas, la encuesta con identificación puede funcionar como un medio de detección temprana de estudiantes que requieran acompañamiento específico). Asimismo, como en otros tantos recursos de Moodle [8], es posible configurar el periodo de tiempo en el que la encuesta estará habilitada (ver parte superior de la Fig. 2).

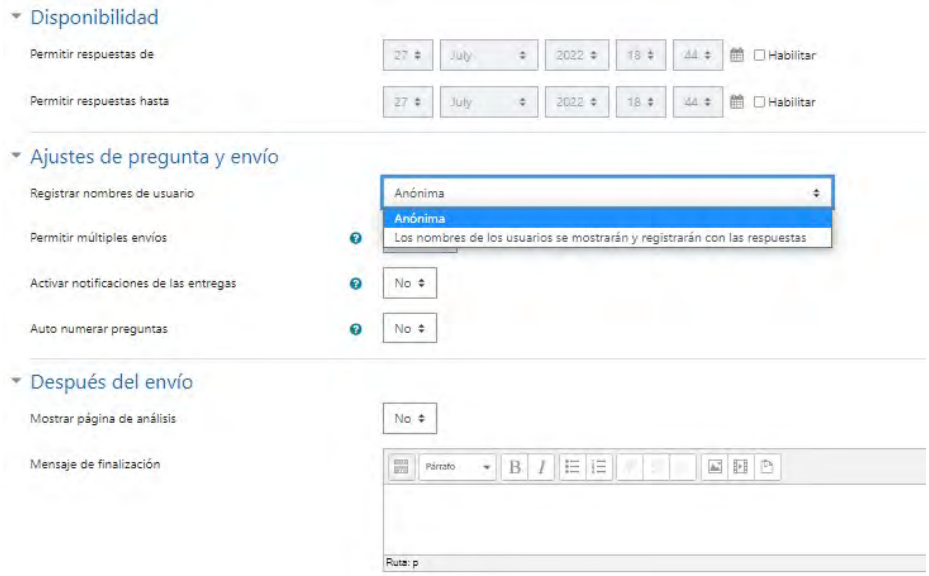


Fig. 2: Algunas opciones de configuración de las Encuestas de retroalimentación de Moodle

A la hora de crear las diferentes preguntas, el recurso ofrece distintos tipos o modalidades (ver Fig. 3) que permiten adecuarse a las necesidades del concepto que se quiera evaluar. Tanto pueden ser preguntas de opción múltiple (sea con única o múltiple respuesta), como preguntas que requieran cierto desarrollo escrito (sea breve o extenso), entre otras. Asimismo, es posible incluir figuras y gráficos utilizando el tipo de pregunta *etiqueta*. Cabe mencionar adicionalmente que las preguntas pueden duplicarse, modificarse, y reordenarse, y pueden configurarse para que sean (o no) de respuesta obligatoria.

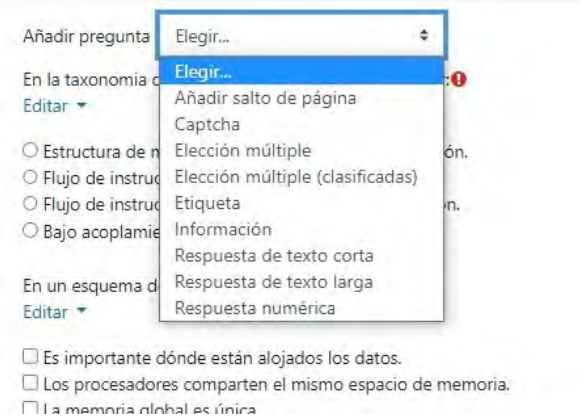


Fig. 3: Tipos de preguntas que pueden crearse (encuestas de retroalimentación de Moodle)

Una vez que la encuesta es puesta a disposición para que los alumnos la completen, es posible visualizar los resultados de diferentes maneras. Tanto pueden analizarse las respuestas individuales, como pueden analizarse porcentual y gráficamente, lo cual ofrece información instantánea y una visión general de cuáles son las tendencias mayoritarias, ya sea por respuestas correctas o erradas. A efectos ilustrativos, en la Fig. 4 se muestra un ejemplo de la forma en la que se presentan gráficamente los resultados para las preguntas de la encuesta, en este caso correspondiente a la primera pregunta de la encuesta de la Fig. 1. En la sección 3 se ahondará en el análisis de los resultados y de la información que brindan las encuestas.

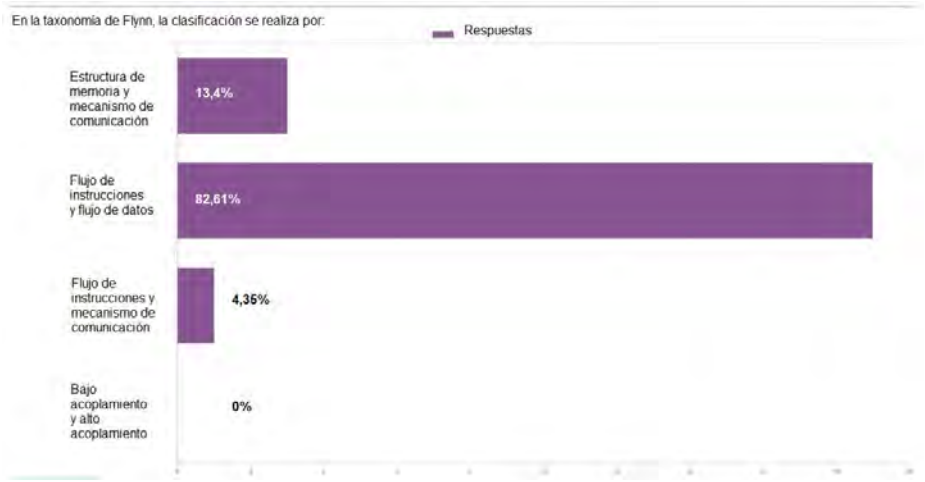


Fig. 4: Ejemplo de visualización de análisis de respuestas. Corresponde a la primera pregunta de la Encuesta de retroalimentación para la Unidad 3 de Computación Paralela 2020 (Moodle)

2.2 Metodología propuesta para la auto/heteroevaluación formativa

Teniendo en cuenta la necesidad de contar con herramientas para mediar el diagnóstico y seguimiento formativo y continuo de los estudiantes en su conjunto, y considerando las características ofrecidas por las encuestas de retroalimentación de Moodle presentadas en la sección 2.1 como una posibilidad para su implementación, hemos incorporado su utilización sucesiva a lo largo del cursado. La experiencia que hemos implementado en ambas asignaturas propone que cada encuesta comprenda entre 5 y 8 preguntas clave a fin de tener un seguimiento de los alumnos y detectar los temas que requieren un tratamiento adicional, antes de su aplicación y/o puesta en práctica. Constituye una novedad introducida con la modalidad virtual la cual fue implementada durante los ciclos 2020 y 2021, ante el cambio producido por la pandemia de COVID-19 y el paso de un esquema de educación presencial a uno completamente virtual/remoto. A diferencia del intercambio verbal que previo al aislamiento por la pandemia se desarrollaba directamente en el aula, las encuestas de Moodle constituyeron el instrumento que nos permitió realizar un seguimiento del

nivel de comprensión alcanzado en la modalidad de aprendizaje asíncrono, para cada una de las unidades temáticas, aun cuando los alumnos se encontraran inhibidos por la distancia y las pantallas para asumir una participación activa y exponer sus inquietudes. Dado que los análisis de los resultados de las encuestas permitieron descubrir a los docentes una manera de cuantificar estadísticamente el nivel de dominio colectivo de los temas (y no solo el del grupo reducido de alumnos que suele participar en las clases), aun en el anonimato de las mismas, es que se propone continuar con su utilización ante la vuelta a la presencialidad, por lo que en el actual ciclo 2022 se continúa utilizando este recurso para el esquema presencial.

En las siguientes viñetas se resume la metodología propuesta para la utilización de las encuestas de retroalimentación para el seguimiento y diagnóstico continuo:

- Al finalizar cada unidad temática, presentar someramente la siguiente unidad, y brindar pautas y el material de estudio necesario para el desarrollo de la misma. Para el caso del inicio de la asignatura, las pautas se brindan en la primera clase y/o a través del aula virtual.
- Los docentes diseñan la encuesta correspondiente a cada unidad temática, incluyendo algunas preguntas clave de fácil respuesta, que permitan sondear, a grandes rasgos, el nivel de dedicación, comprensión y dominio alcanzado por los estudiantes en relación a dicha unidad.
- Cada alumno, por su parte, es responsable y protagonista de administrar en forma asíncrona los recursos brindados para regular el ritmo y nivel de aprendizaje de los distintos tópicos, antes del siguiente encuentro (bibliografía, apuntes, videos, etc.).
- En el siguiente encuentro, se comenzará con la encuesta de seguimiento, con una duración total de aproximadamente 5 minutos para su resolución (entre 5 y 8 preguntas, diseñadas para sondear los saberes primordiales de la temática en cuestión).
- Una vez finalizada la encuesta, se llevará a cabo el análisis de los resultados, conjuntamente por el docente y los estudiantes, valorando la información condensada de forma gráfica, como muestra el ejemplo de la Fig. 4, a fin de destacar las fortalezas y revisar los motivos por los que pueda haber falencias en las respuestas. A su vez, permite a los estudiantes identificar cuáles son los conceptos clave que necesariamente debe asimilar a lo largo del proceso de aprendizaje.
- De acuerdo a las falencias identificadas, o a otras inquietudes que puedan surgir durante el análisis conjunto, los docentes profundizarán y/o reformularán la explicación o tratamiento del tema o concepto que corresponda, previo a su puesta en práctica.

3 Resultados

Los resultados obtenidos a partir del uso de las encuestas para el seguimiento de los estudiantes están ligados a las características de cada uno de los grupos de aplicación. En este sentido, el efecto que el seguimiento tuvo sobre los estudiantes de Computación Paralela (electiva de 3º, 4º, y 5º nivel) fue sumamente positivo, pues

para los estudiantes y cada grupo de trabajo las encuestas funcionaron como indicadores de la necesidad de revisar sus saberes, solicitar asistencia, plantear sus dudas a los docentes, y asistir a clases de consulta. Se trata de un grupo de estudiantes más avanzados en la Carrera, con hábitos de estudio más afianzados, y a la vez con un interés más genuino al cursar la asignatura, por ser de naturaleza electiva. Así, el grupo en su mayoría obtuvo resultados muy buenos, llegando a las instancias de evaluación sumativa con plena conciencia de las fortalezas y falencias que cada uno presentaba. Por su parte, para los alumnos de Matemática Discreta, la utilidad de las encuestas estuvo supeditada al grado de compromiso y dedicación de cada estudiante con su formación. Para una buena proporción del alumnado, las encuestas funcionaron también como indicadores, y si bien la cátedra ofrece más de 8 horas reloj de consulta semanales, lamentablemente los alumnos de primer año no están habituados a recurrir a ellas, y las falencias en muchos casos persisten a lo largo del tiempo, e impactan en las evaluaciones sumativas. No obstante, desde el punto de vista del docente, las encuestas siguen brindando información muy valiosa, e indicadores de tópicos a reformular o reforzar, y así potenciar las capacidades y posibilidades de la porción de estudiantes que tienen mayor iniciativa para aprender y superarse, como así también señalar los aspectos a profundizar y focalizar para los estudiantes con un desempeño más pasivo.

A modo ilustrativo, en la Fig. 5 se muestran los resultados obtenidos en la asignatura Matemática Discreta, en respuesta a la afirmación ***21 es congruente módulo 12 de 9.*** De 27 respuestas, 8 estudiantes respondieron Falso, 9 respondieron que no saben o no recuerdan cómo determinalo, y solo 10 de ellos dieron la respuesta correcta. Este tipo de situaciones enciende rápidamente una luz de alarma para los

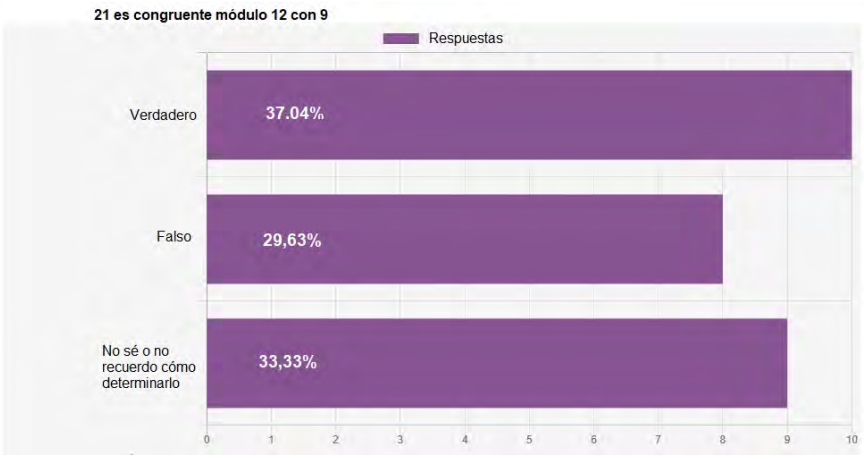


Fig. 5: Respuestas obtenidas para una pregunta particular (27 respuestas en total). Encuesta de retroalimentación para la Unidad 2 de Matemática Discreta 2022 (Moodle).

docentes, que pueden detectar fácilmente cuáles conceptos básicos no han sido estudiados con la profundidad suficiente, o no han sido bien asimilados. Ello brinda la oportunidad de retomar el tema y evacuar las dudas antes de pasar a su aplicación práctica.

En un intento de cuantificar el efecto o la influencia que pudieron tener como resultado las encuestas de seguimiento en la asignatura Matemática Discreta, conjuntamente con otros recursos y modalidades incorporadas para el proceso de enseñanza aprendizaje como el uso de cuestionarios de autoevaluación, y la disponibilidad de videos de la asignatura [2], podemos tomar en consideración el porcentaje de estudiantes que lograron regularizar la asignatura entre los ciclos 2019 y 2022. En la tabla 1 se condensa la descripción del contexto educativo para cada uno de los antes mencionados ciclos lectivos, y el porcentaje de estudiantes que regularizó la asignatura.

Tabla 1: Características y recursos de cada año lectivo, y porcentaje de estudiantes que alcanzaron la regularización de la asignatura Matemática Discreta.

	Modalidad Presencial	Modalidad Virtual	Encuestas seguimiento	Cuestionarios autoevaluación	Videos	%Regula-rización
2019	X			X		**59,32%**
2020		X		X	X	**67,21%**
2021		X	X	X	X	**71,88%**
2022	X		X	X	X	**76,19%**

La comparación es ciertamente sesgada, pues las variables en juego son muy diversas: la modalidad, el contexto y los recursos utilizados en cada año lectivo no fueron los mismos. Sin embargo, rescatamos el aporte que hayan podido tener las encuestas conjuntamente con los demás recursos incorporados, en el incremento del porcentaje de estudiantes que logró la regularidad. Incluso cabe también mencionar la opinión de los estudiantes en ambas asignaturas acerca de la utilidad de las mismas.

4 Conclusiones

El proceso de enseñanza-aprendizaje requiere la continua reflexión y evaluación a fin de que constituya una experiencia transformadora y motivante para los estudiantes, los futuros profesionales. A su vez, las nuevas tendencias nos orientan hacia un mayor protagonismo de los estudiantes en la regulación y asimilación de sus aprendizajes. Este cambio de paradigma que propone centrar el aprendizaje en el estudiante con el docente como guía, requiere además de elementos que permitan obtener información para el seguimiento de los aprendizajes en la etapa formativa, y así orientar o reorientar los aprendizajes de forma más efectiva previo a las evaluaciones sumativas. En este artículo hemos presentado una experiencia en la utilización de encuestas de retroalimentación de Moodle para implementar el seguimiento y diagnóstico continuo de los estudiantes a lo largo del cursado. Se propone este tipo de recurso, dado que Moodle es la plataforma utilizada por la institución, y que las encuestas constituyen el recurso que reúne las características requeridas por los docentes para la actividad. Indudablemente, este tipo de actividades de seguimiento podrían implementarse con otros recursos o tecnologías, incluso a mano alzada o en dinámicas de ronda. Indistintamente de la forma de implementación, su utilización resultó de gran utilidad tanto para los docentes como para los estudiantes a la hora de identificar fortalezas, debilidades, y focalizar en los conceptos que requieren mayor afianzamiento.

Asimismo, resulta una forma enriquecedora de evaluación, ya que le permite al estudiante reflexionar sobre sus aprendizajes y la implicancia que tienen los diferentes temas en su formación codo a codo con sus pares y los docentes, que además de oficiar de guías u orientadores para identificar debilidades para reforzarlas, constituyen una guía para la reflexión y la consideración de diferentes puntos de vista para razonar sobre los conceptos y temáticas en estudio.

Agradecimientos. Se agradece a la UTN y al Proyecto con código TEUTIME0007658TC, financiado por la UTN, por su aporte para la publicación de este trabajo.

References

1. Caymes Scutari, P., Bianchini, G.: Los alumnos detrás de la pantalla: de la presencialidad a la experiencia educativa virtual y remota en el estudio del paradigma paralelo. Actas Congreso Argentino y Latinoamericano de Ingeniería 2021: CADI CLADI CAEDI 2021. Fernández Luco, Luis, ed. Buenos Aires. ISBN 978-987-88-1872-6. p. 200. (2021).
2. Caymes Scutari, P.: Recursos didácticos en la docencia remota: experiencias en Matemática Discreta. V Congreso Internacional de Ciencias de la Computación y Sistemas de Información (CICCSI 2021). Mendoza, Argentina. https://sites.google.com/view/ciccsi-2021-posters. (2021).
3. CONFEDI: Marco conceptual y definición de estándares de acreditación de las carreras de ingeniería. Oro Verde. CONFEDI. https://confedi.org.ar/wp-content/uploads/2021/07/MARCOC1.pdf. (2017).
4. De Miguel Díaz, M.: Modalidades de enseñanza centradas en el desarrollo de competencias: orientaciones para promover el cambio metodológico en el espacio europeo de educación superior. Oviedo: Ediciones de la Universidad de Oviedo. (2006).
5. Edu Trends: Aprendizaje Invertido. Observatorio de Innovación Educativa del Tecnológico de Monterrey. (2014). https://observatorio.tec.mx/edutrendsaprendizajeinvertido (Accedido en 2020).
6. Giordano Lerena, R., González Araujo, L., Larrondo Petrie, M., Páez Pino, A.: Reflexiones de Académicos Latinoamericanos en Pandemia. GEDC-ACOFI-CONFEDI-LACCEI. Bogotá, Colombia. LACCEI Ediciones. (2020).
7. Kowalski, V., Erck, I., Enriguez, H., et al: ¿Cómo vamos a evaluar y cómo vamos a planificar las asignaturas?. Laboratorio MECEK. Universidad Nacional de Misiones. (2020).
8. Moodle. https://moodle.org/ (Accedido en abril 2021)
9. Moodle – Encuestas de retroalimentación. (Accedido en abril 2021). https://docs.moodle.org/all/es/29/Actividad_de_retroalimentaci%C3%B3n (2021).
10. Pimienta Prieto, J.: Evaluación de los aprendizajes: Un enfoque basado en competencias. México: Pearson Educación. (2008).
11. Pimienta Prieto, J.: Las competencias en la docencia universitaria: preguntas frecuentes. México: Pearson Educación. (2012).
12. Tobón Tobón, S., Pimienta Prieto, J., García Fraile, J.: Secuencias Didácticas: Aprendizaje y Evaluación de Competencias. México: Pearson Educación. (2010).
13. Torrano, F., Fuentes, J. L., y Soria, M.: Aprendizaje autorregulado: estado de la cuestión y retos psicopedagógicos. Perfiles educativos, 39(156), 160-173. (2017).

Una Máquina de Turing en la Escuela

Jorge Rodríguez[1], Gerardo Parra[1], Gabriela Gili[1], Susana Parra[1], and Daniel Dolz[1] Hernán Roumec[2]
j.rodrig@fi.uncoma.edu.ar, gparra@fi.uncoma.edu.ar, gabriela.gili@est.fi.uncoma.edu.ar, susana.parra@fi.uncoma.edu.ar, ddolz@fi.uncoma.edu.ar, hroumec@yahoo.com

[1] Grupo de Investigación en Lenguajes e Inteligencia Artificial
Departamento de Teoría de la Computación - Facultad de Informática
Universidad Nacional del Comahue
Buenos Aires 1400, Neuquén, Argentina
[2] Consejo Provincial de Educación
Ministerio de Gobierno y Educación de la Provincia de Neuquén
Belgrano 1300, Neuquén, Argentina

Abstract. Actualmente las Ciencias de la Computación en los niveles obligatorios del sistema educativo están adquiriendo un papel significativamente importante. Por lo tanto, es prioritario trabajar en la producción de recursos educativos en el área. Este artículo presenta un recurso educativo desenchufado, construido en el marco del diseño participativo, destinado a favorecer la enseñanza de conceptos relacionados a las Máquinas de Turing.

Keywords: Educación en Ciencias de la Computación, Escuela Secundaria, Teorías de la Computación, Máquinas de Turing.

1 Introducción

La inclusión de conceptos relacionados con las Ciencias de la Computación en los niveles obligatorios del sistema educativo está adquiriendo un papel significativamente importante durante los últimos años, actualmente se realizan reformas curriculares en esta dirección en numerosos países. Las iniciativas emergentes en los últimos años tienen como objetivo que la población estudiantil, de todos los niveles educativos, tenga acceso a los conceptos centrales de la disciplina [10,11].

En la República Argentina la situación es dispar, mientras las iniciativas desarrolladas por un conjunto de Universidades Nacionales, el Consejo Federal de Educación, Program.ar y los gobiernos provinciales [2,3,4] impulsan reformas curriculares en algunas jurisdicciones, en la mayor parte del país la computación aún está poco representada en la educación obligatoria [12].

Las tendencias curriculares actuales para la educación secundaria promueven desarrollar un recorrido amplio por las áreas de conocimiento. En este marco se consideran conocimientos relacionados al área de Teoría de la Computación [16].

Si bien en la República Argentina se observan iniciativas con cierta preponderancia del área de Algoritmos y Programación, es claro que progresivamente se tiende a un recorrido más amplio [13].

El primer contacto entre los estudiantes y las Ciencias de la Computación suele ser desafiante. En este sentido, los recursos educativos desenchufados demuestran ser una opción adecuada y muy interesante. Esto se debe, fundamentalmente, a que no se requiere aprender programación ni hacer uso de un dispositivo digital y que, por lo general, el ambiente en el que se desarrollan tiene un enfoque lúdico que plantea desafíos para el estudiante [1,15].

Fueron diseñados como una forma de comunicar conceptos computacionales en espacios de la educación no formal, sobre todo como apoyo a la divulgación científica. No obstante y en forma creciente, las escuelas adoptan recursos educativos desenchufados como manera de ofrecer los primeros contactos con conceptos abstractos sobre computación. Sin embargo, existe escasa investigación sistemática sobre su efectividad en el ámbito escolar o acerca de la forma que deben adoptar para adecuarse a estos contextos institucionalizados [1].

Por otra parte, si bien existe un abanico amplio de tópicos disciplinares cubiertos por este tipo de recurso, existen otros que integran las propuestas curriculares y aún no están cubiertos por estas colecciones de recursos. Esta es la situación de un conjunto de conceptos fundamentales sobre Teoría de la Computación, como los relacionados a las Máquinas de Turing[9].

Este trabajo se enmarca dentro de la línea de investigación y desarrollo destinada a la producción de recursos didácticos para enseñar Ciencias de la Computación y evaluar su efectividad en el ámbito de la educación secundaria. En particular, plantea trabajar sobre el desarrollo y evaluación de una colección de recursos educativos desenchufados orientados a facilitar la enseñanza de conceptos relacionados a las Máquinas de Turing y a nociones introductorias de computabilidad[7,9].

El resto de este documento está organizado de la siguiente manera. La siguiente sección describe el modelo propuesto para facilitar la producción de recursos didácticos para enseñar Ciencias de la Computación, la sección 3 está dedicada a presentar los resultados del estudio. Finalmente, se cierra con las conclusiones elaboradas por el equipo de investigación.

2 Modelo Propuesto

El enfoque presentado en este trabajo se sustenta en cuatro perspectivas didácticas: CSUnplugged, Aprendizaje Experiencial, Diseño Participativo e Investigación Acción Participativa. La convergencia de estas perspectivas apunta a la elaboración de nuevos recursos didácticos disciplinares validados en las aulas [1,5,8].

En este contexto se trabaja en el marco de los enfoques metodológicos basados en la investigación y el diseño participativos definidos específicamente por esta línea de investigación y desarrollo [6]. Están basadon en el modelo Participatory Design Framing, un marco de trabajo de trabajo innovador para educación en

computación, donde los docentes de las escuelas se involucran activamente en el proceso de elaboración de recursos educativos [14].

El ciclo en que se organiza este modelo define un marco metodológico estructurado en cuatro etapas.

- **Producción de conjeturas**. El proceso comienza con la definición de algunas conjeturas acerca de cómo apoyar el proceso de enseñanza y de aprendizaje. En esta etapa se adoptan algunas opciones metodológicas para el diseño de los dispositivos didácticos.
  - Promover la colaboración. La utilización de recursos físicos actúa como facilitador de la actividad grupal.
  - Distribuir la complejidad. Cada persona asume la responsabilidad de una pieza de la máquina. La complejidad conceptual se distribuye en el grupo.
  - Aprender jugando. Se traslada la mecánica de juego al ámbito del aprendizaje para lograr mejores resultados en términos disciplinares y de desarrollo de habilidades blandas.
  - Aprender de la experiencia. Centrado en producir conocimiento abstracto y conceptual a partir de reflexionar sobre experiencias concretas.

  Por otra parte se define una mecánica de base para el juego basada en iteraciones sucesivas y se selecciona una colección de Máquinas de Turing que resulten accesibles para estudiantes sin formación previa.
- **Diseño específico**. El objetivo es instanciar la propuesta a la situación concreta de enseñanza y de aprendizaje para producir los ajustes necesarios y avanzar en el diseño de la experiencia.

  En esta instancia, se convoca a un par de docentes de informática, que desempeñan sus labores en la escuela secundaria en la que se desarrolla el trabajo de campo, con intención recuperar sus percepciones acerca del recurso didáctico. En este contexto se desarrolla una sesión piloto con el propósito de evaluar el recurso, e informar sobre las características que se valoran positivamente y acerca de las que requieren ajustes.

  En estas sesiones preparatorias, se sitúa el trabajo de campo sobre tres años de estudio de la escuela secundaria, se define que en todos los casos se trabajará con estudiantes sin formación previa en el área de conocimiento. Se define una afectación de dos horas cátedra para el desarrollo de la actividad, es decir 80 minutos.

  Se producen tres ajustes a la propuesta inicial, se trabaja sobre la complejidad de las Máquinas de Turing a utilizar sumando una máquina de complejidad ligeramente superior. Se define con mayor precisión algunas mecánicas de juego y se modifica la forma en que se representan las máquinas adoptando un modelo más próximo al que se utiliza en el ámbito de la disciplina.
- **Práctica mediada**. Este momento transcurre en las aulas de las escuelas seleccionadas, se realiza una práctica mediada por el juego diseñado que expresa la colección de conjeturas iniciales. Esta instancia contribuye a validar o sugerir ajustes sobre la propuesta inicial. En la Sección 2.2 se describe este proceso.

– **Recuperación de conocimiento**. Finalizada la práctica mediada se recuperan resultados de la experiencia sucedida. Estos se utilizan para producir los ajustes necesarios para el mejoramiento del recurso diseñado y así ofrecer a la comunidad docente nuevos recursos desenchufados para enseñar Ciencias de la Computación en sus aulas. En otro sentido se utiliza para confirmar, ajustar o rechazar conjeturas acerca de las posibilidades reales de enseñar conceptos fundamentales sobre Teoría de la Computación en el ámbito de la educación secundaria. En la Sección 3 se describen los resultados obtenidos.

2.1 Una Máquina de Turing en la Escuela

"*Una Máquina de Turing en la Escuela*" forma parte de una colección de recursos educativos desenchufados diseñados para facilitar la enseñanza de conceptos relacionados a las Máquinas de Turing y a nociones introductorias de computabilidad. Esta colección busca utilizar mecánicas de los juegos de mesa para lograr mejores aprendizajes. Las reglas de los juegos que componen la colección están definidas por los mecanismos de funcionamiento de las Máquinas de Turing.

Esta propuesta está destinada a grupos de estudiantes que transitan la educación secundaria, sin saberes disciplinares previos en el área de conocimiento. "*Una Máquina de Turing en la Escuela*" está delineada para trabajar con varios equipos que estén conformados por 3 a 4 integrantes y la idea fundamental es presentarla como un *juego* con determinadas reglas. La duración estimada para la actividad es de 80 minutos.

Preparación. El material necesario para realizar la actividad incluye una cinta de papel dividida en celdas y un triángulo de papel para indicar la posición actual de la cinta donde apuntaría la cabeza lecto-escritora de la máquina. El cabezal se debe poder desplazar sobre la cinta en ambas direcciones. Además, es necesaria una ficha que tendrá las reglas previamente establecidas que determinan el funcionamiento de la máquina. Las reglas indican, dado el estado actual y el símbolo en la posición de la cabeza lectora, a qué estado debe pasar la máquina, qué símbolo se escribe en la posición actual de cinta y qué desplazamiento (a izquierda o a derecha) debe realizarse.

Funcionamiento. Luego de haber organizado los grupos, entregado el material y haber presentado las reglas del juego, la partida comienza dando inicio al funcionamiento de la máquina. El objetivo principal de la actividad es entender el funcionamiento de una máquina que no es creada con tecnología pero que ejecuta instrucciones de la misma manera que una computadora actual.

Cada grupo ejecutará su máquina siguiendo las reglas que les fueron entregadas. Si la cinta que recibe el grupo contiene la entrada *"9453"* y la cabeza lecto-escritora apunta al "9", una regla posible a aplicar sería *(A,9,derecha)*. Esta regla podría interpretarse informalmente de la siguiente manera: *"la máquina pasa al estado A, escribe 9 en la posición actual y se mueve un lugar a la derecha"*. Entonces ahora los valores de la cinta serán *"9453"*, pero en este caso, el estado actual será A y la cabeza lecto-escritora estará señalando la posición donde está el símbolo 4. En este punto, la regla que debemos aplicar está en la tabla correspondiente al estado A.

Se proyectan tres iteraciones, en cada oportunidad se utiliza una entrada diferente. Cada grupo que acierta con la salida producida por la máquina para la entrada en juego, acumula 2 puntos. La próxima iteración acumula 3 puntos y consiste en descubrir qué realiza la máquina que se está usando, por ejemplo "*Determina si un número es par o impar*" o "*Suma uno al número*". La iteración final suma 4 puntos y consiste en explicar cómo la máquina logra resolver el problema planteado.

Reflexión y formalización. Luego de realizar la actividad se realiza un análisis sobre la experiencia en el aula. Se busca construir conceptos abstractos a partir de entender el funcionamiento de una máquina formal. Es importante mostrar cómo la máquina de Turing, aún cuando pueda resultar a primera vista un dispositivo formal casi rudimentario, tuvo un impacto determinante y fundamental en la aparición de la primera computadora y en el funcionamiento de las computadoras actuales. Además, permite formalizar la noción intuitiva de procedimiento computacional, procedimiento efectivo o algoritmo.

2.2 Práctica mediada: Trabajo de campo con estudiantes secundarios

Este recurso educativo fue llevado a las aulas con la intención de ajustarse progresivamente a partir de la consideración de comentarios, sugerencias y revisiones realizadas sobre el trabajo de campo. La población que participó de la experiencia se compone de 26 adolescentes, que cursan diferentes años de estudio de la educación secundaria, 12 estudiantes asisten al primer y segundo año de estudios y 14 adolescentes al último año. En todos los casos, se trata de estudiantes sin formación previa en el área de conocimiento. El rango de edad en general, es de 13 a 15 años en el primer grupo y de 16 a 18 años en el segundo.

La experiencia se realizó en tres encuentros diferentes de 80 minutos cada clase, con dos docentes, concretando actividades desenchufadas destinadas a la enseñanza y el aprendizaje de conceptos abstractos, que facilitara la comprensión del funcionamiento de la Máquina de Turing.

La actividad buscó que los estudiantes comprendieran aspectos particulares del formalismo. Se consideraron los procesos reflexivos sobre la experiencia, la construcción de conceptos abstractos a partir de vincular las máquinas utilizadas durante la actividad, con los conceptos formales sobre Máquina de Turing.

La experiencia se organizó en cuatro etapas: explicación y demostración del uso del material, experiencia concreta, elaboración de conceptos abstractos y aplicación.

Etapa 0: Explicación y demostración del uso del material. En una primera instancia se realiza una breve explicación sobre la actividad y demostración de uso del material (cintas de entrada, cabeza lectora-escritora, fichas con reglas) para la experiencia.

Etapa 1: Experiencia concreta. En esta etapa se desarrolla la fase experiencial de "*Una Máquina de Turing en la Escuela*". Se trabajó con dos máquinas: la primera para descubrir si la entrada de valores representaba un número par o impar y la segunda máquina para sumar 1 a un número de entrada. Se dividió

el aula en grupos, a cada grupo se les entregaron varias cintas, una cabeza lecto-escritora y las fichas con las reglas escritas.

En cada grupo, los estudiantes se distribuyeron los elementos para seguir la actividad y cumplir un rol (Estado A- Estado S- Cabezal). Quien tenía el Cabezal, realizaba la lectura del símbolo considerado el estado en el que se encontraba y según el estado que correspondiera, el compañero de grupo recuperaba la regla, indicaba la acción a seguir y se establecía el nuevo estado. Al llegar al estado H, finalizada la secuencia, se analizó en forma grupal el resultado.

Se inició con la máquina de par-impar. Luego de probar con dos entradas, los cursos de 1º y 2º año lograron seguir la secuencia dentro de la interacción grupal, obteniendo correctamente la salida, pero fue necesaria la tercera partida y algunas guías por parte de los docentes para que lograran comprender lo que realizaban. En la experiencia con la segunda, lograron descubrir fácilmente en la primera partida y con pocas entradas que el objetivo era sumarle 1 al número representado en la entrada.

Para el grupo del último año del secundario, rápidamente en la primera partida lograron descubrir qué hacía cada una de las máquinas, pero, de todas maneras, continuaron con varias pruebas usando entradas de diferentes longitudes y símbolos, dado que los estudiantes manifestaron resultar entretenida y motivadora la actividad.

Etapa 2: Elaboración de conceptos abstractos Se realiza un análisis retrospectivo sobre la experiencia vivida recuperando algunos aspectos particulares tales como que una máquina tiene diferentes estados y alfabetos y que a cada estado corresponde una lista de acciones de acuerdo con el símbolo de entrada. Se construyen conceptos abstractos a partir de vincular las máquinas utilizadas durante la actividad con los conceptos formales sobre Máquina de Turing.

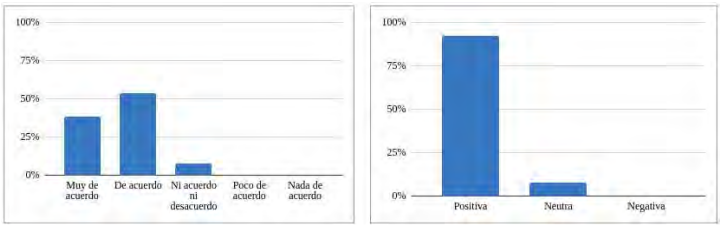
Etapa 3: Aplicación Finalmente, en plenario se analizó de manera conjunta diferentes situaciones donde estos saberes se ponen en juego. Se expuso acerca de la importancia de estos conceptos en el ámbito de las Ciencias de la Computación para ayudar a reconocer el poder computacional de las Máquinas de Turing. La actividad resultó muy motivadora, entretenida y sencilla de seguir. Los estudiantes de los últimos años del secundario lograron apropiarse los conocimientos disciplinares explorados durante la actividad sin dificultad. En los primeros años lograron realizar las secuencias, obteniendo así los resultados esperados, aunque atravesaron algunas dificultades para la comprensión de los conceptos abstractos desarrollados.

3 Resultados

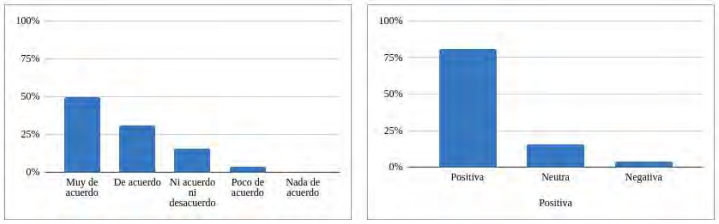
Completada la experiencia se consultó a la población estudiantil participante de la actividad acerca de su percepción en relación a la utilización de este tipo de dispositivos y sobre la efectividad de este recurso didáctico en términos de aprendizaje construidos. Se utilizó una encuesta que no tuvo de carácter obligatoria, tampoco es un componente del proceso de acreditación de la materia y estuvo dirigida al total de la población participante de la experiencia.

La encuesta se estructuró en dos secciones, la primera de estas destinada a recuperar percepciones en relación a a) la utilidad, es decir, en qué grado los estudiantes aprecian que la actividad desarrollada resulta útil para ayudar a comprender conceptos sobre Máquinas de Turing; b) impacto, si ellos consideran que se trata de una experiencia placentera y c) organización, si consideran que desarrollar una experiencia lúdica antes de la exposición formal es una forma adecuada de organizar actividades de aprendizaje.

Esta sección se compone de tres preguntas con las siguientes posibles respuestas: muy de acuerdo, de acuerdo, ni acuerdo ni desacuerdo, poco de acuerdo y nada de acuerdo. A continuación, se presentan los resultados obtenidos en esta sección:



(a) Categoría Utilidad.



(b) Categoría Impacto.

(c) Categoría Organización.

Fig. 1: Sección percepciones.

Categoría: Utilidad

Pregunta: El juego “Una MT en la escuela” me parece una herramienta útil para aprender conceptos sobre Máquina de Turing.

Se observa que el 38% la población consultada manifiesta estar muy de acuerdo con que la herramienta es útil para aprender conceptos sobre Máquinas de Turing, mientras el 54% expresa estar de acuerdo con la misma afirmación y el 8% responde que no está de acuerdo ni en desacuerdo. Es decir que que el 92% comunica una apreciación positiva en relación a la utilidad del dispositivo didáctico. Fig. 1a.

Categoría: Impacto

Pregunta: El juego me parece entretenido y sencillo de utilizar.

En esta categoría, el 50% dice estar de muy acuerdo y 31% estar de acuerdo con la afirmación acerca de que la herramienta es amigable y sencilla de utilizar, el 15% prefiere una respuesta neutra y el 4% responde que está poco de acuerdo. En este caso las respuestas afirmativas acumulan el 81% del total de respuestas. Fig. 1b.

Categoría: Organización

Pregunta: Jugar primero y que después nos cuenten sobre Máquina de Turing me parece una forma adecuada de enseñar.

En relación a la organización, el 54% manifiesta estar de acuerdo con que la organización de la actividad resulta adecuada para aprender, el 8% de los estudiantes se inclinan por la opción muy de acuerdo ante la misma afirmación. Por otra parte el 19% dice estar una poco en descuerdo y el 4% muy en desacuerdo, el restante 15% ofrece una respuesta neutra. Considerando porcentajes acumulados, se observa que el 62% tiene una apreciación positiva hacia la organización de la secuencia de aprendizaje, mientras el 23% no la considera apropiada. Fig. 1c.

La segunda sección, busca indagar acerca de los conocimientos construidos por la población estudiantil a partir de su participación en la experiencia. En esta dirección se plantea estudiar la efectividad para con la enseñanza y el aprendizaje de prácticas y conceptos relacionados a las Maquinas de Turing, en el contexto de la escuelas secundaria, a estudiantes sin formación previa en el área de conocimiento.

Esta sección se organiza en tres categorías, a) Comprender, tiene que ver con comprender e interpretar conceptos desarrollados, como también asignar significado a diferentes elementos, b) Aplicar, está relacionada a la utilización de practicas y conceptos aprendidos a la resolución de casos particulares y concretos y c) Evaluar, está vinculada a establecer en que medida resulta pertinente aplicar prácticas y conceptos aprendidos a situaciones particulares.

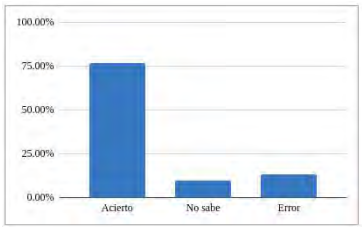
Esta sección se compone de cinco preguntas, cada una ofrece las siguiente respuestas posibles: Muy seguro, Creo que si, No sé, Creo que no, No es así. A continuación se presentan los resultados obtenidos en esta sección:

Categoría: Comprender

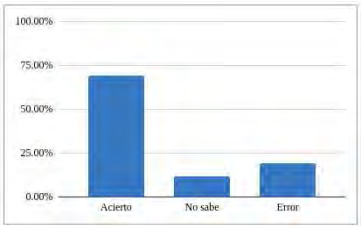
Pregunta 01: La Máquina de Turing de la figura tiene dos estados: A y S.

Pregunta 02: La Máquina de Turing de la figura puede trabajar con letras como “A”, “W” o “T”.

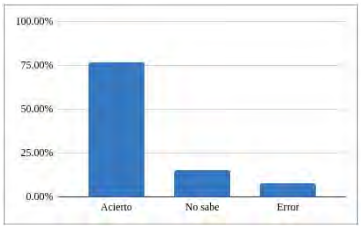
Se observa que el 76% seleccionó una opción acertada como respuesta, mientras que el 13.46% optó por una una opción incorrecta y el 9.62% responde que



(a) Categoría Comprender.



(b) Categoría Aplicar.



(c) Categoría Evaluar.

Fig. 2: Sección conocimientos construidos.

no conoce la respuesta. Los resultados agrupan las repuestas a las dos preguntas de la categoría y acumulan como acierto las opciones "Muy seguro" y "Creo que si", como error a las opciones "Creo que no" y "No es así" y como respuesta neutra "No sé".

Categoría: Aplicar

Pregunta 01: En la posición del cabezal en la figura, si el estado actual es S, se pasa al estado A y se mueve a la derecha.

Pregunta 02: En la posición del cabezal en la figura, si el estado actual es A, se pasa al estado S y se mueve a la derecha.

En la categoría Aplicar, el 69.23% aporta una respuesta correcta, 11.54% manifiesta no conocer la respuesta y el 19.23% responde en forma equivocada. En este caso, se acumula como acierto a las respuestas "Muy seguro" y "Creo que si" para la Pregunta 01 y "Creo que no" y "No es así" para la Pregunta 02.

Categoría: Evaluar

Pregunta 01: Podría utilizar una Máquina de Turing, por ejemplo para saber cuántas cifras tiene un número.

En la categoría Evaluar, el 76.92% aporta una respuesta correcta, 15.38% manifiesta no conocer la respuesta y el 7.69% responde en forma equivocada. En este caso, se acumula como acierto a las respuestas "Muy seguro" y "Creo que si" para la Pregunta 01 y "Creo que no" y "No es así" para la Pregunta 02.

4 Conclusiones y Trabajo Futuro

Este artículo presenta un recurso didáctico que facilita la enseñanza de conceptos fundamentales sobre Teoría de la Computación a estudiantes de las escuelas secundarias. La práctica mediada realizada en el contexto de este estudio aporta fuertes indicios alentadores acerca de que este tipo de recursos educativos dispone de amplias posibilidades de favorecer la efectiva construcción de conocimiento disciplinar de carácter abstracto. Por otra parte, el modelo para el diseño participativo definido en el ámbito de esta línea de investigación y desarrollo ofrece un marco metodológico válido para producir recursos educativos consistentes y situados.

References

1. T. Bell and J. Vahrenhold. Cs unplugged—how is it used, and does it work? In *Adventures Between Lower Bounds and Higher Altitudes*.
2. Consejo Federal de Educación. Res 263/15. *Resoluciones CFE*, 2015.
3. Consejo Federal de Educación. Res 343/18. *Resoluciones CFE*, 2018.
4. Consejo Provincial de Educación. Res 1463/18. *Resoluciones CPE*, 2018.
5. B. DiSalvo, J. Yip, E. Bonsignore, and D. Carl. Participatory design for learning. In *Participatory design for learning*, pages 3–6. Routledge, 2017.
6. D. Dolz, R. Martínez, G. Parra, J. Rodríguez, and N. Ginez. Recursos educativos desenchufados para la enseñanza de las ciencias de la computación en la escuela secundaria. In *XV TE&ET*, 2020.
7. J. Hopcroft, R. Motwani, and J. Ullman. *Introduction to Automata Theory, Languages and Computation*. Addison Wesley, 2006.
8. A. Y. Kolb and D. A. Kolb. Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of management learning & education*, 4(2):193–212, 2005.
9. H. Lewis and C. Papadimitriou. *Elements of the Theory of Computation*. Second Edition. Prentice Hall, 1998.
10. O. McGarr and K. Johnston. Curricular responses to computer science provision in schools: current provision and alternative possibilities. *The Curriculum Journal*, 31(4):745–756, 2020.
11. G. Ottestad and G. B. Gudmundsdottir. Information and communication technology policy in primary and secondary education in europe. *Second handbook of information technology in primary and secondary education*, pages 1–21, 2018.
12. J. Rodríguez and M. Cortez. La posición de las ciencias de la computación en el diseno curricular para la escuela secundaria argentina: Una revisión sistemática. *Electronic Journal of SADIO (EJS)*, 19(2).
13. J. Rodríguez, M. Cortez, and S. Boari. Explorando el lugar de las áreas de conocimiento de las ciencias de la computacion en la escuela secundaria argentina: Una revisión sistemática. *Electronic Journal of SADIO (EJS)*, 21(2), jul. 2022.
14. W. Sandoval. Conjecture mapping: An approach to systematic educational design research. *Journal of the learning sciences*, 23(1):18–36, 2014.
15. R. Taub, M. Armoni, and M. Ben-Ari. Cs unplugged and middle-school students' views, attitudes, and intentions regarding cs. *ACM Transactions on Computing Education (TOCE)*, 12(2):1–29, 2012.
16. M. Tissenbaum and A. Ottenbreit-Leftwich. A vision of k— 12 computer science education for 2030. *Communications of the ACM*, 63(5):42–44, 2020.

Diseño Participativo de Secuencias Didácticas basadas en el Desarrollo de Aplicaciones Móviles en la Escuela

Jorge Rodríguez, Pablo Kogan, Guillermo Guerrero, Guillermo Pereyra, and Fabio Torrico

Grupo de Investigación en Lenguajes e Inteligencia Artificial
Departamento de Teoría de la Computación - Facultad de Informática
Universidad Nacional del Comahue
Buenos Aires 1400, Neuquén, Argentina
{j.rodrig,pablo.kogan,guillermo.guerrero,guillermo.pereyra,fabio.torrico}@fi.uncoma.edu.ar

Resumen Las Ciencias de la Computación es uno de los contenidos de la escolaridad obligatoria que está mediando e influenciando, de manera más intensa, la vida de todas las personas. El aprendizaje de conceptos sobre esta disciplina, mejora las posibilidades de comprender e intervenir el mundo y estar preparados para los empleos del futuro. Elaborar propuestas didácticas con una participación temprana de la docencia en el proceso de diseño, mejora las posibilidades de que la propuesta de aprendizaje este situada y se logre construir una transposición significativa para la población estudiantil.
En este trabajo se describe la experiencia de aplicación del Modelo Zewmayayiñ *haremos juntos* de Diseño Participativo de Secuencias Didácticas, utilizado en un Taller de Desarrollo de Aplicaciones Móviles en Escuelas Secundarias de la provincia del Neuquén.

Keywords: Diseño Participativo, Aplicaciones Móviles, Enseñanza de la Computación.

1 Introducción

El campo de la computación impulsa la innovación en las ciencias, la industria, la economía, el arte, el entretenimiento y los gobiernos. Una parte importante de la vida cotidiana de las personas y las sociedades está mediada por artefactos computacionales, en este contexto se pronostica que la influencia de la informática resulte cada vez más intensa [1,6,10].

En los escenarios actuales, fuertemente influidos por tecnologías computacionales, conocer sobre Ciencias de la Computación se ubica como pieza fundamental para el ejercicio de la ciudadanía, en tanto mejora las posibilidades de comprender e intervenir el mundo, de participar de debates de la sociedad y

de estar preparados para los trabajos del futuro. Aprender sobre computación contribuye al empoderamiento de diferentes grupos sociales [6,10,14].

Durante los últimos años las iniciativas que buscan ampliar la participación de las Ciencias de la Computación en el ámbito de la educación obligatoria han ganado una importancia significativa. En este sentido, con intención de atender estas necesidades, muchos países están redefiniendo sus diseños curriculares adoptando diferentes modalidades y enfoques curriculares [13,10,6].

El Diseño Participativo, un paradigma novedoso en la investigación en el contexto de la educación en informática, busca involucrar tempranamente a la docencia en los procesos de diseño de propuestas educativas. Tiene el propósito de elaborar construcciones metodológicas que resulten más consistentes y situadas. El Diseño Participativo puede contribuir a la construcción de las condiciones pedagógicas sólidas para integrar con éxito y en forma sostenida las Ciencias de la Computación en las propuestas de enseñanza destinadas a la Escuela Secundaria[3,8].

Este enfoque permite incorporar las perspectivas de la docencia a la producción de recursos educativos, mejorando notablemente las posibilidades de que resulten relevantes, significativos y factibles para las comunidades educativas. En este sentido, se asume a la docencia como autor colectivo de las propuestas de enseñanza, esto implica co-construir planes de acción, situarlos en espacios singulares, probarlos, refinarlos y reinventarlos. Es decir articular una síntesis de opciones metodológicas, epistemológicas y pedagógicas en el proceso de elaboración de propuestas de enseñanza [4,3,7].

La mayor parte de los estudios describen qué es el Diseño Participativo, sin profundizar sobre la definición de los marcos de acción que orienten los procesos, es decir, acerca de cómo hacerlo. No abundan enfoques metodológicos consolidados en este campo [11].

En este artículo se presentan resultados preliminares obtenidos a partir de analizar el desarrollo de un taller de co-diseño de secuencias didácticas dirigido a docentes de escuelas secundarias. En este contexto, los docentes conforman duplas integradas por un educador del área computación y uno de otra área de conocimiento, que se unen para conformar un equipo docente. El taller ha sido preparado con el objetivo de co-diseñar secuencias didácticas destinadas a enseñar conceptos básicos sobre algoritmos y programación en las escuelas secundarias.

En el marco de este trabajo el Diseño Participativo se expresa en dos dimensiones, A) que la docencia desarrolle una experiencia de co-creación de propuestas de enseñanza para mejorar su propia comprensión acerca de las prácticas educativas relacionadas a la enseñanza de la computación y B) incorporar el enfoque de Diseño Participativo y sus principios pedagógicos en las unidades curriculares que se diseñan.

El resto del artículo está organizado de la siguiente manera. En la Sección 2, se describe el lugar del Diseño Participativo en el ámbito de la educación. La Sección 3, presenta el modelo Zewmayayiñ para Diseño Participativo. En la Sección 4, se describe la experiencia desarrollada con docentes de escuelas secundarias. En la

sección 5, se discuten los resultados de la experiencia. Finalmente, se presentan las conclusiones del artículo.

2 Diseño Participativo en Educación en Ciencias de la Computación

El Diseño Participativo es un enfoque metodológico para la construcción de artefactos que busca incorporar tempranamente puntos de vista de las partes interesadas a los procesos de desarrollo de productos o procesos que afectan sus vidas o sus trabajos [2,3,11].

En el campo del Diseño Participativo es posible identificar las perspectivas argumentales que de conjunto explican la concepción adoptada en este trabajo. Una de las perspectivas, próxima al concepto de Ciencia Ciudadana, está centrada en utilizar métodos tendientes a democratizar los procesos de diseño y construir espacios de trabajo con mayores grados de horizontalidad, empoderamiento y emancipación de los destinatarios. De esta forma de busca robustecer los resultados incorporando la experiencia, el conocimiento y los puntos de vista de las personas destinatarias [5,2,12].

Una de las discusiones abiertas en el ámbito del Diseño Participativo está centrada en situar la participación de los destinatarios en el proceso de desarrollo. En este sentido el desafío está ubicado en construir artefactos disciplinariamente robustos y consistentes que al mismo tiempo estén influidos por los destinatarios. Estas discusiones participan de la definición de nuevas perspectivas, por una parte, se ubican las centradas en el perfeccionamiento de la calidad de los productos, y por otra, las que prestan mayor atención en mejorar el contexto en los que esos artefactos se ponen en juego [9].

En las perspectivas centradas en el producto, se busca recuperar conocimiento y experiencias que contribuyan a mejorar la calidad de los artefactos, que tengan carácter general más que particular. Esta perspectiva es muy utilizada en la industria del vídeo juego, donde los jugadores participan activamente del desarrollo de un juego para mejorar la jugabilidad, no se busca que el juego se ajuste a las singularidades de un jugador en particular.

En el ámbito de la enseñanza de las Ciencias de la Computación, el Diseño Participativo es un enfoque emergente y novedoso que se conecta con tendencias pedagógicas y didácticas clásicas que ubican al docente como principal autor de las construcciones teóricas y metodológicas que orientan su praxis [4,8].

La computación es una disciplina académica de reciente incorporación en la educación obligatoria, la docencia se encuentra ante el desafío de enseñar temas para los que no existe suficiente disponibilidad de recursos educativos y dispositivos didácticos que se ajusten a las singularidades de cada realidad. El problema de producir materiales educativos destinados a favorecer las enseñanzas y los aprendizajes de conceptos fundamentales sobre Ciencias de la Computación es un campo especialmente susceptible a ser abordado desde los enfoques elaborados en el marco del Diseño Participativo [8].

Los escenarios donde esos procesos de enseñanza suceden resultan especialmente singulares y dinámicos si se enfoca la atención sobre las características de la infraestructura tecnológica escolar, las particularidades de la población estudiantil y la lógica disciplinar. En este contexto facilitar el desarrollo de habilidades para que la docencia produzca secuencias didácticas que logren articular consistentemente estas particularidades, con intención de ofrecer propuestas de enseñanza situadas, resulta un apoyo importante para la incorporación de forma rigurosa y sostenida de la computación en la educación obligatoria.

3 Zewmayayiñ, modelo para el Diseño Participativo de Recursos Educativos

Para el armado del Taller se utilizó el modelo "Zewmayayiñ", que significa *haremos juntos* en mapuzungun, para el Diseño Participativo de Recursos Educativos. Se trata de un marco metodológico formulado por los autores de éste artículo y está definido específicamente para favorecer los procesos de producción de recursos educativos en el ámbito de la enseñanza de las Ciencias de la Computación.

Zewmayayiñ está basado en el modelo Participatory Design Framing, un marco de trabajo innovador para educación, donde los docentes de las escuelas se involucran activamente en el proceso de elaboración de recursos educativos [11].

La Figura 1 muestra el ciclo en que se organiza la producción de recursos educativos en el marco metodológico definido para "Zewmayayiñ".

- **Producción de conjeturas.**
  El proceso comienza con la definición de algunas conjeturas acerca de cómo es posible apoyar los procesos de enseñanza y de aprendizaje.
  En el proceso de Diseño Participativo, definido para esta Línea de Investigación y Desarrollo, el equipo de investigación diseña una primer versión a partir de explorar el área de conocimiento, determinar posibles enlaces curriculares para la actividad y el conocimiento disponible acerca de la enseñanza de la computación en el ámbito de la Escuela Secundaria. Son el resultado de un análisis inicial del problema.
  En general se trata de formas, modelos y recursos provisionales sobre cómo enseñar. Se busca adoptar perspectivas didáctico disciplinares novedosas que tengan la posibilidad de ser situadas a la enseñanza de las Ciencias de la Computación, pero no son comunes en este contexto disciplinar específico o no están suficientemente probadas en las aulas.
- **Diseño específico.**
  En segunda instancia, se convoca a un grupo de docentes de informática de escuelas secundarias a una sesión piloto que busca recuperar sus percepciones acerca de las aproximaciones elaboradas en el momento anterior. Los docentes evalúan el recurso y luego informan sobre las características que valoran positivamente y acerca de las que interpretan que requieren ajustes.

El objetivo es instanciar una conjetura a situaciones concretas de enseñanza y de aprendizaje a través de la elaboración de un diseño específico. Se trata de describir cómo se espera que opere dentro de un contexto particular y producir los ajustes necesarios que logren articular perspectivas teóricas con la práctica docente.

– **Práctica mediada.**
Un tercer momento transcurre en el aula, donde se realiza la práctica mediada que pone en juego las conjeturas iniciales. En esta instancia es posible que que las actividades y conjeturas diseñadas se validen, se ajusten o se descarten.
El concepto de práctica mediada refiere a que las experiencias son mediadas por los recursos diseñados en la fase anterior, los que son un refinamiento de las conjeturas. La práctica mediada tiene un doble propósito, por una parte tiene una dimensión práctica ayudando a mejorar la situación contextual del lugar donde ocurre. Por otra una dimensión teórica produciendo información tendiente a refinar conjeturas.
– **Recuperación de conocimiento.** Producidos los ajustes, la docencia dispone del recurso producido para enseñar Ciencias de la Computación en sus aulas. Un nuevo ciclo de adecuación del recurso se produce al revisar los resultados obtenidos en el trabajo de campo.

Figura 1. Ciclo del Diseño Participativo.

4 Experiencia

En el marco del Taller se concretaron cuatro encuentros de Diseño Participativo con duplas compuestas por un docente del área informática y un docente de otra área de conocimiento. El Taller se desarrolló durante cuatro semanas, en cada encuentro los docentes se reunieron con el equipo de la Facultad para co-diseñar secuencias didácticas basadas en el desarrollo de aplicaciones móviles utilizando el entorno AppInventor. En este contexto se ponen en contacto con conceptos sobre Algoritmos y Programación, como también con aspectos metodológicos relacionados a la enseñanza dela computación.

Entre los encuentros, las duplas trabajaron en forma independiente en la formulación de sus propuestas didácticas. Éstas buscan vincular las áreas de conocimiento a través de que sus estudiantes desarrollen una aplicación móvil.

El Taller comenzó con un conjunto de conjeturas teóricas y prácticas articuladas en la propuesta de trabajo que se comparten con los docentes. La conjetura de base consiste en asumir que los estudiantes secundarios, mientras aprenden a programar, tienen la capacidad de desarrollar aplicaciones móviles que les resulten significativas y tengan posibilidad de incidir positivamente en su contexto social. En este sentido, se estima que el desarrollo de aplicaciones está conectado con los intereses de los estudiantes secundarios.

Por otra parte se presupone que la docencia está en condiciones elaborar las secuencias didácticas en duplas y que lo único que necesitan es que se les ofrezcan los escenarios que les permitan hacerlo. Finalmente, se conjetura que el desarrollo aplicaciones es una estrategia didáctica apta para articular diferentes áreas de conocimiento.

4.1 Participantes

Ocho docentes de Informática y ocho docentes de otras áreas de conocimiento de instituciones educativas públicas del nivel secundario y terciario ubicadas en diferentes localidades de la provincia del Neuquén participaron del *Taller para el Diseño Participativo de Secuencias Didácticas basadas en el Desarrollo de Aplicaciones Móviles*. Trabajaron junto a un equipo de extensión universitaria de la Universidad Nacional del Comahue integrado por tres estudiantes avanzados y dos docentes investigadores.

Algunos de los docentes de informática habían participado previamente de iniciativas de formación docente coordinadas por la Facultad de Informática, pero ninguno había sido parte de una experiencia de Diseño Participativo.

4.2 Casos de Estudio - Secuencias co diseñadas

Los trabajos presentados a continuación involucran el diseño de experiencias educativas elaboradas por las duplas y expresan una variedad de propuestas de carácter singular y situado. Para construir estas secuencias didácticas, se articulan las perspectivas didácticas expuestas en el Taller, la singularidad de cada situación educativa y las lógicas disciplinares de los campos de conocimiento que participan de la propuesta.

Taller de App's para el aula - CeRET Esta propuesta ha sido creada por la dupla cuyos docentes pertenecen al Centro Regional de Educación Tecnológica, es una institución dependiente del Consejo Provincial de Educación dedicada a apoyar pedagógicamente a las Escuelas Técnicas de la provincia.

La propuesta consiste en brindar talleres en modalidad virtual para que los estudiantes desarrollen aplicaciones para dispositivos móviles que les faciliten el aprendizaje de contenidos abordados en las aulas de las escuelas técnicas. Además, con esta propuesta se busca que el estudiante asuma el rol de protagonista como desarrollador de aplicaciones que le permitan tratar y solucionar situaciones problemáticas de su entorno cotidiano y profesional de manera multidisciplinaria. Se lo aparta del simple rol de usuario de programas informáticos. En realción a las aplicaciones a desarrollar se ofrece una propuesta flexible y de aproximación gradual que va desde aplicaciones simples con modalidad preguntas y respuestas a las que involucran cálculos específicos.

AgroApp Propuesta enmarcada en área de Agropecuaria en el ciclo básico y en el ciclo superior de la especialidad Agropecuaria Animal y Vegetal de la Escuela Secundaria Agronómica "Nuestra Señora de la Guardia". La propuesta consiste de dos momentos, un primer momento cuando los estudiantes utilicen una aplicación creada por los docentes como forma de evaluación de los contenidos trabajados en clase. Un segundo momento en donde los estudiantes desarrollan aplicaciones para jugar con sus compañeros.

El proyecto propone, repensar el lugar de las aplicaciones en la educación, ubicándolas como facilitadoras de dinámicas académicas, además buscan otorgar un rol activo a los estudiantes en el desarrollo de nuevas tecnologías.

Profe, ¿qué le debo? Las integrantes de la dupla autoras de esta propuesta pertenecen a las áreas de Computación y Pedagogía. Se presenta el desarrollo de una aplicación por parte de los estudiantes que sirva para llevar un registro personalizado de los trabajos prácticos adeudados, evaluaciones, materias en proceso y materias adeudadas en condición de previa. La intención es que los estudiantes estén involucrados desde el primer momento, brindando ideas originales y creativas. La propuesta busca atraer a los estudiantes a la informática, utilizando como medio la creación de una aplicación simple, que les resulte útil en este momento de su trayectoria académica, como también a lo largo de toda su vida. Esta propuesta presta atención al desarrollo de las habilidades blandas que se ponen en juego al momento de participar del desarrollo compartido de una aplicación móvil como lo son la expresión oral y escrita, la colaboración, la construcción de consensos y la construcción colectiva.

La música en juego Esta propuesta está elaborada por una docente del área Lenguaje Musical en colaboración con un docente de Computación que se desempeñan en la Escuela Superior de Música. Se plantea aprovechar la heterogeneidad de instrumentos que tocan los estudiantes, para el desarrollo de aplicaciones

simples que vinculen sonidos con la ejecución de cada instrumento. Esta propuesta está enmarcada en un entorno de formación de formadores, visibilizando la transdisciplina de una forma simple, claramente situada al contexto escolar. Por otra parte la organización de la secuencia permite la construcción gradual de la autonomía complejizando progresivamente el producto en desarrollo.

Quien se leyó todo Docentes de Computación y Letras crearon una propuesta que propone a los estudiantes desarrollar una aplicación de preguntas y respuestas con asignación de puntos por aciertos. La obtención de puntos determina qué tanto se sabe un jugador sobre un libro determinado. Se busca que los estudiantes se diviertan asumiendo el rol de desarrolladores de una aplicación, proponiendo las preguntas y que se entusiasmen con la lectura. Además, se busca que se impregnen de las posibilidades que brinda el desarrollo de una aplicación para desplegar formas lúdicas de comprobación de lectura.

Frac/App La propuesta tiene como propósito, compartir el análisis y la reflexión acerca del desarrollo de aplicaciones móviles en la enseñanza de las fracciones mediante el trabajo coordinado de las Áreas de Informática y Matemática. El proyecto, que se llama Frac/App, busca transformar la forma de enseñar y aprender brindando la posibilidad de acceder a nuevos espacios de interacción que se llevan a cabo dentro del aula. Como hilo conductor, se enfoca el desarrollo de las aplicaciones utilizando la fracciones para resolver problemas y situaciones de la vida cotidiana.

Appturismo Docentes pertenecientes a las áreas de Computación y Turismo, dentro de la materia "Interpretación de la Naturaleza I" del CPEM 68 de Villa la Angostura, armaron un proyecto que pretende contribuir desde la escuela a la comunidad cercana. Se trata de una ciudad turística receptora de visitantes que cada vez presentan una mayor exigencia y buscan una mayor interacción con destinos que visitan.

Las nuevas tecnologías aplicadas al turismo permiten un mayor conocimiento entre oferta y demanda, una personalización de las mismas y una transformación del sector de un modo más sostenible y sustentable. En este sentido, el objetivo principal es fomentar el pensamiento computacional en pos de desarrollar un prototipo de aplicación enfocada en tecnologías móviles que aumente el conocimiento de las actividades turísticas de zonas.

Calculadora IMC Proyecto que vincula las áreas de Computación, Educación Física y Educación Sexual Integral. Se plantea que los estudiantes desarrollen una aplicación para el cálculo del Índice de Masa Corporal. Se espera que la aplicación solicite algunos parámetros personales y devuelva el valor del IMC con ciertos consejos según el valor obtenido. La complejidad de la calculadora podría ir en aumento solicitando nuevos parámetros para hacer más específico y personalizado el resultado.

La modalidad de trabajo es una adaptación del "Aula invertida" en la que los estudiantes desarrollen los contenidos de forma autónoma en una instancia asincrónica, por ejemplo investigan en sus casas cómo agregar una imagen a la pantalla y luego en el encuentro presencial lo ponen en práctica.

5 Discusión

Las secuencias didácticas producidas por las duplas de docentes confirman las conjeturas iniciales en cuanto a que los docentes de Informática están en condiciones de crear secuencias didácticas para el desarrollo de aplicaciones móviles en articulación con diferentes áreas.

Las primeras experiencias en la práctica mediada, dentro del aula, han mostrado interés por parte de los estudiantes por el desarrollo de aplicaciones móviles. Las experiencias aportan fuertes indicios acerca de que que los estudiantes son capaces de desarrollar aplicaciones, porque existen aplicaciones creadas por estudiantes vinculadas a las secuencias generadas en el marco del Taller. Sin embargo, al momento de la elaboración del presente documento, no se encuentran datos suficientes como para analizar el grado de afinidad de los estudiantes para con el proceso de desarrollo y que estas aplicaciones han incidido positivamente en sus medios.

6 Conclusiones

Una participación activa de la docencia en el diseño y elaboración de secuencias didácticas, acompañada de una práctica mediada, como primeros pasos, permiten por un lado comprometer a los docentes el la formulación de sus propuestas académicas y por el otro validar o rechazar presupuestos en etapas muy recientes. El desarrollo de Aplicaciones es un recurso muy maleable que favorece la interdiciplina y trabajo en duplas o equipos de docentes.

Referencias

1. CC2020 Task Force. *Computing Curricula 2020*. Association for Computing Machinery, New York, NY, USA, 2020.
2. M. Coenraad, J. Palmer, D. Eatinger, D. Weintrop, and D. Franklin. Using participatory design to integrate stakeholder voices in the creation of a culturally relevant computing curriculum. *International Journal of Child-Computer Interaction*, 7 2021.
3. B. DiSalvo, J. Yip, E. Bonsignore, and D. Carl. Participatory Design for Learning. In Betsy DiSalvo, Jason Yip, Elizabeth Bonsignore, and Carl DiSalvo, editors, *Participatory Design for Learning. Perspectives from Practice and Research*, chapter 1. Routledge, New York, 1 edition, 2017.
4. G. Edelstein. Un capítulo pendiente: el método en el debate didáctico contemporáneo. In *Corrientes Didácticas Contemporáneas*, chapter 3, pages 75–89. 1996.

5. B. Gros and E. Durall. Retos y oportunidades del diseño participativo en tecnología educativa. *Edutec. Revista Electrónica de Tecnología Educativa*, (74):12–24, 12 2020.
6. K-12 Computer Science Framework Steering Committee. K-12 Computer Science Framework. Technical report, ACM, New York, NY, USA, 2016.
7. J. Kelter, A. Peel, G. Anton, and S. Dabholkar. Seeds of (r)Evolution: Constructionist Co-Design with High School ScienceTeachers. Dublin, Ireland, 5 2020. Constructionism 2020.
8. B. Naimipour, M. Guzdial, and T. Shreiner. Engaging Pre-Service Teachers in Front-End Design: Developing Technology for a Social Studies Classroom. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2020.
9. T. Robertson and J. Simonsen. Challenges and opportunities in contemporary participatory design. *Design Issues*, 28(3):3–9, 2012.
10. Royal Society. After the reboot: Computing education in UK schools. *Policy Report*, 2017.
11. W. Sandoval. Conjecture Mapping: An Approach to Systematic Educational Design Research. *Journal of the Learning Sciences*, 23(1), 1 2014.
12. Socientize Consortium - European Commission. Citizen Science for Europe. Towards a better society of empowered citizens and enhanced research. Technical report, The Socientize Consortium of the European Commission, 2013.
13. M. Tissenbaum and A. Ottenbreit-Leftwich. A Vision of K-12 Computer Science Education for 2030. *Communications of the ACM*, 63(5), 2020.
14. M. Tissenbaum, J. Sheldon, and H. Abelson. From computational thinking to computational action. *Communications of the ACM*, 62(3):34–36, 2019.

Metodologías innovadoras en el desarrollo y la evaluación de competencias digitales de docentes y estudiantes universitarios

Innovative Methodologies in the Development and Evaluation of Digital Skills of university teachers and students

Claudia Russo [0000-0002-0345-4783], Mónica Sarobe [0000-0001-5987-2696], Tamara Ahmad [0000-0002-9197-266X], Natalia Sinde [0000-0002-2007-2082].
Educación Digital e ITT-CIC, UNNOBA, Sarmiento 1169, Junín (C6000), Argentina
claudia.russo, monica.sarobe, tamara.ahmad, natalia.sinde {@itt.unnoba.edu.ar}

Resumen

Tras garantizar la continuidad académica y la capacitación para el buen funcionamiento del entorno virtual en marzo de 2020, el equipo de Educación Digital y el Sistema Institucional de Educación a Distancia y Digital (SIEDD) de la Universidad Nacional del Noroeste de la Provincia de Buenos Aires (UNNOBA) crearon un *modelo de evaluación para medir la calidad del e-learning.* Se partió de la definición de dimensiones que, indirectamente, observaron competencias digitales. Asimismo, entre las estrategias para incrementar competencias digitales docentes y estudiantiles se hallaron microcharlas, cursos autoguiados y capacitaciones asíncronas con instancias sincrónicas destinadas a la formación virtual continua. Además, se dispuso de una red de tutores digitales y soporte técnico-administrativo. El modelo de evaluación de la calidad *e-learning* brindó al profesorado la posibilidad de tomar decisiones para optimizar el diseño del aula virtual en sus aspectos pedagógico-didácticos, organizativos, navegables, etc. La coherencia necesaria entre *metodologías de evaluación* y *de enseñanza* requirió un cambio en la función docente hacia el rol de facilitador educativo, generando experiencias a ser compartidas en el *Workshop de Innovación y Transformación Educativa* anual. Hoy se auspicia la aplicación de salas virtuales inmersivas, aulas híbridas, holografías y RA.

Palabras clave: Competencias digitales; innovación educativa; metodología de evaluación; metodología de enseñanza; educación digital.

Introducción

El modelo de evaluación de la calidad e-learning se enmarca en un proyecto de investigación cuyos primeros avances fueron publicados en marzo del corriente año (Russo, Sarobe y Ahmad, 2022). La iniciativa promueve estándares internacionales de educación digital partiendo de la innovación en educación en informática para docentes y estudiantes. Puntualmente, procura aplicar estrategias, metodologías y herramientas innovadoras en enseñanza-aprendizaje a través del desarrollo de experiencias novedosas y capacitaciones en competencias digitales. Con tal fin, desde en marzo de 2020, el área de Educación Digital y el SIEDD de la UNNOBA pusieron a disposición de estudiantes, docentes y no docentes entornos virtuales para actividades de grado, posgrado y extensión. A la par, se dio seguimiento a las necesidades del profesorado, se creó el *Programa de Becas de Conectividad* y contactó al estudiantado sin actividad en la plataforma a fin de asesorar sobre el funcionamiento de los espacios digitales. Antes de la coyuntura internacional, en febrero-marzo del 2020, Educación Digital dictó un curso

de posgrado para capacitar docentes en saberes digitales educativos. Tras el aislamiento ofrecieron numerosas microcharlas, capacitaciones, videos y tutorías de acompañamiento académico y crearon dos diplomaturas universitarias de posgrado. Luego de asegurar continuidad académica y formación para el buen funcionamiento del entorno virtual, Educación Digital y SIEDD crearon un *modelo de evaluación para medir la calidad del e-learning.* Esto fue posible gracias a la construcción de indicadores sobre variables con influencia directa en la experiencia educativa, su aplicación y su puesta a punto mediante la red de tutores. La mayor virtud de este modelo fue prever escenarios y transformarlos con reorientación laboral docente y no docente. Además, hizo posible un seguimiento constante del nivel de suficiencia digital desde 2020 hasta la actualidad, facilitando la adecuación de las propuestas de capacitación a las necesidades reales del profesorado y del estudiantado, extendiéndose al ámbito no docente.

Definición y evaluación de la competencia digital docente

Diseñado para su aplicación a cursos y asignaturas virtuales o con un porcentaje de horas virtualizadas, el *modelo de evaluación de la calidad de las aulas virtuales* solicitó la definición de 4 dimensiones observadas mediante 8, 10, 4 y 3 indicadores cada una. La primera dimensión fue la *Presentación del Aula Virtual y* refirió a la estructura del aula, la información disponible para estudiantes sobre generalidades del curso y herramientas básicas destinadas a la comunicación. Sus indicadores fueron la presencia de: 1.1. una *imagen en la descripción del resumen del aula* para su fácil identificación, 1.2. un *texto en la descripción del resumen del aula de no más de 100 caracteres* para su correcta visualización y lectura, y 1.3. una *sección de bienvenida* detallando el equipo docente. También se tuvo en cuenta la existencia de 1.4. un *programa de la asignatura* con objetivos, contenidos, recursos y demás información de relevancia, 1.5. un *cronograma de actividades* para optimizar la organización estudiantil y 1.6. un *documento de información* sobre la modalidad de evaluación, aprobación, comunicación y tutoría. Además, se consideró la disponibilidad de un 1.7. *foro de avisos* para la comunicación unidireccional de docentes hacia estudiantes y un 1.8. *foro de consultas* reservado a un intercambio de mensajes más amplio entre docentes-estudiantes y estudiantes entre sí.

La segunda dimensión observada fue la *Organización Didáctica y Pedagógica del aula* cuyo objetivo era analizar la propuesta pedagógico-didáctica, los materiales utilizados y la organización de los contenidos teniendo en cuenta los recursos ofrecidos por la Plataforma ED UNNOBA. Tuvo como indicadores la existencia de: 2.1. un *buen diseño visual del aula* para evaluar la organización de la propuesta, 2.2. un *mínimo de 3 tipos de materiales* para accesibilidad y motivación estudiantil, 2.3. un *mínimo de 3 actividades de la plataforma* en función de sus objetivos pedagógicos, 2.4. un *mínimo de 2 recursos de la plataforma* para facilitar el material del curso y 2.5. un *orden y concordancia* entre materiales teóricos, actividades prácticas y consignas. Se estimó 2.6. la *utilización de recursos externos* para gamificación e interactividad, 2.7. *retroalimentaciones y/o devoluciones de las actividades* para promover la participación y 2.8. *consignas para cada actividad* donde se explicitasen propósitos y expectativas, carácter individual o grupal y otros detalles. Finalmente, se consideró la disponibilidad de 2.9. una *encuesta final* para la evaluación estudiantil del curso y futuros ajustes y 2.10. la evidencia de actividades que *estimulen trabajos colaborativos*.

La tercera dimensión se ocupó del *Seguimiento de estudiantes* pues el proceso educativo no sólo se trata de aprender contenidos sino también de establecer interacciones entre componentes y vínculos académicos. Fueron indicadores: evidencia de 3.1. *intercambio*

de mensajes por los canales de comunicación activos para una comunicación eficiente y 3.2. *estrategias docentes para estudiantes con participación asincrónica* como correlato de lo sincrónico. El indicador 3.3. apeló a un *bajo porcentaje de estudiantes que no ingresaron nunca al aula* y permitió relevar en un rango de tres opciones (menor a 20%, entre 20%-30%, mayor a 30%) a estudiantes con dificultades para iniciar el curso virtual. Finalmente, la 3.4. *activación y configuración de la herramienta de progreso en finalización de actividades* fue una herramienta proporcionada por la plataforma para facilitar el análisis de datos y la toma de decisiones sobre deserción/desgranamiento.

La cuarta y última dimensión se centró en la *Evaluación* al examinar presencia de pautas claras, calificación y aprobación en el curso virtual. Se consideró importante verificar la existencia de actividades para la autoevaluación, la coevaluación y la heteroevaluación con indicadores, criterios, puntuaciones, rúbricas y demás características siempre explicitadas claramente y disponibles para el estudiantado. Sus indicadores fueron la definición de: 4.1. una *calificación de las actividades planificadas,* que puso de manifiesto si en el curso virtual se calificaron las tareas y la incidencia de cada una de ellas en su aprobación o desaprobación; 4.2. una 4.2. *metodología de evaluación* conocida por los estudiantes desde el comienzo del curso; y 4.3. una *definición clara de los criterios de aprobación* para mejor organización en el desarrollo de las actividades. Con el objetivo de evaluar los indicadores apuntados, se definieron los valores 2 para *Si* y 0 para *No* en casos que admitían dos respuestas posibles, y 2 para SI, 1 para *Puede mejorar-Incompleto* y 0 para *No*, en aquellas situaciones con tres posibles respuestas. La sumatoria de valores se realizó por dimensión y de modo general arrojando una valoración del aula completa y asignando un color según el valor alcanzado: rojo para valoración general baja, amarillo para valoración general media y verde para valoración general alta. Junto a su aplicación por la red de tutores, el modelo se instrumentó como autoevaluación docente y heteroevaluación estudiantil anónima y permanente. (Russo, Sarobe y Ahmad, 2022).

Como estrategias específicas para el incremento de las competencias digitales de docentes y estudiantes, desde 2020 se ofrecen microcharlas, cursos autoguiados y capacitaciones asíncronas con instancias sincrónicas para formación virtual y continua. Además, se redactaron documentos sobre buenas prácticas educativas en el uso de espacios virtuales y protocolos de tutores, docentes y estudiantes sobre metodologías de evaluación para cursos, mesas de exámenes finales y otras instancias virtuales. Las microcharlas versaron sobre creación y curaduría de contenidos digitales educativos, desarrollo de consignas claras, diseño básico de aulas iconográficas, estimación de carga horaria en la virtualidad, guía didáctica sobre *aula invertida* y aplicación de herramientas Moodle como foros, tareas y cuestionarios. También se discurrió sobre acciones tutoriales en el seguimiento estudiantil, creación de videoconferencias y metodologías de bosquejo de lecciones, entre otros temas. En cuanto a las capacitaciones, ahondaron en la aplicación de herramientas y metodologías del entorno virtual y procuraron alfabetizar digitalmente al profesorado para mayor fluidez digital en el uso de equipos, software y comunicación emergente. Los cursos fueron optimizados en dos diplomaturas: la *Diplomatura Universitaria en Diseño de Aulas Virtuales* y la *Diplomatura Universitaria en Diseño de Contenidos Educativos Digitales*; van por su segunda cohorte. La primera se orientó al desarrollo de habilidades de enseñanza en entornos virtuales, la promoción de actividades digitales con recursos de la plataforma ED UNNOBA, la planificación de cursos virtuales, diseño de consignas y reflexión sobre prácticas pedagógico-didácticas con TIC. La segunda se enfocó en el concepto y ejercicio del diseño, la producción y curaduría de contenidos digitales y la utilización de herramientas complementarias a las del entorno (Educación Digital, 2021).

Resultados

En su carácter transformativo, el *modelo de evolución de la calidad e-Learning* brindó la posibilidad de tomar decisiones para la mejora del diseño del aula virtual tanto en sus aspectos didáctico-pedagógicos como en su organización, orden, navegabilidad y demás características generales. El modelo fue compartido con integrantes de la Asociación de Universidades Latinoamericanas - Campus Virtual Latinoamericano (AULA-CAVILA). Quienes representan a la Universidad de Extremadura (UEx) dentro de AULA-CAVILA, trabajaron en una encuesta de autoevaluación similar a la de la UNNOBA y recabaron información luego compartida para su comparación analítica. Desde aquel intercambio inicial, UNNOBA, UEx y otras universidades que participan de AULA-CAVILA trabajan conjuntamente para unificar el modelo de evaluación y cruzar datos en un estudio común. La posibilidad de que el estudiantado acceda a esta herramienta en forma de encuesta final de curso permitió brindar una devolución detallada y sistematizada de su experiencia en las aulas virtuales capitalizando la comunicación estudiante-docente (Russo, Sarobe y Ahmad, 2022). Hoy se aguardan los resultados de las encuentas a docentes y estudiantes del primer cuatrimestre de 2022, cuyo análisis será publicado en presentaciones futuras.

En cuanto al incremento de la competencia digital docente y estudiantil, la técnica FODA presenta la articulación SIEDD-ED como mayor fortaleza institucional, mientras que su vínculo con las demás unidades académicas fue un punto débil que ha sido trabajado hasta alcanzar una mejoría sustancial. Si bien se identificó como principal amenaza la reticencia de parte del cuerpo docente al cambio de paradigma, las entrevistas tras las capacitaciones mostraron alto grado de satisfacción con las actualizaciones propuestas. Se vislumbraron nuevas oportunidades asociadas a las múltiples aplicaciones del *modelo de evaluación del e-learning* y a herramientas digitales de utilización incipientes en la UNNOBA como salas virtuales inmersivas y aulas híbridas (Russo, Sarobe y Ahmad, 2021). En cuanto al soporte técnico-administrativo, el acompañamiento de SIEDD-ED fue cuantificado para el año 2020 en 1171 tickets respondidos, 673 espacios generados, 866 cursos creados en plataforma ED, 1516 expedientes remitidos y 6255 espacios en plataforma ED finales. Para 2021 los valores fueron similares. La *Oficina virtual* sigue recibiendo un promedio diario de cinco consultas de lunes a viernes de 9 a 12 horas (Educación Digital, 2021).

Además, se dio lugar a una nueva normativa en la que se definió la *evaluación virtual* como componente de la propuesta digital formativa presencial y a distancia para observar, recoger y analizar información destinada a mejorar el proceso de enseñanza-aprendizaje. La noción introdujo técnicas e instrumentos alternativos para una *evaluación formativa* (Russo, Sarobe y Ahmad, 2021). La debida coherencia entre *metodologías de evaluación* y *metodologías de enseñanza* requirió del desempeño del profesorado como facilitador y se propuso a cada docente ejercitar al estudiantado a través de rutinas de estudio, debates y planteamientos de dudas, indicación de errores y ofrecimiento de soluciones oportunas. La evaluación formativa implicó el peritaje constante de propuestas, certificación brindada y operacionalización en el entorno LMS de Moodle. Los instrumentos de la Plataforma ED UNNOBA fueron combinados con recursos externos de Youtube, H5P, Genially y Prezi, entre otros. De modo general, se planteó la necesidad de desarrollar ofertas mayormente asincrónicas y siempre accesibles guardando un registro del progreso y de la finalización de tareas a partir de informes de *Actividad del curso*, *Participación en el curso* y *Finalización de la actividad.* (Russo, Sarobe, Ahmad y Sinde, 2022). Se plantearon retos a futuro en torno a cambios en los requisitos de aprobación como la asistencia estudiantil exigida con la vuelta a la presencialidad.

Además, el *Workshop de Innovación y Transformación Educativa* (WITE), las *Jornadas de Estudiantes* y el *Concurso de Materiales Educativos Digitales Innovadores* colocaron

al área de Educación Digital y al SIEDD como líderes transformacionales en la institución y referentes regionales. Puntualmente, el WITE es un evento virtual anual organizado por la UNNOBA junto a la Universidad Nacional de San Antonio de Areco (UNSAdA) y la Universidad Provincial de Ezeiza (UPE). Inaugurado en el año 2020, el espacio posee una importancia fundamental dentro de un contexto en el que la educación a distancia debió ofrecer, desde las particularidades de lo virtual parámetros de calidad tan válidos como los de la educación presencial. Esto presentó un gran desafío y destacó la importancia de la formación profesional continua y de la construcción de espacios de encuentro a nivel intra e inter institucional. El WITE se orientó especialmente al diseño de cursos virtuales y a la producción original de contenidos digitales, impulsó propuestas tutoriales y evaluaciones alternativas, motivó la alfabetización digital y el hacer digital crítico y previó el intercambio de estrategias de enseñanza y tácticas de aprendizaje innovadoras para la creación de respuestas flexibles ante situaciones emergentes. Además, gracias a sus aportes reales para el acceso universal a la educación, a la permanencia académica de calidad y al egreso estudiantil propios de una sociedad genuinamente democrática, el WITE pudo ser estimado como un evento fundamental para el avance hacia un aula sin fronteras. Las ediciones 2020, 2021 y 2022 culminaron con la publicación de libros digitales de acceso gratuito (Educación Digital, 2021).

Por su parte, las *Jornadas de Estudiantes* se desarrollaron en 2020 y 2021 organizando al estudiantado en tres grupos: uno inicial, constituido por estudiantes de primer año; otro de segundo y tercer año; y por último un tercer grupo con cursantes de cuarto y quinto año de las carreras de grado. Las preguntas que dispararon el intercambio fueron: *¿Cómo te adaptaste a la modalidad y las herramientas de la plataforma? ¿Cómo consideras que fue el diseño de los cursos en cuanto a materias, consignas y actividades)? ¿Cómo fue la comunicación con docentes y pares? ¿Cuál fue tu participación en los cursos UNNOBA en movimiento* y en las *Tutorías Disciplinares*? Por último, ¿*cómo fue tu experiencia de evaluación en la virtualidad?* (Educación Digital, 2021).

Además, UNNOBA, UNSAdA y UPE convocaron al primer *Concurso de Materiales Educativos Digitales Innovadores* orientado a reconocer y promover el desarrollo de estrategias docentes para el impulse de buenas prácticas y experiencias transferibles y replicables. El público votó ganadores y menciones especiales (Educación Digital, 2021).

Actualmente, ED-SIEDD trabaja en la inclusión de la realidad aumentada, la utilización de hologramas para la presentación de docentes externos y la creación de simulaciones para aprendizaje visual, entre un sinfín de posibles aplicaciones vinculadas al incremento de la calidad educativa y a la mayor accesibilidad al ámbito universitario.

Referencias

Russo, C.; Sarobe, M.; Ahmad, T. (2022). Definición de indicadores. Calidad en cursos virtuales, *Revista Iberoamericana TEYET*, 31, p. e3, marzo 2022.

Russo, C.; Sarobe, M.; Ahmad, T. (2021). Formación virtual permanente de docentes en contexto de pandemia. Experiencia 2020, *II Workshop de Tutorías en la Educación Superior*, La Plata: GITBA.

Russo, C.; Sarobe, M.; Ahmad, T.; Sinde, N. (2022). "Evaluación continua mediante actividades asincrónicas en los campus virtuales" [aprob.]*, JICV'22*, Arequipa: UCSM.

Educación Digital (2021). Ofertas de pregrado, posgrado y otras propuestas, *Informe de Emergencia 2020-2021*, Junín: UNNOBA.

Estrategias Didácticas para el Aprendizaje y la Enseñanza del Pensamiento Computacional en el Nivel Académico Universitario

Natalia Colussi[1] y Natalia Monjelat[2]

[1] Licenciada, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, UNR, Rosario, Argentina
[2] PhD, Instituto Rosario de Investigaciones en Ciencias de la Educación (IRICE, CONICET-UNR), Rosario, Argentina
colussi@fceia.unr.edu.ar, monjelat@irice-conicet.gov.ar

Resumen. El siguiente artículo presenta los avances de una línea de investigación en desarrollo que busca contribuir a mejorar el proceso de enseñanza y aprendizaje de la programación en carreras afines a las Ciencias de la Computación (CC). Puntualmente, se propone alcanzar este objetivo a partir de una revisión de las estrategias didácticas, empleando la Metodología de Enseñanza Basada en Proyectos (ABP), las actividades grupales, y las estrategias de resolución de problemas basadas en el Pensamiento Computacional. De esta forma, se han realizado numerosos cambios e innovaciones en el desarrollo y puesta en obra del cursado. A partir de las experiencias implementadas, es posible señalar múltiples aprendizajes y fortalezas de este enfoque tales como la posibilidad de seguimiento del progreso en el aprendizaje, la rápida adaptación al cambio de contexto presencial-virtual, la disminución del plagio al tratarse de producciones originales y creativas, entre otras. Estos resultados ponen en valor la estrategia del ABP para la enseñanza de conceptos de CC, particularmente en este caso, desde contenidos disciplinares de programación y el pensamiento computacional y las técnicas de resolución de problemas.

Keywords: Primer Curso de Programación en la Universidad, Aprendizaje Basado en Proyectos y Problemas, Pensamiento Computacional.

1 Introducción

Desde hace ya varios años, tanto Argentina como en el resto del mundo, existen miles de puestos de trabajo sin cubrir en el área de las Ciencias de la Computación, por la falta de recursos humanos con la formación y los conocimientos necesarios. Se observa a su vez, que las carreras afines a éstas ciencias presentan en el último tiempo un ingreso masivo, pero al mismo tiempo una gran deserción en los primeros años [1,2]. Según diversos autores, en esta problemática intervienen varios factores. Entre ellos se señalan al desconocimiento inicial de las disciplinas, la falta de adaptación al

nivel universitario, así como también las dificultades propias de las materias disciplinares impartidas en los primeros años.

A partir del reconocimiento de estas problemáticas, y particularmente de la última, la línea de investigación que aquí se presenta tiene por objetivo mejorar el proceso de enseñanza y aprendizaje de la programación en carreras afines a las Ciencias de la Computación a través de la Metodología de Enseñanza Basada en Proyectos, las actividades grupales, y las estrategias de resolución de problemas basadas en el Pensamiento Computacional. El contexto en el que se desarrolla la propuesta, es el segundo dictado de las materias de programación de los primeros años de las carreras Licenciatura en Ciencias de la Computación (LCC), Licenciatura en Matemática (LM), y Profesorado en Matemática (PM).

La propuesta del redictado que se propone desde la línea de investigación, intenta recurrir a estrategias de enseñanza y aprendizaje que se han mostrado eficaces para la construcción de saberes disciplinares en diferentes contextos y áreas disciplinares [3,4]. De esta forma, el ABP aparece como una opción innovadora para la enseñanza de la programación, haciendo del redictado una experiencia diferente al dictado inicial [5]. Asimismo, los proyectos están sustentados en las premisas del pensamiento computacional [10] y las estrategias de resolución de problemas de Polya [11] adaptadas por Thompson [12] para el desarrollo de programas, estableciendo pilares disciplinares sólidos para el futuro en la carrera académica y profesional de los estudiantes

Para estudiar el impacto de la propuesta se han diseñado encuestas que se administran al finalizar cada cursada. Asimismo se cuenta con registros de actas de cursado y de regularización, que permiten observar la cantidad de estudiantes que inician y finalizan el curso. En este documento se presentarán algunos resultados parciales de estos análisis, considerando la extensión del artículo.

Las acciones desarrolladas se enmarcan en un proyecto de investigación radicado en la Facultad de Ciencias Exactas, Ingeniería y Agrimensura (FCEIA) dependiente de la Universidad Nacional de Rosario (UNR). La investigación conforma un proyecto bienal, extendido por pandemia al año 2022 resolución CS 436/2022 y que dispone como fecha de finalización 31/12/2022.

2 Motivación

En la cátedra de Programación I y Programación se imparten conocimientos referidos al Pensamiento Computacional, haciendo fundamental hincapié en los principios de programación [6, 7]. La adquisición y construcción del conocimiento disciplinar para los estudiantes de primer año de las carreras LCC, LM y PM no se produce para todos al mismo tiempo. Los estudiantes provienen de diferentes especialidades y contextos en su nivel medio, conviviendo así, en un mismo espacio áulico, alumnos con experiencias, intereses y expectativas dispares.

En los últimos años las cátedras de programación, en el primer cursado, presentan ingresos promedios superiores a los trescientos inscriptos, y heterogeneidad de saberes y niveles entre los cursantes. Los registros internos de actas de examen y planillas de regularización muestra que una gran cantidad de estudiantes, aproximadamente un 50% en promedio, no logran aprobar o regularizar este primer

curso de programación. De estos estudiantes libres, aproximadamente un 60% abandona la carrera, pero el 40% que persevera asiste al segundo dictado que se ofrece al siguiente cuatrimestre [15]. Este nuevo dictado, conocido en FCEIA como redictado, representa una segunda oportunidad para incorporar habilidades y/o competencias no alcanzadas en la primera instancia y es esta población de estudiantes sobre la cual se trabaja con el ABP desde el año 2017.

Para llevar adelante la propuesta basada en el ABP en la cátedra del redictado se desarrolló un "Plan de Trabajo Didáctico para el Aula" (PTDA) [8] redefiniendo el proceso de enseñanza-aprendizaje utilizado hasta el momento para brindar un mejor marco de contención y solución a los problemas que presentaba el estudiantado del curso. Esta iniciativa requirió de la reelaboración de la presentación, ejercitación y evaluación de contenidos propios de la materia. Se realizaron dos proyectos grupales de programación a lo largo del cuatrimestre los cuales permitieron a los estudiantes ejercitar y poner en práctica temas disciplinares, y ser evaluados así por los docentes a lo largo de todo el proceso de desarrollo de los mismos.

El trabajo de esta investigación se encuentra centrado entorno a los siguientes aspectos o lineamientos:

- Identificar detalladamente los problemas que manifiestan tener los estudiantes durante el primer curso de programación tanto en la faceta disciplinar como vincular con el medio académico.
- Implementar y adaptar las estrategias de enseñanza y aprendizaje activas, grupales, y colaborativas dentro de las cátedras Programación I y Programación.
- Desarrollar e investigar sobre distintas estrategias didácticas conjuntas y combinadas para abordar las problemáticas técnicas vinculadas al aprendizaje de la programación, como así también, a la inclusión de todos los estudiantes en su heterogeneidad en el proceso.
- Desarrollar e integrar actividades que fomenten aspectos motivacionales en los estudiantes, las cuales brinden el acercamiento entre pares dentro de la carrera, reconocimiento dentro de la comunidad académica, y el ímpetu de continuar y progresar dentro de la carrera; brindando así herramientas que perduren en el tiempo y sean fundacionales para mantener el proceso de aprendizaje y así la permanencia y el avance dentro de la universidad.
- Trabajar con experiencias que permitan desarrollar las denominadas *soft-skill*s [14] (habilidades o competencias no medibles o difíciles de medir, de transmisión fundada en al experiencia) técnicas o disciplinares, como por ejemplo, entender procesos de refinamiento y mejora de diseño de funciones, resolución de problemas, como aquellas *soft-skill* vinculadas a aspectos más profesionales como por el ejemplo las comunicacionales e interpersonales, desarrollo de la creatividad, y la capacidad de imaginar soluciones a problemas originales, desarrollo de capacidad crítica y objetiva ante otras soluciones, valorar respuestas disímiles a la propia, etc.

3 Resultados

Uno de los resultados que nos ha dejado la experiencia es la importancia de acompañar el desarrollo de los proyectos de diferentes formas. La implementación del ABP en los primeros años de la universidad, permite observar claramente cómo los estudiantes progresan en la adquisición de conocimiento y saberes disciplinares, ya que conlleva un seguimiento mayor por parte de los docentes. Este acercamiento provoca que los docentes y estudiantes logren vincularse de una forma más próxima, haciendo que el seguimiento del proyecto por el cual se canaliza el aprendizaje, se obtiene una observación mucho más clara de cómo se logra la comprensión de los temas a través del diseño de la solución (modelo+estrategia+código) que los estudiantes producen como soluciones parciales de un proyecto.

El trabajo grupal es un pilar fundamental en la concreción de los proyectos. Esta forma de trabajo sostiene el trabajo y motiva a finalizarlo, brindando apoyo y contención entre pares tanto en la adquisición conjunta de saberes y competencias como permitiendo la ejercitación de metodologías de trabajo vinculadas a las formas profesionales del trabajo de un programador [5]. Por otro lado, y a raíz de la situación pandémica, se ha podido analizar cómo la estrategia del ABP resulta fácilmente adaptable a la virtualidad. En nuestro caso, esa adaptación ha implicado un rediseño particularmente de la metodología de evaluación y la exposición final de los proyectos, empleando una "Vidriera Virtual de Exposición de Proyectos" ver online (https://sites.google.com/view/vidriera-proyectos-fceia/), que nos permitió llevar adelante un trabajo de evaluación asíncrono entre pares y entre los docentes con excelentes resultados en su dinámica como propuesta de exposición y posterior repositorio de los trabajos [9]. Cabe mencionar por último que la presencia de plagios se ha reducido a cero durante la realización y concreción de los proyectos. Esta problemática es usual en los primeros años y es notable como con un abordaje creativo orientado a la programación grupal favoreció a la presentación de trabajos originales.

3 Trabajos Futuros

A partir de las experiencias realizadas, los aprendizajes alcanzados y las fortalezas detectadas [5], consideramos fundamental continuar con la implementación de esta estrategia en el redictado, ofreciendo a los y las estudiantes propuestas didácticas que les permitan resignificar lo trabajado en sus primeras cursadas. En esta dirección y con la misma filosofía de trabajo, se modificó el cursado del redictado de Programación II para LCC, donde se enseña a programar usando el lenguaje Python y C, para implementar en dicha cátedra también ABP sobre un trabajo integrador en el lenguaje Python [13]. Las experiencias ganadas como docentes sobre el uso y los alcances de la estrategia didáctica y el dominio de la técnica del ABP han permitido lograr nuevas propuestas pedagógicas con resultados igualmente de prometedores, que serán también documentados para su posterior análisis y difusión.

Referencias

1. Benotti, L., Echeveste, M., Schapachnik, F.: Despertando vocaciones en computación mediante el uso de autómatas de chat. En 6tas Jornadas de Vinculación Universidad Industria, 41JAIIO, 12-21 (2012).
2. Dapozo, G., Greiner, C., Pedrozo Petrazzini, G., Chiapello, J:. Vocaciones tic. ¿qué tienen en común los alumnos del nivel medio interesados por carreras de informática? En IX Congreso de Tecnología en Educación & Educación en Tecnología, 128-139 (2014).
3. Sánchez, P., Blanco, C.: Implantación de una metodología de aprendizaje basada en proyectos para una asignatura de Ingeniería del Software. En *Actas XVIII JENUI 2012*, Ciudad Real, Universidad Nacional de Cantabria, España (2012).
4. García Martín, J., Perez Martínez, J. : Aprendizaje basado en proyectos: método para el diseño de actividades. Revista Tecnología, Ciencia y Educación, 5, 37-63 (2018).Martinez Lopez, P.: Las Bases de la Programación. Publicado electrónicamente por la Universidad Virtual de Quilmes, La Plata, (2013).
5. Colussi, N., Viale, P., Monjelat, N.: Proyectos Grupales de Programación. Experiencias del ABP en el Aula Universitaria. En: Sgreccia, N. (ed.) Jornadas de experiencias innovadoras en Educación en FCEIA, 136-153. UNR, Rosario (2021)
6. Martinez Lopez, P.: Las Bases de la Programación. Publicado electrónicamente por la Universidad Virtual de Quilmes, La Plata, (2013).
7. Felleisen, M., Findler, R., Flatt, M., Krishnamurthi, S.: How to Design Programs: An Introduction to Programming and Computing. MIT Press, USA (2001).
8. Colussi, N., Viale, P.:Actividades de Programación Grupales para Primer año de la Licenciatura en Ciencias de la Computación - Experiencias Didácticas en el Aula. Versión Extendida. En Jornadas de CyT de la UNR, (JorCyT), (2019).
9. Colussi, N., Viale, P., Monjelat, N.: Vidriera de proyectos: una modalidad de evaluación posible en tiempos de virtualidad. En Actas de II WITE - TRANSFORMACIÓN DIGITAL. DESAFÍOS DE LA EDUCACIÓN SUPERIOR, 1–10, (2021)
10. Wing, J.M.: Computational thinking. Commun. ACM , 49(3), 33-35, (2006).
11. Polya, G.:How To Solve It: A New Aspect of Mathematical Method. Princeton: University Press, (1973).
12. Thompson S.: Where do I begin? A problem solving approach in teaching functional programming. En Glaser, H., Hartel, P. y Kuchen, H. (eds.) Programming Languages: Implementations, Logics, and Programs. PLILP 1997. Lecture Notes in Computer Science (vol 1292). Springer, Berlin, Heidelberg, (1997).
13. Frydenberg, M., Mentzer, K.:. From Engagement to Empowerment: Project-Based Learning in Python Coding Courses. EDISG Conference, Information Systems & Computing Academic Professionals, (2020).
14. Keogh S., Bradnum J., Anderson E.:Improving professionalism in first year computer science students: Teaching what can't be taught. En Proceedings of the 3rd Conference on Computing Education Practice (CEP '19), 1-4, (2019).
15. Monjelat, N.; Colussi, N.; Viale, P.: Introducción a la Programación en carreras terciarias y universitarias de Computación: estado del arte en el contexto argentino. Jornadas de Ciencia y Técnica de la UNR, (2021).

Track "Gobierno Digital y Ciudades Inteligentes"

Coordinadores

Elsa Estevez (UNS)

Ariel Pasini (UNLP)

Modelado conceptual de ciudades inteligentes: un mapeo sistemático de literatura

Joaquín Cerviño, Lisandro Fernández,
José Luis Gobbe y Marisa Panizzi

Programa de Maestría en Ingeniería en Sistemas de Información. Escuela de Posgrado.
Universidad Tecnológica Nacional. Regional Buenos Aires (UTN-FRBA).
Medrano 951. (C1179AAQ). CABA, Argentina.
cjoackin@gmail.com, lifoernandez@gmail.com,
jlgobbe@outlook.com, marisapanizzi@outlook.com

Resumen. La ciudad inteligente (CI) gestiona de manera eficiente los flujos urbanos a través del proceso en tiempo real de información sobre dispositivos, ciudadanos y activos. Determinar un modelo conceptual de CIs estandarizado es importante para establecer un plan completo, una comprensión concisa sobre el dominio, habilitar estrategias de desarrollo y asegurar que múltiples iniciativas están alineadas.
En este artículo se presentan los resultados de un mapeo sistemático de la literatura (en inglés, *systematic mapping stud*y o SMS) para establecer el estado del arte de las contribuciones de modelado conceptual de las CIs. Se realizó una búsqueda en las librerías digitales IEEE Xplore y ACM desde octubre del 2015 a marzo del 2022 y se analizaron 48 estudios primarios. Se evidenció que del modelado lo más publicado son diagramas de arquitectura. Y dentro de los diagramas UML, los predominantes son los diagramas de clases y diagramas de actividad.

Palabras clave: Modelado conceptual, ciudades inteligentes, mapeo sistemático de la literatura.

1 Introducción

El modelado conceptual busca representar conceptualizaciones y abstracciones relevantes del mundo real de tal manera que sea posible apoyar la comunicación, discusión, análisis y actividades relacionadas [1]. El modelado conceptual y el razonamiento sobre modelos son capacidades humanas para observar, comprender e influir en el entorno. A pesar de innumerables intentos, no existe una definición estricta de uso general de lo que constituye el modelado conceptual. Los intentos de definición son variantes de "Modelado conceptual es modelado con conceptos" e introducir estos conceptos a través de marcos ontológicos más o menos rígidos, o mediante una explicación simple usando lenguaje natural [2]. Los modelos conceptuales son modelos de representaciones mentales que agentes construyen, usan y manipulan durante la actividad cognitiva. Como tales, no son modelos de un dominio dado, sino modelos de cómo concebimos ese dominio. Se los puede caracterizar como artefactos producidos con la intención deliberada de describir una realidad conceptualizada. De esta forma, se puede afirmar que establecen contratos de sentido, con el requisito previo de que se exprese una conexión un modelo que proporcione una semántica conceptual. Estos

artefactos se comprometen con una conceptualización, es decir, la cosmovisión capturada por dicha conceptualización. En definitiva, se puede afirmar que los modelos conceptuales captan y comunican un determinado compromiso ontológico [3].

En la práctica, la Ciudad Inteligente (en inglés, *Smart City*) gestiona de manera eficiente los flujos urbanos a través del proceso en tiempo real de información sobre dispositivos, ciudadanos y activos [4].

Las experiencias tempranas de ciudades inteligentes se remontan a la década de 1970, cuando Los Ángeles realiza el primer proyecto de big data urbana. Cerca del comienzo del siglo XXI el interés aumentó significativamente como consecuencia de la mejora tecnológica y el crecimiento de la población en áreas urbanas, pero a partir de la década de 2010 es cuando este concepto emerge y se comienza a discutir [4].

A pesar de que el concepto ha sido discutido durante varias décadas, todavía no existe una definición del término [5]. La ciudad inteligente es todavía un concepto poco claro sin una nomenclatura estandarizada que pueda ser efectiva describiéndose a sí mismo. La gran mayoría de la literatura define "ciudad inteligente" como infraestructura que cumple las siguientes tres características: (i) el grupo objetivo son las ciudades y comunidades, (ii) se mejora la forma de vivir y trabajar en la región, (iii) se implementan tecnologías de la información y la comunicación (TIC) [4].

De todas formas, la falta de un marco y criterios estandarizados hace que la mayoría de las ciudades inteligentes basen su desarrollo en un marco autorregulado. Los involucrados en este proceso no serán capaces de adecuar correctamente el concepto en sí mismo sin comprender sus fundamentos. Además, de la necesidad de contar con un marco estandarizado es importante establecer un plan completo y una concisa comprensión sobre el dominio [5].

La determinación de un modelo conceptual de ciudad inteligente habilitará a profesionales, políticos y a la academia a establecer mejores estrategias de desarrollo y asegurará que múltiples iniciativas están alineadas [5].

Este artículo se desarrolla en el marco del Seminario de Modelado Conceptual de la Maestría en Ingeniería de Sistemas de Información de la Universidad Tecnológica Nacional, Regional Buenos Aires. La elección del tema ha sido motivada por los tópicos de interés del área de "Aplicaciones avanzadas y multidisciplinarias" propuestas en la 41 edición del Congreso Internacional de Modelado Conceptual (ER 2022) [6]. En este artículo se presenta un mapeo sistemático de la literatura (SMS) para analizar el estado del arte y descubrir las contribuciones que existen en relación con el modelado conceptual de las ciudades inteligentes. Para realizar el SMS se siguieron los lineamientos propuestos por Kitchenham et al. [7].

El artículo se estructura de la siguiente manera: en la Sección 2 se describe la planificación del SMS, en la Sección 3 se describe su ejecución. Los resultados se presentan en la Sección 4. En la Sección 5 se presenta un análisis de las amenazas a la validez y, finalmente, en la Sección 6 se exponen las conclusiones y trabajos futuros.

2 Planificación del SMS

En la presente sección se detalla el protocolo de revisión del SMS: preguntas de investigación (PI), estrategia de búsqueda, criterios de inclusión y exclusión, proceso de selección, estrategia de extracción y síntesis de datos. El objetivo del SMS es dar respuesta a la pregunta de investigación (PI): *¿Cuál es el estado del arte respecto al*

modelado conceptual de ciudades inteligentes? Se considera que dicha pregunta principal puede desglosarse en una serie de sub-preguntas, éstas son detalladas a continuación en la Tabla 1.

Tabla 1. Preguntas de investigación (PI) y motivación.

	Preguntas de investigación (PI)	**Motivación**
PI1:	¿Qué tipos de contribuciones existen en el modelado conceptual de ciudades inteligentes?	Encontrar y comprender qué tipo de aportes otorgan en cuanto modelado conceptual.
PI2:	¿En qué dominios se realizaron contribuciones?	Identificar los dominios de acuerdo con la taxonomía de Wahab *et al*. [5].
PI3:	¿Qué diagramas son utilizados para el modelado conceptual de ciudades inteligentes?	Conocer los diagramas más utilizados para el modelado de las ciudades inteligentes.
PI4:	¿Qué tipos de investigación existen en los artículos?	Identificar los tipos de investigación de acuerdo con la taxonomía propuesta por Wieringa *et al.* [8].

La búsqueda de artículos de congresos y de revistas se realizó en las bibliotecas digitales *IEEE Xplore* y *ACM* por tratarse de bibliotecas de literatura técnica de ingeniería y tecnología de más alta calidad del mundo. El período de búsqueda incluyendo artículos de congresos y de revistas ha sido desde octubre del 2015 hasta marzo del año 2022. Se consideró como fecha de inicio para la búsqueda el año 2015, porque en este año se realizó la primera Conferencia Internacional sobre ciudades inteligentes [9].

Los términos principales que se tuvieron en cuenta para establecer la cadena de búsqueda son “smart city” y “conceptual modelling”, incluyendo términos alternativos la cadena de búsqueda definitiva es:

(("smart city" OR "smart cities") AND ("conceptual model" OR "conceptual modeling"))

Los criterios de inclusión y exclusión utilizados para el proceso de selección de artículos se presentan en la Tabla 2.

El proceso de selección de los estudios consistió en los siguientes pasos: 1) realizar la búsqueda en las fuentes definidas aplicando la cadena en el título y/o en el resumen, 2) eliminar los artículos duplicados, 3) aplicar los criterios de inclusión y exclusión y 4) aplicar los criterios de inclusión y exclusión al texto completo.

Para dar respuesta a cada una de las preguntas de investigación (PI) se definió un esquema de clasificación, que por restricciones de espacio se presenta en un apéndice en [10], junto con el formulario de extracción de datos. Se utiliza una síntesis temática basada en el esquema de clasificación que se representará a través de tablas.

Tabla 2. Criterios de inclusión y exclusión.

Criterios de inclusión.
I1. Dado el caso en que varios artículos de un mismo autor contemplen la misma investigación, se considerará el más completo y reciente.
I2. Artículos en idioma inglés.
I3. Artículos publicados entre octubre de 2015 y marzo de 2022.
I4. Artículos que contengan cadenas candidatas en el título, palabras clave y/o en el resumen.
Criterios de exclusión.
E1. Artículos cuya óptica sea ajena al ámbito de software.
E2. Literatura gris, tesis doctorales, presentaciones en PowerPoint.

3 Ejecución del SMS

En esta sección, se presenta la búsqueda realizada en las bibliotecas digitales, la selección de estudios primarios de acuerdo con lo definido en el protocolo de revisión del SMS. Se aplicó la cadena de búsqueda en las librerías con algunas adecuaciones necesarias en función de las particularidades de cada una que se encuentran en [10]. De un total de 289 artículos encontrados, se analizaron 48 estudios primarios. El listado de los estudios primarios analizados se presenta en [10].

4 Resultados del SMS

En la Tabla 3 se presenta una síntesis de los resultados del análisis de los estudios primarios sobre la base de lo establecido en el esquema de clasificación definido (Ver apéndice, Tabla 1) [10]. A continuación, se pretende dar respuesta a las preguntas de investigación en base al material recolectado.

Tabla 3. Síntesis de los resultados obtenidos.

ID	**Resultados por cada PI**			
	Contribución (PI1)	**Dominio (PI2)**	**Diagrama (PI3)**	**Tipo de Investigación (PI4)**
[EP1]	Modelo	Transporte	Diagrama de arquitectura, Pseudocódigo	Propuesta de solución
[EP2]	Modelo	Infraestructura	Diagrama de arquitectura, Pseudocódigo, Modelo de conocimiento, Modelo usando información geoespacial	Propuesta de solución
[EP3]	Métricas. Modelo	Infraestructura	Diagrama libre	Evaluación
[EP4]	Modelo	Medioambiente	Diagrama de Pila Tecnológica, Diagrama de flujo, Diagrama de arquitectura	Propuesta de solución

ID	Resultados por cada PI			
	Contribución (PI1)	**Dominio (PI2)**	**Diagrama (PI3)**	**Tipo de Investigación (PI4)**
[EP5]	Modelo	Infraestructura	Diagrama de arquitectura, Diagrama libre	Propuesta de solución
[EP6]	Modelo	Infraestructura	Diagrama de arquitectura, Pseudocódigo	Propuesta de solución
[EP7]	Modelo	Seguridad	Diagrama de arquitectura	Propuesta de solución
[EP8]	Modelo	Gobernanza	Diagrama de flujo	Evaluación
[EP9]	Modelo	Transporte	Diagrama de arquitectura, Diagrama de flujo de datos, Modelo usando información geoespacial	Propuesta de solución
[EP10]	Framework	Seguridad	Proceso de decisión de Markov, Cadena de Markov discreta, Pseudocódigo, Diagrama de arquitectura	Propuesta de solución
[EP11]	Framework	Gobernanza	Glosario, Diagrama libre	Propuesta de solución
[EP12]	Framework	Seguridad	Diagrama de Pila Tecnológica, Diagrama de arquitectura	Propuesta de solución
[EP13]	Framework	Seguridad	Diagrama libre	Propuesta de solución
[EP14]	Metodología, Framework	Tecnología	diagrama de comunicación, Diagrama de arquitectura, Diagrama libre,	Propuesta de solución
[EP15]	Modelo	Tecnología	Diagrama de arquitectura, Diagrama de Pila Tecnológica, Diagrama de flujo, Modelo usando información geoespacial	Propuesta de solución
[EP16]	Modelo	Transporte	Diagrama de bloque, Diagrama de flujo, Red de petri, Modelo usando información geoespacial	Propuesta de solución
[EP17]	Modelo, Framework	Infraestructura	Diagrama de Pila Tecnológica, Diagrama de flujo de datos, Modelo usando información geoespacial, Diagrama de arquitectura	Propuesta de solución
[EP18]	Herramienta, Modelo	Transporte	Pseudocódigo, diagrama de clases, Diagrama de arquitectura, Diagrama de flujo de datos	Propuesta de solución
[EP19]	Framework	Agua y desechos	Grafo, Pseudocódigo	Propuesta de solución
[EP20]	Modelo	Hábitat	Metamodelo, Diagrama de arquitectura, Diagrama de bloque	Propuesta de solución

ID	Resultados por cada PI			
	Contribución (PI1)	**Dominio (PI2)**	**Diagrama (PI3)**	**Tipo de Investigación (PI4)**
[EP21]	Modelo	Transporte	Glosario, Diagrama de ontología	Propuesta de solución
[EP22]	Modelo	Tecnología	Diagrama de arquitectura, Modelo usando información geoespacial	Propuesta de solución
[EP23]	Framework	Infraestructura	Diagrama libre	Evaluación
[EP24]	Ontología	Gobernanza	Diagrama libre, Diagrama de ontología	Propuesta de solución
[EP25]	Modelo	Tecnología	Diagrama de arquitectura, Diagrama de Pila Tecnológica, Diagrama de flujo de datos	Propuesta de solución
[EP26]	Modelo	Tecnología	Diagrama de arquitectura, Diagrama de Pila Tecnológica, Diagrama de flujo de datos	Propuesta de solución
[EP27]	Modelo	Tecnología	Diagrama de Pila Tecnológica	Propuesta de solución
[EP28]	Modelo	Tecnología	Diagrama de arquitectura, Diagrama libre	Propuesta de solución
[EP29]	Modelo	Seguridad	Diagrama de arquitectura, Diagrama de flujo de datos	Validación
[EP30]	Herramienta	Medioambiente	Diagrama de arquitectura, Modelo usando información geoespacial	Propuesta de solución
[EP31]	Modelo	Tecnología	Diagrama libre	Propuesta de solución
[EP32]	Modelo	Tecnología	Diagrama de Pila Tecnológica	Artículo filosófico
[EP33]	Framework	Tecnología	Diagrama de flujo, Pseudocódigo, Diagrama de arquitectura	Validación
[EP34]	Modelo	Tecnología	Diagrama de flujo de datos	Propuesta de solución
[EP35]	Modelo	Tecnología	Pseudocódigo, Diagrama de arquitectura, Caso de uso	Propuesta de solución
[EP36]	Herramienta	Tecnología	Diagrama de arquitectura	Propuesta de solución
[EP37]	Herramienta	Tecnología	Diagrama de clases	Evaluación
[EP38]	Modelo, Herramienta	Transporte	Red de Petri, Diagrama de flujo de datos	Propuesta de solución

ID	Resultados por cada PI			
	Contribución (PI1)	**Dominio (PI2)**	**Diagrama (PI3)**	**Tipo de Investigación (PI4)**
[EP39]	Framework	Tecnología	Diagrama de arquitectura, Diagrama de Pila Tecnológica, Diagrama de flujo de datos	Artículo filosófico
[EP40]	Framework	Transporte	Diagrama de Pila Tecnológica, Diagrama de flujo de datos, Modelo usando información geoespacial	Artículo filosófico
[EP41]	Modelo	Tecnología	Modelo de conocimiento	Propuesta de solución
[EP42]	Metamodelo	Tecnología	Diagrama de bloque	Artículo filosófico
[EP43]	Modelo	Tecnología	Diagrama de máquina de estados	Propuesta de solución
[EP44]	Framework	Personas	Diagrama de flujo, diagrama de clases, Caso de uso	Artículo filosófico
[EP45]	Modelo	Tecnología	Modelo usando información geoespacial, diagrama de clases, Diagrama de flujo de datos	Evaluación
[EP46]	Modelo	Tecnología	Diagrama de arquitectura, Diagrama de Pila Tecnológica, Diagrama de flujo de datos, Modelo usando información geoespacial	Propuesta de solución
[EP47]	Herramienta	Gobernanza	Modelo de conocimiento, Grafo	Evaluación
[EP48]	Framework	Infraestructura	Diagrama libre	Propuesta de solución

PI1 ¿Qué tipo de contribuciones existen en el modelado conceptual de Ciudades Inteligentes?

Pribyl *et al.* [EP9] basan su aporte en un acercamiento metodológico para describir y modelar los subsistemas de CI. Fang *et al.* [EP17] se valen de la literatura para poder contribuir con un diseño de un modelo a alto nivel, analizando arquitectura de datos y dominios clave de las CI. Qamar *et al.* [EP24] presentan una ontología a partir de la cual puede clasificarse un amplio espectro de servicios para aplicaciones para CI. Walletzký *et al.* [EP34] proponen una forma de modelar servicios que considera el contexto para poder cumplir con los requerimientos complejos que surgen en el entorno de las CI. La contribución de Muvuna *et al.* [EP37] se basa en la utilización del Systems Engineering Modelling Language para el modelado de las CI. El aporte de Zomer *et al.* [EP42] es un metamodelo para inclusión del comportamiento social y los datos en simulaciones conducidas en el contexto de una CI. Cunha *et al.* [EP44] sugiere a los sistemas sociotécnicos como una metáfora apropiada para modelar CI considerando la naturaleza social y técnica de la implementación de soluciones en la sociedad actual.

Se puede destacar el aporte de Qolomany *et al.* [EP6], Taherkordi y Eliassen [EP14] a partir del modelado de la arquitectura de un servicio que intercambia información de dispositivos IoT con la nube. Robberechts *et al.* [EP36] proponen utilizar una arquitectura llamada Edge to cloud as Service para modelar servicios de redes en TIC a gran escala en CI, utilizando dispositivos IoT. Soltvedt *et al.* [EP35] se basan en IoT para diseñar un modelo de costo para realizar descubrimiento de datos. Nakamura y Bousquet [EP43] se valen de IoT para proponer el modelado de la ejecución de servicios y la integración de ciclo de vida para CI, basándose en el concepto de ciudad como máquina de estados. La tecnología IoT es útil para proveer información sobre el uso y funcionamiento de la infraestructura. Sterbenz [EP12], Wang [EP32] y Patra [EP22], centran su aporte en modelar arquitecturas de IoT con particular énfasis en la obtención de información útil a distinto nivel para diversos dominios y en la resiliencia de las mismas. El aporte de Bhasin *et al.* [EP33] se vale de IoT para proponer un framework para la iluminación de CI. Latif *et al.* [EP19] postula otro para la gestión de cloacas inteligentes. Labib [EP30], aplica el mismo enfoque pero introduce los sistemas de información geográfica para el modelado de una aplicación para la gestión de recolección de residuos.

El transporte es un dominio en el que se cuentan numerosos aportes. Hariz *et al.*[EP1], Matyakubov y Rustamova [EP27], Boreiko y Teslyuk [EP16], [EP38], [EP46] desarrollan sus contribuciones a través de realizar un modelado del sistema de transporte público Abberley *et al.* [EP21], basa su aporte en la creación de una ontología para el análisis de congestiones vehiculares. A su vez Kuklová y Pribyl [EP40] basan el suyo en el diseño de una arquitectura para sistemas de control de tráfico vehicular. Zhou et al. [EP18], contribuyen modelando un sistema que pueda manipular una cantidad masiva de datos espacio temporales.

La seguridad es otro tópico abordado en las contribuciones. Pradhan *et al.* [EP7, EP29], Mohammad [EP10] y Wang *et al.* [EP13] proponen modelado de sistemas de defensa, tipos de amenazas y tipos de datos utilizados por entidades militares en CI.

Dentro de los aportes de interoperabilidad entre CI se encuentra el estudio de Hwang *et al.* [EP5] que consiste en el desarrollo de modelos de interconectividad entre las mismas, llamados Inter Working Models. A su vez, Pradhan *et al.* [EP7] tienen como objetivo en su contribución la interoperabilidad de datos, pero en el contexto de una operación militar. Por otro lado, el aporte de Zhao y Wang [EP41] consiste en el desarrollo de un modelo para facilitar la interoperabilidad de conocimiento entre dominios y ciudades a través del intercambio de datos y servicios.

PI2 ¿En qué dominios se realizaron contribuciones?

El enfoque de la mayoría de los estudios primarios es en el dominio de la tecnología, con veinte publicaciones (41.7%). Siete artículos (14.6%) están dedicados a infraestructura y otros siete a transporte. El resto de los estudios, se distribuyen cinco (10,4%) en seguridad, cuatro (8,3%) sobre gobernanza, finalmente personas, vivienda y agua y desechos cuentan con solamente una (2,1%) publicación en cada dominio.

PI3 ¿Qué diagramas son utilizados para el modelado conceptual de Ciudades Inteligentes?

Del total de los estudios analizados, 28 utilizan el lenguaje de modelado UML. El diagrama más empleado que pertenece a dicho lenguaje es el Diagrama de Arquitectura, observado en 25 publicaciones: Hariz *et al.* [EP1], Cabrera y Clarke [EP2], Bharadwaj *et al.* [EP4], Hwang *et al.* [EP5], Qolomany *et al.* [EP6], Pradhan *et al.* [EP7], Pribyl *et al.* [EP9], Mohammad [EP10], Sterbenz [EP12], Taherkordi y Eliassen [EP14], Sinaeepourfard *et al.* [EP15], Fang *et al.* [EP17], Zhou *et al.* [EP18], Lytra *et al.* [EP20], Patra [EP22], Pradhan *et al.* [EP25], Khan [EP26], Anindra *et al.* [EP28], Pradhan *et al.* [EP29], Labib [EP30], Bhasin *et al.* [EP33], Soltvedt *et al.* [EP35], Robberechts *et al.* [EP36], Guinko *et al.* [EP39], Boreiko y Teslyuk [EP46].

PI4 ¿Qué tipos de investigación existen en los artículos?

En relación con el tipo de investigación de los estudios primarios analizados se encontraron que treinta y cinco (72,9%) estudios primarios son propuesta de solución, seis artículos (12,6%) son evaluación en el contexto real, cinco (10,4%) son del tipo filosófico y dos (4,1%) corresponden a validaciones.

5 Amenazas a la validez

Se analizaron las potenciales amenazas a la validez que podrían afectar al SMS, respecto a las cuatro categorías sugeridas por Wohlin *et al.* [11].

- Validez del constructo. Se estableció de forma unívoca la definición de modelado conceptual [1], [2]. [3] y ciudad inteligente [4], [5] de acuerdo con artículos especializados con revisión de pares.
- Validez interna. Se diseñó un protocolo de revisión que ha sido revisado por la última autora, docente del seminario.
- Validez externa. Se tomó la decisión de utilizar dos bibliotecas digitales con amplio reconocimiento internacional (*IEEE Xplore* y *ACM*) para la búsqueda. No se consideró la literatura gris, así como artículos que no estuvieran disponibles de forma íntegra, presentaciones en PowerPoint, tesis doctorales o libros.
- Fiabilidad. La selección de publicaciones se realizó de acuerdo con criterios de inclusión y exclusión definidos en el protocolo de revisión. Para aumentar la confiabilidad, paralelamente los alumnos aplicaron los criterios por realizaron la catalogación de los estudios; se discutieron las discrepancias entre ellos y la docente, con el propósito de determinar si era apropiado incluir un artículo en particular o no, y de ese modo se obtuvo el listado final de estudios primarios. Además, se diseñó un formulario de extracción de los datos con Excel junto con un esquema de clasificación para responder a cada una de las preguntas de investigación.

6 Conclusiones y trabajos futuros

Del análisis de los 48 estudios primarios se concluye que:

- Dada la complejidad de las ciudades inteligentes, la mayoría de los aportes de los autores se enfocan en el modelado de un aspecto específico como, por ejemplo: tecnología, infraestructura o transporte.

- Las contribuciones se enfocan en ámbitos más estrictamente tecnológicos, como el modelado para soluciones de tecnologías de IoT, o similares, o el modelado para interoperabilidad entre CI.
- Otros dominios específicos destacables son el sistema de transporte o movilidad (14.6 % de estudios), la infraestructura (14.6 % de estudios) y la seguridad (10.4 % de estudios).
- Se observa que el más utilizado es el diagrama de arquitectura dado que el mismo permite el modelado de sistemas considerando la dimensión espacial de los distintos dispositivos ya sean de IoT, sensores, computadoras, servidores, etc.
- Cabe destacar que se descubrió el uso de otros lenguajes de modelado, como el Systems Engineering Modelling Language.
- La mayoría de los estudios (72.9%) proponen modelos terminados que corresponden a propuestas de solución.

Los futuros trabajos para desarrollar son: a) cubrir el área de vacancia de modelado de cada uno de los dominios de Ciudad Inteligente y b) realizar un metamodelo de CI comprensivo.

Referencias

1. Delcambre, L. M. L., Liddle, S. W., Pastor, O. y Storey, V. C.: Characterizing Conceptual Modeling Research. Pp. 40--57 (2019).
2. Mayr, H. C., Thalheim, B.: The triptych of conceptual modeling. Software and Systems Modeling, vol. 20, nº 1, pp. 7--24 (2021).
3. Guarino, N., Guizzardi, G., y Mylopoulos, J.: On the philosophical Foundations of Conceptual Models (2019).
4. Stübinger, J. y Schneider, L.: Understanding Smart City—A Data-Driven Literature Review. Sustainability, vol. 12, nº 20, p. 8460 (2020).
5. Wahab, N. S. N., Seow, T. W., Radzuan, I. S. M. y Mohamed, S.: A Systematic Literature Review on The Dimensions of Smart Cities. IOP Conference Series: Earth and Environmental Science, vol. 498, nº 1, p. 012087 (2020).
6. ER 2022: 41st International Conference on Conceptual Modeling, https://er2022web.github.io/ER2022/callForPapers.html.
7. Kitchenham, B. y Charters, S.: Guidelines for Performing Systematic Literature Reviews in Software Engineering. Citeseer (2007).
8. Wieringa, R., Maiden, N., Mead, N. y Rolland, C.: Requirements engineering paper classification and evaluation criteria: A proposal and a discussion. Requirements Engineering, 11, pp. 102–107 (2005).
9. IEEE International Smart Cities Conference (ISC2-2015), 25-28, in Guadalajara, Mexico (2015). https://site.ieee.org/isc2-2015/about.
10. Cerviño, J., Fernández, L., Gobbe, J. L., Panizzi, M.: Apéndice. Modelado Conceptual de Ciudades Inteligentes: Un Mapeo Sistemático de Literatura. (2022). Disponible en: https://doi.org/10.6084/m9.figshare.20324943.v1
11. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., y Wesslén, A.: Experimentation in software engineering: an introduction. The Kluwer International Series in Software Engineering (2000).

Integrabilidad y ecosistemas digitales: problemática, fundamentos y normalización

Patricia Bazán[1][0000-0001-6720-345X], Horacio Luz Clara[2][0000-0003-1534-8184], Jorge Luis Ceballos [3][0000-0003-4279-5094], Gustavo Giorgetti[4][0000-0003-4596-2900], Diego Felipe Ugalde [4-5][0000-0001-9568-4190] and Dante Adalberto Moreno[6][0000-0003-2064-5546]

[1] LINTI - Facultad de Informática - UNLP
[2] Facultad de Ingeniería - Universidad FASTA
[3] Facultad de Ingeniería y Tecnología Informática - Universidad de Belgrano
[4] ThinkNet SA
[5] Facultad Regional del Neuquén - Universidad Tecnológica Nacional
[6] Gobierno de La Pampa, Argentina. Coordinador de la Comisión de Infraestructura y Ciberseguridad en el Consejo Federal de la Función Pública, Argentina.

pbaz@info.unlp.edu.ar
hluzclara@ufasta.edu.ar
jorge.ceballos@comunidad.ub.edu.ar
ggiorgetti@thinknetgroup.com.ar
dugalde@thinknetgroup.com.ar
dmoreno@lapampa.gob.ar

Resumen. Uno de los mayores desafíos que presenta la utilización de diversas aplicaciones o sistemas informáticos a nivel multiorganizacional, es salvar la dificultad para integrarse con otros sistemas. La integrabilidad de los sistemas informáticos es la capacidad de un componente digital para interoperar con otros en entornos ecosistémicos. Un ecosistema digital de integrabilidad (EDI) genera un entorno informático en el cual conviven diversos sistemas y aplicaciones. Un EDI es una plataforma de intercambio de alta seguridad, basada en una arquitectura distribuida, altamente resistente a fallas, independiente de la tecnología, arquitectura y software con los que están desarrollados los sistemas que se interconectan. Este trabajo analiza la problemática de los ecosistemas digitales y presenta un proyecto de normalización que establezca las características y requisitos de un EDI.

Palabras clave: Integrabilidad, Ecosistema Digital, Normalización.

1 Introducción y motivación

La transformación digital tiene lugar gracias a la participación de factores facilitadores y habilitantes tecnológicos que es posible combinar sinérgicamente para arribar a modelos, usos y resultados innovadores. Una vez visualizadas las aplicaciones posibles y sus beneficios, la adopción de nuevas herramientas puede potenciarse con el desarrollo de estándares técnicos, políticas de incentivo y articulación, e instrumentos de inversión y financiamiento que acompañen en forma orgánica.

En estas páginas se presentan fundamentos para la comprensión de los ecosistemas digitales. Se define la integrabilidad, su relación y diferencias con la interoperabilidad y se explicita la aplicación de la integrabilidad a los ecosistemas digitales. Seguidamente, se proyecta el concepto de EDI sobre posibles ámbitos territoriales o jurisdiccionales. Se consideran los aspectos críticos para la elaboración de un estándar para los EDI (Ecosistemas Digitales de Integrabilidad), y para concluir, se reseña la visión y la estructura del proyecto de norma IRAM 17610 Ecosistema digital de integrabilidad, relativo a esta materia.

En este último sentido, la experiencia acumulada al cabo de catorce años en la implementación del ecosistema digital neuquino, a la que se ha sumado recientemente la de otras provincias argentinas que comienzan a transitar similar proceso, han brindado la oportunidad de capitalizar dicho conocimiento a través de una norma técnica, que tiene por objeto plasmar en una única referencia las bases conceptuales, las mejores prácticas que es posible adoptar y los requisitos que ineludiblemente debe cumplir un EDI para allanar su conformación, operación y evolución. Dicho estándar apunta a ofrecer un modelo que guíe la transformación hacia los *ecosistemas digitales,* a partir de crecientes niveles de integración de datos, procesos y servicios. Para ello se respetaron las pautas internacionales de normalización, pero además se preservó en todo momento el espíritu disruptivo de un pensamiento colectivo que busca un bien superior para el país y sus ciudadanos.

2 Conceptos y definiciones

El escenario en el que se ubica este proceso plantea la necesidad de una persona, humana o jurídica, de acceder a un servicio que involucra múltiples organizaciones públicas y privadas, cada una con sus propios requisitos y con distintos niveles de evolución tecnológica. En esta sección se presentan los conceptos subyacentes en dicho escenario.

2.1 **Ecosistema digital**

Un ecosistema se define como un entorno de agentes abierto, débilmente acoplado, agrupado en dominios, impulsado por la demanda y autoorganizado. donde cada especie es proactiva y sensible para su propio beneficio o ganancia [Chang, et al 2006].

Un ecosistema digital es un tipo generalizado de entorno informático ubicuo compuesto por especies ubicuas, geográficamente dispersas y heterogéneas [Dong et. Al, 2011] y los servicios publicados por estas especies reflejan las mismas características.

Los ecosistemas digitales se han convertido en el propósito político, económico y cultural de todos los países desarrollados del mundo moderno y permiten definir procesos interorganizacionales que simplifican la interacción de los productos que generan las organizaciones, soportando modelos de actividad a través del intercambio de información y la distribución y procesamiento de los datos comunes.

En todos los países desarrollados, se han implementado programas estatales para el desarrollo de ecosistemas digitales. Algunos de los acentos clave de estos programas son los siguientes: 1- apoyar el desarrollo de Internet como una red mundial y plataforma de comunicación, comercio e innovación para negocios [Davidson, 2016];

2- integración de sistemas de información estatales y corporativos y acceso comercial en línea a todo el volumen de datos [G20, 2016]; 3- crear una infraestructura segura para vivir y trabajar en línea, apoyando un nuevo nivel de calidad de los servicios en Internet [Cavanillas et al, 2016] y 4- garantizar una interacción más amplia de las personas con las máquinas y aceptación por parte de toda la sociedad de los principios morales, éticos y aspectos económicos de la digitalización.

Los ecosistemas digitales imitan los ecosistemas biológicos que se refieren a sistemas complejos e interdependientes; las infraestructuras de base de todos los constituyentes interactúan y exhiben comportamientos autoorganizables, escalables y sustentables [Li et al 2012].

Los ecosistemas digitales están formados por la interdependencia generada por la conectividad a través de los datos y se componen por ecosistemas de producción y consumo. Los productores se basan en la interdependencia asociada con la cadena de valor, que, si bien es una interdependencia tradicional, gana protagonismo debido a la conectividad de datos. Los consumidores, por otro lado, se generan por interdependencia entre entidades que completan los datos generados por el uso del producto. En resumen, los ecosistemas digitales desafían los procesos organizacionales para generar y utilizar datos. Los datos y su conectividad son, por lo tanto, un hilo común que atraviesa los ecosistemas digitales, ya sea por producción o consumo.

Un aspecto fundamental a revisar a nivel de las organizaciones se refiere a la digitalización de los ecosistemas y comprende un enfoque fuertemente sostenido por los datos más que por los productos en la creación de valor. En este sentido, resulta evidente la necesidad de establecer estándares normativos que soporten a los ecosistemas digitales, considerando tanto aspectos técnicos como organizacionales.

2.2 Integrabilidad e interoperabilidad

La diversidad de sistemas y aplicaciones y su necesidad de interconectarse para poder reutilizar información es lo que impulsa la interoperabilidad entre ellas, originándose el denominado “distanciamiento digital” cuando dicha interoperabilidad no se produce.

La complejidad del Estado y la existencia de sistemas informáticos heterogéneos en la Administración Pública han impulsado la creación de marcos que aborden los problemas de interoperabilidad, que de otro modo impedirían o dificultarían los procesos gubernamentales a nivel local, nacional o internacional. Se busca por esta vía impulsar el flujo de información entre áreas, organizaciones y jurisdicciones de gobierno, establecer estándares que contemplen el uso de productos estables y bien soportados, dar apoyo para el cumplimiento de los estándares y poder contar con una estrategia de largo plazo.

Los marcos adhieren a las tecnologías de Internet e incorporan metadatos para los recursos de información ofrecidos. La adopción de estándares técnicos y especificaciones abiertos, escalables y soportados por la comunidad, coadyuvan, en este contexto, a la interconexión, la integración de datos y el acceso a los servicios digitales.

Un marco clásico para la interoperabilidad de sistemas a gran escala lo constituye el Marco Europeo de Interoperabilidad (MEI) [Bruselas, 2017]. El MEI es un marco genérico para el desarrollo de un ecosistema de servicios públicos europeos que se propone: 1- inspirar servicios públicos integrados: digitales, transfronterizos y abiertos;

2 - guiar a las administraciones públicas a nivel nacional con miras a la interoperabilidad y 3- contribuir a la creación de un mercado digital único.

El propósito es propiciar un entorno de interoperabilidad coherente y facilitar la prestación de servicios que funcionen colaborativamente entre organizaciones y dominios. El MEI está conformado por un conjunto de principios y recomendaciones, un modelo de interoperabilidad estructurado en dimensiones (técnica, organizacional, semántica, legal, gobernanza de los servicios públicos integrados y gobernanza de la interoperabilidad) y un modelo para los servicios públicos integrados.

En línea con el MEI existente desde 2010, en 2021 la Comisión Europea presentó una Propuesta de Marco para la Interoperabilidad Europea para Ciudades y Comunidades Inteligentes. El proyecto busca enfrentar los desafíos del siglo XXI sentando las bases para un fácil intercambio de información entre diferentes plataformas, tecnologías y partes interesadas, para ofrecer mejores servicios al público, no sólo dentro de una ciudad, sino también atravesando dominios y jurisdicciones.

A pesar de los innegables beneficios que conlleva la interoperabilidad, en especial cuando se estructura a través de un marco apropiado, sería desacertado detenerse en ese punto, en lugar de explorar las oportunidades que es posible capitalizar tras haberla alcanzado a nivel interorganizacional e incluso hacia afuera de un ecosistema digital.

En este último sentido, una técnica innovadora es lograr reutilizar datos, complementar procesos organizacionales y mezclar ambos mundos (datos y procesos) para producir nuevos resultados. Esta capacidad de poder integrar/unificar servicios sobre la interoperabilidad da origen al concepto de integrabilidad.

Una buena analogía para el caso de la integrabilidad puede encontrarse en los teléfonos inteligentes que integran/unifican en un dispositivo las funciones de reloj, cámara, GPS, computadora, calculadora, entre otras, ninguna de las cuales fue una innovación de los actuales fabricantes de estos dispositivos, sino que simplemente las integraron.

Un ecosistema digital de integrabilidad (EDI) debe habilitar la innovación mucho más allá de la interoperabilidad y toda norma que la soporte debe ser creada para facilitar y potenciar la innovación.

2.3 Ecosistema Digital de Integrabilidad (EDI)

Un EDI es una comunidad de organizaciones miembros de un mismo ecosistema que: 1- respeta mínimas reglas de convivencia digital, 2- aplica estándares y componentes de software para poder utilizar y reutilizar los servicios comunes del ecosistema.

A partir de la existencia de un EDI, la prioridad para los sistemas legados, en su mayoría monolíticos y cerrados, es poder interoperar evitando la superposición y duplicación de datos y, para los nuevos desarrollos, explotar el nuevo entorno digital donde las aplicaciones desacopladas y los microservicios permiten alcanzar nuevos niveles en sus prestaciones.

La yuxtaposición de los conceptos antes señalados y su posible aplicación a un ámbito geográfico o jurisdiccional determinado, permite visualizar variantes al momento de pensar en su implementación. Resultan de peso, en cada caso, aspectos como la independencia regulatoria, administrativa, de recursos y tecnológica. Más puntualmente respecto de esto último, no pueden estar ausentes del análisis factores como la performance, la continuidad de servicio, la facilidad de gobierno, etc.

En el caso de grandes extensiones geográficas, como podría darse a nivel nacional o continental, los EDI pueden concebirse como federaciones de confianza, mediante las vinculaciones de los EDI de nivel provincial o nacional. Esto demanda, más allá de la compatibilización tecnológica, la armonización de las políticas de seguridad y de los acuerdos de nivel de servicio en forma transfronteriza. La recolección, intercambio y análisis de datos operativos de cada ecosistema resulta relevante. Por el momento, las federaciones de confianza podrían entenderse a partir de la vinculación directa de los ecosistemas parte, sin recurrir a la posibilidad de transitividad.

En el otro extremo, el del nivel local, en el cual podrían incluirse las ciudades y áreas con cualquier grado de urbanización, es necesario incorporar al análisis algunos elementos adicionales a la luz de la difusión de tecnologías IoT (Internet de las Cosas, por sus siglas en inglés Internet of Things) y su aplicación intensiva a la problemática de dichos territorios bajo la forma de soluciones inteligentes.

En los últimos años, las soluciones orientadas a ciudades inteligentes han intentado resolver temáticas específicas de un dominio, ya se trate del medio ambiente, la provisión de un servicio particular, el tránsito, la seguridad, etc. De esta manera, a través del mundo se han multiplicado los desarrollos y las experiencias aplicables a sectores verticales. Considerando que dichos proyectos pueden insumir recursos públicos y que las inversiones deben recuperarse más allá del corto plazo, resulta natural pensar en que la información contenida en un área de problema pueda fluir hacia otra, para poder conseguir soluciones transversales que apalanquen la infraestructura, hardware y software existentes y eviten la formación de silos o islas de información.

De lo anterior se deduce que la conformación de los EDI en el caso de las ciudades puede suponer la hibridación de un enfoque distribuido con uno centralizado. Dicha combinación habilita la utilización de plataformas urbanas de integración horizontal, en donde los sistemas IoT pueden ser conectados y en donde la ingesta masiva de datos puede ser acopiada, preparada, agregada y sintetizada adecuadamente. La comunicación hacia el exterior del ecosistema urbano (con otras ciudades o unidades administrativas de orden superior) puede realizarse a través de compuertas que proporcionan la seguridad necesaria.

Por último, el ámbito regional o provincial proporciona un escenario típico de EDI, que puede vertebrarse a partir de la vinculación de registros base (propiedad inmueble, propiedad automotor, registro de las personas, etc.) para luego incorporar como miembros a una diversidad de organizaciones tanto públicas como privadas. La suma de participantes y sus repositorios de datos debe alentarse, en vistas a crear una masa crítica que multiplique las oportunidades para la creación de valor y la maximización de la calidad.

En todos los casos la existencia de un marco legal estable y abarcativo resulta deseable, así como el intercambio de buenas prácticas, y la existencia de estándares técnicos y operativos. Similarmente, el reconocimiento mutuo de identidades digitales, la adopción de ontologías comunes y la simplificación de procesos favorecen el desarrollo de los EDI.

2.4 El EDI a nivel gubernamental

La gestión de gobierno en la era digital ha ido evolucionando de la mano de la tecnología. Se observan al menos dos claros caminos de aprendizaje: 1- el que

llamaremos “modelo tradicional de transformación digital del Estado” y 2- el “modelo ecosistémico evolutivo” como, por ejemplo, el seguido por Estonia [E-Estonia, s.f.].

El modelo tradicional es un abordaje de adentro hacia afuera, desde el Estado hacia la ciudadanía, en el cual, recién al final se reconoce y asume la necesidad y la importancia de mejorar la interoperabilidad, una de las características clave de los ecosistemas digitales. Sigue la mirada del proceso administrativo convencional, adaptado a la era tecnológica. Comienza por la reingeniería interna de procesos, tiene en cuenta el enfoque de calidad y mejora continua, y redunda en el desarrollo de aplicaciones y portales de internet independientes y no integrados. En general resuelve los problemas de la administración pública y no se enfoca adecuadamente en la ciudadanía (se piensa hacia el ciudadano y no desde el ciudadano).

Este enfoque transita las siguientes fases: a- desde el “paradigma expediente/papel”, en el cual los trámites son presenciales; b- luego incorpora la tecnología con el “paradigma de gobierno electrónico”, donde se digitaliza la gestión administrativa y los expedientes; c- posteriormente asume la forma de “gobierno digital/gobierno abierto”, en el cual los trámites se realizan vía portales web o aplicaciones móviles, y la mirada ciudadana se enfoca desde los principios del gobierno abierto [Naser, 2017]: participación, transparencia, rendición de cuentas e innovación; d- finalmente, aborda la interoperabilidad como una instancia necesaria para la construcción de un Estado digital integrado, intentando solucionar los problemas devenidos de la diversidad institucional/tecnológica entre los actores, que debe ser articulada e interconectada para lograr impacto en la ciudadanía.

El modelo ecosistémico evolutivo es diferente, y resulta de un enfoque holístico y sistémico digital desde el comienzo, diseñado desde el ciudadano, desde sus eventos de vida,. Prioriza la diversidad existente y la interconexión de todos los actores en el estado de madurez tecnológica en el cual se encuentren, fortaleciendo la interoperabilidad por sobre la reingeniería de procesos. Esto implica un cambio en los modelos de comunicación de la ciudadanía con el Estado, desarrollando aplicaciones unificadas sobre la base del principio “una sola vez”, y recién al final se plantea con precisión, los ajustes que requieren los procesos internos de cada institución para poder operar mejor dentro del ecosistema

Las etapas del modelo ecosistémico evolutivo son: a - Gobierno 1.0, donde los trámites son presenciales y en papel; b -Gobierno 2.0, donde los trámites son vía portales web o aplicaciones móviles; c- Gobierno 3.0 (gobierno invisible), donde el objetivo es una gestión digital desde el enfoque ciudadano, atendiendo sus eventos de vida y su tiempo, sobre la base del principio “una sola vez”, recurriendo a un uso masivo de la interoperabilidad; y d- Gobierno 4.0 (gobierno inteligente), donde se minimiza la burocracia y se recurre a “agentes digitales” (bots en inglés) y “sirvientes de inteligencia artificial (IA)” para asistir a los ciudadanos en su relación con el Estado.

El modelo ecosistémico evolutivo no se guía por los avances tecnológicos, sino por una profunda comprensión de los ecosistemas digitales, su operación y las variables que los gobiernan. También supone un cambio de paradigma y la creación de principios que, dando soporte cultural, actúan como brújulas del proceso de transformación digital, establecen prioridades y criterios de decisión. Además, logran tanto un uso muy eficaz de la tecnología existente, así como un rápido aprovechamiento de las nuevas tecnologías que, por su propia evolución, van emergiendo.

3 Proyecto de norma argentina IRAM 17610 de Ecosistema Digital de Integrabilidad (EDI)

Una norma es un documento técnico establecido por consenso y aprobado por un organismo reconocido que proporciona, para usos comunes y repetidos, reglas, directrices o características para las actividades o sus resultados, a fin de garantizar un nivel óptimo de orden en un contexto dado.

El proyecto de norma argentina IRAM 17610 Ecosistema digital de integrabilidad. Parte 1 – Requisitos, está desarrollado con la participación de más de 40 profesionales de distintas disciplinas, representando a distintos sectores de la función pública, de la actividad privada y de la academia.

Este modelo participativo está fundamentado en la necesidad de aportar un equilibrio entre el espíritu normalizador y la necesaria libertad que habilite pluralidad, creatividad e innovación en el cumplimiento de los requisitos. De tal modo, es posible mitigar algunos de los siguientes riesgos que pueden presentarse al abordar la redacción de una norma:

Exceso de detalle - Ocurre en ocasiones que una determinada materia es abordada a un nivel de detalle excesivo, lo cual resulta en proliferación de estándares, y/o textos extensos, complejos y poco entendibles. Esto impacta en la facilidad de cumplimiento, el tiempo y costo insumido en procesos de consultoría y certificación e incluso, en la extensión de su adopción.

Falta de representatividad - La representatividad de los actores respecto de la totalidad de un sector, así como la fundamentación en mejores prácticas, pueden quedar en entredicho si estos aspectos no son cuidadosamente monitoreados por el organismo normalizador que actúa como anfitrión. En este sentido, no resultaría deseable que un proyecto sea impulsado exclusivamente por un único actor o un puñado de actores dominantes en detrimento de voces más débiles que podrían quedar así apagadas. Similarmente, la incorporación de prácticas que no han sido refrendadas ampliamente en la experiencia de campo como las más recomendables, puede inducir en la cultura organizacional vicios sistemáticos luego difíciles de erradicar.

Falta de claridad - La redacción concreta en que finalmente quede expresado un estándar no es menos importante. Así, los textos que no usan expresiones claras abren espacio a subjetividades y ambigüedades. La formulación elegida podría resultar poco comprensible para quienes no hayan formado parte directamente de su desarrollo, con el corolario de no responder en última instancia a las necesidades de las partes.

Los principios perseguidos para la definición de la norma son: 1- basado en la experiencia y elaborado a partir de las necesidades de la actividad, 2- disponible al público y 3- dirigida a la promoción de beneficios óptimos para la comunidad.

El objeto de la norma es establecer de manera clara y consistente las características de un EDI, así como los requisitos que debe cumplir cualquier producto de software o aplicación informática, en relación a su capacidad de integrarse de manera abierta y segura a ese ecosistema.

3.1 Antecedentes y partes de la norma

La presente norma toma como antecedentes al Referencial IRAM No. 14. Requisitos de calidad de las aplicaciones informáticas - INTEGRABILIDAD [Neuquén, 2014] y el enfoque definido en el MEI.

La norma define el conjunto de principios, recomendaciones y requisitos que orientan los esfuerzos políticos y legales, organizacionales, semánticos y técnicos de las organizaciones miembro del EDI, con el fin de facilitar el intercambio de información.

La norma aplica a cualquier EDI, a cualquier componente de software que quiera integrarse con un EDI y también puede utilizarse en casos como: 1- interoperabilidad interorganizacional en general, 2- interoperabilidad interna de una organización, entre distintos sistemas o tecnologías, y 3- interoperabilidad dentro de un grupo de organizaciones del mismo sector o cluster (ecosistema sectorial), por ejemplo: salud, energía, educación, bancario, financiero, entre otros.

Los puntos mencionados acerca de la interoperabilidad inter e intraorganizacional, así como la que se puede producir dentro de un mismo sector, extienden el concepto de interoperabilidad, hacia la capacidad de integrarse (integrabilidad), dada por la capacidad de los miembros de definir colaborativamente procesos de negocio que articulen a distintas organizaciones miembro y sus respectivos subsistemas. Esto permite crear servicios que anticipen o eliminen la necesidad de interacción de los destinatarios del servicio con varios sistemas de las diversas organizaciones involucradas.

Las premisas que delimitan el campo de aplicación de la norma son: 1 - la interacción es entre componentes de software, 2- no se consideran interfaces entre aplicaciones y personas y 3- los beneficiarios finales de la existencia de un EDI son las personas (humanas o jurídicas).

3.2 Metodología de desarrollo y estado de avance actual

El proceso formal se inicia con la presentación ante IRAM de una solicitud de estudio de norma, refrendada por un núcleo inicial de partes interesadas, que acuerdan un conjunto básico de necesidades a satisfacer. Aprobada dicha solicitud por el organismo de estudio pertinente, el proyecto es incorporado a su cronograma y plan de estudio de normas. Los actores inician entonces la recopilación de experiencias y antecedentes normativos, a nivel tanto nacional como internacional, que resultan relevantes para la materia abordada. En función de ello, se esboza la estructura del proyecto completo, resultando en una norma multiparte que cubra requisitos alineados con las necesidades establecidas previamente, un método validado para el ensayo de los requisitos, y un proceso de evaluación de la conformidad conducente a acreditar el cumplimiento de los requisitos. Con esta visión, el equipo de desarrollo se ha abocado a la obtención de un texto normativo para la primera parte, sentando primeramente el marco conceptual y luego un conjunto de requisitos y recomendaciones. Concluido el cuerpo normativo principal, se añaden varios textos informativos bajo la forma de anexos que amplían, ejemplifican y buscan facilitar el cumplimiento de lo prescripto. El proceso continúa con la edición de la totalidad de dicha primera parte, para arribar a una versión que se somete a discusión pública, instancia próxima a producirse, en la cual se recogen observaciones a tener en cuenta con anterioridad a la publicación del texto definitivo. Se prevé, asimismo que, tras dicha publicación, el trabajo continúe hasta completar la visión global del proyecto, según se ha comentado.

4 Conclusiones y trabajos futuros

El vertiginoso ritmo del desarrollo tecnológico y de transformación de las sociedades ofrece oportunidades y desafíos para la transformación digital. Un tratamiento comprensivo y ordenado asegura el éxito en tan importante cometido y proporciona bases adecuadas para cualquier evolución futura. Más allá de tratarse de una construcción transversal, que involucra potencialmente tanto a actores públicos como privados, el Estado ofrece siempre uno de los entornos más complejos para cualquier proceso de modernización. En dicho contexto, dar los pasos correctos y en tiempo oportuno se vuelve especialmente crítico para mantener la capacidad de respuesta a los problemas de la población, maximizar el rendimiento de la inversión y simplificar la vida de los beneficiarios de sus servicios. El sector público y el sector privado pueden intercambiar experiencias en la transformación digital y apalancar sus logros articulándolos.

La interconexión e interoperabilidad de sistemas resulta imprescindible para aumentar la calidad de servicio y agilizar la gestión. Dicha base proporciona la ocasión para la aparición de marcos de interoperabilidad, habilitando así el desarrollo de ecosistemas digitales. Adicionalmente, la aplicación a estos últimos del concepto de integrabilidad, conduce a un creciente intercambio de datos, coordinación de procesos y liberación de servicios, y subsiguientemente a la dinamización de la industria del conocimiento, una de las más importantes en nuestros días para el robustecimiento de cualquier economía.

Un abordaje orgánico de los ecosistemas digitales de integrabilidad como pieza clave de la transformación digital no puede omitir el desarrollo de estándares técnicos, surgidos de un consenso amplio entre actores relevantes y experimentados en la temática. La extracción de mejores prácticas surgidas de distintos precedentes, tanto a nivel nacional como internacional, otorgan certidumbre y afianzan la adopción de reglas comunes para el involucramiento de nuevos interesados. El proyecto de norma argentina IRAM 17610 Ecosistema digital de integrabilidad. Parte 1 – Requisitos encarna hoy esta línea de pensamiento, ubicándose a la vanguardia de las iniciativas de normalización para la transformación digital.

Desde el punto de vista tecnológico, la aparición de nuevos conceptos y herramientas hacen continuo el trabajo de investigación, con el propósito de obtener la mejor combinación de elementos técnicos para conformar ecosistemas con características de misión crítica. Dado este carácter, una proyección estratégica de la problemática puede comprender programas de financiamiento, investigación y transferencia de tecnología, más allá de los esfuerzos normativos, tanto a nivel político como jurídico y técnico. En cualquier caso, un avance equilibrado sobre dichos frentes, que no omita la consideración del largo plazo, proporcionará siempre los mejores resultados.

Referencias

1. Cavanillas, Jos Mara, Curry, Edward, Wahlster, Wolfgang (Eds.) (2016): New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data in Europe. Springer 2016, 303 pp. [Online]. Disponible: http://www.wolfgang-wahlster.de/wordpress/wp-

content/uploads/Industrie_4_0_Mit_dem_Internet_der_Dinge_auf_dem_Weg_zur_vierten_industriellen_Revolution_2.pdf
2. Davidson A. (2016). Commerce Department Digital Economy Agenda 2016, The Digital Economy: Key to Prosperity and Competitiveness. [Online]. Disponible: https://www.ntia.doc.gov/files/ntia/publications/alan_davidson_digital_economy_agenda_deba_presensation_051616.pdf
3. G20. (2016). Digital Economy Development and Cooperation Initiative. Hangzhou Summit. [Online]. Recuperado de: http://en.kremlin.ru/supplement/5111
4. Li, W., Badr, Y., & Biennier, F. (2012, October). Digital ecosystems: challenges and prospects. In proceedings of the international conference on management of Emergent Digital EcoSystems (pp. 117-122).
5. Vorobieva, D., Kefeli, I., Kolbanev, M., & Shamin, A. (2018, November). Architecture of digital economy. In 2018 10th
6. Dong, H., Hussain, F. K., & Chang, E. (2011). A framework for discovering and classifying ubiquitous services in digital health ecosystems. Journal of Computer and System Sciences, 77(4), 687-704. [Online]. Recuperado de: https://www.sciencedirect.com/science/article/pii/S0022000010000231 (Agosto 2022)
7. Bruselas, 2017. Marco Europeo de Interoperabilidad – Estrategia de aplicación. Recuperado de: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52017DC0134&from=LT (Agosto 2022)
8. e- Stonia s.f., We have built a digital society and we can show you how. Recuperado de: https://e-estonia.com/. (Agosto 2022)
9. Naser A, Ramirez Aluja N (2017). Plan de gobierno abierto: una hoja de ruta para los gobiernos de la región. Recuperado de: https://repositorio.cepal.org/handle/11362/36665 (Agosto 2022)
10. Bulao, Jacquelyn (2022). How Much Data Is Created Every Day in 2022?, Techjury.net. Recuperado de: https://techjury.net/blog/how-much-data-is-created-every-day/Di (Agosto 2022)
11. European Commission (2021). Final Study Report - Proposal for a European Interoperability Framework for Smart Cities and Communities (EIF4SCC), Publications Office of the European Union. Recuperado de: https://digital-strategy.ec.europa.eu/en/news/proposal-european-interoperability-framework-smart-cities-and-communities-eif4scc (Agosto 2022)
12. European Commission (2017). New European Interoperability Framework - Promoting seamless services and data flows for European public administrations, Publications Office of the European Union. Recuperado de: https://www.bvkb.gov.lv/sites/bvkb/files/eif_brochure_final1.pdf (Agosto 2022)
13. Givaudant, E., Luz Clara, H., Todorovich, E., (2020). Análisis exploratorio de plataformas para ciudades inteligentes, Universidad FASTA
14. Gobierno de España (2020). Plan España Digital 2025. Recuperado en: https://avancedigital.mineco.gob.es/programas-avance-digital/Documents/EspanaDigital_2025_TransicionDigital.pdf (Agosto 2022).
15. Sirviö, Ville (2022). From connectivity between databases towards an ecosystem of ecosystems, Nordic Institute for Interoperability Solutions. Recuperado de: https://www.niis.org/blog/2022/7/11/from-connectivity-between-databases-towards-an-ecosystem-of-ecosystems. (Agosto 2022).
16. Neuquén, 2014. Referencial IRAM N° 14 “Requisitos de Calidad de las Aplicaciones Informáticas – Integrabilidad” -I y II -. Recuperado de: https://silo.tips/search/Referencial+IRAM+N%C2%BA+14-1. (Setiembre de 2022)

Desarrollo de Interfaces de Programación de Aplicaciones aplicadas en Experticia, un Sistema Experto Jurídico

Osvaldo Spositto[1], Luis Busnelli[2], Viviana Ledesma[1], Gastón Procopio[1], Cecilia Gargano[1], Julio Bossero[1], Gerardo Frega[1], Victoria Saizar[1], Fabio Quintana[1], Laura Conti[2], Sergio García[3], Carlos Colombain[1] y Gustavo Pérez Villar[4]

[1] Universidad Nacional de La Matanza. Departamento de Ingeniería e Investigación Tecnológicas. Florencio Varela 1903. San Justo. La Matanza. {spositto, vledesma, gprocopio, cgargano, jbossero, gfrega, vsaizar, ccolombain}@unlam.edu.ar

[2] Universidad Nacional de La Matanza. Departamento Derecho y Ciencia Política. {lbusnelli, lconti}@unlam.edu.ar

[3] Palacio de Tribunales. Departamento Judicial de Morón. Alte. Brown. Piso 4. Morón. sergiogabriel.garcia@pjba.gov.ar

[4] Subsecretaría de Tecnología Informática del Poder Judicial de la Provincia de Buenos Aires. Palacio de Justicia, avenida 13 entre 47 y 48, primer piso (La Plata). Argentina. gperez@scba.gov.ar.

Abstract. Construir Sistemas Expertos es intentar capturar la experiencia de personas idóneas en un tema e incorporarla en programas de computación. Esta tarea se basa en averiguar de expertos lo que saben y cómo utilizan su conocimiento para resolver problemas. El derecho y el razonamiento jurídico son uno de los nuevos objetivos para los sistemas de Inteligencia Artificial. Experticia, es un prototipo de Sistemas Expertos jurídico, que ayuda a mejorar la resolución de ciertos trámites legales, optimizando los tiempos y colaborando con el trabajo de los funcionarios. Este documento propone la utilización de Interfaces de Programación de Aplicaciones para acceder a datos en formato datos estructurados, a través de un servicio web. En primer lugar, se describen las tecnologías utilizadas. Luego se realiza un estado del arte sobre la tecnología de servicios web REST. Por último, se describe su especificación y diseño. Además, se explican los detalles de implementación y las pruebas realizadas. Los resultados indican la factibilidad de incorporar esta tecnología en la nueva versión del sistema.

Keywords: Inteligencia Artificial**,** Sistema Experto, Interfaces de Programación de Aplicaciones REST, REACT.

1 Introducción

En el año 2020 la Universidad Nacional de La Matanza (UNLaM), a través de investigadores de dos departamentos, el de Ingeniería e Investigaciones Tecnológicas y el de Derecho y Ciencias Políticas, con una estrecha colaboración del Juzgado de

Ejecución Nº 2 del Departamento Judicial Morón, presentó el proyecto PROINCE[1] C236/PII titulado "*Diseño e Implementación de un Sistema Experto como Apoyo al Proceso de Despacho de Trámites de un Organismo Judicial*", el objetivo del mismo fue la construcción del prototipo denominado Experticia. Se trata de un Sistema Experto (SE) cuyo objetivo es la sistematización y optimización de varios de los procesos judiciales que actualmente se realizan en forma manual o semiautomática en el Poder Judicial de la Provincia de Buenos Aires. Dada la importancia del proyecto, en ese mismo año, la Suprema Corte de Justicia de la Provincia de Buenos Aires firmó un convenio de Colaboración Recíproca con la UNLaM, para el desarrollo de Experticia. Se ha estado trabajando en forma interdisciplinaria, el equipo de investigación mencionado previamente junto a especialistas del área jurídica provincial, y técnicos de la Suprema Corte de Justicia de la Provincia de Bueno Aires (SCBA).

Básicamente Experticia permite tomar la experiencia de los "expertos en la justicia" para construir una base de conocimientos, con modelos estandarizados, que luego desde la interfaz de usuario que provee el sistema, pueden ser aplicados por los operadores en los distintos organismos judiciales (Ver Figura 1).

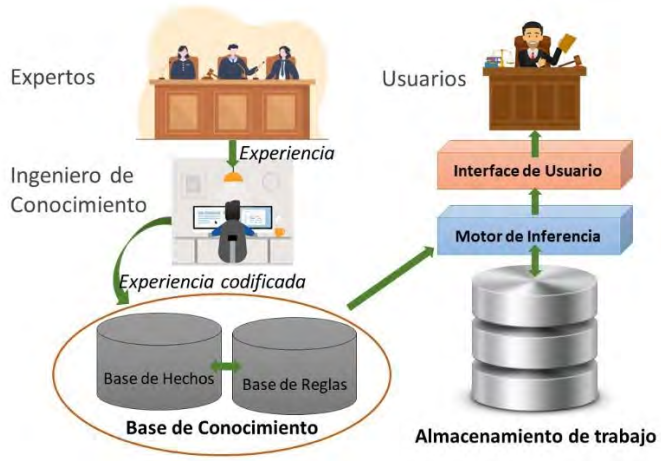


Fig. 1. Estructura básica de un SE. Fuente: Spositto y otros en [5].

Para la etapa inicial se construyó un prototipo, una versión de escritorio o desktop, cuya funcionalidad pretende brindar soporte a los operadores de la justicia en la toma de decisiones para la resolución de una causa, en particular las relacionadas al fuero penal. Entre los potenciales beneficios de Experticia se puede mencionar que permite estandarizar distintos procesos de despacho de trámites, agilizar y reducir tiempos de carga, además de minimizar errores, tanto durante la toma de decisiones, como en el ingreso de datos. También, se comprobó la eficiencia en la capacitación de nuevos agentes. Descripciones de los avances del desarrollo y de las pruebas realizadas han sido descritos y publicados en diversos trabajos [1-5].

[1] Programa de Incentivos para Docentes Investigadores de la Secretaría de Políticas Universitarias, implementado por la Secretaría de Políticas Universitarias del Ministerio de Educación de la Nación.

Algorítmicamente Experticia, se basa en la teoría de la decisión [6], usando árboles de decisión. Estos se constituyen en una serie de decisiones o condiciones organizadas en forma jerárquica [7]. Los primeros procesos, que se desarrollaron, fueron los de pedido de libertad condicional, en [2], se puede observar, la forma en que el proceso se descompone en varios pasos, hasta llegar su resolución. En [3] y [4] se explica, además, cómo se pueden utilizar estos datos con distintos algoritmos de Minería de Datos. Como se comentó anteriormente, inicialmente se implementó, en una versión de escritorio o desktop. En esta etapa de prueba, ha brindado resultados altamente satisfactorios, en el Juzgado de Ejecución Penal Nro. 2 de Morón [5].

Experticia se integra, de forma asincrónica, con el Sistema Informático de Gestión Asistida Multi-fuero y Multi-Instancia, más conocido como Augusta[2], desarrollado por el Departamento de Desarrollo Informático dependiente de la Subsecretaría de Tecnología Informática de la SCBA. Cabe aclarar que, este sistema, se utiliza en todos los juzgados de la Provincia de Buenos Aires.

En este procesamiento, los datos que hacen a la información propia de la causa (en el contexto de Experticia se han denominado "datos esenciales"), se toman de Augusta en forma asincrónica y manualmente, y luego de completar el proceso, se vuelven a actualizar los datos en Augusta, de la misma manera. A partir de los resultados obtenidos en la experimentación, se presentó este año un nuevo proyecto PROINCE bajo el título *"Inteligencia Artificial Jurídica: la Evolución de Experticia hacia un Modelo de Justicia Predictiva"*. En este proyecto, se propone realizar una proyección, del sistema actual, a un sistema web. El nuevo sistema, contará con los mismos módulos, correspondientes a la gestión de la resolución de los despachos asociados a las causas. Uno de los principales cambios propuestos, es que Experticia se comunique o interactúe con el Sistema Augusta en forma sincrónica. De este modo podrá tomar directamente los datos esenciales de una causa, incluso retornar y guardar información en Augusta cuando sea necesario [4]. Siguiendo tal objetivo, se propuso la arquitectura web que se presenta en la Figura 2.

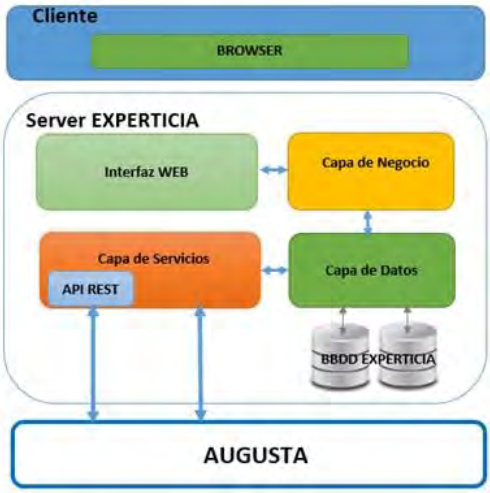


Fig. 2. Estructura del SE Web.

[2] https://www.scba.gov.ar/paginas.asp?id=39889

Como se puede observar, se presenta una arquitectura de software para aplicaciones que requieren comunicación en tiempo real y bidireccional entre servidor y clientes.

Dentro de los distintos paradigmas de programación, uno de los más utilizadas, es la programación por capas, que consiste en dividir el código fuente según su funcionalidad principal [8]. En esta arquitectura, las capas indican la separación lógica de los componentes. El principal beneficio de la arquitectura de tres niveles es que, debido a que cada nivel se ejecuta en su propia infraestructura, cada nivel puede ser desarrollado simultáneamente por un equipo de desarrollo independiente y puede actualizarse o escalarse según sea necesario sin afectar a los otros niveles. Los componentes de cada capa se comunican con los componentes de otra capa mediante interfaces bien definidas [9].

En este trabajo, se aporta un cambio sustancial, en la aplicación Experticia usando esta tecnología, se ha implementado una interfaz web para utilizar una API [10 y 11], que permita el intercambio de información entre Augusta y Experticia.

2. Trabajos relacionados

A partir de una revisión bibliográfica se encontraron varios trabajos relacionados con la Inteligencia Artificial (IA) aplicados a la justicia o algunos procesos que involucran jueces o fallos judiciales. En los trabajos presentados por este equipo de investigación se nombran y describen los mismos [1-5]. A continuación, se citan trabajo de otros autores en relación con la IA y el poder judicial:

- *Corvaláni, J. (2017). Inteligencia artificial: retos, desafíos y oportunidades – Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la Justicia. Revista de Investigações Constitucionais. ISSN 2359-5639 DOI: 10.5380/rinc.v5i1.55334.*

  Resumen: En el año 2019 se presenta Prometea, como el primer sistema de inteligencia artificial predictivo de América Latina, creado en el Ministerio Público Fiscal de la Ciudad de Buenos Aires y actualmente aplicado a la justicia y la administración pública. Entre algunas de sus cualidades podemos decir que predice la solución de un caso judicial en menos de 20 segundos, con una tasa de acierto del 96%. En solo 45 días elabora 1000 dictámenes jurídicos en expedientes relativos al derecho a la vivienda. Sin Prometea el tiempo empleado para la obtención de estos resultados es de 174 días.

- *H Weslei Gomes de Sousa & otros. (2018). Artificial intelligence and speedy trial in the judiciary: Myth, reality or need? A case study in the Brazilian Supreme Court (STF),Government Information Quarterly,Volume 39, Issue 1,2022,101660,ISSN 0740-624X, https://doi.org/10.1016/j.giq.2021.101660.*

  Resumen: Como su título lo dice, es un estudio de caso en el Supremo Tribunal Federal de Brasil (STF) realizado por la Universidad de Brasilia (UnB). El sistema judicial brasileño recibe una cantidad extremadamente alta de casos de demanda todos los días. Estos casos deben analizarse para asociarlos a etiquetas relevantes y asignarlos al equipo adecuado. La mayoría de los casos llegan al tribunal como

archivos PDF únicos que contienen varios documentos. Uno de los primeros pasos para el análisis es clasificar estos documentos. El Sistema que se denomina Victor, es desarrollo utilizando el Procesamiento del Lenguaje Natural (NLP) y emplea un algoritmo de aprendizaje automático del tipo supervisado para la automatización del análisis textual de juicios. Esto es a través de Redes Neuronales Artificiales (ANN).

- *Cinara Rocha, C & Carvalho J. (2022). Artificial Intelligence in the Judiciary: Uses and Threats. Proceedings of Ongoing Research, Practitioners, Workshops, Posters, and Projects of the International Conference EGOV- eDEM-ePart 2022. Disponible en*: https://dgsociety.org/wp-content/uploads/2022/09/CEUR-proceedings-2022.pdf#page=197

  Resumen: Este es un estudio sobre el uso de la IA en el poder judicial. Vincula al volumen creciente de información digital resultante de los procedimientos legales en la mayoría de los países y al uso de la IA parea ayudar a resolver problemas crónicos en organizaciones relacionadas con la justicia, como procesos de justicia lentos y altos costos operativos. Hace una minucioso detalle de la literatura relacionada con los usos de la IA en el Poder Judicial, y menciona ocho categorías de los análisis de contenido considerando el tipo de aplicaciones y funcionalidades. Este artículo presenta y analiza las aplicaciones de la IA en apoyo del trabajo de los jueces y las principales amenazas a los valores de la justicia que plantea su uso en los tribunales.

Otros trabajos en cuanto a temas vinculados con las Interfaces de Programación de Aplicaciones, en idioma español, no se encontraron trabajos. En lengua extranjeras se pueden mencionar las siguientes experimentaciones:

- *Seeam, Preetila & Teckchandani, Nishant & Booneyad, Hansha & Torul, V. & Seeam, Amar. (2018). Employment Law Expert System. 1-6. 10.1109/ICONIC.2018.8601271.*

  Resumen: Presenta un Sistema Experto para ayudar a la población de Mauritania con las consultas que puedan tener sobre la legislación laboral. El sistema utiliza técnicas de aprendizaje automático, reconocimiento/síntesis de voz y procesamiento de lenguaje natural para conversar con los usuarios a través de una interfaz web. Se implementó en HTML5, CSS3 y JavaScript para crear la aplicación de front-end, y se creó un servicio web REST API para que las consultas de los usuarios y las respuestas generadas desde el motor de inferencia.

- *Behzadidoost, R., Hasheminezhad, M., Farshi, M. et al. A framework for text mining on Twitter: a case study on joint comprehensive plan of action (JCPOA)-between 2015 and 2019. Qual Quant (2021). https://doi.org/10.1007/s11135-021-01239-y*

  Resumen: Se trata de un Sistema Experto basado en reglas que utiliza el concepto de huella dactilar en las ciencias judiciales. El sistema toma una huella digital de los tweets de un tema emergente. Para detectar los tweets no etiquetados del tema, utiliza API REST.

3. Consideraciones acerca de las Interfaces de Programación de Aplicaciones

En la programación por capas [8], la capa de servicios (también denominada capa de negocio) consiste en la lógica que realiza las funciones principales de la aplicación: procesamiento de datos, implementación de funciones de negocios, coordinación de varios usuarios y administración de recursos externos como, por ejemplo, el acceso a las bases de datos. Sobre esta capa, operan los Servicio Web (WS, por sus siglas en inglés) y las Interfaces de Programación de Aplicaciones (API) [11]. No es el fin de este trabajo marcar las diferencias entre ambos conceptos, pero algunas de ellas son las listadas en la Tabla 1 [12 y 13].

Tabla 1. Algunas diferencias entre WS y API.

WS	API
• Es una colección de protocolos y estándares de código abierto que se utilizan para intercambiar datos entre sistemas o aplicaciones.	• Es una interfaz de software que permite que dos aplicaciones interactúen entre sí sin la participación del usuario.
• Se basan principalmente en estándares como SOAP (Protocolo Simple de Acceso a Objetos), XML-RPC (abreviatura de *Extensible Markup Language Remote Procedure Call*) y REST (*Respresentational Estate Transfer*), para la comunicación.	• Se usa para cualquier estilo de comunicación.
• Solo admite el protocolo HTTP (*Hypertext Transfer Protocol*).	• Admite el protocolo HTTP/HTTPS (*Hypertext Transfer Protocol) Secure*).
• Admite XML.	• Admite XML y JSON (*JavaScript Object Notation*).

En conclusión, de la comparación, se puede afirmar que todos los WS son API, pero no todas las API son WS [13]. Cuando se construye una API, hay que basarse en un conjunto de definiciones y protocolos que se utilizan para diseñar e integrar el software de las aplicaciones. Suele considerarse, a estas interfaces, como el contrato entre el proveedor de información y el usuario, donde se establece el contenido que se necesita por parte del consumidor (la llamada) y el que requiere el productor (la respuesta). Por ejemplo, el diseño de una API de validación de usuario podría requerir que el usuario escribiera su nombre de usuario y contraseña y que el servidor diera una respuesta en dos partes: la primera, si puede acceder y la segunda, enviando sus permisos.

3.1 Por qué usar la API REST?

La API implementada en Experticia, se ajusta a los límites de la arquitectura REST, este término, es un acrónimo, cuya traducción al español significa *Transferencia de Estado Representacional*. Esta arquitectura de desarrollo web puede ser utilizada en

cualquier cliente HTTP [14]. Además, es más simple que otras arquitecturas ya existentes, como pueden ser XML-RPC o SOAP. Esta simplicidad se consigue dado que emplea una interfaz web que usa hipermedios para la representación y transición de la información [15]. La principal ventaja de esta arquitectura es que ha aportado a la web una mayor escalabilidad, es decir, da soporte a un mayor número de componentes y las interacciones entre ellos [15].

Las implementaciones de la REST también dependen de la noción de un conjunto de operaciones limitadas que tanto el cliente como el servidor entienden totalmente desde el comienzo. En el protocolo HTTP, las operaciones se describen en la "*línea inicial*", y las principales operaciones utilizadas en HTTP son las siguientes [16]:

- GET: devuelve la información que se haya identificado mediante el URI[3] de solicitud.
- PUT: solicita que la entidad adjunta se almacene en el URI de solicitud suministrado.
- POST: solicita que el servidor de origen acepte la entidad adjunta en la solicitud como un nuevo subordinado del recurso identificado por el URI de solicitud.
- DELETE: solicita que el servidor de origen elimine el recurso identificado por el URI de solicitud.

Las primeras tres operaciones son de solo lectura, mientras que las últimas tres son operaciones de escritura [16].

3.2 Sobre el modelado de datos JSON

JavaScript Object Notation (en español Notación de Objetos JavaScript -JSON-) es un formato de datos basado en los tipos de datos del lenguaje de programación JavaScript [17]. Como lenguaje de formato de datos semiestructurados[4], se ha convertido en uno de los principales formatos de intercambio de datos en la World Wide Web en los últimos años y ganó popularidad en la investigación de la comunidad de bases de datos [18]. Como cada objeto JSON, es un conjunto de pares clave-valor, un documento JSON puede ser representado naturalmente como una estructura de árbol de datos llamada árbol JSON. Un valor puede ser un valor atómico como una cadena, un entero, un número, una matriz o un valor nulo. Para capturar la estructura de composición de los datos JSON, cada valor puede volver a ser un conjunto de objetos JSON. Este formato de lenguaje agnóstico, es decir con aspectos de programación que son independientes de cualquier lenguaje específico, se puede utilizar por ejemplo en: Node.js, Python, Ruby, PHP, .NET, Java, etc. [19].

[3] Universal Resource Identifier, o identificador universal de recursos.

[4] Los datos semiestructurados no tienen un esquema definido. No encajan en un formato de tablas/filas/columnas, sino que se organizan mediante etiquetas o "tags" que permiten agruparlos y crear jerarquías. También se les conoce como no relacionales o NoSQL

3.3 Patrón de diseño REACT

Un patrón de diseño que trabaje con API pretende ocultar la complejidad de la implementación interna y presenta una interfaz sencilla a los clientes.

Uno de los distintos patrones de diseño existentes [20-21], es el Modelo Vista Controlador (MVC), es comúnmente utilizado para implementar interfaces de usuario, datos y lógica de control. Enfatiza una separación entre la lógica de negocios y su visualización. Esta "separación de preocupaciones[5]" proporciona una mejor división del trabajo y una mejora de mantenimiento. Las tres partes del patrón de diseño de software MVC se pueden describir de la siguiente manera:

- Modelo: Maneja datos y lógica de negocios.
- Vista: Se encarga del diseño y presentación.
- Controlador: Enruta comandos a los modelos y vistas.

REACT[6] (también llamada React.js o ReactJS) es una biblioteca Javascript de código abierto diseñada para crear interfaces de usuario con el objetivo de facilitar el desarrollo de aplicaciones en una sola página [22]. Esta biblioteca pretende ayudar a los desarrolladores a construir aplicaciones que usan datos que cambian todo el tiempo. REACT sólo maneja la interfaz de usuario en una aplicación; es la Vista en un contexto en el que se use el patrón MVC [21].

5. Evaluación Experimental

Para la realización de esta API experimental, se seleccionó Visual Studio 2019[7] de Microsoft, como entorno de desarrollo integrado (IDE). Este posee numerosas características que respalda varios aspectos del desarrollo de software: editar, depurar y compilar código y, después, publicar una aplicación. Aparte del editor y el depurador estándar que proporcionan la mayoría de IDE, Visual Studio incluye compiladores, herramientas de finalización de código, diseñadores gráficos y muchas más características para facilitar el proceso de desarrollo de software.

La codificación, se realizó en C#[8] , es un lenguaje de programación desarrollado por Microsoft, orientado a objetos, que ha sido diseñado para compilar diversas aplicaciones que se ejecutan en. *NET Framework*.

Por último, para la prueba de las API, se empleó Postman[9], que es una aplicación que permite realizar peticiones y obtener datos de pruebas. Es un cliente HTTP que da la posibilidad de testear *HTTP requests* a través de una interfaz gráfica de usuario, por medio de la cual se obtienen diferentes tipos de respuesta que posteriormente deberán ser validados. En la Figura 3, se grafica la arquitectura propuesta en este trabajo.

---

[5] En inglés *separation of concerns*, es un principio de diseño para separar un programa informático en secciones distintas, tal que cada sección enfoca un interés delimitado.

[6] https://reactjs.org/

[7] https://docs.microsoft.com/es-es/visualstudio/get-started/visual-studio-ide?view=vs-2022

[8] https://docs.microsoft.com/es-es/dotnet/csharp/tour-of-csharp/

[9] https://www.postman.com/

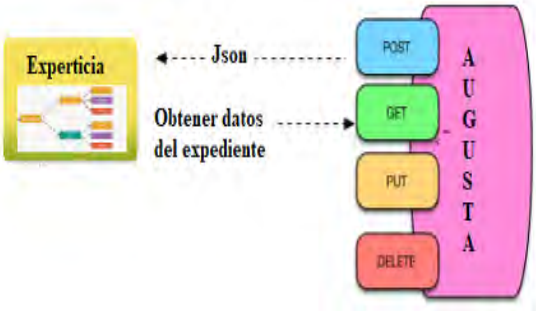


Fig. 3. Estructura del SE Web.

Respecto al modo de comunicar Experticia con Augusta, es utiliza llamados por peticiones como si fuera una URL con parámetros, un ejemplo:

```
…/api/Expediente/ListarBasico?IdOrganismo={idOrganismo}&NroExpediente={nroExpediente}
```

La creación de los objetos JSON implica escribir datos, para ello:

- Los datos están separados por comas.
- los datos se escriban en pares, siendo primero el nombre o atributo del mismo y luego el valor del dato.
- Los objetos JSON están rodeados por llaves "{}".
- Llaves cuadradas "[]" guardan arreglos, incluyendo otros objetos.

En la Figura 4, se muestra un fragmento de los datos en formato Json usados en Experticia:

Fig. 4. Estructura de una Json.

Una ventaja que tiene JSON respecto a XML es que el código resultante es más liviano. Para guardar la misma información utilizando JSON reduce el tamaño ya que no produce redundancia de datos y esto repercute en una mayor velocidad a la hora de transmitir la información [17].

En la Figura 5, se muestra la pantalla con una nómina de varias causas factibles a resolver. Luego de seleccionar la causa, Experticia se comunica con Augusta para solicitar actualización de la causa. El completa los datos esenciales que pudieran faltar

y resuelve el trámite aplicando el modelo proceso correspondiente. Experticia devuelve un resultado como, por ejemplo, un documento electrónico como sería un pedido de libertad condicional. Los datos, que se completaron en Experticia, son almacenados y enviados nuevamente a Augusta por medio de otra API.

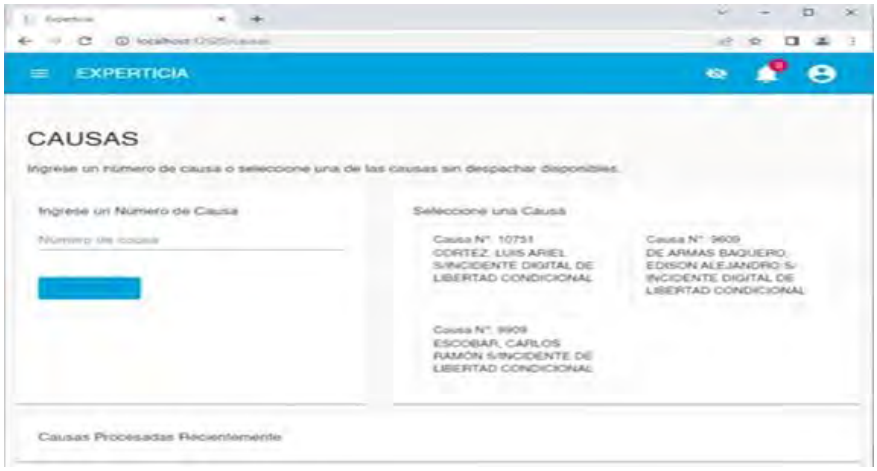


Fig. 5. Pantalla de Experticia para la selección de una causa.

Actualmente, esta nueva versión de Experticia se haya en etapa de pruebas. En la Tabla 2 se muestra los tiempos insumidos por el sistema para la consulta inicial de una causa, cuyos detalles se desglosan en la misma tabla.

Table 2. Detalle de las acciones junto con sus respectivos tiempos.

Proceso	Tiempo
Pedirle a Augusta información de una causa	1202 ms
Traer de Augusta los datos de los involucrados en la causa	143 ms
Actualizar los datos de la causa en Experticia a tratar y enviar tal información al operador	15 ms
Tiempo total insumido para agregar la nueva causa a Experticia con las verificaciones pertinentes	1360 ms

Estas son pruebas de rendimiento muy preliminares, a fin de evaluar la velocidad y capacidad de respuesta. Los tiempos surgen del promedio de 20 causas consultadas. Como se puede observar el mayor tiempo insumido corresponde a la petición de la causa, que ha demandado aproximadamente un 88% del tiempo total, esto a primera vista podría parecer excesivo. Sin embargo, cuando se tiene en cuenta que no se trata de la petición de un dato aislado, sino que en realidad está trayendo el histórico de una causa, podría considerarse un tiempo totalmente aceptable.

5. Conclusiones y Trabajos Futuros

En este artículo se describe parte del trabajo realizado en el desarrollo de Experticia, un SE aplicado a la justicia. Aunque existen algunos sistemas desarrollados para este dominio, Experticia se diferencia en que tiene interacción bidireccional con Augusta, el sistema utilizado en todos los organismos de la SCBA. Para implementar todas las comunicaciones con Augusta se empleó la tecnología de API REST.

Con la implementación de este sistema se espera conseguir una serie de beneficios reales en el quehacer diario de los organismos, en especial mejorando los tiempos que le demanda a los operadores resolver los tramites asociados a las causas en forma manual. Al momento se han realizado las primeras pruebas que integran Experticia con Augusta, obteniendo resultados gratificantes que se esperan mejorar con el avance del desarrollo.

Se planifica avanzar con las pruebas en una población controlada y monitoreada de usuarios finales. Se evaluarán los resultados para detectar además de errores, temas de usabilidad, seguridad, confiabilidad y rendimiento de Experticia.

Como próximo trabajo se espera aplicar estas mismas técnicas sincrónicas para la obtención de los distintos modelos de proceso. Actualmente, dichos modelos se encuentran almacenados en forma de árboles de decisión localmente en Experticia. La ventaja de esta propuesta radica en que con tal implementación se haría posible estandarizar y compartir los diferentes modelos para utilizarlos en los organismos judiciales de los distintos fueros en todo el ámbito provincial.

Por otra parte, se va a estudiar la posibilidad de complementar las funcionalidades de Experticia mediante la aplicación de técnicas de MD. Se va a evaluar la posibilidad de que el operador pueda conocer de modo anticipado el resultado que tendrá la resolución de un trámite, del mismo modo que lo haría aplicando los modelos de proceso con la asistencia que proporciona Experticia.

Referencias

1. Sposito, O.; Ledesma, V.; Procopio, G.; Bossero, J. (2020). Inteligencia Artificial aplicada al Poder Judicial. XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020), U. N. de la Patagonia Austral (UNPA), pp. 7-11, ISBN: 978-987-3714-82-5.
2. Sposito, O.; Busnelli, L.; Conti, L.; García, S.; Pérez Villar, G.; Ledesma, V.; Procopio, G.; Bossero, J. (2020). Sistema Experto para Apoyo del Proceso de Despacho de Trámites de un Organismo Judicial. XIV Simposio de Informática en el Estado (SIE 2020) - JAIIO 49. Facultad de Ingeniería de la UBA. ISSN: 2451-7534, pp. 17-29.
3. Sposito, O. y Otros. (2020). Metodología para evaluar un modelo de Justicia Predictiva. 8vo. Congreso Nacional de Ingeniería Informática y Sistemas de Información (CoNaIISI 2020). Universidad Tecnológica Nacional - Facultad Regional San Francisco. ISBN 978-950-42-0202-8, pp. 527-535.
4. Sposito, O. y Otros. (2021). Experticia. Un Modelo de Sistema Experto aplicada al Poder Judicial. XXIII. Workshop de Inv. Cs. de la Computación (WICC 2021). Univ. Nacional de Chilecito, La Rioja. ISBN: 978-987-24611-3-3; 978-987-24611-4-0, pp. 113-118.
5. Sposito, O. y Otros. (2021). Experticia, un sistema experto para dar apoyo al despacho de trámites asociados al expediente judicial. Suplemento de Derecho de la Alta Tecnología. elDial.com Biblioteca Jurídica Online. ISSN: 2362-3527. Disponible en:

https://www.eldial.com/nuevo/lite-tcd-detalle.asp?id=14162&base=50&id_publicar=&fecha publicar=08/11/2021&indice=doctrina&suple=DAT
6. Russell, S., & Norvig, P. (2004). Inteligencia artificial, un enfoque moderno. Madrid: Perason
7. Shahmin Sharafat, Zara Nasar, y Syed Waqar Jaffry. (2019). Data mining for smart legal systems. Computers and Electrical Engineering 78, 328—342. por la Secretaría de Políticas Universitarias del Ministerio de Educación de la Nación
8. de la Torre Llorente, C. y otros. (2010). Guía de arquitectura en N capas orientadas al dominio con Net 4.0. ISBN -978-84-936696-3-8,2010. Disponible en: https://sistemamid.com/panel/uploads/biblioteca/2018-06-12_04-26-49144688.pdf
9. Three-Tier Architecture. IBM Cloud Learn Hub (2020). Disponible en: https://www.ibm.com/cloud/learn/three-tier-architecture
10. Introducción a los Servicios Web. Invocación de servicios web SOAP. (2014) Universidad de Alicante. Disponible en: http://www.jtech.ua.es/j2ee/publico/servc-web-2012-13/
11. Mestras, Juan Pavón. (2012). Protocolos y arquitecturas de aplicaciones en internet Aplicaciones Web/Sistemas Web. Dep. Ingeniería del Software e Inteligencia Artificial Facultad de Informática. Universidad Complutense Madrid. Disponible en: https://www.fdi.ucm.es/profesor/jpavon/web/10-Introduccion-ProtocolosInternet.pdf
12. Doug Tidwell James Snell Pavel Kulchenko. (2001). Programming Web Services with SOAP Publisher: O'Reilly First Edition. ISBN: 0-596-00095-2b. Disponible en: https://docer.com.ar/doc/5sn10.
13. Sanjna Verma. (2018). APIs versus web services. Disponible en: https://blogs.mulesoft.com/dev-guides/apis-versus-web-services/
14. Beltran, C, (2019). Diferencia entre API y Servicio Web. Disponible en: https://medium.com/beltranc/diferencia-entre-api-y-servicio-web-5f204af3aedb
15. Amodeo, E. (2013) Principios de diseño de APIs REST (desmitificando REST). Disponible en: https://qdoc.tips/introduccionapisrestpdf-pdf-free.html
16. Diseño de API RESTful. (2021). Disponible en: https://www.ibm.com/docs/es/zos-connect/zosconnect/3.0?topic=apis-designing-restful
17. Paiva, R. (2021). Cómo transferir archivos a través de REST para almacenar en una propiedad. Parte I. Disponible en:https://es.community.intersystems.com/post/c%C3%B3mo-transferir-archivos-trav%C3%A9s-de-rest-para-almacenar-en-una-propiedad-parte-1
18. Introducción a JSON (2015). Disponible en: http://www.json.org/json-es.html
19. IBM Business Automation Workflow. (2022). Formato JSON (JavaScript Object Notation) Disponible en: https://www.ibm.com/docs/es/baw/20.x?topic=formats-javascript-object-notation-json-format
20. Gamma E. (203). Patrones de Diseno. Elementos de software orientado a objetos reutilizable. Disponible en: http:////docer.com.ar/doc/sx5s500
21. Stephen W. (2022). Descripción de los modelos, vistas y controladores (C#) Disponible en. https://docs.microsoft.com/es-es/aspnet/mvc/overview/older-versions-1/overview/understanding-models-views-and-controllers-cs
22. ASP.NET MVC 4 Release Notes. Disponible en: https://docs.microsoft.com/en-us/aspnet/whitepapers/mvc4-release-notes

Implantación de GDE en el Municipio de Lobería

María Belén Goyhenespe, Ariel Pasini, Silvia Esponda , Patricia Pesado

Instituto de Investigación en Informática LIDI (III-LIDI)*
Facultad de Informática – Universidad Nacional de La Plata
50 y 120 La Plata Buenos Aires
*Centro Asociado de la Comisión de Investigaciones Científicas de la Pcia. de Bs. As. (CIC)
mbgoyhenespe@hotmail.com
{apasini,sesponda,ppesado}@lidi.info.unlp.edu.ar

Abstract. La Plataforma Integral de Gestión Documental Digital es una de las herramientas propuestas para "Innovar la administración con una visión abierta, eficiente, participativa y transparente" en el eje Calidad de Vida. Esta solución informática se la conoce como GDE (Gestión Documental Electrónica). En el presente trabajo se muestra el proceso de diseño y ejecución del GDE realizado en el marco de un proyecto de modernización en el sector público, en un municipio de la Provincia de Buenos Aires.

Keywords: Gobierno digital, Gobierno Abierto. Metodologías Agiles

1 Introducción

Realizando un recorrido por la línea de tiempo se evidencia que cada 50 años, aproximadamente, se ha producido una innovación tecnológica que provoca un quiebre socio-económico, y que, a su vez, ocasiona un quiebre institucional provocando un cambio en los modelos de Gobierno.

Cada innovación tecnológica ha necesitado de una infraestructura para desplegarse. El ferrocarril, de la construcción de vías, la electricidad de la infraestructura para poder escalar, la informática y las telecomunicaciones requirieron de la construcción de nuevas estructuras de conectividad como por ejemplo la fibra óptica.

Trasladando esto a los modelos de Gobierno actuales, en los últimos años, se precisó de la construcción de un Gobierno Digital, como fruto de la disrupción tecnológica de la informática y las telecomunicaciones, con oficinas virtuales, y readecuación de los servicios a los ciudadanos. Reforzando el concepto de "Gobierno Abierto", el cual ubicaba al ciudadano en el centro de la toma de decisiones de los gobiernos. De esta manera, se abrió un camino para comenzar la modernización del Estado.

Así, surge para los gobiernos la oportunidad de repensar los servicios, construir modelos de gobierno donde realmente el ciudadano sea el centro.

Los paradigmas de Gobierno Electrónico y Gobierno Abierto están en pleno desarrollo. Asimismo, existe hoy la necesidad de innovar, mejorar y simplificar los procesos para agilizar, flexibilizar y transparentar a los ciudadanos las actividades públicas. En esa línea la Plataforma Integral de Gestión Documental Digital más

conocida como GDE (Gestión Documental Electrónica), ha sido una de las herramientas más implantadas en organismos gubernamentales.

El presente trabajo se enmarca en el ámbito de una tesina de grado, desarrollada por la alumna Belen Goyhenespe, bajo la dirección del LIDI.

Se expone la metodología y actividades aplicadas para el proceso de diseño e implantación del GDE en el marco un proyecto de modernización realizado en el sector público, en un municipio de la Provincia de Buenos Aires. En la sección dos se describe el concepto de Gobernanza y Metodologías Agiles. A continuación, la estrategia usada, en el Municipio de Lobería para la aplicación del GDE. Luego, se expone la Plataforma con un relevamiento de sus beneficios en el municipio. Y finalmente, se presentan las conclusiones de este artículo.

2 Conceptos de Gobernanza y Metodologías Ágiles

2.1 Gobierno Electrónico y Gobierno Abierto

Durante la década de 1980 comenzó una etapa de aceleración en la reforma de la administración pública, la cual devino como resultado el comienzo de una crisis del modelo del Estado de Bienestar [1].

En este contexto es donde surge un nuevo modelo de trabajo en la administración pública basado en conocimientos y prácticas administrativas empresariales, un conjunto de herramientas conocidas como Nueva Gestión Pública (NGP). Estas prácticas, conformaron un nuevo modelo de trabajo al que se denominó "posburocrático", enfocado en la transformación de las organizaciones públicas, en cuanto a normas, estructuras y patrones directivos y organizacionales [2]. Concretamente, la NGP incluye el uso de la tecnología en la administración pública, y así surge el nuevo modelo llamado Gobierno Electrónico (GE).

El término se le atribuye la autoría, al entonces vicepresidente de los Estados Unidos, Al Gore, cuando envió el memorando *"E-government directive*", solicitando a las dependencias gubernamentales aplicar las nuevas tecnologías. y señalándoles que "*si se utiliza de forma creativa la tecnología de Internet y la información, pueden ser una herramienta de gran alcance para hacer frente a algunos de nuestros más difíciles problemas sociales*." De esta forma, el GE se posicionó como un nuevo modelo de gestión a nivel gubernamental.

El concepto de Gobierno Abierto (GA) no es nuevo, es tan antiguo como la propia democracia. Sin embargo, dicho concepto cobró relevancia social luego de que Popper escribiera "Sociedad abierta y sus enemigos", durante su exilio a causa de la Segunda Guerra Mundial, con una fuerte crítica a la política rígida y autoritaria del momento. En enero de 2009, el presidente Barack Obama puso nuevamente en escena este concepto, solicitando a su administración abrir la información del Gobierno haciéndola pública, fijando como objetivo trabajar bajo los tres pilares básicos de este concepto: *Transparencia, Participación y Colaboración.* Unos años después, el 20 de septiembre de 2011, en una Asamblea de la ONU fue formalizada la Alianza para

el Gobierno Abierto (AGA), una iniciativa global que intenta asegurar el compromiso de los gobiernos nacionales para promover el gobierno abierto [3].

2.2 Metodologías Agiles (MA)

La filosofía ágil es un movimiento impulsado por valores y principios que surge para la mejora de todo tipo de procesos [4]. Estas metodologías enfatizan una estrecha colaboración entre el equipo de desarrollo del producto y las partes interesadas del negocio, siendo el factor humano especialmente determinante en todas ellas.

Dentro del concepto MA se encuentra una gran heterogeneidad de técnicas y buenas prácticas, que, además, pueden coexistir sin complicaciones. Este esquema ha resultado exitoso para proyectos con características específicas referidas en particular a tiempo de desarrollo y recursos. Cada metodología Ágil tiene sus características propias, sus particularidades y hace foco en algunos aspectos más específicos.

Actualmente, las metodologías ágiles que ganaron protagonismo en el desarrollo de la mayoría de los proyectos son SCRUM, KANBAN y LEAN DEVELOPMENT

Las Metodologías Ágiles son herramientas que han ayudado a los organismos públicos en la implementación de políticas públicas. Con las estrategias que se crean gracias a las metodologías ágiles se puede realizar un buen diseño y una mejor implementación [5] [6].

3 Municipio de Lobería

El municipio de Lobería decidió impulsar un modelo de gobierno con una perspectiva abierta, ubicando al ciudadano como eje central de las políticas para que además de mejorar su calidad de vida, cuente con posibilidades de participación y colaboración.

3.1 Estrategias en GE y GA en el Municipio de Lobería

Se desarrolló en un plan de transformación digital que se dividió en dos etapas: en la primera etapa se trabajó en el reordenamiento de la administración, capacitación del recurso humano y la recomposición de la infraestructura tecnológica. El municipio de Lobería se planteó un Plan de Gobierno con una visión de “Ciudad Amigable”, donde el principal esfuerzo estuvo centrado en la reconstrucción del vínculo entre los distintos actores de la comunidad. A pesar de contar con este esfuerzo, persistían problemas tales como la lentitud de los trámites, la falta de información para el ciudadano sobre la realización de los trámites, el uso constante de papel, la ausencia de interoperabilidad, y la falta de una herramienta tecnológica centralizada. En la segunda etapa, con un enfoque más innovador, se trabajó en un plan de modernización definiendo estrategias de acción en este camino de Gobierno Electrónico y Gobierno Abierto. En esta etapa, se reforzo la inversión en infraestructura tecnológica y capacitación para los empleados del municipio en esta temática.

4 Plataforma integral de Gestión Documental Digital

La *Plataforma integral de Gestión Documental Digital*, ha sido pensada no solo como un elemento digital que aporta eficiencia y eficacia a la administración, sino también, que aporta una mirada más participativa para el recurso humano y transparencia en los trámites y procesos digitalizados.

4.1 Plataforma integral de Gestión Documental Digital: marco normativo y sus funciones

La *Plataforma* tiene como objetivo el reemplazo de documentos y expedientes en formato papel por pares electrónicos integrados por el sistema de gestión documental electrónica para lograr la despapelización y obtener una red de trabajo más dinámica e interactiva.

GDE (Gestión Documental Electrónica), incluye la generación, comunicación, firma individual y firma conjunta, conservación, niveles de acceso, y otras funcionalidades que garanticen la permanente disponibilidad de la documentación oficial. Es un sistema interoperable, puede comunicarse con otros sistemas de la Administración Pública Nacional. En 2016 se aprobó la implementación del GDE como plataforma para la gestión integral del expediente electrónico en el Sector Público Nacional [7].

4.2 Estructura

GDE está compuesto de nueve módulos, *Comunicaciones oficiales (CCOO), Generador electrónico de documentos oficiales (GEDO), Expedientes electrónicos (EE), Legajo único electrónico (LUE), Registro civil electrónico (RCE), Procesos de Compras (COMPR.AR), Gestor de proveedores (GUP), Registro de proveedores (RIP) y Locación de Obras y Servicios (LOyS),* de los cuales, los municipios usan solo cuatro *(CCOO), (GEDO), (EE), (LUE),* Los cinco restantes, están orientados a funcionalidades en el ámbito provincial y nacional [8] .

La plataforma consta de un Escritorio Único (EU) en el que las personas pueden acceder ingresando su usuario y contraseña. El sistema posee un alto nivel de personalización que permite ser configurado o personalizado de forma tal que, se ajuste a la función o necesidad del administrativo o del funcionario.

5 GDE en Lobería

5.1 Implantación de GDE

La implantación de GDE inicio en el año 2020. El objetivo fue generar un cambio en el diseño e implementación de políticas públicas. El proceso se instrumentó en tres etapas consecutivas:

Primera etapa

En primera instancia, la Dirección a cargo sociabilizó con el equipo de gobierno acerca de las bondades que aporta esta herramienta y los posibles obstáculos que se podrían presentar. Además, se comunicó el alcance del GDE y, por último, se realizó un trabajo colaborativo e interdisciplinario en la definición de la meta a lograr.

Luego, se conformó un equipo para poner en conocimiento sobre los detalles organizativos del proceso de implantación, requerimientos y normativas.

Se firmó entonces un convenio entre Municipio de Lobería y el Ministerio de Innovación para la entrega de esta herramienta digital y paralelamente se creó el marco normativo local para su uso a través de una ordenanza.

Segunda etapa

Como metodología de trabajo para el desarrollo de este proyecto se creó un marco basado en la metodología ágil, Scrum. Se pensó en el uso de esta metodología ya que es un proceso incremental que, a corto plazo, logra pequeños avances concretos para el producto final, entendiendo además que los factores de éxito de este tipo de implantación son la colaboración, la participación activa y la aptitud frente a los cambios. Además, se complementó con una metodología que permitiese el involucramiento y la participación de los integrantes de la organización para evitar la resistencia al cambio, y en este sentido se sumó el aporte de los tableros de la metodología KANBAN, que fueron utilizados al momento de evaluar los avances del proyecto.

Instrumentación de Scrum

Se usaron varios elementos de esta metodología como se describe a continuación:

- **Roles**
  P**roduct Owner:** (PO) identifica los requerimientos del producto, los prioriza y determinar cuáles deberían incluirse en el ciclo de trabajo.
  Scrum Master: (SM) es el responsable de que el proceso sea llevado a cabo de manera exitosa. Colabora con el PO y asiste al equipo en todo momento.
  Equipo: integrado por todas las áreas que forman parte del organigrama municipal.
- **Productos**
  Product Backlog: Definido por el PO y el SM. Se conformó por los primeros 25 trámites que ambos consideraron iniciales.
  Sprint Backlog: Se construyó a partir del Product Backlog, durante el Sprint Planning. Se organiza mediante el tablero Kanban. Este tablero permite al SM y al PO, ver cuáles tareas están siendo implementadas, cuáles ya fueron finalizadas y cuáles quedan por implementar.
  Increment: cuando el Sprint finaliza se tiene en ambiente de prueba una porción del producto final, que se suma a los resultantes de los sprint anteriores.
- **Eventos**
  Sprint: Da como resultado un trámite digitalizado. Su duración aproximada fue de 20 días.
  Sprint Planning: este tipo de reuniones se realizó una vez por semana.

Sprint Retrospective: Se realizó cada 3 meses. En este evento se analizó ¿Qué salió bien? ¿Qué no salió tan bien?, ¿Qué podríamos hacer diferente para mejorar?

- **Visión**
  Agilizar los trámites y expedientes municipales mejorando con ello la calidad del servicio que se le presta al vecino, y garantizando transparencia. Despapelizar, eliminar los tiempos perdidos con los pases físicos acortando las distancias, facilitando el trabajo home office, reduciendo al mínimo el archivo físico y asegurando la conservación de este.
- **Criterio Done**
  Un trámite se considera digitalizado cuando:
  - El trámite está probado en plataforma de prueba y es aprobado y validado por el PO.
  - El personal capacitado en su uso.
  - Documentación asociada distribuida entre todo el personal de la municipalidad afectado al uso de GDE.

Aplicación de Metodología Scrum

Se inició el proceso de digitalización de trámites en GDE con una primera reunión entre el PO y el SM, donde se define la visión del proyecto y Criterio Done, es decir las condiciones que un incremento debe cumplir para considerarse trámite digitalizado. Los cuatro momentos de trabajo fueron los siguientes:

1. **Product Backlog**: Se seleccionaron una serie de trámites para incorporar al sistema GDE. La propuesta presentada por el SM al Producto Owner consistió en seleccionar cinco trámites por cada área del municipio para innovar. Conformada esta lista de requerimientos, el SM junto al PO, priorizaron, valorizando los principios de gobierno abierto. Se conformó así el Product Backlog, y se convirtieron en los primeros 25 trámites en incorporarse en GDE.
2. **Sprint:** Una vez definido el Product Backlog, en una primera reunión (Sprint 0), el PO dio inicio al proceso. Se propuso realizar una reingeniería a cada uno de los requerimientos para lograr la optimización de los procesos, respetando el orden de la priorización. Para esto, el equipo acordó la duración de cada entregable, estimó el tiempo que se afectaría a la reingeniería de cada trámite y su implementación en el ambiente de prueba. Se definió entonces la duración de cada Sprint de 20 días. Se realizaron 4 Sprint Planning, es decir cuatro iteraciones.
   El equipo pactó las tareas a ejecutar, conformando el Sprint Backlog las actividades a desarrollar:
   - Describir el proceso actual
   - Detectar las debilidades y fortalezas
   - Proponer un proceso optimizado
   - Plasmar el proceso en un diagrama de flujo
   - Documentar la información asociada al diagrama
   - Capacitación al personal involucrado en ambiente de prueba

En estas reuniones se trabajó con tableros, que aporta la metodología KANBAN, que fueron de máxima utilidad para organizar y visualizar el avance del proceso, donde se plasmaron las tareas a realizar. En principio se utilizó una pizarra física, pero luego se digitalizó con la herramienta Trello.

Cada trámite a digitalizar inició con la descripción del mismo. Una vez compartido se abrió el debate, donde se identificaron las posibles situaciones a optimizar, y como resultado se plasmó ese nuevo proceso en un flujograma que mostraba su nuevo recorrido. Esta documentación se compartió con los usuarios finales. Una vez aceptado el flujograma, se capacitó a todo el personal involucrado en la resolución de ese trámite sobre una plataforma de prueba utilizado hasta el momento de la puesta a punto en el ambiente de producción.

La duración de cada prueba duró 10 días. Cuando se terminó un producto o trámite, automáticamente se planificó el siguiente Sprint tomando del Product Backlog el próximo requerimiento priorizado.

3. **Sprint Retrospective**: Cada 90 días se realizó un proceso de mejora continua. Se analizó ¿Qué salió bien?, ¿Se puede hacer diferente para mejorar? ¿Optimizamos el tiempo de resolución? ¿Las áreas intervinientes trabajaron fluidamente? ¿Qué dificultades encontramos al momento de trabajar con la herramienta tecnológica? A partir de las conclusiones obtenidas en esta reunión de retrospectiva, el PO actualiza la lista priorizada (Product Backlog) y se plasma además en un nuevo flujograma.
4. **Documentación:** Al concluir la resolución de los primeros 25 trámites se diseñó un manual digital que fue distribuido entre todos los involucrados en el uso de GDE. En este proceso continuo de incorporar trámites al sistema GDE, el equipo consensuó reducir el tiempo asignado al sprint, en función de la gimnasia adquirida en la reingeniería de los trámites anteriores. De esta manera, se planificó incorporar un trámite por semana.

Tercera etapa

En esta etapa se configuró y parametrizó la plataforma productiva. En un trabajo colaborativo se generaron las tablas en el ambiente productivo y se setearon los parámetros según la información relevada por el equipo técnico local. Además, se crearon los usuarios del sistema. Simultáneamente, se gestionó el Certificado Productivo del Municipio y la Firma Digital de los funcionarios públicos.

Parametrizado el ambiente productivo, definidos los usuarios y sus perfiles, y probados los trámites, se lanzó una prueba piloto por 30 días en ambos sistemas, GDE y tradicional. Superado este tiempo, se habilitó al municipio el uso de GDE, y se instrumentó como única forma de resolver estos procesos en el municipio.

Integrantes del equipo local, actuaron a partir de ese momento como asistentes en GDE.

Pasados los 90 días de implementación, el equipo, decidió incorporar nuevos trámites a GDE. Se determinó continuar el trabajo con un proceso continuo e incremental de incorporación de trámites, repitiendo el proceso de trabajo que se ejecutó para los primeros 25 trámites. Actualmente, el municipio ha incorporado 66 trámites.

5.2 Resultados Obtenidos

Los resultados respecto de la implantación de GDE se midieron desde dos puntos de vista: respecto de la transformación digital y un análisis de la experiencia.

Relevamiento de transformación digital

Para analizar si la implantación de GDE, había sido una mejora desde el punto de vista tecnológico, económico, medioambiental y social, se propuso autoevaluarse con un autodiagnóstico de Transformación Digital realizado por la Red de Innovación Local.[1] Este autodiagnóstico, tenía como objetivo analizar los componentes claves a la hora de la digitalización de los servicios. Se analizaron nueve dimensiones: 1- *Gobernanza y liderazgo,* 2 - *Visión centrada en el ciudadano,* 3- *Servicios digitales,* 4 – *Procesos,* 5 – *Gestión del cambio,* 6 - *Tecnología y conectividad,* 7 – *Interoperabilidad,* 8 – *Ciberseguridad,* 9 - *Normativa*.

Cada dimensión cuenta con una serie de preguntas que evalúa la situación actual del municipio en esa materia. Lobería, se sometió a este relevamiento en dos ocasiones, para conocer el antes y el después de la implantación de GDE. A partir de los puntajes obtenidos por cada respuesta y dimensión, se pudo hacer una comparación. Tabla1

Tabla 1Autodiagnóstico

	Al 2016	**Al 2021**
1	No contaba con capacitaciones para los empleados y tampoco con métricas o indicadores.	Se delineo su estrategia de transformación digital con métricas e indicadores para medir las iniciativas.
2	Se contaba con indicadores[2] para evaluar la experiencia de la ciudadanía	Se definieron nuevos indicadores y metodologías para mejorar la experiencia del usuario.
3	Solo estaban digitalizados los pagos de tasas, registro de proveedores y el registro de compras y contrataciones.	Se sumaron los servicios de licencia de conducir, turnos de salud, RUB, entre otros.
4	no se priorizó esta dimensión y no había nada hecho en esta materia	se planificó la reingeniería de los procesos, se identificaron y priorizaron los procesos por los que empezar a trabajar.
5	no se había evaluado la madurez de las áreas en relación a transformación digital, tampoco se contaba con una estrategia de comunicación	La estrategia se expuso y se avanzó en la trasformación digital

[1] Asociación civil que trabaja con equipos de gobiernos locales con el objetivo de mejorar sus capacidades de gestión y de transformación de sus ciudades.

[2] Los indicadores se obtienen en un trabajo en conjunto entre la consultora DATEAR y la Universidad Nacional del Centro (UNICEN) que se encarga de medir el índice de satisfacción de las distintas áreas de trabajo del municipio.

6	Era una dimensión aplicada de forma parcial o nula.	Se pasó a contar de forma total con tecnología para generar reportes y hacer análisis a partir del uso de datos, Se aseguró la conectividad en todas sus áreas, e internet en todo el territorio.
7		Se articuló con organismos nacionales y provinciales para vincular información a los servicios digitales, El municipio aplicó la ventanilla única[3].
8	no se contaba con ningún programa de concientización de seguridad	Se implementaron medidas para proteger la seguridad de la infraestructura, el monitoreo continuo para identificar amenazas. Se mejoró el plan de continuidad y respuesta ante incidentes.
9	se contaba con normativa para el proceso de transformación digital.	

De los puntajes obtenidos en cada dimensión, se alcanzó un total de 40 puntos hacia el año 2016 y un total de 110 puntos hacia el año 2021, demostrando que la estrategia de transformación digital trajo consigo una gran mejora en los servicios provistos al ciudadano con una visión más centrada en él, como destinatario de la acción, con un cambio importante en la forma de pensar y ejecutar los procesos. Asimismo, provocó un mayor protagonismo de las personas de la organización, mejorando las relaciones interpersonales, entre áreas y con la ciudadanía. A partir de la implantación de GDE en el Municipio de Lobería se pudo distinguir un antes y un después.

Relevamiento de la experiencia de los involucrados en el proyecto

Con el fin de conocer la situación interna del municipio, a partir de la implantación de GDE, se procedió a realizar una encuesta anónima a empleados de cada área que intervino en este proceso. De los resultados se evidenció que una amplia mayoría (más del 90%) consideró que el uso de GDE agiliza el trámite dentro de la organización, simplifica el trámite a otros organismos públicos y empresas, se redujo el consumo de papel y que la implantación de GDE provocó un cambio en la cultura organizacional del municipio.

[3] Es una herramienta que permite presentar y realizar los trámites de forma electrónica ante una sola entidad.

6 Conclusiones

Este trabajo tuvo como propósito investigar sobre los paradigmas de Gobierno Electrónico y Gobierno Abierto, relacionando el camino recorrido por el Municipio de Lobería. Se describieron las acciones concretas que se implementaron. En este marco, se planteó un problema de vacío tecnológico persistente y, para solucionarlo, se propuso la implementación de la herramienta digital de Gestión Documental Electrónica. Se expuso la experiencia de la implantación de GDE en el Municipio con resultados, que comparan la situación sin la herramienta y con la herramienta, que permiten deducir que se cumplió con el objetivo planteado demostrando que la herramienta de Gestión Documental Digital benefició al Municipio, en lo tecnológico, social, cultural y agilidad.

Finalmente, la incorporación de la herramienta de GDE es un paso hacia la transformación digital, cumplir con los objetivos planteados en la administración, e incorporar el trabajo con metodologías ágiles, lo cual fue un gran beneficio en la administración interna.

7 Referencias

[1] J. Pratts, «De la burocracia al management, del management a la gobernanza,» INAP, Madrid, 2005.

[2] L. F. Aguilar Villanueva, «Gobernanza y Gestión Pública,» Fodo de Cultura Económica, México, 2009.

[3] Á. Ramírez Alujas y N. Dassen, «Gobierno abierto: la ruta hacia una nueva agenda de reforma del Estado y modernización de la administración pública en América Latina y el Caribe,» de *Gobierno Abierto y Transparencia focalizada. Tendencias y desafíos para América Latina y el Caribe.*, Banco Inter-Americano de Desarrollo, 2012,

[4] Lemay, M.. *Agile for everybody.* Sebastopol, CA, United States of America: O'Reilly Media, Inc. (2019)

[5] Mazagatos, J. A. (2018). Metodologías Ágiles y Administración Pública. *Boletic*(82), 42.

[6] Horquin, E. N. (2020). *Aplicación de Scrum en equiposunipersonales.* Tandil: U NiCen. Obtenido de https://docplayer.es/201450710-Aplicacion-de-scrum-en-equipos-unipersonales.html

[7] 561/2016, Sistema De Gestión Documental Electrónica, InfoLEG, http://servicios.infoleg.gob.ar/infolegInternet/anexos/260000-264999/260145/norma.htm, 06/04/2016

[8] Dirección Nacional de Digitalización Estatal, Expediente Electrónico (EE) Manual del usuario, https://www.argentina.gob.ar/jefatura/innovacion-tecnologica/innovacion-administrativa/manuales, 2016

Tecnología y comunicación: herramientas para la transparencia en los Gobiernos Locales

Fabian Gustavo Tisocco[1], Yanina Itatí Dal Molin[1], Marcelo Alberto Colombani[1]

[1] Facultad de Ciencias de la Administración. Universidad Nacional de Entre Ríos, Monseñor Tavella 1424. Concordia. CP (3200). Provincia de Entre Ríos, Argentina
fabian.tisocco@uner.edu.ar, dalmolinyanina@gmail.com, marcelo.colombani@uner.edu.ar

Resumen. La accesibilidad de los ciudadanos a documentos, información y procedimientos de la gestión del gobierno municipal propia del Gobierno Abierto, se funda en principios y condiciones básicas de transparencia y rendición de cuentas y en la participación y colaboración de los mismos en procesos decisionales. Adicionalmente resulta fundamental la innovación en materia de TICs, la sistematización e informatización de procesos críticos de la gestión organizacional y el desarrollo de aplicaciones que soporten dichos procesos; aspectos todos que han tenido un desarrollo notorio en los últimos años y especialmente en el contexto de la pandemia del Covid-19.

El presente trabajo, enmarcado en un Proyecto de Investigación de la Universidad Nacional de Entre Ríos, que pretende desarrollar un sistema de indicadores de responsabilidad social y sustentabilidad para gobiernos locales, tiene por objeto presentar los principales resultados que surgen de un relevamiento efectuado en las Áreas de Sistemas de Municipios del corredor del Río Uruguay, en la provincia de Entre Ríos.

Del mismo surgen como principales resultados, la presencia de asimetrías entre los municipios de mayor tamaño, que cuentan con estructuras internas asignadas a las funciones específicas de desarrollo de software, importante nivel de integración, desarrollos internos y a medida de las necesidades, y en general, un mayor control sobre los datos relevados y almacenados, así como de la información suministrada a ciudadanos y demás grupos de interés, lo que impacta de forma directa en los niveles de transparencia.

Palabras claves: Gobierno local. Innovación tecnológica. Covid-19. Transparencia. Gobierno abierto.

1 Introducción

La utilización masiva de las computadoras y por ende de los sistemas informáticos en las instituciones, ha generado una cultura de "sistematización de los procesos". Sumado a esto, las herramientas de desarrollo han ido permitiendo la generación de software de una manera mucho más fácil, rápida y menos costosa, lo que ha sustentado la cultura de sistematización. Así se logró insertar sistemas software en

empresas o negocios pequeños, que de otra manera no hubiesen podido encarar un proyecto de este tipo [11].

Este avance no ha quedado ajeno en los municipios o comunas, el cual llevó a automatizar y sistematizar sus procesos, incrementándose con el tiempo y logrando que más sistemas administrativos se hayan informatizado y se cuente con más y mejor información para la toma de decisiones.

Esto ha propiciado mecanismos no solo para digitalizar la información, sino también para mejorar la gestión interna de la administración municipal, brindar más y mejores servicios, facilitar el acceso a la misma, asegurar una mayor transparencia, aspectos que impactan en un aumento de la confianza pública y en una mayor participación de los ciudadanos y demás grupos de interés, a través de diferentes canales digitales, logrando de esta manera, lo que se conoce como gobierno abierto.

El gobierno abierto es una doctrina política que establece que los ciudadanos tienen acceso a los documentos y procedimientos del gobierno con el fin de permitir la vigilancia pública efectiva. El gobierno abierto tiene como objetivo que la ciudadanía colabore en la creación y mejora de servicios públicos y en el robustecimiento de la transparencia y la rendición de cuentas [2].

En este marco el presente trabajo expone parte de los resultados provenientes de una investigación que se lleva adelante en la Facultad de Ciencias de la Administración de la UNER, con el objetivo de analizar la estructura, composición y forma de funcionamiento de los sistemas de información que se manejan en los municipios del denominado corredor del Río Uruguay en la provincia de Entre Ríos. El fin último de dicha investigación es desarrollar un sistema de indicadores de medición y comunicación de la Responsabilidad Social y la Sustentabilidad (RSyS) aplicable a los mismos.

Para ello se presentan algunos de los conceptos principales que dan sustento teórico a la indagación, el diseño metodológico específico para esta parte del trabajo de campo y los resultados surgidos de la consulta a los funcionarios identificados como referentes en las Áreas de Sistemas de los gobiernos locales estudiados.

2 Marco teórico

2.1 Gobierno abierto

El concepto de Gobierno Abierto surgió a finales de la década del setenta en Inglaterra, con el objetivo principal de reclamar la apertura del gobierno y la participación ciudadana frente al secretismo con el que se actuaba. Dos décadas después continuó utilizándose el término *open government*, entendiendo por tal, el acceso libre a la información, protección de datos y al conocimiento de las actividades previstas que el Gobierno realizará o está realizando, permitiendo así el ejercicio de la opinión ciudadana [3].

En este sentido el Gobierno Abierto refiere al conjunto de mecanismos y estrategias que contribuye a la gobernanza pública, basado en los pilares de transparencia, participación ciudadana, rendición de cuentas, colaboración e innovación, centrando e incluyendo a la ciudadanía en el proceso de toma de

decisiones, así como en la formulación e implementación de políticas públicas, para fortalecer la democracia, la legitimidad de la acción pública y el bienestar colectivo.

Las políticas y acciones de Gobierno Abierto deben buscar crear valor público teniendo por finalidad la concreción del derecho de los ciudadanos a un buen gobierno, que se traduzca en un mayor bienestar y prosperidad, en mejores servicios públicos y calidad de vida de las personas, para contribuir al fortalecimiento de la democracia y afianzar la confianza del ciudadano en la administración pública [13].

Resumiendo podemos entender el Gobierno Abierto, como una nueva forma de comunicación permanente y transparente entre el Gobierno por medio de sus representantes y los ciudadanos, de manera bidireccional, mediante la participación efectiva en los procesos de decisión, colaboración y control de la administración [5].

2.2 Participación, colaboración y datos abiertos

Un gobierno abierto se asienta en la participación y colaboración ciudadana y en la necesaria apertura de datos, que permite consolidar la transparencia. La participación se busca promoviendo la creación de nuevos espacios de encuentro que favorezcan el protagonismo e implicación de los ciudadanos en los asuntos públicos. En este sentido los gobiernos deben buscar que la ciudadanía se involucre en el debate público, proveyendo los canales apropiados (aportando información y espacios de consultas) y mediante contribuciones que conduzcan a una gobernanza más efectiva, innovadora, responsable y que atienda las necesidades de la sociedad (OEA, 2016).

El desarrollo de estos canales favorece el derecho de la ciudadanía a participar activamente en la conformación de políticas públicas y anima a la Administración a beneficiarse del conocimiento y experiencia de los ciudadanos.

Además de la participación, un gobierno abierto requiere la colaboración, en el sentido de implicación y compromiso de los ciudadanos y demás grupos de interés en el propio trabajo de la gestión gubernamental. La colaboración supone la cooperación no sólo con la ciudadanía, sino también con las empresas, las asociaciones y demás integrantes de la comunidad, y permite el trabajo conjunto dentro de la propia Administración entre sus empleados y con otras Administraciones, aprovechando el potencial y energías disponibles en todos los sectores de la sociedad.

Adicionalmente será necesaria la apertura de datos públicos, es decir, la práctica de poner a disposición determinados datos, de forma libre a todo el que lo requiera acceder. En este sentido, no caben dudas que Internet ha resultado una herramienta fundamental para ofrecer los datos públicos de forma abierta.

Finalmente dos conceptos relacionados son los de Gobierno digital y Gobierno electrónico; el primero, alude al uso de las TICs en la administración pública para mejorar los servicios ciudadanos; mientras que el Gobierno Electrónico se encuentra vinculado con la consolidación de la gobernabilidad democrática, ya que se orienta a facilitar y mejorar la participación de los ciudadanos en el debate público y en la formulación de las políticas en general [13].

2.3 Transparencia en la Administración pública

En el Libro Verde de la Comisión de las Comunidades Europeas (2001), se pone en evidencia la creciente demanda por transparencia y el fomento de la misma desde los medios de comunicación, fuertemente acrecentada por las innovaciones en las TICs. También se menciona como un factor de fomento, a las nuevas expectativas de ciudadanos, consumidores, poderes públicos e inversores, además de la creciente preocupación por el ambiente y su acelerado deterioro, que exigen información clara y oportuna. En la misma línea, García Marzá [5] presenta a la transparencia como parte de los desafíos de la ética en las organizaciones y concretamente, como la primera regla de una gestión ética. Para el autor, la transparencia implica comunicar las intenciones y esfuerzos de la gestión por hacer las cosas bien; la difusión de los actos es parte central de la confianza que se genere en las partes interesadas. Adicionalmente ésta posibilita definir los límites entre el ejercicio del poder y el control social del mismo (OEA, 2016), por lo que podría afirmarse que cuando la gestión no es transparente, su credibilidad se debilita o desaparece.

Un gobierno transparente es entonces aquel que proporciona información sobre lo que está haciendo, sobre sus planes de acción, sus fuentes de datos y sobre lo que puede ser considerado responsable frente a la sociedad. Ello implica que los datos de la administración municipal deban cumplir con parámetros y estándares reconocidos, susceptibles de ser recopilados, clasificados, procesados y utilizados a través de las Tecnologías de la Información y las Comunicaciones (OEA, 2016).

En este sentido Diéguez [4] afirma que "El impacto de las tecnologías de información y comunicación dentro de las burocracias públicas puede convertirse – una vez más – en un factor catalizador de la digitalización de los procesos administrativos fortaleciendo y sofisticando las metodologías de gestión pública, a través de nuevos formatos de trabajo, remotos, más colaborativos y transversales".

Por otra parte, en el Informe Carrots & Sticks, que publican diferentes instituciones, entre las que se encuentra Naciones Unidas y el Consejo Empresarial Mundial para el Desarrollo Sostenible (WBCSD, por sus siglas en inglés) se afirma en la última edición, que ha aumentado un 270% el interés por la rendición de cuentas y la transparencia, en la última década [12]. Según el mismo informe esto demuestra un claro compromiso y esfuerzo de las organizaciones, incluidas las del ámbito público, por mejorar su gestión a través de la divulgación de información social, económica y medioambiental [12].

2.4 El rol de los gobiernos locales

Desde hace varias décadas los gobiernos locales han tenido que afrontar múltiples desafíos debido al proceso de descentralización que se viene desarrollando en nuestro país. Las funciones delegadas por el gobierno nacional ampliaron las competencias y funciones, con el consecuente y paulatino incremento de la complejidad de su administración, obligándolos a generar nuevas estrategias de gestión [14].

Los municipios cuentan con excelentes condiciones de vinculación y relacionamiento con la comunidad, tienen una interacción constante con diversos grupos de interés, que favorece el diálogo y crea escenarios propicios de

participación, ofreciendo herramientas diversas para diseñar políticas cercanas a la ciudadanía y acordes a las particularidades de cada territorio. Es en este punto donde cobran importancia los esfuerzos por dotar a la gestión de mayor transparencia, participación, rendición de cuentas y el rol clave que el Municipio puede desarrollar en este sentido, posibilitando el aumento de la confianza en las instituciones públicas y por ende el fortalecimiento de la democracia [8].

2.5 Digitalización en el contexto de la pandemia

La crisis sanitaria provocada por la pandemia del Covid-19 generó diversas restricciones en la atención dentro de la administración pública en general y de los gobiernos locales en particular; fundamentalmente ante diferentes procedimientos que requerían una continuidad, demandaban soluciones y atenciones para con el ciudadano y otros grupos de interés. Para algunos gobiernos locales, como para diversas organizaciones del ámbito público y privado, la crisis aceleró los procesos de cambios y rediseños de procedimientos y rutinas administrativas y en algunos casos, incluso también, la generación de nuevas políticas que posibilitaran el acceso virtual de reclamos, solicitudes, provisión de información y gestión de expedientes, entre otros [4].

Al respecto el Consorcio Govtech [7] refiere que la disrupción que ha producido la pandemia demuestra la importancia de acelerar la transformación digital en la gestión de los gobiernos, a fin de fortalecer su capacidad de resiliencia ante hechos atípicos como éste, comúnmente denominados cisnes negros. Este escenario ha acentuado la necesidad de que los gobiernos aceleren la digitalización para sus procesos operativos, su relación con el sector productivo, y su relación con los ciudadanos.

En este sentido, las demandas producidas por la crisis sanitaria del Covid-19 tuvieron el potencial de ser el gran acelerador y catalizador de los esfuerzos de digitalización y datización en estos últimos años. Por ello las agendas digitales y los datos continuarán siendo fundamentales para responder a los desafíos de productividad, formalidad, desarrollo sostenible, inclusión social y bienestar [4], [15].

Ante la complejidad de estos desafíos, los especialistas consideran que no se puede seguir pensando en que la digitalización es sólo una cuestión de modernizar, automatizar y digitalizar procesos análogos. La transformación digital del gobierno trata de nuevas maneras de pensar la gobernabilidad, lo que involucra, el diseño, implementación y evaluación del sector público en su totalidad. Adicionalmente esta requiere incorporar políticas y acciones de transparencia, acceso concreto a la información, rendición de cuentas y participación de los ciudadanos, con un fuerte apoyo en las TIC y orientado a lograr niveles de apertura y colaboración que permitan migrar hacia un sector público más ágil e inteligente y paralelamente, restaurar la confianza de los ciudadanos y demás grupos de interés en la gestión de gobierno [10].

En este marco los autores afirman que los gobiernos locales deben definir qué tecnologías implementar, considerando sus recursos económicos y humanos. Esto permitirá aprovechar aún más los innumerables beneficios que trae aparejado la utilización de las nuevas tecnologías por los gobiernos locales, por eso es fundamental que cada uno de ellos reflexionen sobre sus propias características para elegir las mejores opciones o mejorar lo ya existente, permitiéndoles satisfacer tanto las

necesidades internas de gestión, medición y rendición de cuentas, como aquellas de comunicación con sus stakeholders con todos los actores del gobierno [10].

En la misma línea, el informe elaborado por la Red de ISPA (Investigaciones Socioeconómicas Públicas de la Argentina) en el año 2020 plantea la importancia de las inversiones previas en infraestructura que muchos gobiernos habían realizado antes de la pandemia, lo que posibilitó cambios más significativos [15].

En el mismo se destaca la riqueza y potencial del proceso llevado adelante, que en algunos casos, no se ha limitado solo a la digitalización de trámites, sino a la posibilidad de verificar en tiempo real la información de cada ciudadano y vincular los datos con diferentes procesos y también con "cosas", gracias a la tecnología IoT (internet de las cosas).

En otro aspecto, desde la Red se hace hincapié en que a medida que la digitalización avanza, crece también la necesidad de desarrollar mejoras en las comunicaciones 4G y 5G, de manera que las redes móviles soporten las diferentes posibilidades de usos de fabricación inteligente.

Finalmente una cuestión adicional a considerar es la relativa a la protección de los datos de los individuos, lo que constituye un pilar fundamental de la digitalización. Al respecto, los especialistas de la mencionada Red consideran que nuestro país tiene importantes desafíos y cambios culturales que deben producirse. En la misma sintonía Diéguez [4] sostiene que estamos enfrentándonos a un modelo híbrido de administración pública, que combina la digitalización de diversos procesos, con modalidades presenciales; y un fuerte énfasis en la gobernanza de los datos públicos – a través de la promoción del big data y de normas que protejan los datos personales y demás derechos civiles básicos– y la aplicación y regulación de dispositivos de inteligencia artificial.

3 Metodología y Resultados

Se expone aquí el diseño realizado para cumplimentar con uno de los objetivos específicos de la investigación referido a analizar la estructura, composición y forma de funcionamiento de los sistemas de información que se manejan en los municipios bajo análisis. Vale recordar que el presente se enmarca en uno de los aspectos indagados en el marco del PID 7056 que pretende generar un sistema de indicadores para la medición y comunicación de la responsabilidad social y la sustentabilidad en el ámbito de los gobiernos locales del corredor del Río Uruguay en la provincia de Entre Ríos.

3.1 Determinación de la muestra

Para la tarea de relevamiento en los municipios de la provincia de Entre Ríos, comprendidos en el denominado corredor del Río Uruguay, y a los efectos de obtener una muestra representativa se seleccionaron las cuatro ciudades más importantes (Concordia, Concepción del Uruguay, Gualeguaychú y Chajarí), municipios de mediano tamaño, como San José y San Salvador y municipios clasificados como pequeños, entre los que se seleccionaron Ubajay, Villa del Rosario y Santa Ana. De

esta manera, con excepción de Islas, de escasa población y eminentemente insular, se han incluido municipios de todos los departamentos integrantes de dicho corredor.

3.2 Identificación de informantes calificados

A los fines específicos del relevamiento relativo al objetivo del presente, dentro de cada municipio se realizó la identificación de los Referentes de Áreas de Sistemas o funcionarios que cumplieran funciones similares, relacionadas a la responsabilidad de la gestión inherente al desarrollo y mantenimiento del Sistema de Información municipal. Para su identificación, el equipo se valió de los contactos previamente efectuados con cada institución, a partir de relevamientos anteriores tanto del mencionado PID 7056, como del que lo antecede, y se procedió a actualizar los datos de contacto de los funcionarios o asesores que en cada caso cubrían dichas funciones.

3.3 Diseño de la herramienta de recolección de datos

Para efectuar la recolección de la información con los referentes de las Áreas de Sistemas de los municipios se confeccionó una encuesta que cubrió las siguientes variables y ejes temáticos: a) Forma que adopta la estructura en el área de Sistemas; b) Número o cantidad de aplicaciones del Sistema Informático (S.I); c) Nivel de integración de la información entre las diferentes aplicaciones; d) Desarrollo interno o externo del software; e) Experiencias de desarrollo compartido con otros municipios o entidades del gobierno provincial; f) Datos e información social o ambiental relevados, almacenados o producidos por el Sistema de Información; g) Canales de comunicación interna y externa con los que cuenta; h) Tipo de soportes para las aplicaciones o sistemas internos; i) Digitalización de procesos en el contexto de emergencia sanitaria generada por el Covid-19; j) Estadísticas sobre el uso de los S.I online por parte de los ciudadanos; y, k) Funcionamiento del S.I.

En cuanto al canal seleccionado se decidió confeccionar el cuestionario mediante la herramienta "*google form*" y remitir a los referentes vía correo electrónico, previo contacto con los mismos. Vale destacar que además del contacto formal realizado mediante los correos institucionales, en la mayoría de los casos el link del cuestionario se remitió vía *whatsapp* a pedido de los propios referentes. En este sentido debemos resaltar que por la particularidad del trabajo y de la formación de los funcionarios referentes, el contacto se continuó por esta vía, ya sea para complementar información sobre el objetivo de la indagación, aclarar algunos puntos o eventualmente, efectuar algún recordatorio a los mismos para completar el cuestionario.

En forma adicional a la encuesta, se seleccionó el Municipio de Concordia y se profundizó en el análisis, a través de una entrevista a los responsables de la Dirección de Informática, en la búsqueda de conocer con mayor profundidad los procesos y la sistematización desarrollada. También se recurrió a la revisión del sitio web institucional.

La elección de este Municipio dentro de la muestra está fundada en los resultados obtenidos en la encuesta y en la particularidad de haber alcanzado el primer puesto en

la provincia de Entre Ríos del ranking realizado por el IARAF [6] en materia de transparencia.

3.4 Resultados del relevamiento

De las encuestas realizadas, resulta en principio, que el 40 % de los gobiernos locales cuentan con un software único o integral, un 30 % posee diversas aplicaciones integradas y el 30 % restante cuenta con aplicaciones independientes. Si se tiene en cuenta las funcionalidades informadas por los mismos, se puede observar que todos los municipios cuentan con sistema de gestión de Recursos Humanos y Liquidaciones de Sueldos, Sistema Catastral y de Gestión de Tasas.

En general, se puede apreciar que el foco de las aplicaciones está puesto en las tareas administrativas que apoyan el funcionamiento diario de los Municipios.

Al respecto entendemos que contar con un software único integral es muy importante para la gestión del municipio en general, ya que permite un control centralizado de las diversas funciones que tiene a cargo, se facilita la estandarización de la información y adicionalmente se logra un acceso uniforme a la misma.

En cuanto al nivel de integración de los sistemas, en caso de no poseer un software único, se pudo observar que el 80 % de los municipios presentan un nivel de integración por encima del 7 puntos en la escala de 1 a 10 considerada, o sea, que si bien se advierte la presencia de aplicaciones independientes, las mismas presentan una interrelación o conexión entre sí.

Esto es importante ya que ha sido una práctica común, generar aplicaciones (programas) para resolver problemas puntuales, con las consecuentes limitaciones en la gestión posterior de la información, fundamentalmente asociado a la repetición de operaciones y datos y el aumento en la inconsistencia de los mismos. Con un sistema único, o con módulos integrados, estos problemas no se presentan, obteniéndose un mayor control, ya que diferentes áreas validan los datos en diferentes etapas del proceso administrativo, aspecto que también redunda en una mayor transparencia.

Del tercer aspecto indagado, en relación a la influencia de la pandemia del Covid-19 en la digitalización de procesos, surge que el 40% de los municipios no desarrolló ningún cambio, y el 60 % restante, manifestó haber ejecutado cambios menores.

En este sentido, vale aclarar que de acuerdo a la revisión documental previa, efectuada como parte del trabajo de campo del Proyecto de Investigación que da sustento al presente, se pudo conocer que antes de la crisis de la pandemia, algunos de los municipios, habían realizado adecuaciones y mejoras progresivas para la prestación de servicios diversos en base a la tecnología. Entre ellos se pueden mencionar la generación de recibos de sueldo electrónicos, solicitud de licencias del personal y pagos electrónicos de tasas municipales.

Dentro del grupo de gobiernos que manifestaron haber efectuado algunas modificaciones menores, se mencionó haber puesto el énfasis en el uso de VPN (*Virtual Private Network*), para el envió de información a través de correos electrónicos, el pago online de tasas municipales y la gestión de turnos online para algunos trámites.

Otro aspecto indagado refiere a la existencia de estadísticas o datos que den cuenta del uso de los sistemas de información online por parte de ciudadanos y otros grupos

de interés. Al respecto resulta que el 60% de los municipios manifestaron no contar con estos datos y el 40% restante respondió que sí poseen registro de la información.

En cuanto a los resultados que surgen de la indagación específica del Municipio de Concordia, se destaca en principio, la existencia de una sección de transparencia en el portal web institucional, donde se puede encontrar información sobre finanzas públicas, presupuestos y ejecución presupuestaria de ingresos y egresos, deuda pública, como así también un conjunto de indicadores generales de recursos, gastos y solvencia, entre otros.

Por su parte, de la entrevista a los responsables de la Dirección de Informática, surge que la información publicada en la web y las aplicaciones móviles, es obtenida de forma directa de los S.I., sin necesidad de intervención o proceso de exportación adicional, brindando la misma en tiempo real. Además, allí mismo se muestra información de la gestión de compras y contrataciones, donde se pueden encontrar los concursos de precios, las licitaciones públicas y privadas y el listado de proveedores, los boletines oficiales y el digesto, entre otra información; logrando en conjunto, un nivel de transparencia significativo. También se publica información de sueldos, escalas salariales, sueldos de los trabajadores por Secretarías, y además se cuenta con un sistema de verificación de recibos, permitiendo a cualquier interesado poder constatar la veracidad de dicha información, de forma directa de los recibos.

Otro punto a destacar es el presupuesto participativo, el cual consiste en un proceso de intervención directa, permanente, voluntaria y universal en el que la ciudadanía conjuntamente con el gobierno, delibera y decide qué políticas públicas se deberán implementar con parte del presupuesto municipal. En este municipio se efectúa la priorización y elección por medio del sitio web, a través de un sistema de votación electrónica que ha sido diseñado para este fin.

Además, como otro medio de participación ciudadana, posee un sistema de votación electrónica para encuestas a ciudadanos sobre diferentes temas de incumbencia pública, por medio de un sistema de consulta ciudadana online.

Finalmente, surge de este relevamiento específico, que el municipio cuenta con un sistema de botón antipánico para las víctimas de violencia de género conectado con la Policía de Entre Ríos, los Juzgados de Familia y el propio municipio.

4 Conclusiones

Como hemos visto las tecnologías de información y comunicación hacen propicio el gobierno abierto y la transparencia en la administración pública municipal. Diversos han sido los avances recientes en este sentido, generando cambios concretos en la administración y gestión de datos e información, que redundan en beneficios directos para los ciudadanos y para el fortalecimiento de la democracia.

En el análisis de la estructura, composición y forma de funcionamiento de los sistemas de información que se manejan en los municipios bajo análisis, hemos visto que en términos generales los mismos poseen un nivel de desarrollo informático y tecnológico aceptable y en crecimiento, aunque con asimetrías importantes entre los de mayor y menor tamaño. Este será un aspecto que beneficiará a los más grandes y probablemente creará limitaciones e importantes desafíos, en los pequeños, ante

futuras implementaciones como la del sistema de indicadores de sustentabilidad, objeto del proyecto que da marco al presente.

También queda en evidencia que no están utilizando software unificado entre ellos y que en algunos casos, será necesario crear espacios pertinentes en la estructura organizacional de los gobiernos locales, lo que entendemos como un campo de acción y demanda específica para los futuros graduados en Sistemas de Información. Adicionalmente esto puede observarse como una oportunidad de articulación entre la Universidad y los gobiernos locales, fundamentalmente los de menor tamaño, donde se observan mayores dificultades y limitaciones.

En relación a la aceleración de los procesos de digitalización, provocada por la crisis de la pandemia, según el planteo de los especialistas, se observa que los municipios del corredor del Río Uruguay bajo estudio, mantuvieron su estructura y funcionamiento con relativa estabilidad y con un menor nivel de cambios, ya que venían desarrollando procesos similares previos a la pandemia.

Advertimos como un desafío de investigación para dar continuidad a este trabajo, el estudio y aplicación de nuevas tecnologías para dar soporte a la creación de servicios para los ciudadanos y demás grupos de interés, como la certificación electrónica y la gestión documental electrónica mediante firma digital, entre otros.

Referencias bibliográficas

1. Cobo, C.: Gobierno Abierto: de la transparencia a la inteligencia cívica. Info-DF. p. 107. (2013).
2. Lathrop, Daniel; Ruma, Lauren.: Open Government: Collaboration, Transparency and Participation in practice. Sebastopol: O'Really Media. pp. 92-93. ISBN 978-0596804350. (2010).
3. Álvaro V. Ramírez-Alujas. Gobierno abierto y modernización de la gestión pública. tendencias actuales y el (inevitable) camino que viene. Reflexiones seminales. Revista Enfoques. Ciencia Política y Administración Pública. Volumen IX, número 15, pp. 99- 125 (2011).
4. Diéguez, G.: La pandemia del COVID 19: Cuatro reflexiones en torno al rol del Estado y las capacidades de gestión pública. CIPEC (Centro de Implementación de Políticas Públicas para la Equidad y el Crecimiento). Recuperado el 30 de abril de 2022, de https://www.cippec.org/textual/la-pandemia-del-covid-19-cuatro-reflexiones-en-torno-al-rol-del-estado-y-las-capacidades-de-gestion-publica/ (2021).
5. García Marzá, D.: Cuanto más poder tiene una empresa, más responsabilidad. Blog de Alpesa y Upalet. Recuperado el 5 de diciembre de 2021, de https://alpesa.com/blog/domingo-garcia-marza-cuanto-mas-poder-una-empresa-mas-responsabilidad/ (2019).
6. IARAF. Informe Económico Nº 394. Visibilidad fiscal en municipios argentinos. https://www.iaraf.org/index.php/informes-economicos/publicaciones-otras-areas-de-estudio/200-informe-economico-n-394. (2017).
7. Ramirez-Alujas A. Jolias L. & Cepeda J.: GovTech en Iberoamérica. Ecosistema, actores y tecnologías para reinventar el sector público. (1a ed., Vol. 1). Bahía Blanca, Argentina: GovTech Hub. (2021).
8. Grandinetti, R.M. y Miller, E.: Tendencias y prácticas: políticas de gobierno abierto a nivel municipal en Argentina. Recuperado el 10 de julio de 2021, de

https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0719-17902020000010008. (2020).
9. Municipalidad de Concordia: Mapa de los Servicios de la Ciudad. Recuperado el 18 de junio de 2021, de https://www.concordia.gob.ar/servicios/mapa-de-los-servicios. (2021).
10. Municipalidad de Concordia: Sección de Transparencia. https://www.concordia.gob.ar/gestión/transparencia. (2021).
11. Colombani, Marcelo Alberto: Metodologías para el desarrollo de software en PYMES. Requerimientos confusos, incompletos y cambiantes. Universidad Nacional de Entre Ríos. (2017).
12. Breliastiti, R.: Development of mandatory & voluntary instruments of sustainability reporting (SR) according to carrots & sticks 2006—2016. The Indonesian Accounting Review. https://doi.org/10.14414/tiar.v10i1.1931. (2020).
13. Carta Iberoamericana de Gobierno Abierto. Aprobada por la XVII Conferencia Iberoamericana de Ministras y Ministros de Administración Pública y Reforma del Estado. Recuperado el 10 de marzo de 2022, de https://clad.org/wp-content/uploads/2020/07/Carta-Iberoamericana-de-Gobierno-Abierto-07-2016.pdf. (2016).
14. Cravacuore, D.: La Articulación de Actores para el Desarrollo Local en el Área Metropolitana de Buenos Aires. Reflexiones a partir de la Mirada de los Empresarios. VI Seminario Nacional de la Red Nacional de Centros Académicos Dedicados al Estudio de la Gestión en Gobiernos Locales. Recuperado el 10 de abril de 2021, de https://www.researchgate.net/publication/340679602_La_Articulacion_de_Actores_para_el_Desarrollo_Local_en_el_Area_Metropolitana_de_Buenos_Aires_Reflexiones_a_partir_de_la_Mirada_de_los_Empresarios (2004).
15. Red ISPA: La Argentina frente al COVID-19: desde las respuestas inmediatas hacia una estrategia de desarrollo de capacidades, Red de Investigaciones Socioeconómicas Públicas de la Argentina. Buenos Aires, Red ISPA. (2020).

Propuesta e implementación de un sistema basado en servicios de proximidad con BLE Beacons

Juan Fernández Sosa [1], Santiago Medina [1], Maite Segovia, Pablo Thomas [1], Armando De Giusti[1]

[1] Instituto de Investigación en Informática LIDI (III-LIDI). Facultad de Informática – Universidad Nacional de La Plata, La Plata, Argentina.
Centro Asociado a la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC)
{jfernandez, smedina, pthomas, degiusti}@lidi.info.unlp.edu.ar,
maite.segovia@alu.ing.unlp.edu.ar

Resumen. Los servicios basados en la proximidad permiten mejorar la interacción y experiencia de las personas en un determinado espacio. Los BLE beacons son una solución muy popular para brindar este tipo de servicio pudiendo convertir una habitación convencional en un entorno inteligente. En este trabajo se propone un sistema que utiliza dispositivos BLE beacons para mejorar la experiencia de las personas en las actividades que se realizan en el Centro de Innovación y Transferencia Tecnológica de la Facultad de Informática de la UNLP

Keywords: beacons, Bluetooth de bajo consumo, servicios basados en la proximidad, entornos inteligentes.

1. Introducción

El desarrollo de las tecnologías digitales en los últimos tiempos ha permitido avanzar sobre la digitalización e interconexión de cosas, espacios y experiencias. Las Ciudades Inteligentes emplean las últimas tecnologías para prestar soluciones y servicios innovadores a sus residentes. Los procesos de digitalización en este tipo de configuración de ciudad hicieron posible integrar tecnologías y caracterizar como "inteligentes" a diferentes aspectos de la vida cotidiana [1], en el área de transporte, comunicación, salud, infraestructura, edificios, entre otras.

Desde su aparición, el teléfono móvil ha sido generador de grandes cambios sobre aspectos fundamentales de la sociedad. Las personas emplean este tipo de dispositivo para acceder e interactuar con diferentes servicios. Esto genera nuevas formas de relación y comunicación con otras personas y también con su entorno. En su proceso de evolución, este tipo de dispositivo fue incorporando mayor capacidad de procesamiento, almacenamiento, sensado contextual y nuevas tecnologías de comunicación inalámbrica [2]. A partir de la instalación y ejecución de Aplicaciones de Software (apps), estos dispositivos adquieren flexibilidad para explotar su potencial en múltiples dominios brindando innovadores e interesantes servicios.

Los servicios basados en la localización (LBS por sus siglas en inglés) permiten obtener información dependiendo de la ubicación de un dispositivo y de la persona que lo porta [3]. La aplicación de este tipo de servicio responde a preguntas del estilo "¿Dónde estoy? ¿Qué cosas se encuentran en mi cercanía? ¿Cómo puedo llegar a ellas?" [4] pudiéndose aplicar tanto en entornos de exterior como en ambientes de interior. Como ejemplos de este último caso se pueden mencionar centros comerciales, museos, estaciones de transporte, aeropuertos, entre otras. Uno de los objetivos de los servicios en ambientes de interior es el de mejorar la experiencia de las personas en un determinado lugar.

Los *beacons* son dispositivos de pequeño tamaño con capacidad de comunicación inalámbrica que emplean tecnología Bluetooth de bajo consumo (BLE). El funcionamiento de este tipo de dispositivo consiste en transmitir una señal con una pequeña cantidad de información que puede ser captada por otros dispositivos compatibles (smartphones, smartwatches, computadoras de una sola placa -s*ingle-boards computers*-) [5]. La información que se transmite dependerá del dominio donde se requiera utilizar y la configuración del beacon, pudiendo ser un identificador, una dirección de email, valor de temperatura, etc. Los beacons son utilizados para brindar servicios de localización indoor y servicios basados en la proximidad [6]. La instalación de estos dispositivos puede convertir una habitación en un entorno inteligente e interactivo [7].

El Centro de Innovación y Transferencia Tecnológica (CIyTT) [8] de la Facultad de Informática de la Universidad Nacional de La Plata es un espacio orientado al desarrollo de actividades vinculadas a la innovación en temas que corresponden a la disciplina Informática. En este lugar se llevan a cabo diferentes actividades con el fin de acercar este espacio de innovación a toda la comunidad.

El objetivo de este trabajo es presentar un sistema de software que brinde servicios basados en la proximidad y permita a los visitantes del CIyTT a recibir información en sus dispositivos móviles mientras recorren las instalaciones. Este tipo de sistema busca mejorar la calidad de las visitas de la comunidad y la difusión de los proyectos de innovación que se llevan a cabo en este espacio.

El resto de este trabajo se organiza de la siguiente manera: en la sección 2 se realiza una descripción de los dispositivos beacons y de la tecnología Bluetooth de bajo consumo que utilizan para comunicarse. En la sección 3 se introduce el concepto de servicios basados en la proximidad y algunos ejemplos de aplicación. Luego se detalla el nuevo sistema propuesto detallando su motivación, arquitectura y aspectos técnicos. Para finalizar se presentan conclusiones y trabajo futuro.

2. BLE Beacons

Los BLE Beacons, comúnmente conocidos como beacons, son dispositivos de pequeño tamaño que transmiten información de manera inalámbrica a otros dispositivos cercanos, empleando tecnología Bluetooth de bajo consumo (BLE) [6].

La comunicación siempre se realiza desde el beacon hacia otros dispositivos, compatibles con el estándar Bluetooth 4.0 o superior, como por ejemplo los smartphones y/o computadoras de placa única (Raspberry Pi). El beacon transmite una pequeña señal cada cierto intervalo de tiempo y con un cierto rango de transmisión con el objetivo de dar a conocer su existencia. En los dispositivos receptores se ejecuta una aplicación de software particular para detectar dicha señal y en función de su procesamiento, realizar una acción determinada. Para lograr esta interacción entre dispositivos existen diferentes protocolos de comunicación de datos basados en BLE. Los protocolos más populares son: iBeacon, Eddystone, AltBeacon y GeoBeacon [6][9].

El Bluetooth de bajo consumo es una tecnología de comunicación inalámbrica presentada en el año 2010 y diseñada particularmente para ser utilizada en aplicaciones de Internet de Las Cosas (IoT) y Ciudades Inteligentes [10]. Los beacons son dispositivos particularmente populares para implementar soluciones en dichos dominios, debido a su bajo costo, tamaño pequeño y a la autonomía de su batería. Estos dispositivos son utilizados para el desarrollo de aplicaciones de marketing y publicidad, servicios basados en la proximidad, localización indoor, posicionamiento y seguimiento de objetos, entre otras [7] [11].

3. Servicios basados en la proximidad (PBS)

Como se mencionó anteriormente, un tipo de aplicación que se puede implementar utilizando beacons es el servicio basados en la proximidad. Este tipo de servicio tiene como finalidad alertar sobre la proximidad de un objeto o persona a un punto de interés.

Los beacons por su pequeño tamaño pueden ser colocados en casi cualquier lugar sin tener que alterar la infraestructura donde se despliegan. Esto permite convertir una habitación o espacio convencional en un entorno inteligente. Por ejemplo, en [4] para mejorar la experiencia de los visitantes en un museo, se colocaron diferentes beacons en exposiciones o salas de interés. Los visitantes con una app en sus dispositivos móviles y su Bluetooth encendido, reciben notificaciones e información a medida que recorren el lugar y se encuentran próximos a estos puntos de interés. Otro ejemplo son centros comerciales donde se colocan beacons sobre objetos o pasillos para así enviar promociones y atraer la atención de las personas cuando están cerca de ellos, esto se conoce como márketing por proximidad [12] [13].

4. Desarrollo propuesto

4.1 Motivación

El Centro de Innovación y Transferencia Tecnológica (CIyTT) de la Facultad de Informática tiene dos objetivos principales: la creación de un espacio de trabajo para la innovación en temas de la disciplina informática que convoque a graduados, docentes, alumnos e investigadores; y la generación de acciones de transferencia que acerquen los proyectos presentados en el centro a la comunidad y viceversa.

En sus instalaciones, el CIyTT cuenta con múltiples estaciones gestionadas por las diferentes unidades de investigación de la Facultad. Cada una de estas estaciones está conformada por diferentes proyectos innovadores. En las visitas guiadas que se realizan, los visitantes recorren el edificio explorando sus estaciones y conociendo los proyectos que allí se presentan.

En este contexto se propuso el desarrollo de un sistema de software para complementar las diferentes visitas y talleres que se realizan en el edificio, utilizando servicios basados en la proximidad para incrementar el impacto tecnológico en la experiencia de los visitantes.

4.2 Arquitectura del sistema

Para dar soporte a los servicios por proximidad de este nuevo sistema, se propone una arquitectura de tres componentes tecnológicos principales: beacons, una aplicación de software para dispositivos móviles y una plataforma web de gestión, tal como se presenta en la figura 1a. Dichos componentes se describen a continuación.

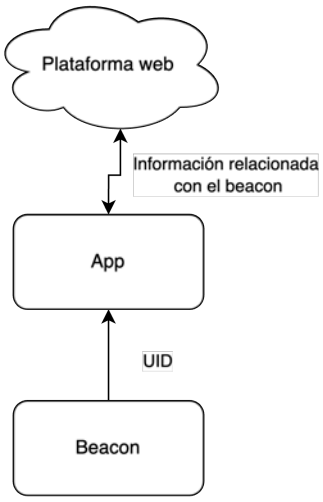


Fig 1a

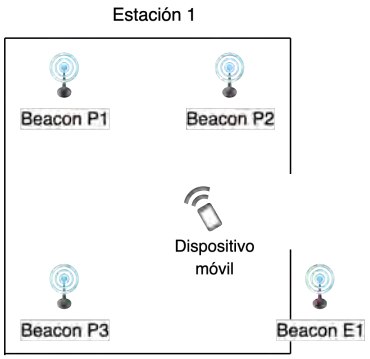


Fig 1b

Figura 1. a) Arquitectura del sistema propuesto. b) Despliegue de beacons en una estación

4.2.1 Beacons

Los beacons estarán ubicados de manera estática en diferentes lugares de interés, relacionados a los proyectos y las estaciones, como se muestra en la figura 1b. Dichos beacons transmitirán periódicamente su código de identificación único (UID).

Para su desarrollo se decidió utilizar placas de desarrollo: Node-MCU ESP32, ESP32-CAM y HM-10 Ble. En cada una de estas placas se realizaron pruebas de comunicación con dos de los protocolos más populares: IBeacon y EddyStone.

4.2.2 Aplicación de Software para dispositivos móviles

Los visitantes del centro deberán descargar la app en sus dispositivos móviles. La app tendrá dos modos de funcionamiento. El primero da soporte al servicio por proximidad por lo que los usuarios deberán activar el Bluetooth para poder interactuar con los diferentes beacons. En este modo, la app se encargará de detectar la señal transmitida por un beacon y calcular la proximidad al mismo a partir de la intensidad de dicha señal (RSSI). El código de identificación del beacon se enviará a la plataforma de gestión, a través de una API-Rest, para conocer el nombre de ese punto de interés. En la aplicación se visualizará esta información permitiendo al usuario obtener mayor cantidad de información relacionada con ese espacio. En caso de estar interesado, la plataforma devuelve más contenido en formato de texto y multimedia (audios, imágenes y videos).

El segundo modo de funcionamiento permite a los usuarios navegar por la interfaz de la aplicación visualizando la información completa sobre las estaciones y proyectos, sin tener que interactuar con los beacons.

Para el desarrollo de esta componente, se eligió utilizar un enfoque multiplataforma, empleando el framework de desarrollo Ionic.

4.2.3 Plataforma web

La plataforma web estará desplegada en la nube y permitirá tener centralizada toda la información para ser consultada por la aplicación móvil. Se podrán dar de alta, editar y relacionar estaciones y proyectos. Cada proyecto está conformado por una descripción en formato texto y archivos multimedia tales como audio, imágenes y videos. Por otra parte, la configuración de los beacons y su relación con una estación o proyecto determinado se realizará con esta misma plataforma.

Para el desarrollo se eligió Node.Js como tecnología para el backend, Vue.js para el frontend y una base de datos PostgreSQL.

5. Conclusiones y trabajo futuro

En este trabajo se presentó la propuesta de un nuevo sistema de software que brinda servicios basados en la proximidad. Este sistema propone la utilización de dispositivos BLE beacons para mejorar la experiencia de las personas en las actividades que se realizan en el Centro de Innovación y Transferencia Tecnológica de la Facultad de Informática de la Universidad Nacional de La Plata.

En esta primera instancia del desarrollo se presentó una arquitectura de tres componentes tecnológicos: beacons, una aplicación de software para dispositivos móviles y una plataforma web. Del primer componente, se realizaron pruebas de comunicación empleando los protocolos iBeacon y EddyStone en diferentes placas de desarrollo. Por otra parte se definieron dos modos de operación de la aplicación móvil, uno que permite interactuar con los beacons y otro que permite acceder al contenido de manera convencional. Por último, se desplegó la primera parte de la plataforma web de gestión que permite la creación de proyectos y estaciones.

Además de concluir con el desarrollo, despliegue e integración de cada una de las componentes de la arquitectura propuesta, como trabajo futuro se plantea agregar funcionalidades para encuestas y votaciones en la aplicación móvil, la medición de performance de la señal emitida por los beacons en cada protocolo utilizado y la generación de reportes de uso dentro de la plataforma web.

Referencias

[1] Gray, J., & Rumpe, B. (2015). Models for digitalization. Software & Systems Modeling, 14(4), 1319-1320.

[2] Fernández Sosa, J. F. (2021). Utilización de dispositivos móviles como herramienta de sensado en aplicaciones de IoT (Tesis, Universidad Nacional de La Plata).

[3] Huang, H., & Gartner, G. (2018). Current trends and challenges in location-based services. ISPRS International Journal of Geo-Information, 7(6), 199.

[4] Steiniger, S., Neun, M., & Edwardes, A. (2006). Foundations of location based services. Lecture Notes on LBS, 1(272), 2.

[5] Lindh, J. (2015). Bluetooth low energy beacons. Texas Instruments, 2.

[6] Spachos, P., & Plataniotis, K. (2018). Beacons and the City: Smart Internet of Things. In Cooperative and Graph Signal Processing (pp. 757-776). Academic Press.

[7] Spachos, P., & Plataniotis, K. N. (2020). BLE beacons for indoor positioning at an interactive IoT-based smart museum. IEEE Systems Journal, 14(3), 3483-3493.

[8] https://ciytt.info.unlp.edu.ar/

[9] Jeon, K. E., She, J., Soonsawad, P., & Ng, P. C. (2018). Ble beacons for internet of things applications: Survey, challenges, and opportunities. IEEE Internet of Things Journal, 5(2), 811-828.

[10] Chang, K. H. (2014). Bluetooth: a viable solution for IoT?[Industry Perspectives]. IEEE Wireless Communications, 21(6), 6-7.

[11] Akinsiku, A., & Jadav, D. (2016, April). BeaSmart: A beacon enabled smarter workplace. In NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium (pp. 1269-1272). IEEE.
[12] Bilolo, A., Boeck, H., Durif, F., & Levesque, N. (2015). The impact of proximity marketing on consumer reactions and firm performance: A conceptual and integrative model.
[13] Muddinagiri, R., Ambavane, S., Jadhav, V., & Tamboli, S. (2020, March). Proximity Marketing Using Bluetooth Low Energy. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 856-861). IEEE.

Short Papers – Alumnos

Coordinadores

Armando De Giusti (UNLP)

Mónica Tugnarelli (UNER)

Marcelo Estayno (UNSAM)

DomoHome: Un Sistema Domótico Inteligente

Delfina Fenocchio[1], Marco Popovich[1], Melisa Kuzman[2], and Raúl Rivera[2]

[1]Universidad CAECE, Departamento de Sistemas, Mar del Plata
[2]Universidad Nacional de Mar del Plata, Departamento de Electrónica y Computación, Mar del Plata
{delfifenocchio,m.popovich20}@gmail.com
{melisakuzman,rrivera}@fi.mdp.edu.ar

Resumen En este trabajo de proyecto final de carrera se muestra el desarrollo de un sistema IoT para aplicaciones en domótica para configurar y monitorear los procesos en forma remota desde una aplicación móvil. El objetivo es aportar una dinámica versátil y escalable al automatizar y controlar diferentes tareas del hogar, permitiendo implementar servicios como el control de iluminación, detección de presencia, nivel del tanque de suministro de agua e información sobre temperatura y humedad en las habitaciones, así como también el registro de estos parámetros en una base de datos para su posterior visualización y análisis. Se describen brevemente las herramientas para crear aplicaciones interactivas y los componentes utilizados para su desarrollo y experimentación, compuestos por un microcontrolador con conectividad WiFi, un servidor Broker MQTT en la nube y una base de datos de tiempo real.

Keywords: IoT, domótica, Internet, microcontrolador, servidor, MQTT

1. Introducción

La cantidad de dispositivos inteligentes que nos rodean crece rápidamente y, con ello, la necesidad de interconectarlos para compartir datos. Cada día surgen nuevos escenarios y nuevas aplicaciones que hacen la vida más sencilla en muchos contextos diferentes. Esta necesidad de conectar los dispositivos crea interés por establecer comunicaciones e interacción entre ellos. Así es como aparece el término "Internet of Things (IoT)", Internet de las Cosas. Esta filosofía busca interconectar digitalmente el mundo de las cosas por medio de dispositivos, programas y plataformas de diversos tipos, conformando redes de intercambio de datos entre objetos para realizar operaciones de manera automatizada. Se trata de mejorar, sofisticar y hacer más eficiente el funcionamiento de dispositivos que originalmente funcionaban sin conectarse con otros aparatos u objetos y centros de control digitales[1].

A partir del auge de estas nuevas tecnologías, se propone diseñar y desarrollar un sistema de módulos IoT para el monitoreo, control y registro de procesos ambientales y de seguridad, mediante el uso de actuadores y sensores, para el

intercambio de información en red que cumplan con los requerimientos de automatización, acceso y seguridad en aplicaciones de domótica de acceso remoto[2]. Permite la implementación de servicios tales como control de iluminación, detección de presencia, nivel del suministro de agua e información de temperatura y humedad en las habitaciones.

2. Arquitectura del sistema

El sistema implementado se encuentra conformado por diferentes componentes que brindan funcionalidades o capacidades al sistema. En la Figura 1 se observa un esquema que representa la arquitectura simplificada de DomoHome. Como punto central de este desarrollo se encuentra el microcontrolador (μC) ESP8266[3], el cual provee la adquisición de información de los sensores y el control de los actuadores, también presentes en el esquema. Una de las características más relevantes de este μC, es la disponibilidad de conectividad WIFI integrada en el chip que facilita su conexión a Internet. Por su parte, los usuarios se encuentran representados por un dispositivo móvil bajo la leyenda *cliente*, quienes a través del uso la aplicación desarrollada pueden monitorear el estado del hogar y controlar los actuadores disponibles. La comunicación que se establece entre estos dispositivos móviles y la plataforma IoT es a través del protocolo MQTT[4], cuyo desarrollo es presentado en 2.3.

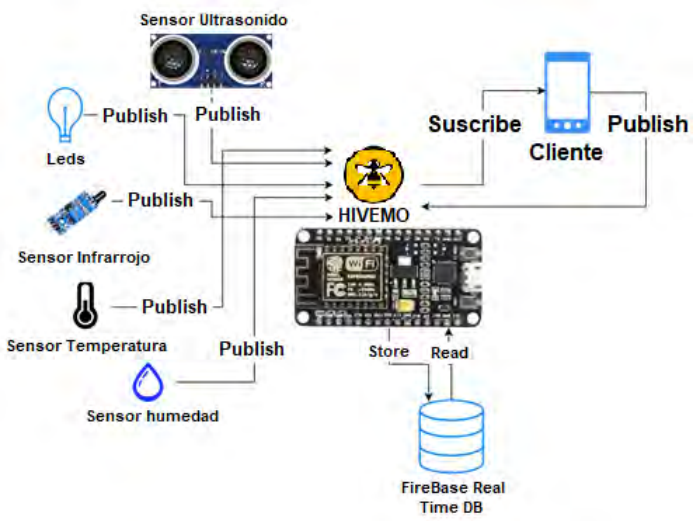


Figura 1. Esquema del sistema implementado

Para este proyecto se utiliza HiveMQ[5], una plataforma que provee un servidor Broker MQTT en la nube para el intercambio de mensajería gratuita hasta

100 clientes. Este protocolo tiene la capacidad para trabajar con un gran número de clientes y establecer comunicaciones cifradas que aportan a la red una capa extra de seguridad. Para la estructura de la base de datos se utiliza FireBase[6], una base de datos en tiempo real NoSQL que almacena los datos de configuración básicos de la aplicación y el registro de los usuarios.

2.1. Plataforma IoT

Para adquirir, procesar y comunicar los diferentes estados del sistema domótico, se elige el μC ESP8266 que posee las interfaces y periféricos necesarios para este proyecto. Su memoria Flash se utiliza para contener el código del programa, pero también para mantener los datos de la configuración del sistema cada vez que se cambian. Como este dispositivo no cuenta con una memoria EEPROM integrada como en la mayoría de los μCs, se utiliza una zona dedicada en Flash para guardar los últimos datos de configuración recibidos, con el fin de que el sistema se encuentre actualizado ante el posible caso de una desconexión.

El *firmware* se desarrolla en lenguaje `C` y `C++`, determinando allí el comportamiento del sistema ante ciertos eventos, pero también contiene el código para configurar las interfaces del chip. Cabe remarcar que el usuario tiene la posibilidad de agregar o quitar sensores a través de la aplicación, pero siempre considerando las limitaciones de la plataforma. En este sentido, la funcionalidad de los puertos de entrada y salida se encuentran definidas y fueron diseñadas tomando como referencia una casa de tres ambientes. Con cada plataforma se pueden agregar hasta:

- Tres sensores de temperatura y humedad (DTH11 o DTH22 con comunicación one wire).
- Cuatro sensores de presencia (o sensores de aberturas).
- Cuatro controles de luminaria con salidas de relay optoacopladas.
- Un sensor para el nivel del tanque de agua.
- Una salida con una intefaz de relay optoacoplada para el control de la bomba.

De necesitar un número mayor de entradas y/o salidas, o que los ambientes sean muy lejanos entre sí, se pueden agregar módulos adicionales con las mismas prestaciones. En cuanto al puerto serie, se reserva para depuración y reprogramación.

En la Figura 2 se presenta la configuración que fue utilizada para el desarrollo del prototipo de la plataforma IoT: una placa de pruebas (también conocida como protoboard), el μC, los sensores y los actuadores. Si bien esta imagen no reproduce la instalación de DomoHome dentro de una casa, permite discernir con mayor facilidad los distintos componentes que conforman al sistema. De izquierda a derecha se pueden observar: un sensor de presencia por infrarrojo, un led que representa el control de un relay para activar la iluminación, un sensor de humedad y temperatura DTH11[7] y un sensor de ultrasonido para detectar el nivel de agua presente en el tanque. En la parte superior, y sobre la derecha de la imagen se encuentra el microcontrolador.

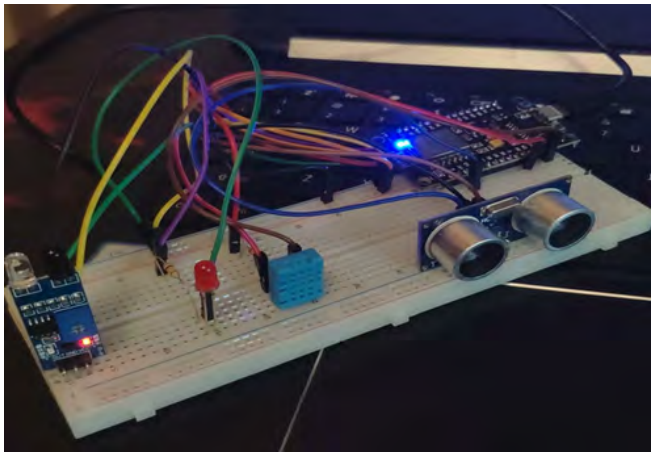


Figura 2. Sistema montado en una protoboard

2.2. Aplicación

Para el diseño de la aplicación se utilizó “Marvel App”, siendo esta una herramienta para crear aplicaciones interactivas de plataformas digitales. Este prototipo es la versión “alpha” de la App, y sus vistas interactivas se encuentra disponibles en línea para su visualización[8]. En la Figura 3 se observa que la aplicación diseñada posee una etapa de inicio de sesión para que cada usuario pueda acceder a sus datos de forma segura. A través de las diferentes vistas que posee la aplicación, cada usuario puede personalizar el nombre del hogar, como así también los sensores y actuadores que quiere utilizar (considerando que están instalados físicamente), o simplemente eliminarlos.

En la Figura 4 se presentan las pantallas más relevantes para el usuario, una vez realizado el ingreso. En primer lugar se encuentran todos los datos disponibles con los sensores instalados en la ”Habitación 1”. Si se presiona el botón + que se encuentra en la parte inferior, la App permite agregar otro sensor según la necesidad del usuario. En la última pantalla se presenta la información que corresponde al tanque de agua instalado.

La información requerida para el monitoreo y control de un tanque de agua es la capacidad en litros, altura y diámetro del mismo (el cálculo es solamente válido para tanques tradicionales, cilíndricos verticales). Aquí el usuario puede definir los valores máximos y mínimos para apagar o encender la bomba, respectivamente, como también una alarma de seguridad en el caso que se supere alguno de esos límites.

2.3. Protocolo de comunicación con el usuario

Para comunicar a la plataforma IoT con los diferentes dispositivos del usuario se utiliza el protocolo MQTT, muy difundio en estas aplicaciones[9]. Es una

Figura 3. Inicio de sesión

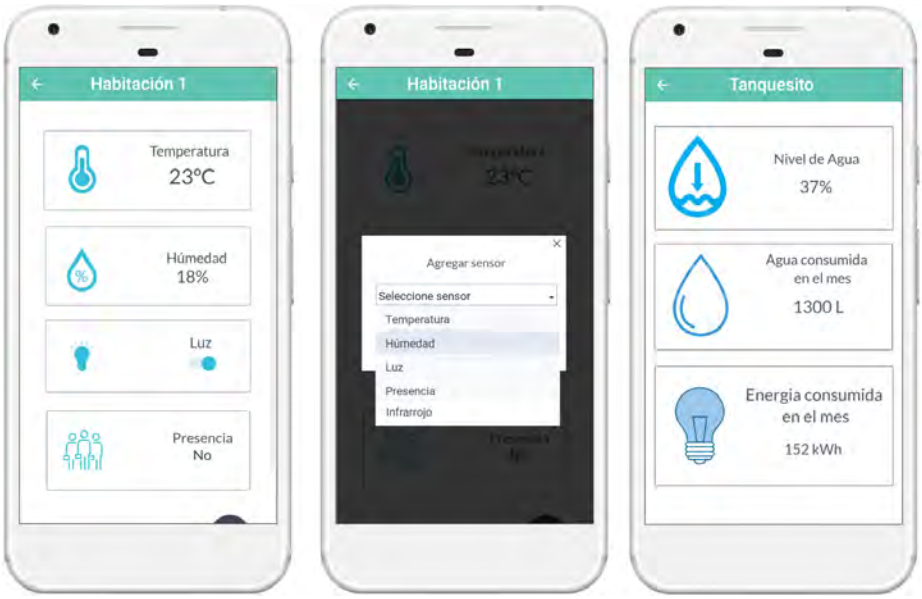


Figura 4. Pantallas más relevantes

comunicación de mensajería de publicación/suscripción, extremadamente simple y liviana, diseñada para dispositivos restringidos y redes de bajo ancho de banda.

En este protocolo existen dos componentes, un servidor llamado *broker* quien es el responsable de recibir y distribuir la información y los clientes que publican y/o se suscriben a diversas etiquetas que se los denominan *topics*. Estas etiquetas deben ser conocidas por los clientes que quieran recibir la información, la cual se envía a los suscriptores conjuntamente con el tema dentro del campo denominado *payload*. La autenticación utilizada es parte de los niveles de seguridad a nivel transporte (TLS) y de aplicación. Por un lado se encuentra el certificado de la validación del cliente al servidor, mientras que a nivel aplicación se requiere de un usuario y su contraseña. Para ello, cada cliente envía un mensaje del tipo CONNECT, y es el broker quién evalua la credencial y responde con otro mensaje del tipo CONNACK: conexión aceptada, conexión rechazada por usuario y contraseña errónea o conexión rechazada por falta de autorización.

Para el broker MQTT en la nube se utiliza HiveMQ debido a su sencillez de implementación. Su plan gratuito permite configurar la instancia broker que se ejecuta en sus servidores, y gracias a eso se puede tener una red en línea rápidamente. Además posee una UI (Interfaz de Usuario) para monitorear los procesos, *topics* de publicación y suscripción.

2.4. Base de datos

FireBase es una plataforma digital que se utiliza para facilitar el desarrollo de aplicaciones web o móviles de una forma efectiva, rápida y sencilla. Su principal objetivo, es mejorar el rendimiento de las apps mediante la implementación de diversas funcionalidades que van a hacer de la aplicación en cuestión, mucho más manejable, segura y de fácil acceso para los usuarios. Una base de datos en tiempo real es un sistema de base de datos que utiliza el procesamiento en tiempo real para manejar cargas de trabajo cuyo estado cambia constantemente, como en este proyecto. Esto difiere de las bases de datos tradicionales que contienen datos persistentes, en su mayoría no afectados por el tiempo.

La estructura de la base de datos diseñado se presenta en la Figura 5. Aquí los clientes se conectan a la base de datos y mantienen una conexión bidireccional abierta a través de websockets. Luego, si algún cliente envía datos a la base de datos, se activa y en este caso, informa a todos los clientes conectados los datos recién guardados. Esta forma de trabajo es similar al broker MQTT y cómo reacciona cuando recibe un mensaje de un editor y lo envía a todos los suscriptores. La diferencia esta vez es la adición de la parte persistente de datos, que es la base de datos. Por lo tanto no es necesario enrutar los mensajes utilizando otros protocolos: Firebase Realtime Database se encarga de eso además de realizar su función de base de datos normal.

De esta manera se puede conectar el dispositivo a la Base de datos en tiempo real de Firebase y enviar datos periódicamente a la base de datos. En la otra parte del sistema, existe una aplicación web que se conecta al mismo servicio que el dispositivo y recibe nuevos datos cada vez que haya un cambio en la base de datos.

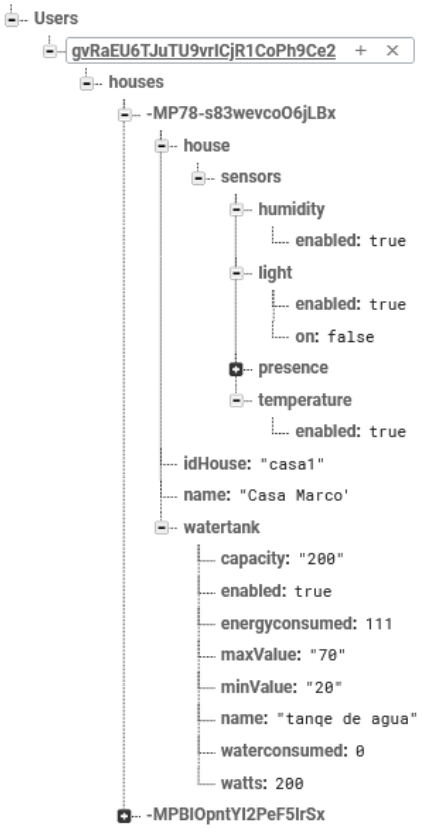


Figura 5. Estructura de la Base de datos implementada

3. Conclusiones

El Sistema Domótico con Módulos Inteligentes IoT desarrollado en este proyecto ofrece una aplicación accesible para usuarios que dispongan con dispositivos basados en Android y que tengan interés en un sistema de acceso remoto IoT. Esta tecnología posee librerías de conexión en distintos S.O. por lo cual es fácilmente adaptable para dispositivos móviles Apple, o a través de un servidor web accesible desde cualquier navegador. El trabajo consistió en la elección y desarrollo del hardware, software, protocolo de comunicación y la estructura de la base de datos de tiempo real, esta última seleccionada en base a la carga de trabajo variable en el tiempo que característica a este aplicación. Con este sistema se automatiza un hogar de una manera accesible e intuitiva, recopilando datos de temperatura, humedad, presencia, accionamiento de luces y monitoreo remoto del suministro de agua.

A partir de este desarrollo surgieron nuevos proyectos con el objetivo de incorporar nuevas funcionalidades dentro de un hogar o edificio, adaptando al

sistema a los requerimientos del usuario tales como, control de apertura de persianas, acceso al hogar o aquellas orientadas a la seguridad como detección de monóxido para distintos ambientes y monitoreo del consumo eléctrico para el uso eficiente de la energía.

DomoHome permitió además contar con una plataforma para experimentar con la interacción entre objetos y arquitectura escalable que caracterizan a IoT.

Referencias

1. V. Miori and D. Russo, "Domotic Evolution towards the IoT,"2014 28th International Conference on Advanced Information Networking and Applications Workshops, 2014, pp. 809-814, doi: 10.1109/WAINA.2014.128.
2. V. Miori, D. Russo and L. Ferrucci, Ïnteroperability of home automation systems as a critical challenge for IoT,"2019 4th International Conference on Computing, Communications and Security (ICCCS), 2019, pp. 1-7, doi: 10.1109/CCCS.2019.8888125.
3. ESPRESSIF SYSTEMS, "Hoja de datos ESP8266", https://www.espressif.com/en/products/socs/esp8266.
4. MQTT.org, "Protocolo MQTT", https://mqtt.org/
5. HiveMQ GmbH, "HiveMQ broker nativo en la nube", https://www.hivemq.com/
6. "Base de datos Firebase", https://firebase.google.com/?hl=es
7. Hoja de datos Sensor DHT11, http://www.datasheet.es/PDF/792210/DHT11-pdf.html
8. "Aplicación prototipo", https://marvelapp.com/ef0gh30/screen/70598433
9. B. Mishra and A. Kertesz, "The Use of MQTT in M2M and IoT Systems: A Survey,ïn IEEE Access, vol. 8, pp. 201071-201086, 2020, doi: 10.1109/ACCESS.2020.3035849.

Análisis Visual para Datos Abiertos Enlazados vinculados a las Ciencias del Mar

Gustavo Nuñez[1], Carlos Buckle[1[0000−0003−0722−0949]], Marcos Zárate[1,2[0000-0001-8851-8602]]

[1] Laboratorio de Investigación en Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco (LINVI-UNPSJB), Puerto Madryn, Argentina.
[2] Centro para el Estudio de Sistemas Marinos, Centro Nacional Patagónico, Consejo Nacional de Investigaciones Científicas y Técnicas (CESIMAR-CENPAT-CONICET), Puerto Madryn, Argentina.
gnunez@ing.unp.edu.ar, cbuckle@unpata.edu.ar, zarate@cenpat-conicet.gob.ar

Resumen: El propósito de la exploración y visualización de datos (DV) es ofrecer formas de percibir y manipular información, así como extraer e inferir conocimiento. En este breve artículo presentamos avances en representaciones visuales y técnicas de interacción intuitivas basadas en inteligencia artificial. Esto contribuye significativamente a la exploración y comprensión de la información relacionada con las ciencias marinas representada por ontologías y datos enlazados. Esta investigación preliminar llevada adelante desde 2018 entre LINVI-UNPSJB y CESIMAR-CENPAT-CONICET permitirá a los científicos y usuarios no expertos analizar conjuntos de información relacionada con la oceanografía, meteorología y parámetros ambientales, con el fin de promover el conocimiento científico y la innovación productiva en el océano Atlántico Sur utilizando Datos Abiertos Enlazados (LOD por siglas en inglés).

Palabras clave: Visualización de Datos · Datos Abiertos Enlazados · Ciencias Marinas.

1 Introducción

El propósito de la DV es ofrecer formas de percibir y manipular la información, así como extraer e inferir conocimiento [1,2]. La DV proporciona a los usuarios una manera intuitiva de explorar el contenido, identificar patrones de interés e inferir correlaciones y causalidades, además de brindar un gran aporte a las actividades de construcción de significado. Uno de los enfoques más prometedores para abordar la problemática asociada con la integración y la representación gráfica es almacenar los datos de forma estructurada y reproducirlos mediante gráficos. La Web Semántica (SW) [3] ofrece soluciones a estas necesidades al utilizar LOD [4], en donde las entidades se identifican de forma única y las relaciones entre ellas se especifican explícitamente. LOD es un enfoque potente que permite difundir y consumir datos

científicos de varias disciplinas [5,6,7,8]. Implica publicar, compartir y conectar datos en la Web.

En los últimos años, esta forma de publicar datos ha sido adoptada en un gran número de disciplinas LOD [9]. Esto ha hecho que la visualización y exploración de información sea una tarea crucial para la mayoría de los consumidores LOD. Científicos de datos, expertos del dominio y usuarios no experimentados buscan maneras intuitivas y visuales de interactuar con estos recursos. En el campo de las ciencias marinas, la visualización de datos en disciplinas como Oceanografía, Meteorología y la Biodiversidad enfrentan grandes desafíos, ya que existe un aumento exponencial de su volumen debido al crecimiento de las tecnologías y la multiplicidad de plataformas, además de la demanda de conocimiento para contribuir globalmente a los distintos modelos que buscan combatir el cambio climático [10]. Además, existe una gran diversidad en el tipo de registros que deben mostrarse adecuadamente, las características físicas, químicas, geológicas, meteorológicas y los valores biológicos deben integrarse correctamente, y los productos de análisis/información deben basarse en todos ellos para que el usuario pueda hacer una interpretación correcta [11].

El resto de este documento breve está estructurado de la siguiente manera: la Sección 2 presenta diferentes iniciativas basadas en LOD para las ciencias marinas. La sección 3 presenta una prueba conceptual de una plataforma desarrollada para visualizar información relacionada a las ciencias marinas en el Atlántico Sur. Finalmente, en la sección 4, presentamos algunas conclusiones basadas en experiencias y la planificación de trabajos futuros.

2 Antecedentes y trabajos relacionados

En los últimos años se han introducido una gran cantidad de herramientas de visualización de LD, la mayoría provenientes del ámbito académico. Las herramientas de DV en datos enlazados proporcionan representaciones gráficas de un conjunto de datos o partes de él, con el fin de facilitar su análisis y generación de conocimientos a partir de información compleja e interrelacionada en tiempo y espacio. Las técnicas pueden variar según el dominio, el tipo de registro, la tarea que el usuario está tratando de realizar, así como las habilidades del usuario.

Son varias las iniciativas que se llevan a cabo en el contexto argentino para publicar datos de ciencias marinas como LD, entre ellos podemos destacar: [12] que presenta la publicación de metadatos de campañas oceanográficas como LD. OceanGraph [13] define un prototipo de gráfico de conocimiento oceanográfico para gestionar información de expediciones, publicaciones científicas y variables ambientales, mientras que en [14] se propone la explotación de OceanGraph con ejemplos concretos de posibles usos por especialistas.

A nivel internacional también existen iniciativas para la publicación de información marina de datos científicos como LD, entre los principales podemos mencionar GeoLink [15], un proyecto financiado por la iniciativa EarthCube, que ha aprovechado la principios de LOD para crear una base de datos que permite a los usuarios realizar consultas en algunos de los repositorios de geociencias más destacados de los Estados Unidos. El conjunto de datos de GeoLink incluye información tan diversa como escalas

en puertos realizadas por cruceros oceanográficos, metadatos de muestras físicas, financiación de proyectos de investigación y personal, y autoría de informes técnicos. Los datos han sido publicados de acuerdo con las mejores prácticas para LOD [16], y están disponibles públicamente a través de un endpoint SPARQL[1] que actualmente contiene más de 45 millones de tripletas RDF.

3 Visualización LD en Ciencias Marinas

En el contexto de las ciencias marinas, la exploración visual es un enfoque prometedor para explorar y analizar datos y comprender mejor la dinámica de los complejos procesos oceánicos. Aunque la publicación de datos como LD tiene varios casos de éxito [15,17], la visualización sigue siendo un problema porque es una tarea que difiere del clásico DV, principalmente debido a las características de LD. Los usos de vocabularios comunes (dominios cruzados) para la descripción de los registros, o el uso de propiedades tipificadas para capturar las relaciones entre los recursos dentro un conjunto o entre diferentes conjuntos, difieren de las formas tradicionales de visualización que son incapaces de captar las complejas relaciones posibles. Para las pruebas descritas a continuación, utilizamos información pública sobre especies marinas y variables capturadas en el Atlántico Sur a través de un endpoint SPARQL cuya URL es http://linkeddata.cenpat-conicet.gob.ar/snorql/ . La metodología utilizada para la creación y publicación se detalla en [17].

3.1 Casos de estudio

Nuestro enfoque está puesto en el front-end basado en la Web, más precisamente en herramientas de consultas y visualización. Hemos desarrollado una prueba de concepto para la visualización interactiva de información oceanográfica, ambiental y de biodiversidad marina. La plataforma permite la representación y visualización de mapas interactivos con trayectorias de buques oceanográficos, y la recuperación de esquemas gráficos de la relación entre variables ambientales y especies. Para ello, se llevó a cabo una selección de herramientas de código abierto compatibles con la visualización de tipos específicos de información, por ejemplo, datos geoespaciales, distribución de especies, trazabilidad y registros relacionados con el medio ambiente. La Figura 1 muestra dos visualizaciones utilizadas para interpretar información sobre una especie específica, en este caso Mirounga Leonina (elefante marino del sur). El mapa muestra la información de los viajes realizados por varios individuos durante sus viajes de alimentación en el mar, superponiéndose adicionalmente capas con información ambiental y especial. La otra visualización muestra información bibliográfica asociada a la especie.

[1] http://data.geolink.org/sparql

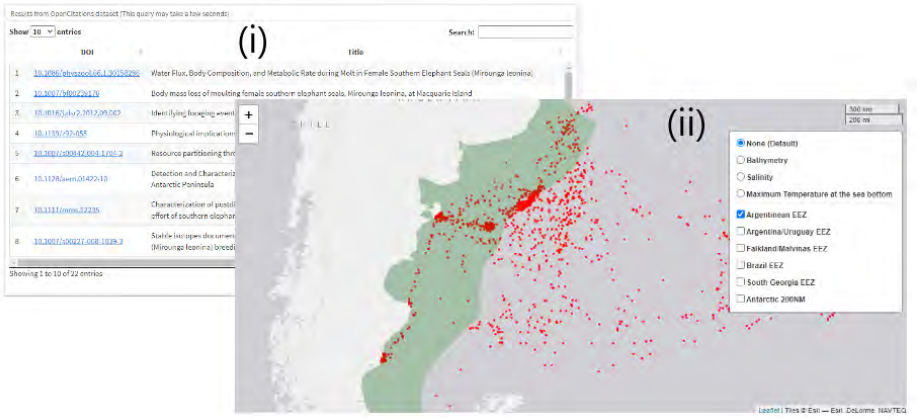


Fig. 1. Visualizaciones utilizadas para relacionar: (i) especies marinas con información bibliográfica (ii) información geoespacial de especies con variables ambientales y regiones marinas.

4 Conclusiones y trabajos futuros

De experiencias anteriores, podemos concluir que resulta necesario desarrollar sistemas que sean capaces de gestionar visualmente la información para usos integrales y secundarios, tanto de los colectivos participantes como de usuarios externos que requieren información. Los resultados de esta investigación preliminar constituyen un aporte sustancial, no solo para las ciencias del mar, sino también como aporte metodológico a visualizaciones científicas usando LD.

Como trabajo futuro, se plantea la necesidad de formalizar la prueba de concepto. Para ello es necesario profundizar en los siguientes aspectos: a) Estudio e investigación de visualizaciones científicas típicas de las ciencias marinas. b) Desarrollar una plataforma de base de datos en línea para la visualización basada en modelos de predicción, con el fin de proporcionar herramientas visuales analíticas y permitir consultas y análisis interactivos de diferentes capas de información. c) Ampliar la plataforma o escalar los resultados a otros espacios marinos, en particular a las Áreas Geográficas Prioritarias (AGP) de la iniciativa Pampa Azul.

Referencias

1. Jeffrey Heer y Ben Shneiderman. Interactive dynamics for visual analysis. *Communications of the ACM*, 55(4):45–54, 2012.
2. Stratos Idreos, Olga Papaemmanouil, y Surajit Chaudhuri. Overview of data exploration techniques. En *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, páginas 277–281, 2015.
3. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

4. Christian Bizer, Tom Heath, y Tim Berners-Lee. Linked data: The story so far. En *Semantic services, interoperability and web applications: emerging concepts*, páginas 205–227. IGI Global, 2011.
5. Richard K Lomotey y Ralph Deters. Terms extraction from unstructured data silos. En *System of Systems Engineering (SoSE), 2013 8th International Conference on*, páginas 19–24. IEEE, 2013.
6. Syed Ahmad Chan Bukhari, Mate Levente Nagy, Paolo Ciccarese, Michael Krauthammer, y Christopher JO Baker. icyrus: A semantic framework for biomedical image discovery. En *SWAT4LS*, páginas 13–22, 2015.
7. Syed Ahmad Chan Bukhari. *Semantic enrichment and similarity approximation for biomedical sequence images*. PhD thesis, University of New Brunswick (Canada), 2017.
8. Roderic D.M. Page. Ozymandias: a biodiversity knowledge graph. PeerJ, 7:e6739, April 2019.
9. The open linked data cloud. https://lod-cloud.net/, 2021. [Online; accessed 3-May-2021].
10. Tanu Malik y Ian Foster. Addressing data access needs of the long-tail distribution of geoscientists. En *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, páginas 5348–5351. IEEE, 2012.
11. Alex Hardisty y Dave Roberts. A decadal view of biodiversity informatics: challenges and priorities. *BMC ecology*, 13(1):16, 2013.
12. Marcos Zárate, Pablo Rosales, Pablo Fillottrani, Claudio Delrieux, y Mirtha Lewis. Oceanographic data management: Towards the publishing of pampa azul oceanographic campaigns as linked data. En *Proceedings of the 12th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW 2018)*, 2018.
13. Marcos Zárate, Pablo Rosales, Germán Braun, Mirtha Lewis, Pablo Rubén Fillottrani, and Claudio Delrieux. Oceangraph: Some initial steps toward an oceanographic knowledge graph. In Boris Villazón-Terrazas and Yusniel Hidalgo-Delgado, editors, *Knowledge Graphs and Semantic Web*, páginas 33–40, Cham, 2019. Springer International Publishing.
14. Marcos Zárate, Carlos Buckle, Renato Mazzanti, Mirtha Lewis, Pablo Fillottrani, y Claudio Delrieux. Harmonizing big data with a knowledge graph: Oceangraph kg uses case. En Enzo Rucci, Marcelo Naiouf, Franco Chichizola, and Laura De Giusti, editors, *Cloud Computing, Big Data & Emerging Topics*, páginas 81–92, Cham, 2020. Springer International Publishing.
15. Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, y Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2018.
16. Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of Linked Data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.
17. Marcos Zárate, Germán Braun, Mirtha Lewis, and Pablo Fillottrani. Observational/hydrographic data of the south atlantic ocean published as lod. *Semantic Web*, 13(2):133–145, 2022.

Hacia el análisis de tesis de grado de carreras informáticas de la UM mediante minería de textos

Gabriel Mariuz[1] Iris Sattolo[1], Marisa Panizzi[1]

[1]Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias Universidad de Morón. Cabildo 134 (B1708JPD), Partido de Morón, Argentina.
gmariuz91@gmail.com, iris.sattolo@gmail.com, marisapanizzi@outlook.com

Resumen. La categorización de documentos de textos es una aplicación de la minería de textos que pretende extraer información de texto no estructurado o semi estructurado. La justificación de su aplicación se debe a que se estima que alrededor del 80% de los datos de las organizaciones son no estructurados. El presente trabajo de tesis de la carrera Licenciatura de Sistemas de la UM pretende analizar los títulos de las tesis realizadas en la cátedra para categorizarlas según su área temática mediante minería de textos y evaluar la eficacia de la técnica utilizada al hacerlo. Antes de comenzar con la construcción de modelos de minería de textos, se construyó el estado del arte mediante un mapeo sistemático de la literatura (en inglés, *systematic mapping study* o SMS). Se presentan los resultados logrados mediante el desarrollo del SMS y se describen las actividades definidas para la finalización del trabajo de tesis.

Palabras claves: Minería de textos, categorización, aprendizaje automático, tesis de grado, carreras de informática.

1 Introducción

La cantidad de documentos de diversos tipos disponibles en una organización o establecimiento es enorme y continúa creciendo cada día. Estos documentos son a menudo un repositorio fundamental del conocimiento de la organización, pero a diferencia de éstas la información no está estructurada. La minería de textos tiene como objetivo extraer información de texto no estructurado, tal como entidades (personas, organizaciones, fechas, cantidades) y las relaciones entre ellas. La categorización de documentos de texto es una aplicación de la minería de texto que asigna a los documentos una o más categorías, etiquetas o clases, basadas en el contenido.

El enfoque tradicional para la categorización de textos en que los expertos en el dominio de los textos definían manualmente las reglas de clasificación ha sido reemplazado por otro basado en técnicas de aprendizaje automático, o en combinaciones de éste con otras técnicas [1].

Actualmente, las cátedras de tesis del área de Informática en la Universidad de Morón cuentan con un archivo en formato xls que contiene los datos referidos a las tesis realizadas en las carreras informáticas desde el año 2004 hasta la actualidad. Este archivo cuenta, entre otros datos, el título de la tesis, su resumen, el área temática a la que

corresponde cada tesis basada en las áreas propuestas en los workshops del Congreso CACIC[1] organizado por la RedUNCI[2]. Con el objetivo de categorizar cada una de las tesis según su área temática se planteó utilizar técnicas de minería de textos para generar patrones, además esto permitirá validar si la categorización automatizada es similar a la realizada manualmente y determinar su eficacia.

Con el fin de conocer el estado del arte con respecto al uso de la minería de textos, en el dominio educativo, es que se ha realizado un mapeo sistemático de la literatura de acuerdo con los directrices propuestas por Kitchenham *et al.* [2].

El artículo se estructura de la siguiente manera: en la Sección 2 se describe la planificación del SMS, en la Sección 3 se describe su ejecución. Los resultados se presentan en la Sección 4. En la Sección 5 se exponen las conclusiones y trabajos futuros.

2 Planificación del SMS

Se presenta la definición del protocolo de revisión del SMS: preguntas de investigación (PI), estrategia de búsqueda, selección de los estudios, criterios y proceso de selección, formulario de extracción y el proceso de síntesis de los datos.
El objetivo de este SMS es responder la siguiente pregunta de investigación (PI):
¿Qué trabajos existen de la aplicación de minería de textos para la categorización de tesis?

Esta pregunta principal se descompone en un conjunto de sub-preguntas (PI1-4), las cuales se presentan en la Tabla 1 junto con su motivación.

Tabla 1. Preguntas de investigación (PI) y su motivación.

Pregunta de investigación (PI)	Motivación
PI1: ¿Qué técnicas y algoritmosson los más utilizados en minería de textos?	Identificar las técnicas y algoritmos más usados en minería de textos para categorizar documentos.
P2: ¿Con qué herramientas y lenguajes de programación se trabaja en la minería de textos?	Identificar las herramientas y lenguajes de programación más utilizados para la categorización dentro de la minería de textos.
PI3: ¿Qué metodologías y procesos son utilizadas en la minera de textos?	Identificar las metodologías y procesos más utilizados en minería de textos.
PI4: ¿Qué tipos de investigación se encuentra en los artículos?	Identificar los tipos de investigación realizada en los estudios primarios de acuerdo con la clasificación propuesta por Wieringa *et al.* [3].

[1] CACIC: Congreso Argentino en Ciencias de la Computación.
[2] RedUNCI: Red de Universidades con carreras informáticas. Disponible: https://redunci.info.unlp.edu.ar/

Se definieron para la búsqueda de artículos las siguientes librerías, plataformas y repositorios digitales: SEDICI[3], *Scielo, Dialnet*, *Google Schoolar* y *ScienceDirect*, considerando artículos de congresos y artículos de revistas. El período comprendido definido ha sido entre julio de 2005 hasta marzo de 2022.
La cadena de búsqueda resultante es:

(Minería de datos) OR (Data Mining) OR (Minería de texto) OR (Text Mining) (Minería de texto en la educación) OR (Text Classification) OR (Educational Text Mining) OR (Clasificación de texto)

En la Tabla 2, se presentan los criterios de inclusión y exclusión utilizados para el proceso de selección de artículos.

Tabla 2. Criterios de inclusión y exclusión.

Criterios de inclusión	Criterios de exclusión
Artículos que respondan a las preguntas de investigación.	Artículos que no estén accesibles para su lectura completa.
Artículos publicados a partir de julio de 2005 hasta marzo de 2022.	Literatura gris, tesis doctorales, presentaciones en PowerPoint.
Artículos preferentemente en español ya que se busca conocer el estado y las investigaciones realizadas en aquellos países de habla hispana.	

El proceso de selección de los estudios consistió en realizar la búsqueda en las fuentes definidas aplicando la cadena en el título y/o en el resumen, para luego eliminar los artículos duplicados y aplicar los criterios de inclusión y exclusión en el título, resumen y palabras clave, después se aplicaron los criterios de inclusión y exclusión al texto completo.

Para dar respuesta a cada una de las preguntas de investigación (PI) se definió un esquema de clasificación que junto con el formulario de extracción de datos se presenta en un apéndice por restricciones de espacio [4].

3 Ejecución y resultados del SMS

Se encontraron 27 artículos de los cuales se analizaron 15 estudios primarios que se encuentran en el apéndice [4]. Los resultados del SMS para dar respuesta a las preguntas de investigación en base a la literatura analizada mediante gráficos.

PI1: ¿Qué técnicas y algoritmos son los más utilizados en minería de textos?

Los algoritmos más utilizados en minería de textos para la categorización de documentos son las redes neuronales artificiales, siendo además de las más precisas, comparadas a las redes bayesianas y a las máquinas de vectores de soporte (SVM). (Ver Figura 1).

[3] SEDICI: Repositorio Institucional de la Universidad Nacional de La Plata. Disponible en: http://sedici.unlp.edu.ar/

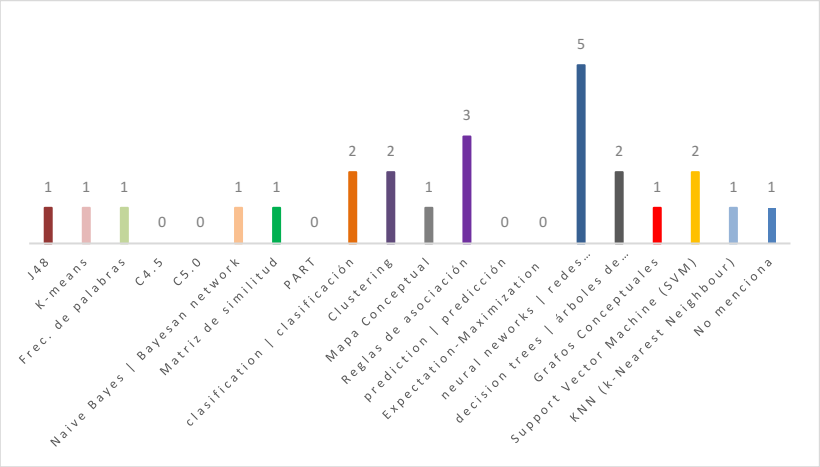


Figura. 1. Técnicas y algoritmos utilizados en la minería de textos.

P2: ¿Con qué herramientas y lenguajes de programación se trabaja en la minería de textos?

Se ha observado que existe variedad en el uso de herramientas y se halló que el uso de una u otra dependerá de cada usuario según ciertos valores como la usabilidad, potencia y versatilidad que pueda ofrecer dicha herramienta. Se evidenció que weka es la herramienta más usada mientras que Phyton y R los lenguajes más utilizados. Sin embargo, existe otra opción poco explorada debido a su reciente aparición como es GPT-3. Es un tipo de red neuronal que emplea aprendizaje automático y está enfocada en producir texto que simula la redacción humana y que, dada la cantidad de información disponible para su entrenamiento, tiene el potencial de ser usada para otras tareas para las que no fue pensada originalmente. Los resultados hallados se presentan en la Figura 2.

PI3: ¿Qué metodologías y procesos son utilizadas en la minera de textos?

En general, se puede evidenciar que la metodología más utilizada en la minería de textos es la metodología KDT (*Knowledge Discovery in Text*), una variante de KDD enfocada en el proceso de descubrimiento de conocimiento en texto. (Ver Figura 3).

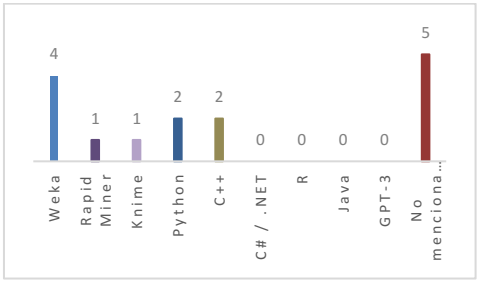


Figura. 2. Herramientas y lenguajes de programación utilizados en la minería de textos.

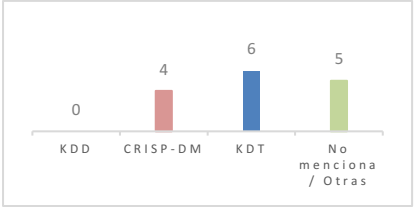


Figura. 3. Metodologías y procesos utilizados en la minería de textos.

PI4: ¿Qué tipos de investigación se encuentra en los artículos?
La mayor cantidad de los estudios analizados presentan la evaluación de los resultados de aplicar minería de texto (6 estudios) e informar las experiencias personales obtenidas al utilizarlo sobre un conjunto de datos (3 estudios). (Ver Figura 4).

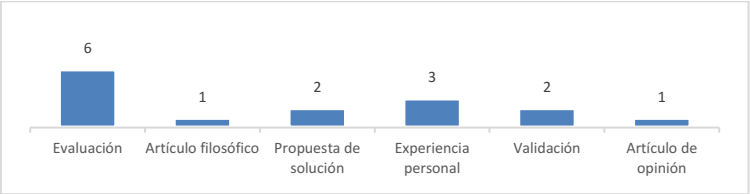


Figura. 4. Tipos de investigación de los artículos [4].

4 Conclusiones y trabajos futuros

Se logró construir el estado del arte respecto a la aplicación de la minería de textos para la categorización de las tesis de grado mediante el desarrollo de un SMS. Se analizaron 15 estudios primarios y se concluye que:

- En la mayoría de los estudios analizados se presentan principalmente propuestas de evaluación y en menor medida de informar una experiencia como tipo de investigación.
- Los algoritmos más utilizados son las redes neuronales.
- Las herramientas o lenguajes de programación más usados son Weka y Rapid Miner, mientras que, en menor medida, para los lenguajes de programación, son R y Phyton.
- La metodología más utilizada es KDT.

Como futuro trabajo para continuar el desarrollo de la tesis, se realizará: 1) Experimentación con la herramienta GPT-3 para conocer el alcance de sus capacidades, 2) Uso de la metodología KDT y, por último, 3) Evaluación de la categorización automática obtenida con la minería de textos para contrastarla con la categorización manual.

Referencias

[1] Abelleira, M., Cardoso, A. Categorización automática de documentos. XII Argentine Symposium on Artificial Intelligence (ASAI), 20-31. (2011).
[2] B. Kitchenham, D. Budgen y P. Brereton, Evidence-Based Software Engineering and Systematic Reviews, USA: CRC Press (2015).
[3] Wieringa, R., Maiden, N., Mead, N., Rolland, C. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. Requirements Engineering, 11(1), pp. 102-107 (2006).
[4] Mariuz G., Sattolo I., Panizzi M. Apéndice. Disponible en: https://doi.org/10.6084/m9.figshare.20514666.v1 (2022).

Buenas prácticas para la Seguridad Informática en PyMES

Tomás Alcántara, Marisa Panizzi[1], Iris Sattolo[1]

[1]Escuela Superior de Ingeniería, Informática y Ciencias Agroalimentarias Universidad de Morón. Cabildo 134 (B1708JPD), Partido de Morón, Argentina.
talcantara1995@gmail.com, marisapanizzi@outlook.com, iris.sattolo@gmail.com

Resumen. Actualmente la seguridad informática es uno de los principales desafíos dentro de las empresas ya que un ataque podría arruinar la reputación de esta ocasionando pérdidas a nivel económico y confiabilidad. A su vez dentro de la economía mundial muchas de las empresas son PyMES, con lo cual juegan un rol muy importante en la economía de cada país, así como también en el área laboral de los ciudadanos. El presente trabajo de tesis de la carrera Licenciatura en Sistemas de la UM pretende desarrollar un conjunto de buenas prácticas de seguridad informática para PyMES. Antes de comenzar con el diseño de la solución, se elaboró el estado del arte respecto a la seguridad informática en PyMES mediante un mapeo sistemático de la literatura (en inglés *systematic mapping study* o SMS). Se presentan los resultados del SMS y las actividades planificadas para la finalización de la tesis.

Palabras claves: Seguridad informática, PyMES, buenas prácticas, SMS.

1 Introducción

La seguridad informática es la disciplina que se ocupa de diseñar las normas, procedimientos, métodos y técnicas destinados a conseguir un sistema de información seguro y confiable [1].

La seguridad informática ha tomado un rol muy importante en el ámbito de la tecnología. La mayoría de las empresas y los organismos estatales comenzaron a incorporar personal especializado en seguridad en sus equipos ya que deben asegurar su infraestructura para disminuir la posibilidad de un ataque que perjudique su imagen corporativa o que sufran un perjuicio económico.

En ese sentido si hablamos de las de las organizaciones uno de los motores principales de las economías en los países son las PyMES. En Argentina según los últimos datos de la SEPYME hay más de 1.633.000 de empresas registradas bajo la modalidad de PyMES [2]. Muchas de estas empresas suelen brindar servicios a otras empresas u organismos, con lo cual en varias ocasiones son los puntos de entrada de ataques ya que al brindar servicios a otras empresas suelen tener ciertos privilegios sobre otras infraestructuras. Muchas veces es más simple y rápido realizar un ataque sobre las mismas PyMES que sobre las empresas de mayor tamaño, esto se debe a que en las empresas

chicas no suelen tener personal idóneo en el área de seguridad y muchas veces ven este área o soluciones de seguridad como un costo y no como una inversión.

La motivación principal de esta investigación es que durante el 2021 se duplicaron los ataques informáticos con respecto al 2020 según el Equipo de Respuesta ante Emergencias Informáticas Nacional (CERT, por su sigla en inglés) de la Dirección Nacional de Ciberseguridad de Argentina [3]. Con respecto a los números que manejan las principales empresas de seguridad se puede observar que según el informe de la empresa Kaspersky [4] al menos el 15,45% de los usuarios de internet fue afectado por malware.

Dada las altas tasas de ataques informáticos informados por organismos y empresas reconocidas, el trabajo de la tesis se focalizó en el diseño de un conjunto de prácticas a nivel seguridad informática que pueda ser utilizado en PyMES de Argentina Antes de comenzar con el diseño de la solución se realizó la construcción del estado del arte respecto a la seguridad informática en PyMES mediante un mapeo sistemático de la literatura (en inglés, *Systematic Mapping Studies* o SMS)

Dada la situación actual de las PyMES y del auge de la seguridad informática se decide realizar un análisis de la situación actual con respecto a la problemática de seguridad informática en PyMES Para realizar el SMS se siguieron los lineamientos propuestos por Kitchenham *et al.* [5] y por Petersen *et al.* [6].

El artículo se estructura de la siguiente manera: en la Sección 2 se describe la planificación del SMS, en la Sección 3 se describe su ejecución. Los resultados se presentan en la Sección 4. En la Sección 5 se exponen las conclusiones y trabajos futuros.

2 Planificación del SMS

En esta sección se presenta la definición del protocolo de revisión del SMS: preguntas de investigación (PI), estrategia de búsqueda, selección de los estudios, criterios y proceso de selección, formulario de extracción y el proceso de síntesis de los datos.

El objetivo de este SMS es responder la siguiente pregunta de investigación (PI): *¿Cuál es el estado del arte respecto a la existencia de un modelo de mejores prácticas de seguridad informática en PyMES argentinas?* Esta pregunta principal (PI) se descompone en un conjunto de sub-preguntas (PI1-3), las cuales se presentan a continuación:

- *PI1: ¿Qué tipo de contribuciones existen respecto a la seguridad informática en PyMES?*
- *PI2: ¿En qué capa de seguridad se realiza la contribución?*
- *PI3: ¿Qué tipos de investigación se encuentra en los artículos?*

Se decide realizar la búsqueda en las siguientes bibliotecas y repositorios digitales: *Google Schoolar, Scielo, Dialnet* considerando publicaciones de congresos y revistas. La búsqueda se realizó en un período comprendido entre enero del 2010 a mayo del 2022.

La cadena de búsqueda utilizada es:

(Seguridad informática) and (PyMES) and (estándar OR modelo OR arquitectura OR esquema OR guía OR procedimiento)

Los criterios de inclusión y exclusión utilizados para el proceso de selección de artículos se presentan en la Tabla 1.

Tabla 1. Criterios de inclusión y exclusión.

Criterios de inclusión	Criterios de exclusión
Artículos vinculados a las PI.	Artículos que no estén accesible para su lectura completa.
Artículos publicados a partir de enero del 2010 hasta mayo del 2022.	Libros, tesis, presentaciones en power point, informes técnicos, artículos que cuenten solo con el resumen.
Artículos en el idioma español.	

Para dar respuesta a cada una de las preguntas de investigación (PI) se definió un esquema de clasificación, que por restricciones de espacio se presenta en un apéndice en [7], junto con el formulario de extracción de datos. Se utiliza una síntesis temática basada en el esquema de clasificación que se representará a través de gráficos.

El proceso de selección de los estudios consistió en los siguientes pasos: 1) realizar la búsqueda en las fuentes definidas aplicando la cadena en el título y/o en el resumen, 2) eliminar los artículos duplicados, 3) aplicar los criterios de inclusión y exclusión en el título, resumen y palabras clave, 4) aplicar los criterios de inclusión y exclusión al texto completo. Este proceso permitió la selección de los estudios primarios que se analizaron para dar respuesta a las preguntas de investigación (PI) formuladas.

3 Ejecución del SMS

Por restricciones de espacio, la cantidad de artículos encontrados en cada uno de las librerías, plataformas y repositorios digitales definidos en el protocolo de revisión se encuentran en un apéndice en [7] junto con el listado de los 10 estudios primarios analizados.

4 Resultados del SMS

A continuación, se pretende dar respuesta a las preguntas de investigación (PI) en base a la literatura analizada mediante la utilización de gráficos.

PI1: ¿Qué tipos de contribuciones existen respecto a la seguridad informática en PyMES?

El estándar más mencionado en los estudios primarios es ISO 27001. En la mayoría de los estudios se encontró que para resolver las problemáticas de seguridad en las PyMES se utilizan soluciones que son para empresas de más envergadura, como por ejemplo, COBIT, ITIL y los estándares de ISO (Ver Figura 1).

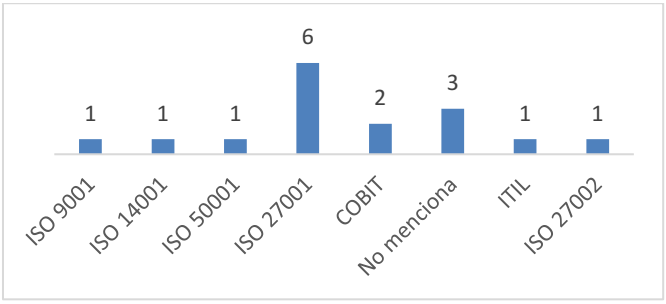


Fig. 1. Estándares utilizados.

PI2: ¿En qué capa de seguridad se realiza la contribución?
En los estudios analizados se encontró que en un mismo artículo se resuelven problemas de diferentes capas de seguridad. Las capas de seguridad a las cuales hacen referencia la mayoría de los estudios son "aplicaciones" y "red" (Ver Figura 2).

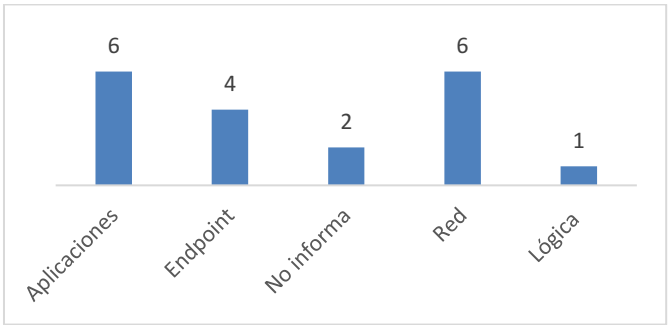


Fig. 2. Capas de seguridad.

PI3: ¿Qué tipos de investigación se encuentra en los artículos?
Dentro de los hallazgos se logró evidenciar que la mayoría de los estudios se corresponden a validaciones (6 en total). Y dos estudios son del tipo de investigación "evaluación" y dos estudios del tipo "propuesta de solución" (Ver Figura 3).

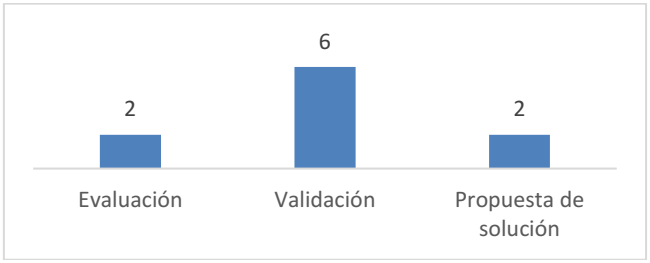


Fig. 3. Tipos de investigación según la clasificación propuesta por Wieringa [8].

5 Conclusiones y trabajos futuros

Se logró construir el estado del arte respecto a la situación de soluciones de seguridad informática utilizada en PyMES mediante el desarrollo de un SMS. Se analizaron 10 estudios primarios recuperados de las fuentes de búsqueda definidas en el protocolo de revisión cuyo período de búsqueda ha sido comprendido entre enero del 2010 y mayo del 2022. Una vez analizados los estudios primarios, se concluye que:

- La mayoría de las contribuciones hacen referencia a la utilización de estándares definidos por ISO siendo la mayoría correspondiente a la ISO 27001.
- Se pudo evidenciar que la mayoría de las investigaciones abordan la problemática de la seguridad informática sobre redes, aplicaciones y endpoint. En ese sentido se puede analizar que los principales vectores de ataques se encuentran sobre estas capas. También es importante destacar que el principal problema de seguridad informática es el usuario, que pueden llegar a ser un punto de ataque por desconocimiento del tema o por un mal uso de la tecnología. Estas tres capas mencionadas anteriormente son los primeros puntos por proteger en una organización, serían como los pasos iniciales a revisar al momento de plantear una arquitectura segura.
- Se logró evidenciar la ausencia de una solución específica para PyMES dado que en las investigaciones se utilizan estándares, prácticas y las contextualizan para resolver problemas de seguridad informática en este tipo de empresas.

Las futuras actividades para continuar con el desarrollo de la tesis son: 1) el desarrollo de un conjunto de buenas prácticas que pueda ser utilizadas en PyMES de Argentina y 2) Validar la solución en diferentes estudios de casos en PyMES de Argentina.

Referencias

1. López Purificación A. Seguridad informática. Editex. ISBN 978-84-9771-657-4 (2010).
2. Ministerio de desarrollo productivo, Más de 16 millones de empresas ya se incorporaron al registro MiPyME. Disponible: https://www.argentina.gob.ar/noticias/mas-de-16-millones-de-empresas-ya-se-incorporaron-al-registro-mipyme (2022).
3. Dirección Nacional de Ciberseguridad. Informe del CERT.ar. Disponible en: https://www.argentina.gob.ar/jefatura/innovacion-publica/ssetic/direccion-nacional-ciberseguridad/informes-de-la-direccion-3 (2020).
4. Kaspersky. Boletín de Seguridad de Kaspersky. Estadísticas 2021. Disponible en: https://securelist.lat/kaspersky-se2021)curity-bulletin-2021-statistics/96099 (2021).
5. Kitchenham B., Chartes, S. Guidelines for performing systematic literature reviews in software engineering, Keele University, EBSE-2007-01 (2007).
6. Petersen K. Wohlin C.: Context in industrial software engineering research. Third International Symposiumm on Empirical Software Engineering and Measurement (2009).
7. Alcantara T., Panizzi M., Sattolo I. Apéndice- Buenas Prácticas para la Seguridad Informática en PyMES (Camera Ready). Disponible en: https://doi.org/10.6084/m9.figshare.20514780.v5 (2022).
8. Wieringa R., Maiden N., Mead N., Rolland C. Requirements engineering paper classification and evaluation criteria: a proposal and a discussion. Requirements Engineering, 11(1), pp. 102-107 (2006).

Automatización del Armado del Repertorio de Aperturas de Ajedrez

Emanuel Brea[1][0000-0002-7848-774X], Maximiliano Dos Santos[1][0000-0002-4505-0648] y Pablo Ezequiel Inchausti [1][0000-0002-8342-1796]

[1] Universidad Argentina de la Empresa (UADE)
{ebrea, maxdossantos, pinchausti}@uade.edu.ar

Resumen. El estudio de la primera fase del juego de ajedrez, conocida como la Apertura, es fundamental para el jugador y su progreso, pues permite conocer los principales planes del juego en base a partidas pasadas, y así evitar cometer errores conocidos. Sin embargo, armar un repertorio de aperturas de ajedrez puede ser una tarea muy compleja, especialmente para los jugadores inexpertos. El presente trabajo plantea una solución para generar automáticamente un repertorio de aperturas para el jugador de ajedrez, considerando su estilo y el nivel de juego. El aporte de simplificar la tarea de armar un repertorio de aperturas permite que el jugador de ajedrez se enfoque en comprender los planes detrás de las variantes que lo componen, para mejorar la calidad de su juego.

Keywords: Automatización, Ajedrez, Aperturas, Arquitectura

1 Introducción

En una partida de ajedrez, durante la apertura ambos jugadores luchan por obtener una posición que les conceda ventaja y les permita jugar el medio juego con mayores probabilidades de obtener la victoria. Por regla general, se enuncia que el primer movimiento le otorga al jugador de piezas blancas una leve ventaja, que el jugador de piezas negras lucha por neutralizar [1].

Para el jugador de ajedrez es imprescindible conocer que se jugó con anterioridad por otros jugadores de alto nivel, para poder reutilizar planes y evitar realizar movimientos que conducen a posiciones inferiores. Este conjunto de conocimiento conforma la teoría de aperturas [2], que evoluciona con el tiempo pues aparecen nuevas ideas como también ciertas variantes son refutadas y pierden popularidad.

La colección de variantes que el jugador decide emplear en sus partidas se conoce como repertorio de aperturas, y son vitales para el progreso en el ajedrez. Poseer un sólido repertorio permite transitar la apertura con mayor seguridad al saber que la variante elegida fue jugada con anterioridad por grandes maestros. Asimismo, le ahorra tiempo en el reloj que puede ser empleado para momentos más críticos de la partida durante el medio juego. Por último, estudiar partidas pasadas le permite al jugador conocer las ideas principales en cada variante y así evitar caer en trampas conocidas.

No obstante, debido a la gran cantidad de variantes en los primeros movimientos [3], la tarea de seleccionar que líneas incorporar al repertorio puede ser una tarea ardua, especialmente para jugadores inexpertos. Y a partir de este problema, el presente trabajo propone desarrollar una solución para simplificar la creación de repertorios de aperturas de ajedrez.

En la Tabla 1 se muestra como ejemplo, las estadísticas de dos posibles movimientos en una posición de la apertura, según los datos de una base de datos de partidas pasadas.

Tabla 1. Estadísticas de dos movimientos de las piezas blancas en una posición en la apertura.

Movimiento	Partidas	% victorias blancas	% empates	% victorias negras	Rating promedio
A	500	30	45	25	2500
B	200	35	50	15	2400

La tabla muestra que, en una posición cuyo turno es del jugador de piezas blancas, sus principales respuestas son los movimientos A y B, empleados en 500 y 200 partidas respectivamente. Para armar el repertorio de aperturas, el jugador deberá elegir que jugada incluir. Esto implicaría que, en caso de llegar a la posición en una partida de ajedrez, puede realizar el movimiento seleccionado con la seguridad de que es, en principio, una jugada buena pues fue empleada con anterioridad por grandes maestros.

Sin embargo, se plantean varios problemas, pues la elección puede no ser trivial. La jugada A fue empleada más veces que la B, pero la opción B logró un porcentaje mayor de victorias.

Adicionalmente, se pueden considerar factores, como el rating promedio de los jugadores de las piezas blancas y sus oponentes. Podría ocurrir que la jugada B tiene mayor porcentaje de victorias porque el nivel de juego de las piezas negras era menor.

También se debe considerar el año en que se jugaron las partidas. Es posible que gran parte de las partidas donde se empleó la jugada A fue antes de descubrir los beneficios de la jugada B, y la jugada A perdió popularidad entre los grandes maestros.

En el ejemplo de la Tabla 1 se consideran solo dos movimientos posibles A y B, pero en una posición existen más posibilidades. Y el análisis de los movimientos debe repetirse numerosas veces para cada posición resultante, conformando el repertorio.

Por todo ello, se observa que la elección de variantes no es una tarea simple y requiere del análisis de cada posición y sus estadísticas. Este esfuerzo se multiplica para jugadores principiantes, pues desconocen las ideas detrás de cada jugada.

2 Motivación

Para validar la existencia del problema, se realizó una encuesta a 351 jugadores de ajedrez, que van desde aficionados, hasta jugadores expertos, federados en la Federación Internacional de Ajedrez (FIDE), con títulos de Maestro y ranking ELO.

La encuesta arrojó que el 75% confirma que se requiere de mucho tiempo para estudiar las aperturas de ajedrez, y también confirmaron que las aperturas mejoran el

nivel de juego. De los encuestados, el 50% no guarda las aperturas en un software, y recurre a la improvisación y la memoria. También consideran como *útil* o *muy útil* una aplicación de armado de repertorios de aperturas de forma automatizada.

La principal motivación del presente trabajo es presentar una solución que ayude a jugadores de ajedrez en todo el mundo, en particular aquellos que recién comienzan a involucrarse en el juego, a preparar su repertorio de aperturas fácilmente. Se espera que esta herramienta les ayude a mejorar el desempeño en torneos y su nivel de juego, al poder visualizar cómodamente qué movimiento realizar en cada posición, las respuestas que se podrían esperar por parte del oponente, y cómo responder a las mismas.

El trabajo se desarrolla durante el año 2022 en el contexto del Proyecto Final de Ingeniería en Informática (PFI) de Emanuel Brea, con Maximiliano Dos Santos y Pablo Inchausti como cotutores, y también docentes de la cátedra PFI.

3 Materiales y Métodos

3.1 Interfaz de usuario

La solución se presenta como una aplicación web, que le permite al usuario acceder a su repertorio desde cualquier lugar. Esto puede ser útil en torneos de ajedrez, por ejemplo, antes del inicio de cada ronda, para que el jugador pueda consultar una variante en que se sienta inseguro o sabe que su rival la emplea con frecuencia.

El repertorio es representado por diagramas con la jugada recomendada en cada posición, y debajo se encuentran las respuestas del oponente. Si el usuario desea revisar una jugada, puede seleccionarla y a su vez, se muestra la siguiente réplica recomendada. El proceso se repite hasta el fin de la variante. Esto facilita el estudio y memorización de las aperturas, y en todos los casos, las jugadas son acompañadas con pequeñas descripciones de los planes, como también de estadísticas básicas de la posición.

Evidentemente, el número de variantes posibles puede crecer exponencialmente [4], y por eso se aplican criterios, como ser, dosificar la cantidad de variantes a incluir en el repertorio, de acuerdo con el nivel de experiencia del jugador.

Por otro lado, la aplicación le permitirá al usuario avanzado ingresar sus propios movimientos en repertorio, y calculará automáticamente las variantes derivadas.

3.2 Bases de Datos

La solución utiliza como fuente de datos más de un millón de partidas de grandes maestros jugadas entre los años 1970 y 2021. En base a estadísticas sobre la base, el algoritmo selecciona que movimientos incorporar al repertorio del jugador. Esta fuente de datos es necesaria para realizar las recomendaciones.

3.3 Arquitectura de la solución

Se propone el despliegue de los servicios en AWS para lograr escalabilidad, siguiendo una arquitectura de tres capas [5]. De acuerdo con la Fig. 1, AWS Amplify se utiliza en el frontend, que invoca a los servicios del backend por medio de un API Gateway.

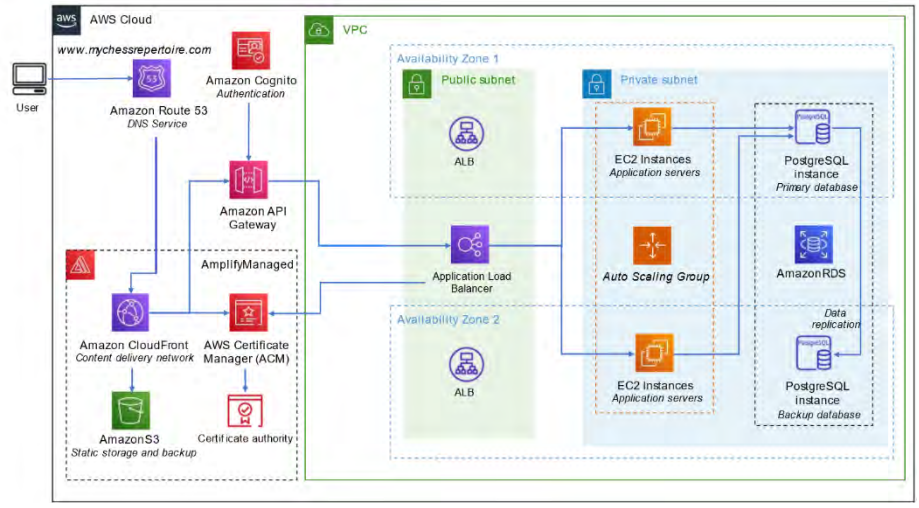


Fig. 1. Arquitectura de la aplicación "*MyChessRepertorie.com*" en AWS

3.4 Algoritmo de selección

Para recomendar una jugada en una posición dada, e incluirla en un repertorio se implementó un algoritmo de selección que toma en cuenta los siguientes factores:

- Cantidad de veces que se realizó dicha jugada en esa posición
- Cantidad de victorias, empates y derrotas
- Ranking Elo del jugador, y ranking Elo del jugador rival
- Año en que se jugó la partida

A estos factores, se le suma el estilo de juego y un factor aleatorio, para garantizar que dos jugadores con el mismo estilo no siempre obtengan el mismo repertorio.

Dada una posición de inicio, el algoritmo calcula y asigna un peso a cada jugada que impacta en la probabilidad de ser elegida. Para identificar unívocamente cada posición, se utiliza la técnica de Hash de Zobrist [6]. Y para hacer el hash legible al usuario, se le asocia la notación FEN, de Forsyth-Edwards, utilizada para diagramas de ajedrez.

4 Aportes del trabajo

El aporte de simplificar el armado de un repertorio de ajedrez le permite al jugador concentrarse en comprender las ideas estratégicas detrás del repertorio y las variantes que lo componen, para mejorar la calidad de su juego.

Como no todas las variantes tienen la misma importancia y popularidad, el jugador puede identificar las líneas principales para priorizar su tiempo dedicado al estudio.

Por otro lado, la solución propuesta debe complementarse con las alternativas existentes para el estudio especializado en la teoría de aperturas. Por ejemplo, el usuario puede usar el sitio web del presente trabajo, para revisar su repertorio y los principales

planes de cada jugada, pero si necesita profundizar con mayor detalle ciertas variantes, podrá recurrir a sitios especializados, que también explican con tutoriales o videos las aperturas de ajedrez [7][8][9]. Los libros de aperturas también son un recurso en donde se analizan partidas de jugadores famosos empleando las mismas líneas. Es decir, la solución propuesta no planea reemplazar las herramientas existentes, sino ser un complemento para el desarrollo del jugador de ajedrez.

5 Posibles líneas de investigación futura

En futuras versiones, la fuente de datos utilizada para recomendar jugadas en base a millones de partidas puede complementarse con motores de recomendación diseñados a partir de las preferencias de las variantes favoritas en jugadores de estilo similar.

Estos datos pueden ser obtenidos agregando botones de 'me gusta' o 'agregar a favoritos' en cada movimiento. Y disminuyendo el nivel de recomendación en variantes que fueron descartadas por los jugadores.

El modelo de sugerencias mencionado es similar al empleado por Spotify [10], donde las canciones son recomendadas a partir de los gustos similares de los usuarios. De esta manera, la sugerencia de variantes se basará en la experiencia y preferencias de otros jugadores, y no solo en la colección de partidas pasadas.

Referencias

1. Emms, John: Discovering Chess Openings: Building Opening Skills from Basic Principles. 1ª ed. London: Everyman Chess, 2006. ISBN 1857444191
2. Van Der Sterren, Paul: FCO: Fundamental Chess Openings. 1ª ed. London: Gambit Publications, 2009. ISBN 1906454132
3. Matanovic, Aleksandar: Encyclopaedia of Chess Openings. 1ª ed. London: Batsford, 1975. ISBN 0713430133
4. Shannon, C.: Programming a Computer for Playing Chess. Philosophical Magazine, 1950.
5. Amazon Web Services, Inc., Web Application Hosting in the AWS Cloud - AWS Whitepaper, (2021)
6. Zobrist, Albert. A new hashing method with application for game playing. The university of Wisconsin, 1970
7. Chess.com. Openings [en línea]. https://www.chess.com/openings (accedido 05/08/2022).
8. Chessable. Opening Explorer [en línea]. https://www.chessable.com/explore/ (accedido 06/08/2022).
9. Lichess.org. Opening Explorer [en línea]. https://lichess.org/analysis (accedido 22/07/2022)
10. Björklund, G., Bohlin, M., Olander, E., Jansson, J., Walter, C.E., Au-Yong-Oliveira, M., "An Exploratory Study on the Spotify Recommender System", en Information Systems and Technologies, Cham, (2022), pp. 366-378. doi: 10.1007/978-3-031-04819-7_36

Aplicaciones Móviles y Salud. Posibilidades para la Promoción de la Higiene Postural

Milagros Salas[1][0000-0002-3058-2967] y Edith Lovos[1][0000-0002-2875-0239]

[1] Universidad Nacional de Río Negro, Sede Atlántica, CIEDIS, Viedma R8500, Argentina.
salasmilagros99@gmail.com, elovos@unrn.edu.ar

Abstract. El uso cotidiano e intensivo de las tecnologías de la información y la comunicación (TIC), y sus efectos en los hábitos posturales, han despertado el interés de especialistas del campo de la salud, generando diversos estudios sobre el tema. En este trabajo se presenta una propuesta de investigación, a desarrollarse a través de una beca CIN, que aborda la temática tecnologías móviles para la promoción de la higiene postural. Se espera, realizar un trabajo de campo que permita identificar las posibilidades, limitaciones y contraindicaciones que los docentes y estudiantes avanzados de la carrera Lic. en Kinesiología y Fisiatría que se dicta en la Sede Atlántica de la Universidad Nacional de Río Negro, encuentran en la experimentación con aplicaciones móviles como recursos complementarios en actividades de promoción y educación postural.

Keywords: Higiene Postural, Promoción, Dispositivos Móviles.

1 Introducción

Este trabajo presenta en forma resumida una propuesta de investigación que buscará identificar las posibilidades, limitaciones y contraindicaciones que los docentes y estudiantes avanzados de la Lic. en Kinesiología y Fisiatría de la Universidad Nacional de Río Negro (UNRN), encuentran en la experimentación con aplicaciones móviles que tienen como objetivo la promoción y educación postural. El trabajo se llevará adelante a través de una beca de Estímulo a las Vocaciones Científicas (CIN 2021) en el marco de un proyecto de investigación (PI-UNRN-40C-876) acreditado por la institución que se vincula a tecnologías en educación y en el cual se han realizado estudios previos sobre inclusión de tecnologías en el campo de la kinesiología [15,16]

1.1 Tecnologías e Higiene Postural

En el campo de la fisioterapia, la higiene postural se puede definir como la capacidad del individuo de mantener una postura adecuada, mientras realiza actividades de la vida diaria, con la intención de evitar lesiones provocadas por malos hábitos posturales [1]. Actualmente, y cada vez con más frecuencia las tecnologías de la información y la comunicación (TIC) están presentes en la vida diaria, en particular aquellos dispositivos más accesibles como los teléfonos celulares. Y sus efectos en los hábitos posturales, han despertado el interés de especialistas del campo de la salud, generando diversos

estudios sobre el tema. En algunos casos, los estudios dan cuenta de patologías específicas como la cervicalgia [2,3], en otros se aborda el diseño de programas terapéuticos destinados a tratar el dolor [4], así como también a la prevención [5].

Por otra parte, durante el contexto de pandemia, el tiempo de exposición frente a pantallas se vió incrementado en todos los grupos etarios, aunque principalmente en los más jóvenes, así lo demuestra un estudio realizado en la provincia de Córdoba al inició del contexto de aislamiento[6]. Sumado a esto, los datos aportados por el INDEC, acerca del acceso y uso de tecnologías de la información y la comunicación en Argentina, registró en el cuarto trimestre de 2021, que el 64,2% de los hogares urbanos tiene acceso a computadora y el 90,4%, a internet. Asimismo, los datos indican que 88 de cada 100 personas usan teléfono celular y 87 de cada 100 utilizan internet, resultando el teléfono móvil como la tecnología de uso más extendida para la población joven y adulta. Y específicamente para la región patagónica donde se inserta la propuesta de investigación que se describe en este trabajo, se presentan los valores con mayor uso de telefonía móvil [7].

En relación a los ámbitos de uso, los dispositivos móviles adquieren cada vez más presencia en propuestas educativas de diferentes niveles, propiciando otras formas de interacción y acceso a los contenidos con efectos positivos para el aprendizaje [8], a la vez que habilitan la inclusión de otras tecnologías, como el caso de la realidad aumentada o los juegos educativos móviles, que pueden apoyar el aprendizaje de temas complejos o de difícil acceso [9,10,11]. Estas posibilidades, se han extendido a otros campos, entre ellos la salud, donde se conoce como Mobile Health (mHealth), definida por el Observatorio Mundial de la Salud Electrónica (GOe), como la práctica médica y de salud pública apoyada en dispositivos móviles como celulares, dispositivos de monitorización de pacientes, asistentes digitales personales (PDA) y otros dispositivos inalámbricos. Específicamente sobre el uso de dispositivos móviles en fisioterapia, una revisión bibliográfica de los últimos años[12], indica que los mismos permiten incrementar el acceso al servicio de salud, favorecer la promoción de la salud, así como detectar en forma precoz determinadas deficiencias y evaluar aspectos vinculados a la marcha y equilibrio. En el caso de Basiratzadeh [13], presentan un estudio que incluye realidad aumentada, donde mediante el uso de marcadores fiduciales y un dispositivo móvil, pueden medir en tiempo real la postura y la amplitud de movimiento (ROM) permitiendo una evaluación clínica. Otras investigaciones, presentan el desarrollo de aplicaciones móviles para la prevención de alteraciones posturales, como el caso de PostureUp [14] destinado a personas que realizan actividades con ordenadores.

Así, en esta propuesta de investigación, se buscará identificar las posibilidades, limitaciones y contraindicaciones que los docentes y estudiantes avanzados de la Lic. en Kinesiología y Fisiatría de la Universidad Nacional de Río Negro, encuentran en la experimentación con aplicaciones móviles destinadas al área, como recursos complementarios en actividades de promoción y educación postural.

2 Metodología y Plan de Tareas

Se propone trabajar siguiendo una metodología de investigación de tipo cuantitativo con aspectos cualitativos, combinando investigación teórica con trabajo de campo empírico, que permita realizar un análisis principalmente exploratorio - descriptivo en relación a la temática de estudio, con la intención de entender y comprender los temas de investigación que se abordan. Se propone realizar un muestreo probabilístico, donde todos los docentes y estudiantes de la población de estudio, tienen la misma posibilidad de formar parte de la muestra. Para alcanzar el objetivo propuesto, la investigación se llevará a cabo en etapas. A continuación se describen en forma resumida las etapas y sus tareas. En la primera fase se realizará una revisión de referencias bibliográficas sobre el tema estudio, que permitan analizar el grado de aceptación de las tecnologías móviles por parte de profesionales del campo de la kinesiología, y diseñar una experiencia de uso y evaluación de las aplicaciones del tipo PostureUp [14] y SmartPosture [17] con docentes y estudiantes avanzados de la Lic. en Kinesiología y Fisiatría de la UNRN. Asimismo, se diseñarán los instrumentos necesarios para recuperar las posibilidades, limitaciones y contraindicaciones que los participantes de la experiencia, encuentran en las aplicaciones móviles incluidas en la misma. En la segunda etapa, se llevará adelante la implementación de la experiencia diseñada en la etapa anterior, se aplicarán los instrumentos de recolección de datos, y a partir de allí se procederá a analizar la información teniendo en cuenta los aportes teóricos trabajados y así construir las conclusiones sobre el tema de investigación. Es importante señalar que la selección de las aplicaciones priorizará la disponibilidad para dispositivos móviles con sistema operativo Android, por ser este el de mayor penetración en el contexto de aplicación, así también condiciones de acceso a las mismas y los dispositivos que demande su uso.

3 Aportes

A través de esta propuesta de investigación se generará conocimiento sobre el uso aplicaciones móviles destinadas a la higiene postural. Así mismo, se espera que los resultados puedan ser aportes al diseño, producción y evaluación de materiales digitales móviles, que puedan incluirse como complemento en actividades educativas destinadas a promover buenos hábitos posturales de los individuos. Por otra parte, el desarrollo de la propuesta posibilitará la vinculación entre estudiantes y docentes de carreras del ámbito de la salud y las ciencias informáticas que se dictan en la Sede Atlántica de la UNRN y su posterior transferencia de conocimientos al medio.

4 Líneas de Investigación Futuras

Teniendo en cuenta que el grupo etario que utiliza tecnologías con más frecuencia, son niños y jóvenes, sería interesante analizar si el uso a largo plazo de aplicaciones móviles como el caso de PostureUp [14], mejoran los hábitos posturales de los usuarios.

5 Referencias

1. Gómez Conesa Antonia (2002). Higiene postural y ergonomía. Elsevier. Fisioterapia.2002;24: 1-2.
2. Freire Nolivos, P. E. (2020). Incidencia de la cervicalgia asociada al uso del teléfono celular en los estudiantes de 15 a 17 años de la Unidad Educativa Particular San Fernando, durante el periodo octubre 2019-febrero 2020 (Bachelor's thesis, Quito: UCE).
3. Mejia, C. P., & Melani, S. (2018). Relación entre la alteración postural de la columna torácica y el uso excesivo de dispositivos móviles en estudiantes de tecnología médica de la Universidad Privada Autónoma del Sur, Arequipa. 2018.
4. García Amor, B. (2019). Eficacia de un programa de ejercicio terapéutico en adolescentes con dolor musculoesquelético asociado al empleo de dispositivos móviles.
5. Bocanegra Padilla, J. S., & Calderón Moreno, M. D. P. (2021). Protocolo de medidas básicas para la prevención de riesgo ergonómico en estudiantes de educación media.
6. Liviero, B., Favalli, M., Macció, J. P., Aguirre, T., Verzini, J. R., & Endrek, M. S. (2020). Pantallas y síntomas de la superficie ocular en cuarentena por COVID-19. Oftalmología Clínica y Experimental, 13(4).
7. INDEC (2O22). Acceso y uso de tecnologías de la información y la comunicación. EPH. Cuarto trimestre de 2021.
8. Lagunes-Domínguez, A., Torres-Gastelú, C. A., Angulo-Armenta, J., & Martínez-Olea, M. Á. (2017). Prospectiva hacia el aprendizaje móvil en estudiantes universitarios. Formación universitaria, 10(1), 101-108.
9. Loa Barrientos, L. S. (2017). Influencia de un Software con Realidad Aumentada para el Proceso de Aprendizaje en Anatomía Humana en la Educación Primaria IEIP Pitágoras Nivel A, Andahuaylas.
10. Tamami Dávila, C. A. (2017). La realidad aumentada y el proceso de enseñanza-aprendizaje de Anatomía en los estudiantes de la carrera de Enfermería de la Facultad de Ciencias de la Salud de la Universidad Técnica de Ambato (Bachelor's thesis, Universidad Técnica de Ambato. Facultad de Ciencias Humanas y de la Educaciòn. Carrera de Docencia en Informática).
11. Ruiz, C.S. (2018). Enseñanza de la anatomía y la fisiología a través de las realidades aumentada y virtual. Innovación Educativa 19(79),57-76.
12. Angarita Rodríguez, Diana Cristina, & Castañeda Giaimo, Jorge Nicolás. (2017). Uso de dispositivos móviles en fisioterapia. Revista Cubana de Información en Ciencias de la Salud, 28(2), 1-13.
13. Basiratzadeh, S., Lemaire, E. D., & Baddour, N. (2020, July). Augmented Reality Approach for Marker-based Posture Measurement on Smartphones. In 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 4612-4615). IEEE.
14. Fardoun, H. M., Alghazzawi, D. M., & Romero-López, S. (2019, September). PostureUp: a system to rehab and prevent postural issues at the office. In Proceedings of the 5th Workshop on ICTs for improving Patients Rehabilitation Research Techniques (pp. 175-178).
15. Ponce Cévoli, Ismael , Lovos Edith. (2020). Tecnologías aplicadas a la Kinesiología. El uso de la Realidad Virtual en la rehabilitación post ACV. En “IV Jornadas de investigadores noveles, becaries y tesistas: #InvestigaEnCasa”. Centro Interdisciplinario de Estudios sobre Derechos, Inclusión y Sociedad (CIEDIS) de la UNRN.
16. Ponce Cévoli, Ismael (2021). Intervención terapéutica post ACV. Neurorehabilitación integrando realidad virtual. Trabajo final de grado. Universidad Nacional de Río Negro.
17. https://smartposture.net/

StopFire: Alertas de Incendios Forestales en Argentina Usando IoT y Machine Learning

Alejandra Curbelo[1] [0000-0003-1708-6624], Juan Cruz Alric[1] [0000-0002-8854-9844] y Pablo Ezequiel Inchausti[1] [0000-0002-8342-1796]

[1]Universidad Argentina de la Empresa (UADE), Instituto de Tecnología (INTEC). Buenos Aires, Argentina
{alcurbelo, jalric, pinchausti}@uade.edu.ar

Resumen. Los incendios forestales son devastadores para un ambiente y sus efectos se miden en cientos de miles de hectáreas. Para contribuir en la prevención de incendios forestales, se desarrolla una solución de monitoreo por imágenes de las zonas en riesgo basada en IoT, y la utilización de modelos de Machine Learning con Redes Neuronales Convolucionales para identificar en las imágenes la presencia del fuego. La solución se completa con un tablero de monitoreo de los dispositivos de IoT y una aplicación móvil para reportar el riesgo de incendio enviando imágenes geolocalizadas.

Keywords: Incendios Forestales. Machine Learning, IoT, Redes Neuronales

1 Introducción

Un incendio forestal se define como un fuego que se propaga rápidamente y de forma descontrolada a través de la vegetación de una zona, y, si bien se pueden originar por causas naturales, como las tormentas eléctricas, más del 95% de los incendios forestales en Argentina son provocados por el hombre, ya sea de forma intencional, o no [1].

Los incendios forestales suelen producir efectos devastadores, y de acuerdo al Servicio Nacional de Manejo del Fuego (SNMF), en el 2021 más de 302 mil hectáreas en Argentina fueron afectadas por el fuego, con la provincia de San Luis encabezando la lista [2]. En febrero del 2022, los incendios forestales de Corrientes afectaron más de 900 mil hectáreas, un equivalente al 11% de la superficie provincial [3].

El objetivo del presente trabajo es contribuir a la prevención de incendios forestales en Argentina, y desarrollar un sistema de alertas y monitoreo por imágenes. Se utiliza IoT para el monitoreo y se determina la presencia del fuego con modelos de Machine Learning (ML). Se complementa la solución con una aplicación móvil para reportar el riesgo de incendio en una zona, enviando imágenes con información geolocalizada.

El trabajo inicia en el 2022 cómo Proyecto Final de Ingeniería Informática (PFI) de Alejandra Curbelo y Juan Cruz Alric, siendo Pablo Inchausti, docente en PFI, el tutor.

También se integra en el INTEC, el Instituto de Tecnología de UADE, a la línea A21T03: *Aplicaciones de Machine Learning para el uso de Recursos Naturales.* [4] y continúa el trabajo de *AQUA* [5] sobre prevención de incendios forestales en Pinamar.

2 Materiales y Métodos

El sistema StopFire está compuesto por una aplicación web, con un módulo de detección de incendios forestales que integra un modelo de Machine Learning para analizar imágenes de zonas forestales capturadas por dispositivos de IoT.

Los dispositivos de IoT se conectan a servicios desplegados en el Cloud Provider AWS, y reciben imágenes cada 30 minutos, para procesarlas y generar una alerta en caso de que el modelo entrenado detecte que existe un posible incendio en la zona.

El sistema de monitoreo se presenta como un tablero que permite ver las cámaras conectadas a la red, con su ubicación geográfica. En el módulo de alertas se muestra la foto que el modelo utilizó para determinar el potencial foco de incendio, con información geolocalizada de la cámara y el día y la hora de la captura de la imagen. En la **Fig. 1**. se describe el modelo de solución de StopFire a alto nivel:

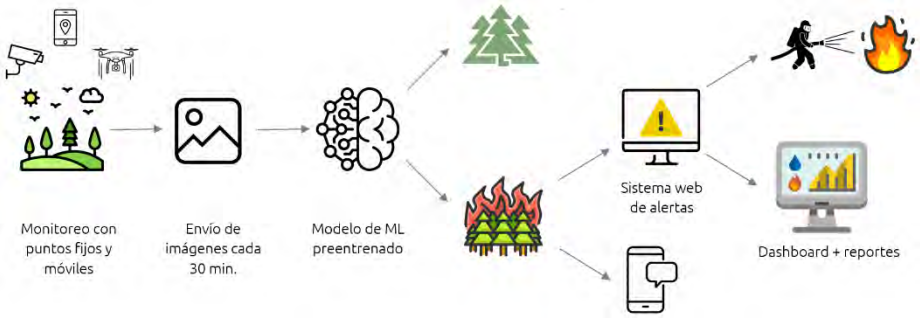


Fig. 1. StopFire: modelo de solución

Respecto a la aplicación móvil, permitiría que, de forma voluntaria, las personas presentes en zonas forestales puedan reportar incidentes con riesgo de incendios. Desde la aplicación se permite el envío de imágenes al sistema de monitoreo y analizarlas con los modelos de detección de incendios, y también obtener la información geolocalizada del lugar y el tiempo de la captura de la imagen para identificar la zona de riesgo.

Teniendo en cuenta que el factor humano es responsable de más del 95% de los incendios forestales [1], la aplicación móvil permite involucrar de forma directa a nuestra sociedad, y le permite contribuir en la prevención reportando incidentes desde la zonas de riesgo, y así reducir el nivel de responsabilidad en estos desastres naturales.

De esta forma, también el envío de imágenes mediante las aplicaciones móviles permite escalar la red de monitoreo más allá de los dispositivos de IoT desplegados en la zona. Y el análisis de las imágenes enviadas, también sirven tanto para mejorar el entrenamiento de los modelos, como para identificar lugares propensos a incendios con el objetivo de ampliar la red de dispositivos.

La tecnología para el desarrollo de los modelos de Machine Learning (ML) está basada en Redes Neuronales Artificiales (RNA), técnicas de visión por computadora y Redes Neuronales Convolucionales (CNN) que son especialmente efectivas en campos como la visión artificial.

2.1 Redes Neuronales Convolucionales (CNN)

La Red Neuronal Convolucional (CNN) es un tipo de red neuronal multicapa o arquitectura de aprendizaje profundo inspirada en el sistema visual de los seres vivos. Tienen una arquitectura *feed forward*, que significa que las neuronas de cada capa se conectan con todas las neuronas de la capa siguiente, pero no con neuronas de la misma capa. Las primeras capas de la red aprenden y extraen las características de alto nivel, mientras que las capas más profundas, extraen las características de bajo nivel. [6]

En las CNN, hay tres tipos de capas intermedias: la capa convolucional, que es el componente más importante de las arquitecturas CNN porque es la que genera un mapa con las características de las imágenes a partir de las imágenes de entrada. La capa de agrupación, que toma las características de mayor tamaño y las reduce en mapas de características de menor tamaño. Y finalmente, la capa completamente conectada, que se utiliza como clasificador, y que, a partir de la última capa de agrupación, se pasa por la red *feed forward* para generar la salida. [6]

Para el desarrollo de la CNN se utilizó la librería *fast.ai* [7], que se especializa en Deep Learning, para creación de las redes neuronales.

2.2 Preparación de los datos

Para el entrenamiento de la red neuronal, se usaron conjuntos de datos públicos de incendios forestales provenientes de plataformas libres como Kaggle. Por ejemplo, *Wildfire Detection Image Data* [8] es un conjunto de datos con imágenes de uso libre de incendios forestales para entrenar modelos de Machine Learning y Deep Learning.

El procedimiento inicia con la carga de las 1.832 imágenes etiquetadas de Kaggle, y adicionalmente, para aumentar la muestra disponible para el entrenamiento, se utilizó la librería *fast.ai* y su función *aug_transforms* para crear una lista de transformaciones a partir de rotaciones, acercamientos, deformaciones e iluminación sobre los datos de entrenamiento. Adicionalmente se utilizaron imágenes de incendios reales en Argentina para probar el rendimiento del modelo con datos locales de Argentina.

En la **Fig. 2** se pueden ver algunas de las imágenes que se utilizan en la etapa de entrenamiento, posterior al proceso de preparación de los datos.

Fig. 2. Imágenes procesadas y listas para ser utilizadas en el entrenamiento de la Red Neuronal.

2.3 Entrenamiento de la Red Neuronal Convolucional (CNN)

Una vez que se prepararon los datos, se debe entrenar la red neuronal. Se probaron 2 modelos que reciben los nombres de *ResNet* y *ConvNeXt*.

Se crearon y probaron 3 versiones de *ResNet*: *ResNet34*, *ResNet50* y *ResNet101*. En los 3 casos, el modelo arrojó falsos positivos y negativos en imágenes de bosques donde había neblina, mucho humo, o imágenes donde se veía el atardecer de fondo, como se observa en la **Fig. 3**

Fig. 3. Ejemplo de falsos positivos y negativos obtenidos usando *ResNet*

Para mejorar los resultados obtenidos con *ResNet*, se creó una red CNN denominada *ConvNeXt* que obtuvo una precisión del 100%, que significa que no hubo falsos positivos o negativos, y el modelo pudo clasificar correctamente las imágenes con y sin fuego. En la **Fig. 4** se observa la matriz de confusión asociada al modelo *ConvNeXt*

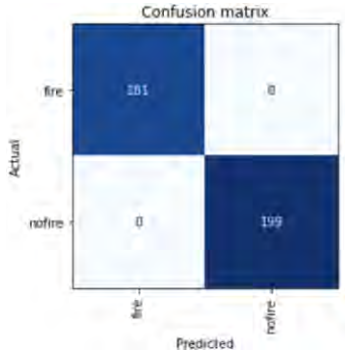


Fig. 4. matriz de confusión del modelo *ConvNeXt*

Después de generar los 4 modelos y comparar los resultados de la matriz de confusión de cada uno, se decidió integrar el modelo *ConvNeXt* en el tablero de la aplicación de monitoreo. Es decir, se priorizó un modelo que no genera falsos negativos, porque se desea que el modelo no falle cuando se necesite identificar si en una imagen se está produciendo un incendio, que se traduce en el envío de alertas de incendios en fases tempranas del incidente.

Con la información provista por el modelo, en el tablero de monitoreo se genera una alerta que le permite a los equipos de bomberos reaccionar a tiempo, y así reducir las probabilidades de que el fuego se convierta en un desastre difícil de contener.

3 Posibles Líneas de Investigación Futura

Como línea de investigación futura, proyectada a desarrollar dentro del INTEC, se propone extender al presente monitoreo por imágenes, con una red de dispositivos IoT con sensores de variables meteorológicas y atmosféricas. Este tipo de monitoreo por sensores, que incluiría variables climáticas, como la humedad relativa y la temperatura, y del estado el aire, como la presencia de partículas de humo y de gases combustibles, [9] permitiría obtener indicios de la presencia de incendios en la zona.

Para validar el aporte del presente trabajo y sus líneas de investigación, se entrevista a una autoridad de la Federación Mendocina de Bomberos Voluntarios. Respecto a StopFire, el entrevistado indica que mejoraría el esquema actual de prevención de incendios, ya que dependen principalmente de notificaciones vía telefónica y del análisis de imágenes satelitales de sitios gratuitos y en tiempo real como *FireMap* [10].

En cuanto al aporte de la solución de monitoreo integral, el entrevistado la considera *muy buena,* y agrega que es ideal detectar la columna de humo 'cuanto antes' porque las dos primeras horas son cruciales para controlar un incendio y evitar el caos.

Agradecimientos. Al Instituto de Tecnología (INTEC) de la Universidad Argentina de la Empresa (UADE) por integrar el PFI en la línea de investigación de *Aplicaciones de Machine Learning para Mejorar el uso de Recursos Naturales* (A21T03).

Referencias

1. Argentina.gob.ar, «¿Qué es y cómo funciona el Servicio Nacional de Manejo del Fuego?», (2020). https://www.argentina.gob.ar/ambiente/fuego/servicio-nacional (accedido 31/7/22).
2. Argentina.gob.ar, «Reportes diarios del Servicio Nacional de Manejo del Fuego», (2021). https://www.argentina.gob.ar/ambiente/fuego/diciembre-2021 (accedido 31/7/22).
3. Kurtz, D.B., Perucca, A.R., Saucedo, G., «Al 21 de febrero de 2022, la superficie quemada fue de 934.238 hectáreas | INTA». https://inta.gob.ar/noticias/al-21-de-febrero-de-2022-la-superficie-quemada-fue-de-934238-hectareas (accedido 31/7/22).
4. Inchausti, P.E., Martinez Saucedo, A.C., Amet, L., Blanco, P., Nievas, G., Giusto, L., «Aplicaciones de Machine Learning para el uso Sustentable de Recursos Naturales». https://wicc2022.tk/workshop/6256d0d67c76870009464c77/post/6260a4135fedd100097c5c73 (accedido 31/7/22).
5. Martínez Saucedo, A.C., «AQUA: Desarrollo de un Modelo de Machine Learning para Prevenir Incendios Forestales en Pinamar», Tesis, Universidad Argentina de la Empresa, (2021). https://repositorio.uade.edu.ar/xmlui/handle/123456789/14106 (accedido 5/9/22)
6. Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., De, D., «Fundamental Concepts of Convolutional Neural Network», (2020), pp. 519-567. doi: 10.1007/978-3-030-32644-9_36.
7. fast.ai, «fast.ai - Welcome to fastai». https://docs.fast.ai/ (accedido 31/7/22).
8. Dincer, B., «Kaggle» (2021) https://www.kaggle.com/datasets/brsdincer/wildfire-detection-image-data Wildfire Detection Image Data (accedido 31/7/22).
9. Argentina.gob.ar, «¿Cuáles son las variables y qué factores las afectan?», (2018). https://www.argentina.gob.ar/ambiente/fuego/conocemas/variables (accedido 3/9/22).
10. Robert E. Wolfe, «NASA-FIRMS», (2022). https://firms.modaps.eosdis.nasa.gov/map/ (accedido 22/8/22).

Comparación de Herramientas de Accesibilidad Web

Mag. Pablo Pandolfo[1], Gonzalo Fuentes[1], Rodrigo Lema[1]

[1]Instituto de Tecnología (INTEC) de la Universidad Argentina de la Empresa, UADE, Buenos Aires, Argentina
{ppandolfo, gonfuentes, rlema}@uade.edu.ar

Resumen. Las herramientas de accesibilidad web son aplicaciones software que permiten identificar problemas de accesibilidad web de forma automática de acuerdo con determinadas normas. Este artículo presenta la comparación de las herramientas de accesibilidad web Axe-Core, Pa11y y Lighthouse. Todas ellas detectan errores de accesibilidad web, sin embargo, cada una presenta sus resultados en forma diferente.

Palabras clave: accesibilidad web, WCAG 2.0, W3C, WAI, herramientas de evaluación automática de accesibilidad, Axe-Core, Pa11y, Lighthouse

1 Introducción

La accesibilidad web se asocia con la idea de diseñar y desarrollar páginas web que puedan ser utilizadas por las personas con discapacidad. Sin embargo, la accesibilidad web no está orientada exclusivamente a las personas con discapacidad. La accesibilidad web es única porque no diferencia entre dispositivos desde donde se visualiza contenido Web y es universal porque prevé la utilización de la web independientemente de las características del usuario. La accesibilidad web es el acceso universal a la Web, independientemente del tipo de hardware, software, infraestructura de red, idioma, cultura, localización geográfica y capacidad del usuario. [REVILLA MUÑOZ, 2013]

El análisis de la accesibilidad web tiene por finalidad analizar, estudiar y validar las páginas web para verificar el cumplimiento de las pautas y directrices de accesibilidad existentes. Una herramienta de evaluación de la accesibilidad web es un programa informático que detecta problemas de accesibilidad de una página web. La evaluación de la accesibilidad web se puede realizar de forma automática o manual.

La evaluación automática permite realizar una evaluación rápida, ayuda a tener una primera impresión de la accesibilidad de una página web, pero no proporciona un análisis definitivo, ya que puede no detectar errores o señalar falsos positivos. Se requiere un análisis manual complementario. [SEGOVIA, 2006]

Word Wide Web Consortium (W3C) es la organización internacional que trabaja en el desarrollo de estándares web [W3C, 1994]. Uno de los grupos de trabajo desde el año 1997 es la Web Accessibility Initiative (WAI) dedicada a promover soluciones de accesibilidad en la Web [WAI, 1998]. La WAI publicó en el año 2008 la versión 2.0 de las Pautas de Accesibilidad al Contenido Web (WCAG 2.0), las cuales son consideradas estándares internacionales de accesibilidad web.

La WCAG 2.0 se organiza en cuatro principios [WCAG, 2008]:

- *Perceptible*: la información y los componentes de la interfaz de usuario deben ser presentados a los usuarios de modo que ellos puedan percibirlos.
- *Operable*: los componentes de la interfaz de usuario y la navegación deben ser operables.
- *Comprensible*: la información y el manejo de la interfaz de usuario deben ser comprensibles.
- *Robusto*: el contenido debe ser robusto como para ser interpretado de forma fiable por una amplia variedad de aplicaciones de usuario.

Los principios están conformados por pautas, las pautas por criterios de conformidad y los criterios por técnicas que pueden ser suficientes o recomendables.

2 Herramientas para la validación de la accesibilidad web

Las herramientas de evaluación son aplicaciones de escritorio que se pueden descargar e instalar en la computadora del usuario o aplicaciones web que se pueden acceder y usar a través de un navegador. [LUJAN MORA, 2006]

Algunas herramientas permiten realizar una prueba de las aplicaciones una vez en producción, otras permiten realizar pruebas de accesibilidad web como parte del proceso de desarrollo. En caso de utilizar un proceso de integración continua (CI/CD), es posible realizar pruebas de accesibilidad para detectar vulnerabilidades. Además, se podrán utilizar los resultados de las pruebas automatizadas, para crear informes para el equipo de producto y desarrollo. Por tal motivo, se realiza un análisis de tres herramientas para generar tests automatizados de accesibilidad:

- Pa11y
- Axe-core
- Lighthouse

Se realiza un análisis de las tres herramientas, evaluando siete ítems que se observan en la tabla:

Tabla1. Tabla comparativa de herramientas de accesibilidad.

Características	Pa11y	Axe Core	Lighthouse
Configuración	Excelente	Excelente	Excelente
Usabilidad	Excelente	Muy Buena	Muy buena
Recursos	Excelente	Muy buena	Mala
Feeback	Excelente	Excelente	Bueno
Costo	Excelente	Bueno	Excelente
Limitaciones	Pocas	Pocas	Media
Fortalezas	Muchas	Muchas	Media

Facilidad de configuración

Pa11y requiere el uso de una interfaz de línea de comandos para la instalación, las pruebas y los informes. Para instalar Pa11y es necesario node.js previamente instalado. Pa11y se instala ejecutando el comando "npm install -g pa11y".

Axe-core, requiere node.js y la ejecución del comando "npm install axe-core". Además, Axe-core dispone de una extensión para navegadores web, disponible para Chrome, Firefox y Edge, pudiendo administrar su instalación desde la sección de administración de extensiones de su navegador.

Lighthouse está disponible automáticamente en Chrome. Se puede acceder a Lighthouse abriendo Chrome Developer Tools.

Usabilidad

Para usar Pa11y, se puede ejecutar una auditoría básica con un comando usando la URL del sitio web que se desea auditar, como "pa11y http://ejemplo.com". Se puede optar para que el informe se imprima en la pantalla, o pueda ser usado el parámetro "--reporter", para que se genere un archivo csv, cli, html, json o tsv.

Para realizar una auditoría con Axe-core por línea de comando, se ejecuta "axe http://ejemplo.com". En cuanto al informe, puede ser mostrado en pantalla o generar un archivo formato json.

Lighthouse en Chrome Devtools tiene una interfaz muy sencilla de usar. Le permite elegir ejecutar solo una auditoría de accesibilidad o ejecutar auditorías adicionales, que incluyen Rendimiento, Aplicación web progresiva, Mejores prácticas y SEO.

Recursos

Pa11y utiliza fuentes WCAG. El informe enumerará las pautas de las WCAG para el error en cuestión. Los usuarios pueden usar esos comentarios y seguirlos directamente hasta la documentación estándar de las WCAG.

La extensión de Axe utiliza reglas axe-core que cubre WCAG 2.0 y 2.1 para A y AA. Al ver cada problema dentro de la pestaña Herramientas de Desarrollo, se observarán las etiquetas WCAG a la derecha del panel que hacen referencia a una guía en particular.

Lighthouse, por su parte, posee una puntuación de accesibilidad, que se encuentra en la parte superior del informe. Estas evaluaciones de impacto del usuario se basan directamente en WCAG 2.0 y WCAG 2.1 para nivel A y AA.

Feedback

Pa11y con la configuración predeterminada, provee un exhaustivo análisis de accesibilidad para la WCAGAA, informando los problemas, y los comentarios sobre los mismos de forma precisa. También ofrece opciones de configuración para auditar, lo que permite brindar un informe más completo.

Axe, detecta de forma correcta los errores de accesibilidad, afirmando tener 0 falsos positivos. Una vez completa la auditoría, al hacer clic en un problema se

muestra un área de descripción que explica el mismo en relación con la guía WCAG y tiene información sobre cómo resolverlo. Junto con la descripción, cada problema muestra una calificación de impacto codificada por colores de "menor", "moderado", "grave" o "crítico".

Lighthouse genera una lista de todos los problemas de accesibilidad que detecta y una puntuación general de accesibilidad usando una escala de 0 a 100. Lighthouse posee documentación sobre cómo se calcula el puntaje de accesibilidad.

Costo

Pa11y es gratuito, de código abierto y no requiere la creación de ninguna cuenta.

La extensión del navegador Axe es de uso gratuito tanto para el navegador Chrome como para Firefox. En caso de querer integrar Axe con un proceso CI/CD, hay una versión paga llamada axe DevTools que ofrece más funciones.

Lighthouse es de uso gratuito en todas sus formas.

Limitaciones y Fortalezas

Las limitaciones de Pa11y radican en el nivel de comodidad del usuario con la línea de comandos. Configurar esta herramienta con todo lo que tiene para ofrecer, puede llevar mucho tiempo. La flexibilidad y escalabilidad de Pa11y son sus mayores fortalezas. Se puede utilizar con diferentes opciones para auditar una o más páginas con criterios específicos. Además, la documentación proporcionada es completa.

La extensión del navegador Axe es fácil de usar y para realizar una auditoría rápida. La desventaja es que se debe analizar manualmente una página a la vez.

La fortaleza de Lighthouse es que tiene la capacidad de ejecutar auditorías en una variedad de métricas en un solo lugar. Si bien implica una configuración más complicada, ejecutar Lighthouse mediante programación resolvería muchos de los desafíos que encuentra un usuario en la versión de DevTools.

3 Motivación

Para validar la existencia del problema, se tomó como punto de partida el informe de investigación llamado “El acceso a los servicios de la información y la comunicación y las personas con discapacidad”, el cual fue publicado en agosto de 2019.

Esta investigación fue realizada en Argentina, Chile y Uruguay. En Argentina, la prevalencia de población con “alguna dificultad” de 6 años y más es 10,2% (del total de la población argentina). En términos absolutos, se corresponde con una estimación de 3.571.983 personas, según datos del INDEC, 2018. Cabe resaltar, que se estima que más de mil millones de personas viven con algún tipo de discapacidad; es decir, alrededor del 15% de la población mundial, según la OMS en su informe en 2022.

Como conclusión, los informes de los tres países señalan falencias en cuestiones de accesibilidad, indicando incumplimientos específicos sobre los derechos de las personas con discapacidad en general. En lo que a Argentina refiere, y sabiendo de la existencia de la Ley de Accesibilidad de la Información en las Páginas Web (Ley 26.653), en el informe se puede observar el incumplimiento de la misma.

La principal motivación del presente trabajo, es presentar una solución que ayude a los desarrolladores a probar la accesibilidad web durante el proceso de desarrollo y así realizar páginas accesibles llegando al mayor número de personas posibles.

El trabajo se desarrolla durante el año 2022 en el contexto de Proyecto Final de Ingeniería en Informática (PFI) de Rodrigo Lema y Gonzalo Fuentes, con Pablo Pandolfo como tutor.

4 Aportes del trabajo

Al comparar las tres herramientas de análisis de accesibilidad web, se observa, que Pa11y es la mejor opción. Es una herramienta gratuita y su configuración predeterminada encuentra una gran cantidad de errores y es sencilla con su informe. Sin embargo, un requisito previo para Pa11y, es el uso de la línea de comandos. Si eso no es una preocupación, Pa11y tiene muchas configuraciones y funciones adicionales disponibles para personalizar. Aunque Pa11y tiene algunos desafíos con sus requisitos previos y configuración para pruebas más sólidas, su flexibilidad lo convierte en una herramienta muy robusta en su caja de herramientas de accesibilidad.

El uso de una herramienta automatizada como parte de su flujo de trabajo de accesibilidad, es el primer paso para desarrollar aplicaciones web que todos pueden usar. Si bien la extensión del navegador axe detecta la mayoría de los problemas, no reemplaza las pruebas manuales para detectar problemas de accesibilidad con elementos como la navegación con el teclado y el texto de enlace no específico.

Lighthouse en DevTools tiene mucho potencial como herramienta de auditoría, pero se siente subdesarrollado en muchos aspectos ya mencionados. Sin embargo, para un usuario que esté interesado en realizar una auditoría de accesibilidad y no necesite entrar en detalles (como las especificaciones WCAG), esta sigue siendo una buena manera de detectar muchos problemas de accesibilidad.

Luego de este análisis, se utilizará Pa11y como parte de la creación de una librería que pueda ser integrada a un proceso de integración continua (CI/CD), y que la misma realice pruebas de accesibilidad de acuerdo a la normativa argentina vigente.

5 Posibles líneas de investigación futura

A futuro, vemos necesario el análisis de estas tres herramientas en el contexto de WCAG 3.0, pronto a ser publicada, e incorporar nuevos ítems de análisis.

Referencias

1. LUJAN MORA, Sergio. Accesibilidad en la Web. [en línea]. 2006.
2. REVILLA MUÑOZ, Olga. WCAG 2.0 de forma sencilla. 1a. ed.: Itákora Press, 2013.
3. SEGOVIA, Claudio. Accesibilidad en Internet. 1a. ed.: Creative Commons, 2006.
4. W3C. [en línea]. 1994. <https://www.w3.org/>
5. WAI. Web Accessibility Initiative. [en línea]. 1998. <https://www.w3.org/WAI/>
6. WCAG.. [en línea]. 2008. <https://www.w3.org/TR/WCAG20>

Realidad Virtual por alumnos y para alumnos de UTN FRBA

Franco Cortínez[1], Gabriel Montenegro[1], Cinthia Vegega[1][0000-0002-5382-7875] y María F Pollo-Cattaneo[1][0000-0003-4197-3880]

[1] Grupo de Estudio de Metodologías para Ingeniería en Software (GEMIS)
Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Argentina

fcortnez@frba.utn.edu.ar, gmontenegroaguilar@frba.utn.edu.ar,
cinthia.vegega@gmail.com, flo.pollo@gmail.com

Abstract. En el marco de las becas de investigación a alumnos que brinda la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional y, como continuación del proyecto comenzado en el año 2020, llamado "Aula Virtual en Realidad Virtual" (realizado dentro de las actividades del grupo GEMIS), el presente trabajo tiene como objetivo obtener una beta funcional de la plataforma donde se encuentra alojada el aula virtual de forma que alumnos y docentes de la facultad puedan conectarse y compartir recursos. Tras el desarrollo, se incorporan diferentes funcionalidades nuevas a la aplicación, y se mejoran las preexistentes, logrando generar un mejor entorno para el usuario.
La Realidad Virtual es la generación de un entorno a través de una simulación computarizada. Para la creación de dicho entorno se utiliza una amplia variedad de tecnologías y dispositivos de inmersión.

Keywords: Realidad Virtual, ChatBot, Plataforma Online, Educación

1 Motivación

En esta sección se exponen las diferentes causas que motivaron al desarrollo del trabajo actual y que hicieron posible su continuidad. En la subsección 1.1 se detalla el contexto académico que promueve la realización del proyecto, en la subsección 1.2 se presentan los antecedentes del proyecto junto con las metas alcanzadas previas al desarrollo del presente trabajo, en la subsección 1.3 se define el concepto de Realidad Virtual y se presentan diferentes trabajos que hacen uso de esta tecnología aplicada a la educación.

1.1 Contexto Académico

Dentro del ámbito de la Universidad Tecnológica Nacional, Facultad Regional Buenos Aires (UTN FRBA), en el marco de las actividades del Grupo de Estudio de Metodologías para Ingeniería en Software (GEMIS) [1], bajo la dirección de la Dra. María Florencia Pollo Cattaneo y la coordinación de la Mg. Cinthia Vegega, se lleva a cabo desde el año 2020 el desarrollo de un Aula Virtual, utilizando la Realidad Virtual

como tecnología principal. El equipo de trabajo del presente año se encuentra formado por once alumnos de la carrera de Ingeniería en Sistemas de Información en la UTN FRBA y cuatro integrantes que pertenecen a una institución técnica secundaria. El grupo GEMIS tiene como objetivo la obtención de nuevos conocimientos y la motivación para que sus miembros asciendan dentro del escalafón de la carrera de investigadores. Los miembros pertenecientes al grupo GEMIS poseen una firme vocación de trabajo en las áreas de informática, sistemas de información, metodología y buenas prácticas, ingeniería en software, sistemas inteligentes y su vinculación con la explotación de información y tecnología educativa, de manera tal que canalizan y proveen una base sustentable de aporte a los proyectos en desarrollo.

1.2 Antecedentes del Proyecto

Durante el año 2020, en un contexto de pandemia global, un grupo de estudiantes de la asignatura Sistemas y Organizaciones (materia integradora de primer nivel) perteneciente a la carrera Ingeniería en Sistemas de Información de la UTN FRBA, se dispusieron a desarrollar un prototipo funcional basado en Realidad Virtual (RV) [2], logrando el objetivo de crear un espacio áulico donde puedan interactuar docentes y alumnos. Dicho proyecto fue llamado Aula Virtual de Realidad Virtual (AVRV). En octubre del mismo año, se presenta AVRV como prototipo en la octava edición del Congreso Nacional de Ingeniería Informática y Sistemas de Información (CONAIISI), contando con el modelo 3D de todo el entorno áulico y permitiendo que se conecte un usuario de manera local a la simulación.
Posteriormente, en el año 2021, otro grupo de estudiantes, de la asignatura Análisis de Sistemas de la carrera de Ingeniería en Sistemas de Información de la UTN FRBA, se decide dar continuidad al proyecto [3], agregando nuevas funcionalidades que permiten generar una experiencia mejorada al usuario (apertura de archivos ".pdf" en el cuaderno del estudiante y permitir al docente habilitar cámara, siendo presentada la misma en el pizarrón, entre otras). Por otra parte, el mismo año, se incorporan dos alumnos investigadores becarios, que desarrollan un ChatBot, que pudiera interactuar con los estudiantes, para luego incorporarlo al proyecto AVRV [4]. En octubre de 2021, se presentan dos artículos del proyecto: "AVRV - Una nueva forma de aprender" y "AVRV - Una nueva forma de aprender a través del ChatBot", en CONAIISI 2021, logrando obtener el segundo puesto entre más de 180 trabajos presentados con el desarrollo del ChatBot.

1.3 Realidad Virtual

La Realidad Virtual (RV) consiste en la creación de un entorno a través de una simulación computarizada, utilizando dispositivos de inmersión, con la finalidad de que el usuario experimente una interacción lo más cercana posible a la realidad con el entorno creado [5]. La RV resulta una herramienta útil para la enseñanza, gracias a la incorporación de los sentidos, que refuerzan el aprendizaje ya que el 95% de la información es obtenida a través de los sentidos del tacto, vista y oído, que son los sentidos sobre los que trabaja principalmente la RV. Es por ello que es una herramienta eficaz para el aprendizaje [6]. Un ejemplo de esto es el estudio realizado por la Universidad Nacional Tecnológica de Lima [7], en el que se utiliza RV con el objetivo

de explicar Geometría Fractal (temática que no está dentro del programa educativo nacional de Perú) a estudiantes secundarios, arrojando resultados positivos con respecto a la asimilación por parte de los alumnos de dicha temática. Por otra parte, la Universidad de Alicante implementa una aplicación que hace uso de RV para utilizar como herramienta en la educación secundaria llamada VR Student [8], la cual destaca por su utilidad para el aprendizaje de los estudiantes y el apoyo para la resolución de ejercicios. Por último, la Universidad Tecnológica Nacional, Facultad Regional de Mendoza realizó una investigación [9], en la cual afirma que la RV es una herramienta eficaz para las capacitaciones laborales, en especial aquellas que acarrean un alto riesgo.
La variedad de beneficios mencionados en múltiples investigaciones son la motivación de la creación del proyecto AVRV que busca innovar en los métodos de educación.

2 Objetivos y Aportes

El objetivo original del proyecto es desarrollar un espacio áulico utilizando la Realidad Virtual como tecnología principal que permita que alumnos y docentes se conecten e interactúen dentro de la plataforma, produciendo una sensación de naturalidad y presencialidad en relación a la interacción con el entorno, facilitando así el aprendizaje en la educación a distancia.
Para el presente año se plantea dar continuidad al desarrollo de la plataforma iniciada en el año 2020 y continuado en 2021, con el fin de obtener una Beta funcional (Versión 1.0). La consecución del objetivo definido en el proyecto propuesto dispone de seis sub proyectos, independientes entre sí, los cuales son realizados por distintos grupos coordinados por los autores del presente trabajo y se describen a continuación:

- Modo multi-usuario: permitir que más de un usuario se conecte a la plataforma de manera simultánea, utilizando una arquitectura de red cliente/servidor. La generación del servidor es proveída por la herramienta “Photon Unity Networking” (PUN) utilizando el servicio “Photon Cloud” o, realizando un servidor dedicado en el ordenador del usuario que desee generar un aula.
- Comunicación Voz/Texto: con la finalidad de permitir que los usuarios puedan interactuar entre sí, se produce dicho sub proyecto cuyo objetivo es realizar un canal de comunicación vía voz utilizando el micrófono o, vía texto mediante el teclado. Debido a compatibilidad y adaptación a las necesidades del proyecto se decide utilizar la herramienta PUN, que provee las herramientas necesarias para cumplir con dicho objetivo.
- Rediseño: diseñar nuevas texturas para los modelos 3D ya existentes en la plataforma, implementar una interfaz gráfica de usuario para el menú, crear pantallas de conexión de aulas y agregar "avatares" con texturas con el objetivo de obtener una interfaz más agradable para el usuario.
- Actualización del ChatBot: a través del uso de Inteligencia Artificial y la herramienta DialogFlow [10], se busca desarrollar una actualización del ChatBot realizado durante el año 2021, cuya finalidad es interactuar con los usuarios y proporcionar respuestas e información de ayuda acerca del uso de la plataforma, versión actual, datos de los creadores y contacto con soporte. La actualización consiste en agregar

una mayor variedad de respuestas e interacciones que resulten de utilidad para el usuario.

- Sistema de Herramientas: desarrollar una aplicación dentro de la plataforma la cual ofrece herramientas útiles para los usuarios, como el ChatBot, calculadora, bloc de notas, tabla periódica, mapas, pasaje entre sistemas numéricos. La realización de dicho desarrollo y las herramientas propuestas se realizan utilizando Unity y C#, para lograr una compatibilidad con el proyecto base.
- Reconocimiento de Gestos: mediante la utilización de Inteligencia Artificial y Machine Learning realizar un sistema de reconocimiento de gestos para que el usuario, mediante la utilización de una webcam, pueda interactuar con ciertas funciones de la plataforma realizando gestos frente a la cámara. Por ejemplo: cuando el usuario deja su mano abierta por 3 segundos frente a la cámara, en el aula aparece un mensaje informando que dicho usuario se encuentra levantando la mano.

Para la realización de cada sub proyecto presentado se planifica y establece un tiempo de ejecución el cual se representa en el diagrama de Gantt (ver Fig. 1).

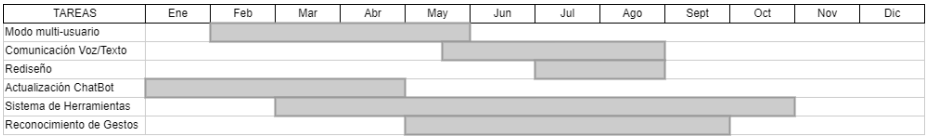


Fig. 1. Diagrama de Gantt representativo de tiempos de cada sub proyecto.

3 Estado Actual y Trabajo Futuro

El presente trabajo se encuentra en proceso de desarrollo. A continuación, se detalla el estado actual de cada sub proyecto y se exponen los grupos que participan de cada uno:

- Modo multi-usuario: sub proyecto finalizado. El mismo fue desarrollado por dos alumnos de la carrera de Ingeniería en Sistemas de Información y se encuentra implementado en la versión 0.3 del proyecto, el cual está disponible para su descarga en la sección versiones de la página web oficial del proyecto [11].
- Comunicación Voz/Texto: se ha implementado el plugin Photon Voice y configurado correctamente los componentes Photon Voice View y Speaker, posibilitando la comunicación de hasta 4 usuarios en simultáneo. Actualmente el sub proyecto se encuentra en su etapa final de desarrollo donde se busca implementar dicho desarrollo en la plataforma. Este sub proyecto es realizado por los mismos alumnos que el Modo multi-usuario.
- Rediseño: la texturización de los modelos 3D se encuentran incluidas al proyecto y se cuenta con 3 avatares disponibles para su uso, Se han finalizado los diseños para la interfaz gráfica de usuario, sin embargo, la misma, no se encuentra implementada aún en la plataforma. Dicho sub proyecto es coordinado por una estudiante de la carrera de Ingeniería en Sistemas de Información y cuenta con el apoyo de cuatro estudiantes de una institución técnica.
- ChatBot: actualmente se encuentra en una fase de actualización y mantenimiento por un estudiante de la carrera, se han agregado 29 intents nuevos y se han establecido 4 grupos nuevos para dichos intents, “Como apoyar al proyecto”, “Contacto”, “Quienes

somos" y "Redes sociales". Además, se han agregado 16 respuestas para posibles dudas del usuario. La implementación de este sub proyecto en la plataforma se realiza como una herramienta más en el Sistema de Herramientas.

- Sistema de Herramientas: se encuentra en desarrollo por un grupo de seis alumnos de la asignatura Sistemas y Organizaciones de la carrera de Ingeniería en Sistemas de Información. A la fecha, se ha desarrollado correctamente una calculadora, una tabla periódica interactiva y un buscador online (actualmente se encuentra en desarrollo un mapa político del mundo y un bloc de notas). Finalmente, resta por desarrollar un "App Center", sistema que contiene todas las aplicaciones para su uso, e integrarlo al Aula Virtual.
- Reconocimiento de Gestos: se encuentra en etapa de desarrollo por los coordinadores del proyecto. Actualmente se cuenta con un sistema que realiza una detección y seguimiento de manos mediante Inteligencia Artificial, además, se desarrolla un clasificador de imágenes en tiempo real utilizando Inteligencia Artificial y Machine Learning. Por último, se deben unir ambos desarrollos para cumplir correctamente con este sub proyecto entrenando el clasificador de imágenes con distintos gestos y así poder ser implementado en la plataforma.

En general, las actividades se encuentran en desarrollo y monitorización constante de parte de los alumnos responsables y los dos coordinadores, investigadores formados. Como futuras líneas de trabajo se prevé cumplir con las actividades previstas hasta fin de año para luego establecer los nuevos desafíos previstos para el año 2023. Los mismos estarán relacionados con finalizar los desarrollos pendientes luego de la realización de la Beta funcional en el presente año.

Referencias

[1] GEMIS, https://grupogemis.com.ar, último acceso 23/07/2022.

[2] Cortinez, F. M., Del Campo Kenny, F., Kalinin, A., Mariano, M. G., Montenegro, L. N., De la Torre, M., Vega, L. G.: "Aula Virtual en Realidad Virtual". En: 8vo Congreso Nacional de Ingeniería Informática y Sistemas de Información, pp. 918 (2020).

[3] Cortinez, F. M., Fernández, D., Porzolis Requena, A., Carrasco, C. T., Risberg, M. E., Ruiz, F. E., Corbalán S. G.: "AVRV - Una nueva forma de aprender". En: 9no Congreso Nacional de Ingeniería Informática y Sistemas de Información, pp. 700 (2021).

[4] Cortinez, F. M., Afonso, M., Corbalán, S.: "AVRV - Una nueva forma de aprender: Chatbot". En: 9º Congreso Nacional de Ingeniería Informática y Sistemas de Información, pp. 657-662 (2021).

[5] Levis, D. "¿Qué es la realidad virtual?", https://bit.ly/3dR7VPX, Último acceso 23/07/2022.

[6] "La realidad virtual como recurso y herramienta útil para la docencia y la investigación", Zapatero, D. https://bit.ly/3AcGCXD, último acceso 23/07/2022.

[7] Chavil M., Dante. Romero, I., Rodríguez, J. "Introducción al concepto de fractal en enseñanza secundaria usando realidad virtual inmersiva", https://bit.ly/3pEKKL6, último acceso 23/07/2022.

[8] Romero, D. "VR Student Desarrollo de una aplicación de Realidad Virtual para el refuerzo en educación secundaria obligatoria", https://bit.ly/3wKUBDp, último acceso 28/07/2022.

[9] Pérez, S., Muñoz, A. Stefanoni, M., Carbonari, D. "Realidad virtual, aprendizaje inmersivo y realidad aumentada:
Casos de Estudio en Carreras de Ingeniería", https://bit.ly/3TdeSe7, último acceso 28/07/2022.

[10] DialogFlow, https://dialogflow.cloud.google.com, último acceso 31/07/2022.

[11] AulaVirtualRV, https://aulavirtualrv.com.ar/versiones, último acceso 20/08/2022.

Implementación en SHACL de reglas de verificación de consistencia semántica para gestión de requisitos

Luciana Tanevitch[1][0000−0002−5322−9314], Diego Torres[1,2][0000−0001−7533−0133], Leandro Antonelli[1][0000−0003−1388−0337], and Alejandro Fernández[1][0000−0002−7968−6871]

[1] LIFIA, CICPBA-Facultad de Informática, UNLP
{nombre.apellido}@lifia.info.unlp.edu.ar
[2] Departamento de Ciencia y Tecnología, UNQ

Resumen Los sistemas de gestión de proyectos permiten administrar los requerimientos que serán desarrollados. A medida que los proyectos crecen en complejidad y comienzan a intervenir más personas, es más probable que se generen inconvenientes a causa de errores en las especificaciones. La Web Semántica dispone de tecnologías para la formalización de conceptos y validación de datos que pueden aplicarse en la Ingeniería de Requerimientos para mitigar estos inconvenientes. El objetivo de este trabajo es implementar un conjunto de reglas de verificación de consistencia, completitud y calidad de requerimientos usando el lenguaje SHACL. El método propuesto es aplicado sobre un conjunto de requerimientos para mostrar la aplicabilidad y usabilidad. Este trabajo se enmarca dentro del proyecto I+D+I con alumnos "Soporte semántico para mejorar la calidad de los requerimientos" y el trabajo final de la materia Tecnologías para la Web Social Semántica, ambos desarrollados en la Facultad de Informática, Universidad Nacional de La Plata.

Keywords: Ontología de Requerimientos · Grafos de Conocimiento · Shapes

1. Introducción

El Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) define un requerimiento como la capacidad o condición que un sistema debe poseer para satisfacer un documento formal de especificaciones [1]. Las herramientas de gestión de proyectos son ampliamente utilizadas en la industria del desarrollo de software ya que permiten la creación de tareas que describen requerimientos, tareas para la gestión de errores de implementación, asignación de personas encargadas de esas tareas, manejo del estado de desarrollo de una tarea (si está sin asignar, en curso o resuelta), gestión de riesgos, prioridades y costos. Sin embargo, el uso de una herramienta con estas características no garantiza la correctitud de los requerimientos, ya que los mismos pueden sufrir inconsistencias, no estar relacionados con las metas del desarrollo, o poseer problemas de calidad asociado

a riesgos innecesarios. En este sentido, la herramienta se limita a cumplir la funcionalidad que las personas ejecutan, sin verificar qué es lo que ejecutan, y esto podría conducir a un producto defectuoso.

Por su parte, la Web Semántica [3] introdujo estándares para la representación de contenido para que sea fácilmente comprensible por autómatas y así manipular y derivar información compleja de obtener para una persona. La creación y uso de taxonomías y ontologías pueden ser la base para la obtención, estructuración y gestión de la información relevante de los requerimientos [11]. Con el motivo de agilizar el desarrollo y reducir costos, el reúso de ontologías resulta una buena práctica [9]. A partir de una ontología se pueden construir bases de conocimiento representadas en un grafo de conocimiento que integre diversos conceptos. Los grafos de conocimiento describen entidades del mundo real y sus relaciones, organizándolas en forma de grafo [10] usualmente modelado en RDF para facilitar la interoperabilidad en la Web. Además, permiten derivar nuevo conocimiento a partir de la información que contienen [5].

Existen diferentes enfoques en el área de la Web Semántica que permiten representar y validar requerimientos. Diversos trabajos proponen el uso de ontologías para la conceptualización de requerimientos. En el marco de SoftWiki, un programa para la obtención, organización y gestión de requisitos, se construyó la ontología SWORE (SoftWiki Ontology for Requirements Engineering) para describir un pequeño subconjunto de los aspectos de Ingeniería de Requerimientos tales como Requerimiento y Stakeholder, pero también está alineada a FOAF y SIOC para agregar discusiones y comentarios [12], funciones muy utilizadas en las herramientas de gestión de proyectos. Antonelli et al. [2] propone el uso de una ontología para la representación de requerimientos escritos como Escenarios, y un conjunto de consultas en lenguaje SPARQL para identificar ciertos atributos de calidad de los requerimientos, como la consistencia y completitud. Siegemund et al. [14] propone un enfoque para capturar y validar la completitud y consistencia de un conjunto de requerimientos de proyectos de desarrollo usando ontologías y razonamiento basado en inferencias. En su tesis doctoral [13], detalla una ontología (Requirements Ontology) para la documentación de requisitos acorde a la especificación IEEE 830, y el conjunto de reglas basadas en axiomas para realizar verificaciones automáticas de diferentes criterios de calidad.

Si bien los lenguajes RDFS y OWL permiten definir restricciones y axiomas para realizar validaciones, éstas están limitadas a detectar inconsistencias lógicas. Shapes permite definir un conjunto de restricciones aplicables a un grafo para realizar validaciones sobre éste [7]. SHACL es un lenguaje para validar grafos de datos mediante un conjunto de condiciones representadas como shapes. La contribución de este trabajo es describir las reglas para la verificación de consistencia, calidad y completitud en requerimientos, a partir de las definidas por Siegemund et al. [13], utilizando el lenguaje SHACL, y aplicarlas sobre un grafo de conocimiento construido a partir de información obtenida de un programa de gestión de proyectos del estilo que maneja el producto Jira. Este trabajo se desarrolla en el contexto del proyecto I+D+I con alumnos “Soporte semántico

para mejorar la calidad de los requerimientos" y el trabajo final para la materia Tecnologías para la Web Social Semántica, en la Facultad de Informática de la Universidad Nacional de La Plata.

El artículo está organizado de la siguiente manera: la sección 2 introduce el uso de técnicas de Web Semántica en proyectos relacionados a requerimientos, en la sección 3 se desarrolla la contribución presentada en este trabajo y finalmente la sección 4 resume los resultados y menciona posibles trabajos futuros en relación al propuesto.

2. Contexto y trabajos relacionados

Van den Bersselaar et al. [4] proponen un prototipo para la validación de requerimientos en proyectos de construcción. En su trabajo genera un grafo con datos de un modelo de información para construcción (BIM) y un grafo de shapes a partir de los requerimientos que deberían satisfacerse. Lin et al. [8] presenta una ontología de requerimientos que da soporte a un proceso de gestión de requisitos en diferentes aspectos como trazabilidad, completitud, consistencia, satisfactibilidad, etc. Utiliza axiomas lógicos para la definición de restricciones e inferencia de nuevo conocimiento. Happel y Seedorf [6] mencionan diferentes enfoques de aplicación de las ontologías en Ingeniería de Software. En particular, destaca que es importante asegurarse que todos los participantes de un proyecto comprendan de la misma manera el problema de dominio para evitar inconvenientes en las implementaciones a causa de ambigüedad e inconsistencia, y para esto, las ontologías tienen ciertas ventajas respecto a las especificaciones semi-estructuradas debido a que automatizan las validaciones de ambigüedad y consistencia.

Existen trabajos que describen ontologías específicas del dominio en el que se posiciona esta contribución. Tappolet et al. [15] define EvoOnt, una ontología para conceptualizar la evolución del software que tiene como objetivo simplificar las tareas relativas al análisis, aplicando técnicas de la Web Semántica. SEON [16] es una familia de ontologías para el análisis semántico de la evolución del software, una de sus componentes es una ontología para representar conceptos de Jira. El vínculo de estos trabajos con el presente está dado porque las herramientas de gestión están centradas en la descripción de funcionalidades a través de *issues*, que inherentemente describen requerimientos, y conocer las relaciones entre requerimientos permitirá detectar luego los inconvenientes planteados en el dominio de estas herramientas.

3. Contribución

La contribución de este trabajo se desarrollará a lo largo de cuatro etapas que se detallan a continuación.

Definición de la ontología. Se requirió determinar los conceptos que permitirían describir los requerimientos de un proyecto. Con el objetivo de que la conceptualización sea apropiada para un dataset de un sistema de gestión de proyectos, se reusaron las ontologías SWORE y Requirements Ontology (RO),

alineando aquellos conceptos que semánticamente son equivalentes. Por ejemplo SWORE:Requirement se alineó a RO:Requirement. Se realizaron modificaciones mínimas para reemplazar las clases RO:LevelOfCost y RO:LevelOfRisk por dos propiedades análogas a las que define SWORE para *priority* y *status*.

Implementación de las reglas de verificación. En base a las reglas planteadas en el trabajo de Siegemund et al. [13], se adecuaron aquellas que resultan acordes al contexto utilizado y permitan validar la consistencia, completitud y calidad de los datos. El autor propone escoger un subconjunto de requerimientos para realizar las verificaciones llamado *configuración de requerimientos*. En este trabajo se aplican sobre la totalidad de un proyecto. Se seleccionaron y tradujeron al lenguaje SHACL las siguientes reglas. Para la característica de Completitud se escogieron las reglas: (i) Al menos una meta debe ser especificada, (ii) Cada requerimiento debe estar asociado a al menos una meta, (iii) Para cada requerimiento, se debe definir si es obligatorio u opcional. Respecto a Consistencia, se seleccionó: (iv) No debe haber requerimientos conflictivos. Y para Calidad, (v) No debe haber requerimientos opcionales con un riesgo o costo alto. A modo de ejemplo, se detalla la implementación de las reglas (i), (iv), y (v). La totalidad de la implementación puede accederse en el repositorio del proyecto [3].

```
ex:GoalCounterShape
a sh:NodeShape ;
sh:targetNode ro:Goal ;
sh:property [
  sh:path [ sh:inversePath rdf:type ] ;
  sh:minCount 1 ;
  sh:message "Al menos una meta debe ser especificada"@es;
] .
```

Regla (i) Al menos una meta debe ser especificada.

```
ex:RequirementShape
a sh:NodeShape ;
sh:targetClass swore:Requirement ;
sh:property [
  sh:path ro:isInConflictWith;
  sh:maxCount 0 ;
  sh:message "No debe haber requerimientos conflictivos"
      @es;
];
```

Regla (iv) No debe haber requerimientos conflictivos.

```
ex:RequirementShape
a sh:NodeShape ;
sh:targetClass swore:Requirement ;
sh:sparql [
   a sh:SPARQLConstraint ;
```

[3] https://github.com/tanevitch/SHACL4J

```
    sh:message "No debe haber requerimientos opcionales con
        un riesgo o costo alto"@es;
    sh:select """
        SELECT $this WHERE {
          $this a swore:Requirement .
          $this ro:isMandatory false .
          {$this sw:cost "high"} UNION {$this sw:risk "high"}
        }""";
] ;
```

Regla (v) No debe haber requerimientos opcionales con un riesgo o costo alto.

Construcción de un grafo de requerimientos. Se construyó un programa en Python que a partir de un archivo de requerimientos y la ontología definida genere un grafo RDF. Esta tarea requirió analizar cómo debían ser vinculados los datos disponibles a los conceptos que define la ontología.
Validación del grafo. A partir del grafo de datos y las reglas generadas anteriormente, el programa permite ejecutar la validación, emitiendo un informe con los inconvenientes detectados.

4. Conclusiones y trabajo futuro

En este trabajo se propone la utilización de técnicas de la Web Semántica para la verificación de requerimientos, a partir de un sistema de gestión de proyectos. El enfoque propone incorporar a estos sistemas la asistencia en la detección de problemas en la definición de los requerimientos. Jira, Trello y Pivotal Tracker son algunas de las alternativas populares que los equipos de desarrollo ágil pueden escoger para su proyecto. En general, estas herramientas permiten exportar los datos en un formato tabulado, lo que resulta práctico para procesar. Además, este artículo presenta un prototipo que permite la construcción de un grafo de conocimiento con datos obtenidos de un proyecto gestionado con el producto Jira. La propuesta de este artículo puede ser aplicada en diferentes etapas de desarrollo de un proyecto. Si el mismo se encuentra aún en desarrollo, la propuesta permite detectar en forma temprana incongruencias y minimizar los costos que produciría tratar estos errores en etapas más avanzadas. Por otra parte, si el proyecto ya ha finalizado, la validación del grafo de conocimiento permitiría detectar problemas relacionados a la calidad en la organización de los requerimientos en modo de retrospectiva general. Por ejemplo, analizar problemas en la carga de las tareas, inconsistencias cometidas o fallos en la carga de especificaciones. Queda pendiente para una futura contribución detallar el proceso de extracción de información de requerimientos escritos en lenguaje natural, y cómo se vincularían esos conceptos con las clases que define la ontología. Otra posible rama podría ser el desarrollo de una producto que pueda obtener automáticamente la información de un sistema de requerimientos, extraer la información y detectar el concepto más adecuado, definido por una ontología, con el que debe vincularse.

Referencias

1. IEEE Standard Glossary of Software Engineering Terminology. IEEE Std 610.12-1990 pp. 1–84 (1990). https://doi.org/10.1109/IEEESTD.1990.101064
2. Antonelli, L., Torres, D., Hozikian, M., Hernandez, J.E.: Semantic Support for Scenarios to Improve Communication in Agribusiness. In: Camarinha-Matos, L.M., Afsarmanesh, H., Antonelli, D. (eds.) Collaborative Networks and Digital Transformation, vol. 568, pp. 447–456. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28464-0_38, http://link.springer.com/10.1007/978-3-030-28464-0_38
3. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5), 34–43 (2001), publisher: JSTOR
4. van den Bersselaar, E., Heinen, J.J., Chaudron, M., Pauwels, P.: Automatic Validation of Technical Requirements for a BIM model using Semantic Web Technologies: 1st 4TU/14USA research day on Digitalization in the Built Environment (2022)
5. Ehrlinger, L., Wöß, W.: Towards a Definition of Knowledge Graphs
6. Happel, H.J., Seedorf, S.: Applications of ontologies in software engineering. In: Proc. of Workshop on Sematic Web Enabled Software Engineering"(SWESE) on the ISWC. pp. 5–9. Citeseer (2006)
7. Hogan, A.: The Web of Data. Springer International Publishing (2020), https://books.google.com.ar/books?id=4CPlzQEACAAJ
8. Lin, J., Fox, M.S., Bilgic, T.: A requirement ontology for engineering design. Concurrent Engineering **4**(3), 279–291 (1996)
9. Nogueira, G.G., Barcellos, M.P., Souza, V.E.S.: Towards a characterization to aid in ontology reuse. In: ONTOBRAS. pp. 138–143 (2021)
10. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic web **8**(3), 489–508 (2017)
11. Riechert, T., Lauenroth, K., Lehmann, J.: Swore - softwiki ontology for requirements engineering. pp. 111–118 (01 2007)
12. Riechert, T., Lauenroth, K., Lehmann, J., Auer, S.: Towards semantic based requirements engineering. In: Proceedings of the 7th International Conference on Knowledge Management (I-KNOW). Citeseer (2007)
13. Siegemund, K.: Contributions to ontology-driven requirements engineering. PhD Thesis, Citeseer (2015)
14. Siegemund, K., Thomas, E.J., Zhao, Y., Pan, J., Assmann, U.: Towards ontology-driven requirements engineering. In: Workshop semantic web enabled software engineering at 10th international semantic web conference (ISWC), Bonn (2011)
15. Tappolet, J., Kiefer, C., Bernstein, A.: Semantic web enabled software analysis. Web Semantics: Science, Services and Agents on the World Wide Web **8**(2), 225–240 (Jul 2010). https://doi.org/10.1016/j.websem.2010.04.009, https://www.sciencedirect.com/science/article/pii/S1570826810000338
16. Würsch, M., Ghezzi, G., Hert, M., Reif, G., Gall, H.: SEON: A Pyramid of Ontologies for Software Evolution and its Applications. Computing **94**, 1–31 (Nov 2012). https://doi.org/10.1007/s00607-012-0204-1

Detección de somnolencia utilizando técnicas de visión artificial en entornos móviles.

Autores: Macarena Quiroga, Emilio Melo
Director: Martín Bilbao

Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Comodoro Rivadavia, Chubut. {meloemilio.prog,msinf.quiroga}@gmail.com, martinbilbao@ing.unp.edu.ar

Resumen Este trabajo tiene como objetivo presentar el desarrollo e implementación de un dispositivo inteligente de detección de somnolencia en conductores empleando principalmente técnicas de visión artificial y una Raspberry Pi 4, con el fin de alertar al conductor y prevenir la ocurrencia de potenciales accidentes de tránsito. En primera instancia, se realizará una introducción a la temática abordada, junto los conceptos claves necesarios para el desarrollo del trabajo. Posteriormente, se hará mención sobre el desarrollo de los datasets, las comparativas y estudios de escalabilidad usando los modelos preentrenados MobileNetV2 e InceptionV3, el proceso de detección de somnolencia empleando como técnicas de visión artificial los modelos de CNN mencionados y los clasificadores basados en cascadas de Haar, finalizando con las pruebas en la Raspberry Pi en un entorno real de conducción.

Keywords: Detección de somnolencia · Raspberry Pi · Aprendizaje por transferencia · Vision artificial.

Contexto

El trabajo forma parte de la tesina de grado realizada para la obtención del título "Licenciatura en Informática", presentada en Agosto de 2022 en la Universidad Nacional de la Patagonia San Juan Bosco, sede Comodoro Rivadavia. El objetivo principal ha sido desarrollar e implementar un dispositivo inteligente que permita detectar, a través de una cámara y en tiempo real, si un conductor presenta indicios constantes de somnolencia. La motivación para su desarrollo fueron los altos índices de mortalidad en Argentina y en el mundo ocasionados por los accidentes de tránsito, siendo la somnolencia una de sus causantes.

1. Introducción

La somnolencia puede ser definida como la necesidad de conciliar el sueño o una tendencia a quedarse dormido. Es un proceso, resultado del ritmo biológico humano normal, el cual consiste en ciclos que involucran contraer sueño y estar despierto[1].

Las Redes Neuronales Convolucionales (CNN) y el empleo de técnicas de visión artificial, permiten retratar el comportamiento del ojo humano, posibilitando así el reconocimiento de objetos a través de la captura de imágenes, con el objetivo de realizar algún tipo de inferencia y actuar en consecuencia.

Las técnicas de visión artificial están logrando avances significativos en diversos ámbitos siendo ampliamente utilizadas en diferentes tipos de aplicaciones, ofreciendo grandes ventajas que conllevan a generar un impacto favorable en la sociedad, contrarrestando las problemáticas existentes en la actualidad. Una de ellas, es la alta tasa de accidentes viales, causada en gran medida por conductores somnolientos.

El 30 % de los accidentes graves están relacionados a la somnolencia y/o fatiga. Al conducir cansado o con sueño, por consumo de medicamentos contraproducentes, ingesta de bebidas alcohólicas, contar con trastornos del sueño, la conducción nocturna, entre otros factores, hacen que el riesgo de un accidente aumente[2]. Es por ello, que el desarrollo de un dispositivo inteligente portable para detectar la somnolencia, permitiría disminuir la probabilidad de ocurrencia de accidentes viales, reduciendo por ende, las pérdidas humanas y/o lesiones.

2. Desarrollo

La implementación del dispositivo inteligente involucró una serie de desarrollos, los cuales han sido la elaboración de datasets de imágenes de ojos abiertos y cerrados de rostros de personas, desarrollo de modelos basados en la arquitectura MobilenetV2 e InceptionV3, empleando la técnica transfer learning (o aprendizaje por transferencia); el ensamble del prototipo utilizando como componente principal la Raspberry Pi 4 modelo B y un script para realizar el proceso de detección de somnolencia.

2.1. DATASETS

Los datasets empleados para el entrenamiento de los modelos preentrenados seleccionados, han sido elaborados utilizando capturas de ambos ojos, pertenecientes a rostros de 5 (cinco) personas. Para realizar dichas capturas, se desarrolló un script en Python el cual utiliza la librería OpenCV para la interacción con la cámara OV5647 de la Raspberry Pi, junto con los clasificadores basados en cascadas de Haar ***haarcascade_righteye_2splits.xml*** y ***haarcascade_lefteye_2splits.xml***, obtenidos del repositorio oficial de OpenCV en GitHub, para la detección de ambos ojos. En el cuadro 1 se refleja la composición de los dos datasets elaborados.

Dataset	Cantidad de imágenes						Total
	Entrenamiento		Validación		Prueba		
	Abiertos	Cerrados	Abiertos	Cerrados	Abiertos	Cerrados	
1	500	500	200	200	14	14	1428
2	200	200	50	50	14	14	528

Cuadro 1: Composición de los dataset elaborados.

2.2. TRANSFER LEARNING

Aprendizaje por transferencia es una técnica que permite transferir conocimiento adquirido utilizando modelos preentrenados que poseen un gran conjunto de datos, para que puedan ser personalizados en una tarea determinada. Esto permite hacer uso de arquitecturas que están altamente probadas, permitiendo desarrollar modelos de manera rápida y eficiente, brindando soluciones a problemas complejos como son los de visión artificial. Las arquitecturas basadas en CNN para la detección de objetos que han sido utilizadas en el trabajo fueron MobilenetV2 e InceptionV3.

MobileNet es una arquitectura propuesta por Google. Fue diseñada para obtener una máxima eficiencia en la precisión, teniendo en cuenta los recursos limitados, para una aplicación de visión artificial, en los dispositivos móviles e integrados. Pueden utilizarse para detección, clasificación, incrustación y segmentación. Por otro lado, InceptionV3 es un modelo muy utilizado para el reconocimiento de imágenes. El mismo fue entrenado con el conjunto de datos de ImageNet, y ha demostrado lograr una exactitud mayor al 78.1 %[4].

Para cada modelo entrenado utilizando aprendizaje por transferencia, se realizó su correspondiente estudio de escalabilidad. La escalabilidad, en términos de un algoritmo de deep learning, hace referencia a la capacidad del mismo de adaptarse, manteniendo o mejorando su eficiencia cuando se modifica el tamaño o el volumen de entrada de los datos.

La metodología empleada para realizar el estudio de escalabilidad consistió en evaluar los modelos propuestos sobre los dos datasets presentados en la sección 5.1, aplicando tamaños de imágenes de entrada de 165x165 píxeles y 96x96 píxeles. Para cada tamaño, se realizó el entrenamiento partiendo de 4 épocas, aumentando este número hasta alcanzar una precisión igual o superior al 90 %, o hasta ver una decadencia de ésta u otras métricas que indiquen que el modelo no está mejorando. Finalmente, los dos modelos seleccionados, basados tanto en MobileNetV2 como en InceptionV3, fueron aquellos que obtuvieron un 100 % en exactitud, precisión, recall y F1 en las pruebas previas.

2.3. DISPOSITIVO INTELIGENTE

La construcción del prototipo partió de la utilización de una Raspberry Pi 4 con 4GB de memoria RAM, a la cual le fueron conectados una serie de disposivos.

Uno de ellos fue una cámara modelo OV5647 para realizar las correspondientes capturas del rostro del conductor, permitiendo además su utilización durante la noche gracias a los módulos infrarrojos incorporados. Adicionalmente, también se hizo uso de un buzzer activo de 5V y 90 decibeles para la alerta sonora, y un cooler de 5V para la refrigeración de la Raspberry. Por otra parte, se utilizó una memoria de clase 10 de 32 GB para el sistema operativo Raspberry Pi OS, el sistema operativo oficial.

Los modelos desarrollados fueron convertidos a un formato apto para poder ser ejecutados en la Raspberry Pi. El proceso de detección de somnolencia se llevó a cabo mediante un script desarrollado con el lenguaje de programación Python, utilizando los clasificadores basados en cascadas de Haar para detectar las regiones de interés (ojos), de las cuales se realizan capturas que luego son enviadas al modelo a fin de que el mismo prediga el estado de los ojos (abiertos o cerrados). La alarma de detección de somnolencia, entonces, es emitida si los ojos del conductor se detectan como cerrados durante 3 o más cuadros, emitiendo así la alerta sonora e incrementando un contador de somnolencia, el cual indica la cantidad de veces que la alarma fue disparada. En la figura 1 se puede observar el proceso de detección de somnolencia.

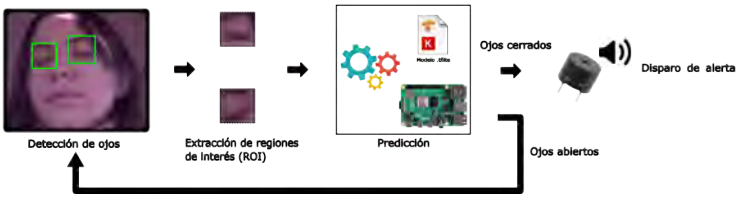


Figura 1: Esquema simplificado del proceso de detección de somnolencia.

3. Resultados

Se realizaron pruebas con ambos modelos durante el día y la noche, en el cual el modelo MobileNetV2 presentó un mejor desempeño en comparación a los tiempos de retardo en la detección de ojos y la emisión de alerta sonora de InceptionV3. Los tiempos correspondientes al retardo de captura y emisión de alerta utilizando MobiNetV2 fueron de 2.45 y 3.50 respectivamente, mientras que en caso de InceptionV3 fueron de 3.28 y 4.49 segundos. En vista de los tiempos antes mencionados fue que se decidió utilizar como modelo final en el dispositivo aquel basado en MobileNetV2, el cual ofrece una diferencia de 830 milisegundos con respecto al retardo de captura de la detección y 1 segundo en la emisión de la alerta.
En la figura 2 se puede observar la detección de somnolencia durante la noche, y en la figura 3 se visualiza la detección de ojos abiertos durante el día con la presencia de un falso positivo, detectando la boca como un ojo.

Figura 2: Detección de somnolencia durante la noche.

Figura 3: Detección de ojos abiertos durante el día.

Los scripts, datasets y modelos entrenados se encuentran disponibles para su acceso, descarga e instalación en el respositorio de GitHub [5].

4. Conclusiones y trabajos futuros

Se ha logrado desarrollar un dispositivo inteligente capaz de detectar indicios de somnolencia y alertar mediante una alarma sonora al conductor, haciendo uso de clasificadores Haar y un modelo entrenado basado en MobileNetV2 con una precisión del 100 %, demostrando así el beneficio de utilizar las técnicas de aprendizaje por transferencia y visión artificial en aplicaciones de tiempo real.

El dispositivo desarrollado, además de brindar una efectividad satisfactoria, sienta un precedente para mejoras y funcionalidades futuras, como incorporar nuevas técnicas y librerías que permitan ampliar los tipos de indicios de somnolencia detectados por el dispositivo, tales como el bostezo y cabeceo, permitir la detección de los ojos abiertos y cerrados con cascadas de Haar cuando el conductor utiliza anteojos, entrenando nuevos detectores que permitan alcanzar este objetivo y añadir un sistema de autenticación, ya sea por reconocimiento facial o tarjeta RFID, de tal forma que el conductor pueda registrarse al ingresar al vehículo y obtener así datos estadísticos por usuario.

Referencias

1. Aleksandar Colic, Design and implementation of driver drowsiness detection system, 2014.
2. Organización civil Luchemos por la vida, http://luchemos.org.ar. Ultimo acceso:10-06-22.
3. Mark Sandler and Andrew Howard and Menglong Zhu and Andrey Zhmoginov and Liang-Chieh Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, 2019.
4. Google Cloud, Advanced Guide to Inception v3, https://cloud.google.com/tpu/docs/inception-v3-advanced. Ultimo acceso: 10-03-22.
5. Repositorio Drowsiness Pi, https://github.com/DevTeamCR/DrowsinessPi.

AlfaDatizando: análisis de opciones para login unificado

Scopel Iván[1[0000-0002-1564-690X]], Gómez Lucia[1[0000-0002-2783-916X]], Lliteras Alejandra Beatriz [1,2[0000-0002-4148-1299]], Gardey Juan Cruz[1,3[0000-0002-1765-8189]], Grigera Julián[1,2,3[0000-0002-7962-4312]]

[1] UNLP, Facultad de Informática, LIFIA. 50 y 120. La Plata. Bs.As. Argentina
[2] CICPBA Bs.As. Argentina
[3] CONICET. Argentina
{iscopel,lgomez,lliteras, jcgardey, julian.grigera}@lifia.info.unlp.edu.ar

Abstract. En este trabajo se presenta un análisis de redes sociales usadas en la actualidad y segmentadas por grupos etarios. El objetivo es seleccionar un subconjunto de ellas para incluirlas, justificadamente, en el desarrollo de login unificado en la plataforma AlfaDatizando. Ésta es una plataforma de visualización de datos para trabajar Pensamiento Computacional en Humanidades Digitales diseñada e implementada para ser usada con estudiantes de colegios secundarios. Para establecer las redes sociales a considerar, se llevó a cabo un proceso en tres etapas, primero relevando de la bibliografía las redes sociales más usadas, luego descartando aquellas redes que no son posible usarlas en una implementación y por último, una breve descripción de las APIs de las redes sociales candidatas a ser implementadas. A partir de este estudio, se espera lograr una identificación, tanto por parte de los docentes, como de los estudiantes con las redes consideradas en esta plataforma para favorecer su apropiación.

Keywords: Login Unificado, Visualización de Datos, Pensamiento Computacional, Humanidades Digitales, Ciencias Sociales

1 Motivación

El presente trabajo se enmarca en el Proyecto de Innovación con Alumnos (2022), de la Facultad de Informática, UNLP, llamado "Aprendo con Datos. Plataforma para la visualización de datos con fines educativos en nivel secundario para Ciencias Sociales y Humanidades". La temática abordada es parte del proyecto de doctorado de la profesora Lliteras Alejandra.

El pensamiento computacional es una de las habilidades requeridas para el siglo XXI [1] y una de las maneras de desarrollarlo es mediante la visualización de datos [2]. Si bien existen diversas plataformas que permiten la visualización de datos (por ejemplo, Tableau[1] y SocioViz[2]), pocas son creadas con fines educativos (por ejemplo, CODAP[3]y entre ellas no consideran la posibilidad de generar y reusar actividades de visualización y administrar la resolución de las actividades por parte de los estudiantes y considerar el feedback a partir de esto entre docente y estudiantes.

[1] https://www.tableau.com/
[2] https://socioviz.net/
[3] https://codap.concord.org/

A raíz de lo anterior, surge el diseño e implementación de AlfaDatizando[4] [3], una plataforma para la visualización de datos para desarrollar Pensamiento Computacional con estudiantes secundarios, en las áreas de las Humanidades Digitales y Ciencias Sociales.
Una primera versión de la plataforma AlfaDatizando, implementa registro de usuarios y login propio. En una siguiente versión se espera incluir el login unificado, funcionalidad ampliamente incorporada en las diferentes plataformas disponibles. Sin embargo, dado que se trata de una plataforma educativa, se decidió investigar respecto a las redes sociales que usan las personas de diferentes rangos etarios, ya que se espera una identificación tanto por parte de los docentes como de los estudiantes con las redes consideradas en esta plataforma para favorecer su apropiación.

2 Aportes

Se realizó un relevamiento bibliográfico, a partir del cual se identificaron aquellas redes sociales que son más populares entre distintos rangos etarios. A partir del mismo, es posible decir que, a nivel mundial, las redes sociales más utilizadas son Facebook, YouTube, WhatsApp, Instagram y WeChat. [4]. Mientras que puntualmente en Argentina las redes que ocupan los primeros lugares, en cuanto a porcentaje de usuarios, son WhatsApp, Facebook e Instagram [10]. En este estudio no consideraremos la red WeChat por estar su uso limitado a China.
Teniendo en cuenta que la plataforma AlfaDatizando está diseñada para ser usada por docentes y estudiantes secundarios, se decidió dividir el uso de las redes sociales según grupos etarios. El primer grupo consiste en personas que usan las redes sociales y su edad no supera los 19 años, el segundo grupo está formado por personas entre los 20 y 34 años [8] [9], y, por último, el tercer grupo, compuesto por personas mayores a 35 años [4] [9]. En base a lo anterior, las personas de hasta 19 años inclusive, tienen a Instagram, WhatsApp, YouTube y Facebook como redes sociales preferidas [5] [6] [7] [8], luego, las personas de entre 20 y 34 años eligen con mayor frecuencia YouTube, Facebook e Instagram [8] [9]. Por último, las personas a partir de los 35 años presentan una preferencia hacia Facebook y YouTube [4] [9].
A continuación, la Tabla 1 presenta los datos relevados.
En base a los resultados obtenidos en la Tabla 1, se analiza para cada red social si es pertinente su inclusión para el desarrollo de login unificado.
El segundo paso del proceso consiste en analizar cada una de las redes sociales presentadas en la Tabla 1, justificando si es pertinente incluirla en el login unificado.
Cuando se usa la red social YouTube, el acceso a la misma se hace usando la cuenta de Gmail, por tal motivo no se considera como una red social a incluir para implementar el login unificado.
Al analizar WhatsApp como red social, se llega a que la API no contiene funcionalidades que permitan la utilización de WhatsApp como medio de login o registro, por lo que también se la descarta.

[4] http://www.alfadatizandonos.okd.lifia.info.unlp.edu.ar/

Tabla 1: Preferencias de redes sociales por grupo etario

Redes Sociales	**Edades**		
	hasta 19 años	de 20 a 34 años	35 o más años
Youtube	x	x	x
Whatsapp	x		
Facebook	x	x	x
Instagram	x	x	

Facebook es una red social de amplia popularidad y su API, provee opciones para trabajar con el login. Adicionalmente, Instagram al ser fusionada con Facebook, dejó de tener en su API, la opción de login propio.
Por último, si bien WeChat aparece como una de las redes sociales más usadas, no es considerada para su inclusión como parte del login unificado, ya que responde a una red social usada en China.
En la Tabla 2, se muestra el resumen de las redes sociales y la decisión de inclusión o exclusión.

Tabla 2: Redes Sociales incluidas y excluidas

Red social	Inclusión
Youtube	no
Whatsapp	no
Facebook	si
Instagram	no
WeChat	no

Por otro lado, se agrega al estudio la opción de iniciar sesión mediante el usuario de Gmail (API de Google), en general, la mayoría de las personas cuenta actualmente con una cuenta de Gmail, ya que los dispositivos Android, la piden para acceder al sistema. Adicionalmente, durante la pandemia, plataformas como Google Classroom brindaron apoyo a los docentes, sumando los requerimientos que estos solicitaban para sostener

el vínculo pedagógico [11]. Lo anterior nos hace suponer que las cuentas de google son ampliamente usadas.
Como consecuencia del análisis de la etapa dos, Facebook y Google quedan como candidatos a ser usados para el login unificado.
A continuación, analizaremos las APIs de Facebook y Google para login unificado.
Para poder utilizar la API de Google en la implementación del login unificado, el requisito básico es tener una cuenta de google, registrar nuestro proyecto (o producto) y crear credenciales. En este caso, creando una clave de API y un ID de cliente OAuth alcanza.
En cuanto a la API de Facebook, ésta cuenta con una integración específica para distintos SO (iOS, Android, web, u otros dispositivos tales como smartTV. Los requisitos técnicos son similares a los mencionados anteriormente, se necesita una cuenta de desarrollador en Facebook y registrar una aplicación. Además, cuenta con un SDK, que debe ser utilizado para poder desarrollar el registro en nuestro sistema. La Tabla 3, resume lo antes descripto.

Tabla 3: Análisis de requerimientos técnicos

Redes S.	Requerimientos técnicos
Google	-Crear proyecto (registrar producto) -Crear client id
Facebook	- App de Facebook registrada [12] - Facebook SDK for JavaScript [13]

Se destaca además que, tanto Google como Facebook proveen una implementación del estándar OAuth 2.0

3 Líneas de Investigación Futura

En el futuro cercano, se espera incluir en el desarrollo de AlfaDatizando el login unificado justificado a partir del análisis presentado en este trabajo. Posteriormente se realizará un estudio de usabilidad de la Plataforma en general.
Por otro lado, se espera incorporar a la plataforma nuevas fuentes de datos a ser visualizadas, como por ejemplo Wikidata y YouTube.
Trabajar en una plataforma con fines educativos, requiere que los estudiantes y los docentes puedan apropiarse de ella, aún más, requiere que los directivos den apoyo a los docentes para llevar adelante nuevas e innovadoras prácticas de enseñanza/aprendizaje. Se espera trabajar en consecuencia, con directivos y docentes para generar nuevos espacios de trabajo que involucren la visualización de datos como parte del desarrollo del Pensamiento Computacional en las escuelas secundarias.

Referencias Bibliográficas

1. Hazzan O., Ragonis N., Lapidot T. (2020) Computational Thinking. In: Guide to Teaching Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-030-39360-1_4
2. Özkök, G. A. (2021). Fostering Computational Thinking Through Data Visualization and Design on Secondary School Students. J. Univers. Comput. Sci., 27(3), 285-302.
3. Lliteras A., Artopoulos A., Fernandez A. & Huarte J. AlfaDatizando: a Data Visualization Platform to work Computational Thinking in Digital Humanities. Laclo 2022. In press.
4. https://wearesocial.com/es/blog/2022/01/digital-2022/
5. http://publicservicesalliance.org/wp-content/uploads/2018/06/Teens-Social-Media-Technology-2018-PEW.pdf
6. https://www.onlinescientificresearch.com/articles/impact-of-social-media-on-teenagers-nigerian-experience.pdf
7. https://dialnet.unirioja.es/servlet/articulo?codigo=7778049
8. Adame, A. (2019, April 29). Redes sociales más usadas en el mundo hispano: estadísticas y tácticas. Social Media Marketing & Management Dashboard. https://blog.hootsuite.com/es/redes-sociales-mas-usadas/
9. Auxier, B., & Anderson, M. (2022, May 11). Social Media Use in 2021. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/
10. Statista. (2021, July 2). *El uso de internet en Argentina – Datos estadísticos*. https://es.statista.com/temas/7678/el-uso-de-internet-en-argentina/#dossierContents__outerWrapper
11. Artopoulos, A., & Huarte, J. (2022). Continuidad educativa durante la pandemia en Argentina. Políticas, pedagogías y plataformas. Revista de Ciencias Sociales, 35(51), 107-130.
12. https://developers.facebook.com/docs/development#register
13. https://developers.facebook.com/docs/javascript

Necesidades de Comunicación Complejas: Desarrollando una Aplicación SAAC Móvil para el Hospital Zonal de Caleta Olivia

Hernán SOSA, Adriana MARTIN y Viviana SALDAÑO

Grupo de Investigación y Formación en Ingeniería de Software (GIFIS),
Instituto de Tecnología Aplicada (ITA), Universidad Nacional de la Patagonia Austral (UNPA), Unidad Académica Caleta Olivia (UACO)
{sosahernanmisael // adrianaelba.martin // vsaldanio }@gmail.com

Resumen: Los profesionales de la salud necesitan disponer de un soporte que permita crear situaciones de comunicación con aquellos pacientes, cuya discapacidad o trastorno, les limite de forma temporal o permanente el uso del lenguaje oral. Frente a esta problemática, establecer una comunicación efectiva entre las partes, es una condición fundamental para que estas personas reciban la atención adecuada por parte de los especialistas. En este trabajo, se propone el desarrollo de un Sistema Aumentativo y Alternativo de Comunicación (SAAC) para dispositivo de tipo Tablet que brinde soporte a necesidades comunicativas específicas de los sectores de guardia e internación de un hospital.

Palabras clave: Sistemas Aumentativos y Alternativos de Comunicación (SAAC) | Necesidades de Comunicación Complejas (NCC) | Pacientes con Discapacidad Comunicativa | Sistema y Profesionales de Salud.

1. Introducción

Actualmente las Tecnologías de la Información y las Comunicaciones (TIC) están muy presentes en la vida diaria y han facilitado el acceso a la información, transformado la forma de comunicarnos y relacionarnos. Sin embargo, muchas personas encuentran barreras a diario, siendo una de ellas el acto comunicativo. Quienes padecen un impedimento físico, cognitivo o mental para hablar, tienen dificultad para comunicar sus pensamientos y necesidades a la sociedad.

Una persona con Necesidades de Comunicación Complejas (NCC) no se puede comunicar por sus propios medios, y por lo tanto, no podrá tomar decisiones que afecten su vida, lo que generará dependencia de Sistemas Alternativos y Aumentativos de Comunicación, como el sistema Braille o la Lengua de Señas, la asistencia de otras personas, o en el peor de los casos el aislamiento [1]. Esta situación genera una profunda frustración tanto para la persona que la padece, y se agrava en establecimientos que prestan servicios a los ciudadanos, como un hospital o centro de salud.

2. Contexto y Motivación

Este trabajo surge como una iniciativa de Profesionales de Fonoaudiología y Educación Especial para derribar las barreras de comunicación de cierto sector de la población que asiste al Hospital Zonal de Caleta Olivia (HZCO). Se detecta que las personas con una discapacidad que les impide usar el lenguaje verbal, o pacientes con NCC, no pueden comunicarse con los especialistas para ser atendidos adecuadamente.

Muchas organizaciones han establecido convenciones y leyes [2] [3] [4], acerca de la necesidad de disponer de servicios públicos apropiados y accesibles a todos sus

ciudadanos, para: (i) permitirles participar en el mundo, (ii) disponer de las herramientas adecuadas para poder hacerlo y, (iii) acceder a un soporte de comunicación alternativo, a los efectos de contribuir a la prestación de sus respectivos servicios a ciudadanos con discapacidad.

3. Objetivos

Ante esta problemática, el presente trabajo propone realizar un análisis del problema y el desarrollo de una herramienta de tipo Sistema Aumentativo y Alternativo de Comunicación (SAAC) basada en pictogramas que permita la interacción del personal de salud con las personas que no pueden comunicarse a través del lenguaje oral (y teniendo en cuenta el compromiso físico, la discapacidad motora y el aspecto cognitivo), en los sectores más importantes de un centro de salud (guardia, internación, turnos).

Figura 1: Uso de Aplicación - Interacción Profesional-Paciente

Si bien, existen en el mercado productos destinados específicamente a brindar una herramienta de soporte a la persona con discapacidad comunicativa [5] [6] [7], se requieren soluciones más integrales, que consideren las características y situaciones propias del dominio de aplicación y sus actores. Ante esta particularidad, y alineados a los requerimientos a los que debe dar soporte la Aplicación SAAC Móvil en desarrollo, se establecieron los siguientes objetivos:

- Reemplazar las tarjetas de cartón impresas con símbolos, para dar soporte específico a la comunicación Profesional-Paciente a través de la aplicación que ofrece un conjunto de pictogramas del Centro Aragonés para la Comunicación Aumentativa y Alternativa (ARASAAC) [5], entidad que distribuye el material bajo Licencia Creative Commons BY-NC-SA[1].
- Permitir a los profesionales de la salud que a través de la aplicación puedan gestionar Categorías, Preguntas, Rutinas y Frases frecuentes, relacionadas a la atención de pacientes en el contexto del hospital y para derribar barreras de comunicación existentes al atender pacientes con NCC. La Figura 1, ilustra los diferentes escenarios de interacción en los que prestará soporte la aplicación.
- Permitir la creación y la gestión de los recursos sin conexión, es decir no dependerá del servicio de internet. Este punto es clave, debido a que las falencias de conexión resultan contraproducentes en cuanto al uso de este tipo de sistemas.
- Proveer una interfaz apta para dispositivos de tipo Tablet, aunque es deseable que sea de tipo *responsive* y se adapte a diversos tamaños de pantalla.

[1] CC Creative Commons <https://creativecommons.org/licenses/by-nc-sa/3.0/es/>

4. Enfoque de Desarrollo

De acuerdo al dominio del problema y las características de los usuarios, se usa un enfoque customizado y centrado en el usuario basado en Lean UX [9]. La Figura 2, ilustra el enfoque que se ha estado aplicando.

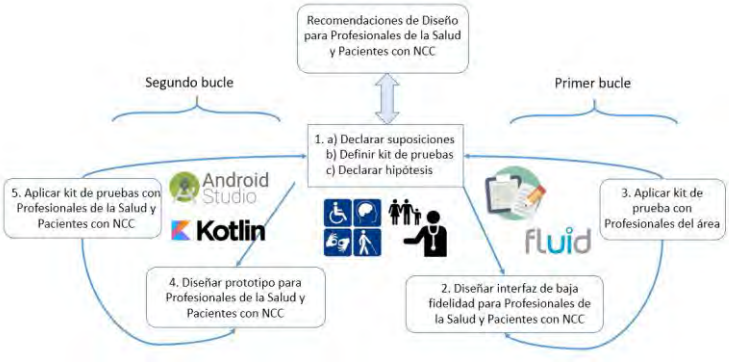


Figura 2. Enfoque de Desarrollo

Para definir las características del sistema y suposiciones de negocio, se realizaron entrevistas de forma personal y mediante video llamadas con fonoaudiólogas y docentes de Educación Especial, recopilando información relevante para detectar las dificultades de comunicación entre los profesionales de la salud y los pacientes con NCC. A su vez, para diseñar el prototipo, se indagó en las Pautas de Usabilidad propuestas para dispositivos móviles [10], guías de Accesibilidad y recomendaciones de diseño del Sistema Operativo Android[2] [11] [12]. De acuerdo a las Recomendaciones de Diseño, se generaron "kits de prueba" para realizar las validaciones, permitiendo de esta manera la participación permanente de los profesionales de salud y los pacientes con NCC. Estos "kits de prueba" incluyen las tareas más relevantes a las que debe brindar soporte la aplicación que son evaluadas por los usuarios en cada iteración:

- Primer Bucle: permite obtener un prototipo de baja fidelidad usando Fluid UI[3], este prototipo es testeado por los usuarios de la aplicación; si es aceptado, inicia el Bucle 2; caso contrario, se regresa a revisar las suposiciones.
- Segundo Bucle: permite obtener el prototipo final, el cual se implementa en aplicación nativa utilizando el lenguaje Kotlin[4].

5. Resultados

En relación a la adquisición de conocimientos acerca del dominio y de la discapacidad comunicativa, a lo largo del desarrollo de este trabajo, se contó con un aporte importante de fonoaudiólogas y profesionales de la salud (médicos y enfermeros), para proponer y diseñar soluciones. Debido a esta valiosa devolución de los especialistas involucrados y la participación en varios proyectos GIFIS, el autor de este trabajo ha adquirido conocimientos en áreas tales como: Accesibilidad, Usabilidad y Experiencia de Usuario (UX), Interacción Humano-Computadora (HCI), Interfaces de Usuario (UI), Diseño Centrado en el Usuario (UCD). Estos antecedentes son fundamentales a los efectos de poder desarrollar e implementar un sistema personalizado, es decir, que sea a la medida de las necesidades del HZCO para que dispongan del soporte adecuado a la atención de los pacientes con NCC. Es importante resaltar que se pretende, además, establecer y formalizar vínculos con los profesionales de las Escuelas de Educación Especial, que

[2] Sistema Operativo de dispositivos móviles. Sitio Oficial <https://developer.android.com/>
[3] Herramienta para crear prototipos y diseños de interfaces. Sitio oficial: https://www.fluidui.com/
[4] Lenguaje de programación oficial de Android. Sitio Oficial <https://kotlinlang.org/>

asisten a personas con discapacidad comunicativa, para conocer también sus necesidades y requerimientos.

Desde el aspecto de la adquisición de conocimientos acerca de la tecnología involucrada, el proyecto ha demandado una alta carga horaria de esfuerzo técnico sobre la arquitectura del Sistema Operativo Android y el lenguaje de programación Kotlin. Para esta tarea se ha usado como base la documentación oficial, siguiendo los lineamientos de diseño de aplicaciones nativas.

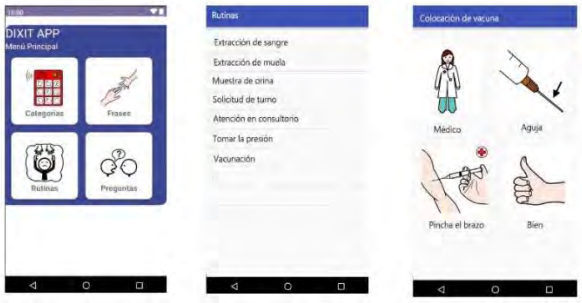


Figura 3. Navegación de la vista Rutinas

En [9], se presentan los avances del proyecto, como así también los eventos en los que el autor ha participado[5,6], el cual actualmente se encuentra alcanzando la etapa final de desarrollo e implementación de la primera versión del producto. La Figura 3, ilustra el menú principal de la aplicación y las interfaces de navegación de la vista Rutinas, que como su nombre lo indica, se utilizará para llevar adelante las prácticas rutinarias en pacientes con NCC.

6. Trabajo Futuro

Este trabajo se enmarca en el Proyecto de Investigación UNPA N°29/B285 (2022-2025), denominado "*Desarrollo y Evaluación de Proyectos Web y Móviles Centrado en la Experiencia de Usuario*", dirigido y codirigido por la Dra. Adriana Martín y la Mg. Gabriela Gaetán, respectivamente, del "*Grupo de Investigación y Formación en Ingeniería de Software (GIFIS)*". Particularmente, el presente trabajo se corresponde con un Proyecto Final de Desarrollo del AdeS Hernán Misael Sosa, integrante de GIFIS, para obtener el título de Ingeniero en Sistemas UNPA y está dirigido y codirigido por Dra. Adriana Martin y la Mg. Viviana Saldaño, respectivamente.

Como trabajo futuro se proyectan los siguientes objetivos:

- Ampliar el número de pruebas con expertos (de educación y de fonoaudiología) y de usuarios (profesionales de la salud y pacientes).
- Aportar en repositorio *open source* para que el código de esta aplicación pueda servir de base para otros desarrolladores, ajustándolo a sus necesidades.
- Extender el alcance de la aplicación a otras regiones del país, además de los alrededores de Caleta Olivia.
- Modificar y extender la aplicación a otros ámbitos como Bomberos, Policía y Centros de Atención donde se requiera un soporte similar para brindar servicios a los ciudadanos con NCC.

[5] II Jornadas de Discapacidad y Tecnologías <https://www.uaco.unpa.edu.ar/las-ii-jornadas-de-discapacidad-y-tecnologias-se-realizaran-el-29-y-30-de-noviembre>

[6] La Noche Mágica del Labtem después de la Pandemia <https://www.lavanguardianoticias.com.ar/nota/40504-se-realizo-la-muestra-anual-la-noche-magica-del-labtem-de-la-unpa-uaco-en-el-sum-del-centro-cultural/>

Referencias

[1] D. Abadín, C. Santos Delgado y Á. Cerrato Vigara, «“Comunicación Aumentativa y Alternativa”. Centro de Referencia Estatal de Autonomía Personal y Ayudas Técnicas (CEAPAT),» mayo 2010. [En línea]. Available: http://riberdis.cedd.net/handle/11181/3425. [Último acceso: 29 5 2020].

[2] Cámara de Senadores y Diputados de la República Argentina. , «Ley 26.378. “Apruébese la Convención sobre los Derechos de las Personas con Discapacidad y su protocolo facultativo”. Aprobada mediante resolución de la Asamblea General de las Naciones Unidas el 13 de diciembre de 2006”.,» 6 junio 2008. [En línea]. Available: http://www.infoleg.gob.ar/infolegInternet/anexos/140000-144999/141317/norma.htm.

[3] Organización de las Naciones Unidas, «“Convención Internacional sobre los Derechos de las Personas con Discapacidad". Aprobada mediante Asamblea General, por Resolución 61/106.,» Diciembre 2006. [En línea]. Available: http://www.un.org/spanish/disabilities/convention/qanda.html.

[4] Senado y Cámara de Diputados de la Nación Argentina, «“Apruébese una Convención Interamericana para la Eliminación de Todas las Formas de Discriminación contra las Personas con Discapacidad.”,» 6 julio 2000. [En línea]. Available: http://www.infoleg.gov.ar/infolegInternet/anexos/60000-64999/63893/norma.htm. [Último acceso: Sancionada el 6 de julio del 2000].

[5] AssistiveWare, «Aplicación ProloQuo2Go Comunicación Aumentativa Alternativa,» [En línea]. Available: https://www.assistiveware.com/es/productos/proloquo2go. [Último acceso: 2022].

[6] OTTA Project, «OTTA Project: Sistema Aumentativo Alternativo de Comunicación,» [En línea]. Available: https://www.ottaaproject.com/. [Último acceso: 2022].

[7] ¡Háblalo!, «¡Háblalo! La App para Comunicarse sin Barreras,» [En línea]. Available: https://hablalo.app/. [Último acceso: 2022].

[8] Centro Aragonés de la Comunicación Aumentativa y Alternativa (ARASAAC), «Centro Aragonés de la Comunicación Aumentativa y Alternativa,» 2019. [En línea]. Available: http://www.arasaac.org/.

[9] C. Cardozo, A. Martin, V. Saldaño y G. Gaetán, «Una propuesta para mejorar la experiencia de los adultos mayores con las redes sociales,» *Revista Tecnología, Ciencia y Educación. Centro de Estudios Financieros (CEF) /Universidad a Distancia de Madrid (UDIMA),* nº 16, pp. 113-142, Mayo-Agosto 2020.

Propuesta de sistema de registro y generación de pulseras de identificación de pacientes

María de la Concepción Pérez de Celis Herrero y Saúl Maldonado Navarro

Benemérita Universidad Autónoma de Puebla, Puebla, México
saul.maldonado@alumno.buap.mx
maria.perezdecelis@correo.buap.mx

Abstract. Los centros sanitarios y hospitales suelen contar con un sistema propio de identificación de pacientes, estos pueden ser variados, desde implementar el número de cuarto y cama en hojas o señalización, hasta el uso de pulseras con los datos escritos a mano por el personal o en su defecto impresas. El cuidado de este sistema y su buen funcionamiento es imprescindible para la correcta identificación y por ende correcto tratamiento de los pacientes, ya que, sin esto el personal médico puede cometer errores graves en diversos momentos de la estadía del paciente en el centro sanitario e incluso fuera del mismo, con prescripciones erróneas de medicación o dosis. Es por esto, que se presenta una propuesta de software, un sistema que permita llevar un correcto registro de los pacientes, su información personal, de salud y poder generar pulseras de identificación con la información necesaria para la correcta identificación del paciente.

Palabras Clave: Paciente, Identificación, Pulsera, Personal, Sistema.

1 Motivación del proyecto

El presente articulo expone los objetivos y resultados de la investigación, análisis, diseño y programación de un proyecto de investigación realizado en el programa de servicio social de la BUAP en la Facultad de Ciencias de la Computación, con el fin de generar una propuesta de sistema que hace uso de las tecnologías de la información para ayudar y dar solución problemas reales en el apartado de registro e identificación de pacientes en un centro sanitario.

1.1 Sistemas de identificación de pacientes

La seguridad del paciente es una dimensión fundamental de la calidad de la atención y se ha convertido en una estrategia prioritaria en los sistemas de salud a nivel mundial [1]. El sistema de identificación de pacientes en un hospital o centro sanitario es un aspecto clave e indispensable para el buen funcionamiento y buena atención de este. El sistema debe asegurar siempre una correcta identificación de pacientes. Cada vez que un miembro del personal del centro sanitario tiene interacción con un paciente, desde

la recepción hasta en el consultorio y todo el transcurso que hay de por medio, el paciente debe ser identificado por cada miembro del personal que tenga algún contacto o atención con el mismo.

1.2 Causas y consecuencias de los principales errores en la identificación de pacientes

Durante la asistencia sanitaria, la mala o incorrecta identificación de pacientes es una causa importante de problemas y complicaciones asociados a descuidos e inexactitudes de medios como características físicas o psicológicas, errores de dedo al momento de capturar datos, desatenciones de parte del personal a causa de fatiga, cansancio, falta de voluntad o estados del paciente tales como: estado de shock, barreras del lenguaje, edad o estado del paciente (podría estar inconsciente o sedado) [2].
Todo lo anterior puede provocar grandes riesgos en la administración de medicamentos e intervenciones invasivas y no invasivas [3]. De acuerdo con análisis hechos por Kim T. [4] a 11,898 reportes de seguridad de pacientes revelan que el 37.9% de los pacientes recibieron una *"inapropiada administración de medicamentos"* mientras que en el apartado de las *"órdenes y prescripciones"* existen errores en alrededor del 41.9%.

2 Sistema de registro y generación de PIP

Para la identificación de un paciente, se debe dar de alta en el sistema del centro sanitario, comenzando por capturar datos relevantes, como:

- Nombre completo.
- Fecha de nacimiento.
- Sexo.
- Dirección.
- Números de contacto.
- Contacto(s) de emergencia.
- Alergias
- Cuidados o condiciones especiales.

Esta información dependiendo la situación puede ser proporcionada por el paciente, o por medio de sus familiares o acompañantes.

Esta propuesta de sistema es capaz de manejar perfiles de tipo administrador, personal y paciente con sus respectivos privilegios, así como tener la opción de registrar infraestructura como cuartos, consultorios, áreas el hospital, camas, ver en tiempo real las estadísticas el hospital como: camas ocupadas, libres, pacientes bajo procedimientos o consultas, editar, actualizar o eliminar información y claro, generar las PIP de los pacientes. Se pude observar en la Fig.1 la interfaz que muestra los datos de un paciente.

Fig. 1. Perfil de un paciente

3 Implementación de PIP para la identificación de pacientes, estándares y recomendaciones

3.1 Implementación de tecnologías y diseño:

Una propuesta para agilizar la identificación de pacientes y reducir los errores es la implementación de pulseras de identificación de pacientes o PIP, ya que se ha observado una disminución de alrededor de un 50% en los errores asociados a la inadecuada identificación de los pacientes con el uso de estas en los centros sanitarios [5].
De la misma forma uso de las tecnologías de la información toma un papel fundamental si queremos que los métodos de registro y generación de PIP sea de una forma ágil y óptima. Una manera para implementar las TI en el diseño de las PIP es incluir fotografías digitales del paciente y el uso de códigos de barras o QR. Cuando son implementadas estas tecnologías de acuerdo con accesor [5] se puede observar una disminución del 33% en los errores del "fármaco inadecuado", de un 52% en la "omisión de la dosis" y de un 47% en los errores de transcripción.

Para el diseño de una PIP hay datos sumamente importantes que debe contener [6], hablamos de información como: nombre completo del paciente, fecha de nacimiento, número de seguridad o identificador (depende del centro sanitario), sexo y tipo de sangre. También tenemos información opcional, pero de gran ayuda: fecha de ingreso, fotografía del paciente o área de especialidad donde va a ser tratado el paciente.
El diseño implementado se puede ver en la Fig.2.

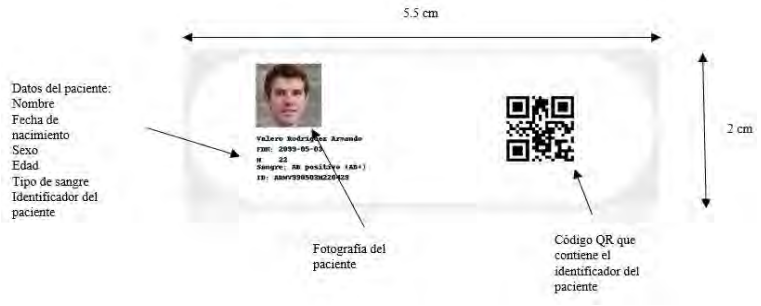


Fig. 2. Diseño de la PIP implementada en el sistema

4 Conclusión y posibles líneas de investigación futura

Es necesario saber que para la implementación de este sistema en un centro sanitario, se requiere de la autorización de la alta gerencia del lugar, así como una inversión económica para adquisición de equipo de cómputo, software, capacitación para el personal y un cambio en la metodología de trabajo y protocolos de ingresos y altas de los pacientes, así como posiblemente una migración de la información y registros actuales del hospital o centro sanitario en cuestión, no obstante, son claras las ventajas y mejoras a los antiguos sistemas:

- Seguimiento e historial de movimientos y procedimientos realizados a un paciente, esto permitirá saber dónde está, dónde estuvo y que tratamientos fueron aplicados, en qué fecha y hora exactas.
- Tener un registro del personal del hospital y de los pacientes.
- Será más fácil para el personal poder identificar a los pacientes y acceder a su información personal.
- A comparación con los sistemas dónde las PIP son llenadas a mano, la implementación de este sistema agilizaría la generación de estas y sus posibles actualizaciones de datos.

Por su parte, las posibles líneas de investigación y mejora a implementar mejoras en el Sistema son claras: haciendo una búsqueda del estado del arte actual, el uso de tecnología RFID para tener información en tiempo real del paciente y también de la infraestructura, además de las posibles mejoras del apartado gráfico de interfaz de usuario.

Referencias

1. Pérez De Celis Herrero, M. Seguridad del paciente: Propuesta de mejora para la reducción de errores en la identificación del paciente. Reporte Técnico. UPAEP (2019).
2. The Importance of Patient Identification. Identification Systems Group, https://www.identificationsystemsgroup.com/the-importance-of-patient-identification/, Recuperado el 2022/06/07
3. Patient Identification: Why It's Important and How Best Ways to Implement It, https://www.m2sys.com/blog/guest-blog-posts/patient-identification-why-its-important-and-how-best-ways-to-implement-it/, Recuperado el 2022/06/06
4. Kim, T. Health Information Technology–Related Wrong-Patient Errors: Context is Critical | Patient Safety. Patient Identification Errors: A Systems Challenge. (2020)

5. Identificación de paciente y Estrategias de mejora, https://www.accesor.com/identificacion-paciente-estrategias-mejora/, Recuperado el 2022/06/06
6. Australian Commission on Safety and Quality in Health Care, https://www.safetyandquality.gov.au/sites/default/files/migrated/Specs-PatID-Band.pdf, Recuperado el 2022/01/21

Vinculación de portales abiertos mediante API

Juan Ignacio Torres, Ariel Pasini, Silvia Esponda

Instituto de Investigación en Informática LIDI (III-LIDI)
Facultad de Informática – Universidad Nacional de La Plata 50 y 120 La Plata Buenos Aires
Centro Asociado Comisión de Investigaciones Científicas de la Pcia. de Bs. As. (CIC)
{jitorres,apasini,sesponda}@lidi.info.unlp.edu.ar

Abstract. El presente trabajo se desarrolla en el marco de la Práctica Profesional Supervisada de la carrera de ATIC (Analista en Tecnologías de la Información y la Comunicación) de la Facultad de Informática - UNLP. En los últimos años, los gobiernos buscan generar herramientas que involucren a los ciudadanos a participar de forma activa en las decisiones de gobierno. Una forma de incrementar la participación ciudadana es brindar, a la comunidad, acceso a la información y permitir que ellos mismos sean capaces de generar nuevos aportes que asistan a las decisiones del gobierno. Los diferentes niveles de gobierno ponen al alcance de los ciudadanos una gran cantidad de información en formatos abiertos. Sin embargo, pueden existir incompatibilidades a la hora de acceder y analizar esta información. Por lo que se propone desarrollar una herramienta que permita compatibilizar diferentes fuentes de datos disponibles mediante APIs, de forma sencilla para el ciudadano, facilitando la comparación, el procesamiento y análisis de la información de manera generar valor agregado.

Keywords: Datos Abiertos, Gobierno abierto, API, vinculación de datos.

1. Introducción

En los últimos años se popularizó un nuevo paradigma de gestión pública, llamado gobierno abierto, el cual se basa en el acceso a la información por parte del ciudadano permitiendo controlar y/o generar nuevos aportes al gobierno. Un ejemplo común en los diferentes niveles de gobierno es la búsqueda de transparencia en las cuentas públicas, poniendo a disposición de los ciudadanos información tal como balances, contrataciones, licitaciones, etc. para que ellos puedan ver el destino de los fondos públicos. Por otro lado, los ciudadanos podrían tomar dicha información y analizarla desde diferentes puntos de vista generando aportes constructivos al gobierno. Este proceso genera mayor transparencia en la gestión y en consecuencia mayor confianza del ciudadano.

Para lograr mayor transparencia ante sus ciudadanos, los diferentes niveles de gobierno ponen una gran cantidad de datos públicos en formatos abiertos a disposición de los ciudadanos de manera tal que éstos analicen y procesen la información para

generar valor agregado, apoyando el proceso de toma de decisiones de manera inteligente.

Sin embargo, al obtener información de distintos proveedores (ya sean agencias gubernamentales, organizaciones internacionales u ONGs) o incluso de un mismo proveedor mediante APIs, se pueden encontrar problemas de compatibilidad en los formatos de las respuestas, como por ejemplo diferentes unidades.

Es por eso, que, mediante el uso de una herramienta que vincule las APIs de las fuentes de datos, se buscará dar una solución a estos problemas, de manera tal que se pueda generar un nuevo dataset, en el cual los datos sean comparables y analizables, permitiendo a los usuarios participar en el proceso de generación de valor agregado.

Este trabajo se desarrolla en el marco de la Práctica Profesional Supervisada de la carrera de ATIC (Analista en Tecnologías de la Información y la Comunicación).

2. Problemática

A nivel global, los ciudadanos esperan cada vez más de sus gobiernos y gobernantes. Estas demandas, junto con la revolución tecnológica generaron un cambio drástico en la interacción entre un gobierno y sus ciudadanos. El uso inteligente de la información registrada por los gobiernos en formatos abiertos permite mejorar la calidad de vida de los ciudadanos. Estos datos abiertos, pueden ser accedidos mediante la comunicación con una API.

No obstante, puede suceder que cuando se intenta acceder a esta información, procedente de diferentes proveedores o incluso de un mismo proveedor, el usuario no esté conforme con el formato de los datos recibidos.

Además, pueden existir diferencias en los formatos de respuesta o incompatibilidad de unidades entre las distintas fuentes de datos. Un ejemplo de estas diferencias se evidencia en la extensión de las líneas de subte en Argentina y en los Estados Unidos, donde en Argentina la extensión se mide generalmente en kilómetros, mientras que en los Estados Unidos se mide en millas.

En algunos casos, los usuarios podrían requerir operaciones sobre los datos de cada una de las fuentes, como filtrarlos, agruparlos bajo algún criterio, o simplemente realizar operaciones aritméticas.

Aplicando las operaciones requeridas por quien analice los datos y solucionando las diferencias mencionadas, se podría facilitar la comparación y el procesamiento de la información, permitiendo así la generación de valor agregado por y para parte de los ciudadanos.

3. Desarrollo propuesto

Se desarrollará una herramienta que permitirá brindar una solución a la incompatibilidad de formato y/o unidades entre fuentes de datos accesibles mediante APIs, que el usuario desee comparar y analizar.

Esta herramienta, pretende estandarizar la información a comparar mediante una conversión del formato de respuesta (en caso incompatibilidad) o un factor de

corrección (en el caso de incompatibilidad de unidades), generando un nuevo dataset que contendrá información significativa para el análisis del usuario.

Para lograr este resultado, es necesario pasar por un proceso de 4 pasos:

1. Búsqueda de la información: el usuario selecciona las fuentes de datos disponibles mediante comunicación vía API que sean de su interés a través de portales de datos abiertos.
2. Conexión con la API: Una vez elegidas las fuentes de datos, el usuario cargará en la herramienta las URL correspondientes. La comunicación con las APIs se realizará mediante HTTPS, y en caso de ser correcta la conexión, se obtendrán los datos en formato JSON.
3. Análisis previo: El usuario podrá visualizar en paralelo la información elegida en cada dataset para su análisis. En esta etapa, podrá realizar diferentes operaciones con los datos (como por ejemplo filtrado o agrupamiento). Posteriormente, seleccionará las columnas a comparar y el factor de corrección a aplicar (en caso de que haya problemas de compatibilidad en las unidades).
4. Visualización: una vez seleccionadas las columnas a comparar, las operaciones aplicadas y el factor de corrección en caso de ser necesario, se podría generar un nuevo dataset que contenga la información a comparar de manera estandarizada. El usuario podrá ver en pantalla esta información y podrá descargar un archivo .csv para su futuro análisis y procesamiento.

Al finalizar este proceso, el usuario obtendrá información comparable y procesable, solucionando así la incompatibilidad de formato y/o unidades.

Volviendo al ejemplo de la extensión de las líneas de subte, en caso del usuario haber elegido ver la información medida en kilómetros, obtendrá un dataset con la extensión de las líneas de Argentina y Estados Unidos medidas en kilómetros.

4. Líneas de investigación futura

El principal desarrollo a futuro es la implantación de un sitio de gestión para la vinculación de diferentes portales de datos abiertos que provea datos mediante API, que permita cargar y almacenar la información de conexión con las API de los diferentes portales de datos abiertos para que el ciudadano los pueda vincular de manera sencilla.

5. Bibliografía básica

Attard, J., Orlandi, F., Scerri, S., Auer, S. (2015). A Systematic Review of Open Government Data Initiatives. Government Information Quarterly. 32. 10.1016/j.giq.2015.07.006.

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The state of open data. Limits of Current Open Data Platforms.

Greco, A. O. (2020). “IndiMaker: Una herramienta para la construcción de indicadores personalizados en tableros de control de SGC”. Tesina de Licenciatura en Sistemas, Facultad de informática, UNLP. http://sedici.unlp.edu.ar/handle/10915/118501

J. S. Preisegger, A. Greco, A. C. Pasini, M. Boracchia, and P. M. Pesado, Marco de vinculación de datos abiertos aplicado al contexto de datos medioambientales. Actas del XXVI Congreso

Argentino de Ciencias de la Computación (CACIC 2020), ISBN: 978-987-44-1790-9, págs. 756-766, 2020.
Janssen, M., Charalabidis, Y., Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government, Information Systems Management, 29:4, 258-268, DOI: 10.1080/10580530.2012.716740
Lathrop, D., Ruma, L. (2010). Open government: Collaboration, transparency, and participation in practice. United States of America: O'Reilly Media, Inc.
Naser, A., Ramirez, A., Rosales D: Desde el gobierno abierto al Estado abierto en América Latina y el Caribe. Libros de la CEPAL - Planificación para el Desarrollo No. 144.
O. Government, "Memorandum on Transparency and Open Government," Fed. Regist., pp. 21–22, 2009, [Online]. Available: https://www.archives.gov/files/cui/documents/2009-WH-memo-on- transparency-and-open-government.pdf.
Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22, OECD Publishing, Paris, https://doi.org/10.1787/5k46bj4f03s7-en
Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. Government Information Quarterly, 31(1), 17–29. https://doi.org/10.1016/j.giq.2013.04.003

Análisis de plataformas de Computación en la Nube para implementación de protocolo de comunicaciones con una aplicación móvil 3D

Mauro Santos[1], Diego Encinas[1,2]

[1]Instituto de Investigación en Informática (III-LIDI). Facultad de Informática, Universidad Nacional de La Plata - Centro Asociado CIC. La Plata, 1900, Argentina.

[2]SimHPC-TICAPPS. Universidad Nacional Arturo Jauretche. Florencio Varela, 1888, Argentina.

maurosantos1907@gmail.com, dencinas@lidi.info.unlp.edu.ar

Abstract. Se realizó un análisis e investigación de la performance de comunicaciones en una aplicación móvil orientada a redes de sensores con tecnología Cloud Computing. La aplicación móvil desarrollada en Unity, reproduce un entorno visual 3D vinculado a diferentes dispositivos y sensores que pueden encontrarse en una vivienda u oficina. Se desarrolló una comunicación con la aplicación 3D y diversas plataformas de Computación de la Nube. De las misma, se analizaron distintas métricas de rendimiento en las comunicaciones como latencia y throughput del sistema.

Keywords: Cloud Computing, latencia, domótica, servidor.

1 Introducción

La motivación para llevar a cabo este proyecto fue impulsada por la necesidad de realizar la medición de diversas métricas en el rendimiento de la comunicación en entornos de diferentes plataformas nube con aplicaciones 3D.
El trabajo fue desarrollado dentro del área de Móviles 3D (Desarrollo de Aplicaciones 3D para IoT) como una Práctica Profesional Supervisada, bajo la supervisión del Magister Diego Encinas, integrante del Instituto y docente de la Universidad en varias asignaturas de la carrera de Ingeniería en Computación.

2 Presentación

El análisis inició con la investigación de reconocidos servidores que emplean la tecnología Cloud Computing, logrando que puedan comunicarse mediante el uso del servicio MQTT. Se utilizó este servicio para concretar el pasaje de mensajes entre el servidor y una aplicación móvil de domótica programada y diseñada en la facultad de Informática de la Universidad Nacional de La Plata. La aplicación 3D programada en Unity en la que se trabajaba, sólo se podía comunicar con Amazon Web Services.
La comunicación se utilizaba para que la aplicación móvil encendiera o apagara los dispositivos del hogar, mediante el envío de un mensaje MQTT hacia el el servidor de Amazon. Luego la nube se comunicaba con un dispositivo físico para realizar lo solicitado por el usuario.

Por lo tanto, la motivación de este proyecto fue la búsqueda de servidores Cloud que puedan ser compatibles con la comunicación de la aplicación móvil. Luego de un extenso análisis de las distintas posibilidades que brindaba el mercado IT. Se concretó con la elección de Google Cloud y Microsoft Azure para realizar esta tarea. Ambas nubes han tenido un crecimiento exponencial en el mundo IT estos últimos años.

Con la elección de servidores Cloud ya elegida se comenzó verificando si ambas nubes podían lograr una comunicación exitosa con la aplicación de domótica. Ambas lo superaron de forma positiva, por lo que se prosiguió con un análisis en el performance de la comunicación, enviando una totalidad de 100 mensajes, 50 mensajes de 2 bytes y 50

mensajes de 3 bytes teniendo un envío efectivo de 2000 bytes. Enviando la cantidad de 1,2,4,5 y 10 mensajes por segundo.

Posteriormente se realizó un estudio sobre la latencia de los servidores en un periodo diario de toma de muestras en el lapso estimado de 2 semanas para poder analizar su variación. Por último, para los servidores elegidos se planteó un análisis comparativo entre los costos producidos por el uso de ese servicio y la latencia media que tenía el servidor en una determinada ubicación.

Finalmente, se muestra la Fig. 1 que representa el análisis en el envío de 10 mensajes por segundo en la comunicación entre la aplicación 3D y la plataforma de nube Google Cloud.

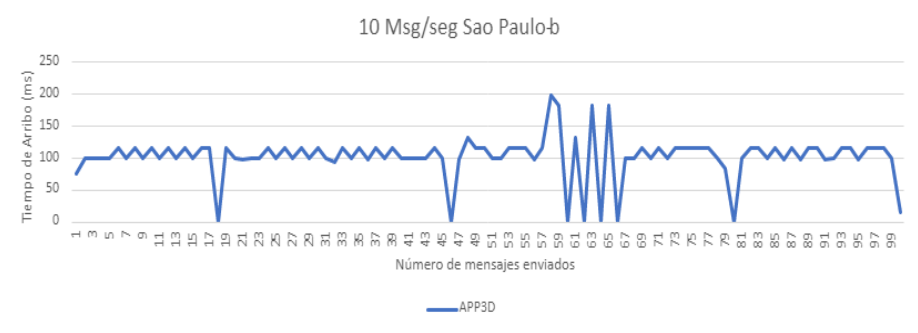


Fig. 1. Intervalo de tiempos entre mensaje y mensaje

2.1 Conclusión

El resultado de este informe, muestra que la aplicación móvil no necesariamente tiene que estar cerrada a la posibilidad de sólo usar la nube de Amazon Web Services. Es decir, como se muestra en el trabajo otras plataformas de arquitectura cloud pueden funcionar de manera confiable con la aplicación.

Aun así, los resultados de las distintas pruebas realizadas en el informe mostraron una performance en la comunicación similar o un poco peor por parte de las otras dos plataformas probadas. Por lo tanto, la elección óptima, para una cantidad baja de mensajes por segundo, sería Amazon Web Services. Mientras que para una cantidad mayor de mensajes se podría usar tanto Amazon Web Services como también Google Cloud.

Además, se observó en las 2 plataformas estudiadas, que en cuestión de los precios por la utilización de la nube se puede analizar según los

requisitos que el usuario requiera. El informe se encarga de buscar un mejor precio mensual por su uso, haciendo que la latencia sea más elevada debido a la ubicación del servidor. Intentando que el usuario note muy poco la diferencia de tiempos, pero que sea un precio más accesible.

El proyecto se logró de manera satisfactoria, en el cual se cumplieron los objetivos planteados. Debido a los motivos sanitarios del momento, el trabajo se desarrolló completamente a distancia. Pero la comunicación con el docente fue fluida y nunca se vio afectada.

3 Proyectos Futuros

En relación con posibles proyectos futuros basados en estas prácticas profesionales, podrían mencionarse:

- La búsqueda de diversas formas de pasaje de mensajes entre Unity y los servidores Cloud, con la finalidad de un mejor performance en la comunicación.
- Comparación con otros servidores Cloud en proceso de crecimiento sobre el performance, latencia y costo sobre las nubes analizadas.
- Estudio sobre el comportamiento del servicio MQTT en la necesidad de envío de mensajes de mayor longitud midiendo el impacto que podría producir en la comunicación.
- La adaptación de los servidores Cloud analizados en diferentes zonas de toda la Argentina para investigar los cambios que pueden ocurrir debido a la ubicación del usuario. Analizando tanto en la performance como en la latencia que puede llegar a tener el pasaje de mensajes entre los servidores y la aplicación.

4 Bibliografía

1. Ailsa Da Conceicao Seco, "¿Qué es la domótica y cómo funciona una casa domótica?" https://blog.caloryfrio.com/que-es-la-domotica-y-como-funciona-una-casa-domotica/ [Accedido 15/8/2022]
2. Amazon Web Services Home Page https://aws.amazon.com/es/ [Accedido 15/8/2022]
3. Arduino IDE Home Page https://www.arduino.cc/ [Accedido 15/8/2022]
4. Claudio Peña," Arduino IDE: Domina la programación y controla la placa" https://books.gogle.es/books?hl=es&lr=&id=Xgv2DwAAQBAJ&oi=fnd&pg=PP1&dq=arduino+ide&ots=vNAXDgTu0X&sig=qdEI7_aYOJBTuM95f_z6HMfroRc#v=onepage&q=arduino%20ide&f=false [Accedido 15/8/2022]
5. Francesc Moreno Cerdà, "Demostrador arquitectura publish/subscribe con MQTT" https://upcommons.upc.edu/bitstream/handle/2117/117782/MQTT_MEMORIA.pdf [Accedido 15/8/2022]
6. Gaston C. Hillar, "MQTT Essentials-A Lightweight IoT Protocol" https://books.google.es/books?id=40EwDwAAQBAJ&printsec=frontcover&hl=es#v=onepage&q&f=false [Accedido 15/8/2022]
7. Google Cloud Home Page https://cloud.google.com/ [Accedido 15/8/2022]
8. Microsoft Azure Home Page https://azure.microsoft.com/es-mx/ [Accedido 15/8/2022]

Construcción de un grafo de conocimiento para un observatorio inmobiliario

Felipe Dioguardi[1][0000−0001−6039−8653], Diego Torres[1,2][0000−0001−7533−0133], Leandro Antonelli[1][0000−0003−1388−0337], y Juan Pablo del Río[3][0000−0002−4031−3007]

[1] LIFIA, CICPBA-Facultad de Informática, UNLP
{nombre.apellido}@lifia.info.unlp.edu.ar
[2] Departamento de Ciencia y Tecnología, UNQ
[3] LINTA-CICPBA, CONICET, UNLP

Resumen Los observatorios inmobiliarios permiten la producción y sistematización de datos provenientes del mercado inmobiliario. En manos de estadistas y expertos del dominio, resultan herramientas invaluables para el estudio de los valores del suelo en un área geográfica determinada. Crear un observatorio inmobiliario requiere la disponibilidad de una gran cantidad de datos, lo que puede resultar un problema si no se cuenta con información extensa, confiable, actualizada, y pública. Para solucionarlo, este artículo presenta una metodología para la extracción de conocimiento proveniente de páginas web dedicadas a la publicación de avisos inmobiliarios, utilizando tecnologías de *web scraping*. Además, propone el almacenamiento de la información inmobiliaria en un grafo de conocimiento estructurado por una ontología acorde al dominio, que dotará los datos externos de valor semántico. Esto posibilitará la inferencia de nuevo conocimiento, y facilitará su manipulación por parte de máquinas y sistemas automatizados. Por último, este artículo ofrece los resultados preliminares de la implementación de una herramienta que sigue con la metodología propuesta, con relación a la capacidad de relevamiento inherente a un proceso manual. Este trabajo se desarrolla en el marco de la tesina de grado titulada "Evaluación de técnicas de detección de duplicados sobre grafos de conocimiento de avisos inmobiliarios", presentada con el proyecto "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano".

Palabras clave: Observatorio inmobiliario · Grafos de conocimiento · Ontologías · Web semántica

1. Introducción

Los observatorios inmobiliarios son herramientas de información que permiten capturar los precios y características de bienes inmuebles en una zona geográfica particular. Los datos que estos brindan pueden resultar un valioso aporte al Estado, pues son la base de investigaciones y estudios estadísticos realizables en el marco de la creación y mejora de políticas sociales.

En Argentina en general y en la provincia de Buenos Aires en particular, existe una carencia estructural en la disponibilidad pública de información sobre los valores del mercado inmobiliario. Por este motivo, en agosto de 2021 el Ministerio de Ciencia, Tecnología e Innovación aprobó el proyecto denominado "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano". Este tiene como objetivo principal resolver la falta de información estratégica de valores de mercado inmobiliario para cuantificar las valorizaciones producidas por la acción del Estado, en el marco de la política de Integración social y urbana de barrios populares [6].

En las instancias iniciales del proyecto, los investigadores realizaron un balance de las diversas fuentes de información disponibles, clasificándolas según la cantidad y calidad de los avisos publicados sobre algunos partidos de interés de la provincia de Buenos Aires. A partir de esa evaluación, seleccionaron diferentes sitios de ofertas inmobiliarias, y recabaron manualmente información acerca de cientos de los inmuebles publicitados en ellos.

Para poder efectuar un estudio completo del valor del suelo en las distintas áreas geográficas, es necesario tener la capacidad de extraer grandes volúmenes de datos de manera sistemática. *Web scraping* es una técnica para obtener información de páginas de Internet, y almacenarla en un archivo o base de datos local para su posterior análisis [9]. A su vez, un *web crawler*, *spider*, o araña, es un agente informático utilizado para la descarga masiva de páginas web [7,8]. Ambos conceptos suelen acompañarse, pues es común querer recolectar en una misma base el conocimiento contenido en un gran conjunto de páginas web.

Una herramienta de *web scraping* con la capacidad de acceder a todos los avisos inmobiliarios útiles de los sitios seleccionados se presenta como una alternativa adecuada para resolver la tarea propuesta. La herramienta deberá además normalizar y estructurar los datos obtenidos, para garantizar que el análisis consecuente pueda llevarse a cabo. Esta necesidad surge del carácter heterogéneo inherente a las diversas páginas de Internet.

Una manera de formalizar el conocimiento recuperado es a través del uso de ontologías [1]. Una ontología es una descripción del conocimiento sobre un dominio de interés, cuyo núcleo es una especificación procesable por las máquinas con un significado formalmente definido [5]. Definir una ontología para avisos inmobiliarios no solo permitiría establecer un esquema más riguroso para representar la información, sino que también la dotará de valor semántico que podrá ser aprovechado para su curado.

Este artículo, elaborado en el marco de la tesina de grado titulada "Evaluación de técnicas de detección de duplicados sobre grafos de conocimiento de avisos inmobiliarios", presenta la construcción de una herramienta que permite extraer de conocimiento de páginas web y almacenarlo en un grafo de conocimiento, para su utilización en un observatorio inmobiliario. La sección 2 explica cómo se normalizó el vocabulario hallado en los distintos sitios web, ahondando en la definición de una ontología inmobiliaria con base en la reutilización y alineando estándares preexistentes. La sección 3 introduce la metodología de extracción de conocimiento de los portales web, junto con la descripción de una

herramienta que la implementa, y sus resultados preliminares. Finalmente, la sección 4 muestra las conclusiones y detalla posibles trabajos futuros.

2. Definición de una ontología inmobiliaria

Para llevar a cabo un estudio estadístico del mercado inmobiliario, es necesario seleccionar los datos relevantes de cada aviso disponible. Esto implica realizar un análisis de los distintos portales web a explorar, determinando que campos o variables contienen el conocimiento deseado en cada uno.

En primer lugar, fue necesario definir un vocabulario común para normalizar los nombres con los que cada página se refiere a los conceptos de interés. Por ejemplo, un aviso del *sitio 1* podría referirse al valor al cual se oferta cierta propiedad como *precio*, mientras que un aviso del *sitio 2* como *price*.

Para conseguir estructurar la información proveniente de avisos inmobiliarios es necesario identificar los conceptos principales que los representan y las relaciones que los vinculan.

En este dominio es necesario centrarse primordialmente en dos aspectos: uno referente al inmueble, y otro al aviso que lo oferta. Para lograr este objetivo, se realizó un estudio del estado del arte en lo que respecta a ontologías inmobiliarias y de publicaciones online. Para modelar el conocimiento relativo al inmueble como los espacios físicos, edificios, y sus habitaciones, se utilizó la ontología RealEstateCore. Del mismo modo, para conceptualizar los avisos inmobiliarios y su información (a qué sitio pertenece y quién lo publicita), se aprovechó la ontología SIOC.

RealEstateCore [4] es una ontología modular y libre que rápidamente se convirtió en un estándar en el modelado de conocimiento sobre edificios inteligentes. De todos los conceptos que incluye, son destacables *rec:Real_Estate* y *rec:Space*. *rec:Real_Estate* es el elemento que representa una propiedad inmueble, pudiendo estar conformada por más de un edificio, terreno, o similar. Por otra parte, *rec:Space* representa un área del mundo físico que puede a su vez contener subespacios, por lo que permite representar regiones, terrenos, edificios, y habitaciones.

Además, RealEstateCore hace uso del concepto de *foaf:Agent* definido en FOAF [3] para referirse a los humanos u organizaciones que realizan una acción, particularmente sobre un *rec:Real_Estate*. La clase *foaf:Agent* resulta útil a la hora de vincular una propiedad en el mercado con la persona o entidad que la oferta. Para representar que un *foaf:Agent* publicita un inmueble, detallar las plataformas en las que publica los avisos, y adjuntar el conocimiento que corresponde a cada oferta, se utiliza la ontología SIOC.

SIOC provee los principales conceptos y propiedades requeridos para describir información sobre comunidades online en la Web Semántica [2]. Define clases para modelar sitios web, los items que contienen, el contenido de cada uno, y el tipo de publicaciones que se realizan. La utilidad de SIOC a la hora de modelar avisos inmobiliarios está determinada principalmente por las clases *sioc:Site*, *sioc:Post*, y *foaf:Agent* (que referencia directamente al recurso definido en FOAF). Por un lado, *sioc:Site* representa un sitio web que actúa como

comunidad online. Por el otro, *sioc:Post* es un artículo que se publica en un *sioc:Site*. Y finalmente *foaf:Agent* representa aquellos actores que cumplen un rol en alguna tarea. El uso de SIOC posibilita modelar los avisos inmobiliarios como *sioc:Post*s, dado que suelen permitir visualizar y filtrar conjuntos de publicaciones, solo que limitando las interacciones entre los usuarios. Además, permite indicar que los avisos pertenecen a una plataforma identificada como un *sioc:Site*, y son publicados por *foaf:Agent*s, que podrían ser inmobiliarias o personas particulares. Además, SIOC presenta la propiedad *sioc:about*, que permite vincular a un *Post* con el recurso principal al que hace referencia, sin importar que estuviera definido por otra ontología.

El hecho que tanto RealEstateCore como SIOC hayan diseñado sus clases teniendo en cuenta la definición de *foaf:Agent* permite que ambas ontologías se alineen fácilmente mediante el anunciante de un aviso inmobiliario, y a través de la propiedad específica que se tiene en consideración. A esto pueden sumarse la definición de nuevas clases y propiedades que permitan mejorar la estructura del modelo para el dominio inmobiliario. Por ejemplo, podría agregarse una clase *RealEstateListing* como subclase de *sioc:Post* para representar ofertas inmobiliarias, y una propiedad *price* que indique el precio del inmueble según ese aviso particular.

El resultado de este alineamiento es una ontología que permite crear una base de conocimiento capaz de inferir las respuestas a preguntas del estilo *¿cuántas inmobiliarias ofertan el departamento en Libertador al 400?* y *¿a qué precios se ofrece la quinta en Gral. San Martín al 1600?*

3. Recolección de datos

En las etapas iniciales del proyecto los investigadores realizaron un relevamiento manual de información inmobiliaria, seleccionando tres sitios web dedicados a la búsqueda de propiedades e inmuebles, y cinco partidos de prueba dentro la provincia de Buenos Aires. Luego realizaron un muestreo de los sitios, que les permitió formar una base de datos de aproximadamente 2.000 clasificados inmobiliarios en el transcurso de entre 2 y 3 meses.

A fin de conseguir una muestra de mayor tamaño, y de aumentar la significancia estadística del análisis a efectuar, se optó por automatizar el proceso utilizando tecnologías de *web crawling* y *web scraping*. Para eso se diseñó un *scraper web* en Python utilizando como base Scrapy, un framework para la navegación de sitios y extracción de datos estructurados.

Antes de utilizar las arañas, fue necesario configurar períodos de demora personalizados entre cada petición realizada, para evitar sobrecargar los servidores de los portales y no generar problemas en su funcionamiento. Una vez confeccionadas las arañas de cada sitio, se ejecutaron en un entorno paralelo en el que en menos de una semana consiguieron recolectar conocimiento de más de 500.000 avisos inmobiliarios de 40 partidos de la provincia de Buenos Aires.

4. Conclusiones y trabajos futuros

En este artículo se presentó una alternativa para automatizar la extracción y almacenamiento de conocimiento de portales inmobiliarios. Primero, se definió una terminología en inglés con los expertos del dominio, para normalizar el vocabulario variable de las distintas páginas de Internet. Se realizó un estudio del estado del arte de los estándares ontológicos para la representación de inmuebles y publicaciones web. A partir de este, se alinearon las ontologías RealEstateCore y SIOC, consiguiendo un modelo estructurado para la descripción avisos inmobiliarios publicados en la red. Partiendo de este modelo, se diseñó una metodología de extracción de conocimiento basada en técnicas de *web crawling* y *web scraping*. La misma se implementó en una herramienta capaz de almacenar la información de las ofertas inmobiliarias en una estructura que respeta el modelo definido. Finalmente, se comprobó que la herramienta fue capaz de aumentar el tamaño del corpus de datos en más de un 1.000 % en relación a los esfuerzos manuales previos.

Como trabajos futuros se pondrá el foco en la detección automática entidades duplicadas en el grafo de conocimiento, especialmente en lo que refiere a inmuebles referenciados por múltiples avisos. Con este fin será necesario analizar y evaluar diferentes técnicas de deduplicación en grafos de conocimiento, para así determinar la estrategia más eficaz para solucionar el problema.

Referencias

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American p. 4 (May 2001)
2. Breslin, J., Decker, S., Harth, A., Bojars, U.: SIOC: An approach to connect web-based communities. IJWBC **2**, 133–142 (Jan 2006). https://doi.org/10.1504/IJWBC.2006.010305
3. Brickley, D., Miller, L.: FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project (2004), http://xmlns.com/foaf/0.1/
4. Hammar, K., Wallin, E.O., Karlberg, P., Hälleberg, D.: The RealEstateCore Ontology. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web – ISWC 2019, vol. 11779, pp. 130–145. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_9
5. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman & Hall - CRC Press (Aug 2009). https://doi.org/10.1201/9781420090512, journal Abbreviation: Foundations of Semantic Web Technologies Publication Title: Foundations of Semantic Web Technologies
6. Observatorio de valores del suelo para fortalecer la política de Integración social y urbana de barrios populares (Aug 2021), https://www.argentina.gob.ar/noticias/observatorio-de-valores-del-suelo-para-fortalecer-la-politica-de-integracion-social-y
7. Olston, C., Najork, M.: Web Crawling. Foundations and Trends® in Information Retrieval **4**(3), 175–246 (2010). https://doi.org/10.1561/1500000017, http://www.nowpublishers.com/article/Details/INR-017

8. Schrenk, M.: Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL. No Starch Press, second edition edn. (2012)
9. Zhao, B.: Web Scraping. In: Schintler, L.A., McNeely, C.L. (eds.) Encyclopedia of Big Data, pp. 1–3. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-32001-4_483-1

AlfaDatizando: Visualización de contenido generado por usuarios de redes sociales

Paladino Jeziel[1][0000-0002-5457-518X], Lliteras Alejandra Beatriz [1,2][0000-0002-4148-1299], Gardey Juan Cruz[1,3][0000-0002-1765-8189], Grigera Julián[1,2,3][0000-0002-7962-4312]

[1] UNLP, Facultad de Informática, LIFIA. 50 y 120. La Plata. Bs.As. Argentina
[2] CICPBA. Bs.As. Argentina
[3] CONICET. Argentina
{lpaladino,lliteras, jcgardey,
julian.grigera}@lifia.info.unlp.edu.ar

Abstract. La visualización de datos para desarrollar la habilidad de Pensamiento Computacional permite trabajar con los estudiantes secundarios diferentes tipos de datos y contenidos provenientes de diversas fuentes. En particular, los contenidos generados por los usuarios de las redes sociales, lo que podría ser usado en materias relacionadas a las humanidades digitales. En este trabajo se presenta un relevamiento de redes sociales que permiten acceder al contenido generado por sus usuarios a través de APIs. Para este trabajo, se realizó un trabajo en dos etapas, primero, el relevamiento bibliográfico de artículos que mencionan la posibilidad de acceder a contenido y luego, la validación analizando la disponibilidad de las APIs con el fin de consumir posteriormente, el contenido desde la plataforma de visualización AlfaDatizando.

Keywords: Redes Sociales, Contenido Generado por Usuarios, API, Visualización de Datos, Pensamiento Computacional, Humanidades Digitales, Ciencias Sociales

1 Motivación

El presente trabajo se enmarca en el Proyecto de Innovación con Alumnos (2022), de la Facultad de Informática, UNLP, llamado "Aprendo con Datos. Plataforma para la visualización de datos con fines educativos en nivel secundario para Ciencias Sociales y Humanidades". La temática abordada es parte del proyecto de doctorado de la profesora Lliteras Alejandra.

Las nuevas formas de comunicación digital han permitido nuevos tipos de interacciones entre los miembros de los movimientos sociales, incluso permitiendo interacciones entre actores que no tenían relaciones previas y que nunca se han visto cara a cara. Los recientes desarrollos en datos, tecnología y métodos de análisis ofrecen oportunidades para que el análisis de redes sociales desempeñe un papel destacado en el nuevo mundo de la investigación de las Ciencias Sociales [1].

Por otro lado, es común el uso de contenido de la red social Twitter [2] mediante plataformas de visualización (por ejemplo, SocioViz[1]) en actividades relacionadas a investigación en el área de Humanidades Digitales y de las Ciencias Sociales.

[1] https://socioviz.net/

AlfaDatizando[2][3] es una plataforma de visualización de datos para desarrollar Pensamiento Computacional en estudiantes secundarios con foco en las Humanidades Digitales y las Ciencias Sociales, que en la actualidad cuenta con visualización de datos provenientes de la API de Twitter y de fuentes de datos en archivos csv. Sabiendo que las redes sociales de mayor alcance son [4]: YouTube, Facebook, Instagram, Pinterest, LinkedIn, Snapchat, Twitter, WhatsApp, TikTok, Reddit y Nextdoor, se analizarán en la bibliografía estudios que muestren que se usó alguna API oficial para acceder a contenido generado por los usuarios de alguna de las redes sociales.
En este trabajo se presenta un relevamiento y análisis de bibliografía para determinar las redes sociales que proveen de una API oficial para acceder al contenido generado por sus usuarios, para luego determinar, si en la actualidad, dichas redes sociales efectivamente cuentan con esa posibilidad, para posteriormente implementar su acceso desde AlfaDatizando.

2 Aporte

El trabajo consiste en dos etapas, por un lado, relevamiento y análisis bibliográfico respecto a las redes sociales sobre las cuales se accede al contenido generado por los usuarios y luego, conociendo las redes sociales de mayor alcance, cuáles de ellas de acuerdo con las APIs disponibles, era posible acceder al contenido generado por sus usuarios
La ***metodología*** empleada para este trabajo consistió en armar el siguiente *string* de búsqueda:

(“social network" and "api” and "development" and "case study" and (“user content” OR “users data”)) -"data security" -"CYBERATTACKS" -"blockchain" -"privacy" -"Domain Specific Language" -DSL

Como motor de búsqueda, se usó *Google Scholar* y se limitó la búsqueda entre los años 2018-2022 ya que es de suma importancia información reciente sobre las redes sociales. Nótese el caso de Meta (antes Facebook) que periódicamente fue añadiendo restricciones en cuanto al uso de sus API 's, provocando que muchos desarrollos e investigaciones, queden obsoletas.
A continuación, en la Tabla 1, se detallan los criterios de inclusión para los trabajos que retorne la búsqueda.

Tabla 1: Criterios de inclusión

#	Criterio
I1	Descripción sobre cómo usar la API de una red social.
I2	Presentación de la API de una red social.

[2] http://www.alfadatizandonos.okd.lifia.info.unlp.edu.ar/

I3	Utilización de los datos de los usuarios de una red social por medio de APIs oficiales provistos por la misma red social.
I4	Artículo en inglés.
I5	Muestra cómo obtener el contenido generado por los usuarios.

Mientras que los criterios de exclusión se detallan en la Tabla 2.

Tabla 2: Criterios de exclusión

#	Criterio
E1	La red social utilizada no se encuentra entre las de mayor alcance [4]
E2	No describe cómo usar la API de una red social.
E3	No presenta la API de una red social.
E4	No utiliza los datos de los usuarios de una red social por medio de APIs oficiales provistos por la misma red social.
E5	No está en inglés.
E6	No muestra cómo obtener el contenido generado por los usuarios.

Una vez ***ejecutada la búsqueda*** en Google Scholar usando el string de búsqueda, se obtuvieron 112 resultados. En base a la lectura del título, abstract y palabras claves de cada trabajo del conjunto de resultado, y aplicando los criterios de inclusión/exclusión, 33 trabajos fueron incluidos, mientras que 79 fueron excluidos. De los 33 trabajos que quedaron al cumplir alguno de los criterios de inclusión a partir de la lectura del título, abstract y palabras claves, se procedió a la lectura de trabajo completo y nuevamente se aplicaron los criterios de inclusión/exclusión. Luego de esta segunda fase, de los 33 trabajos, quedan 9 incluidos. La Fig. 1 muestra el proceso de análisis de los resultados.

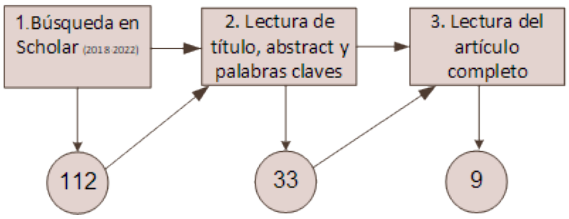


Fig. 1: Proceso de selección de artículos

De los 33 trabajos que quedaron al aplicar por primera vez los criterios de inclusión y exclusión, 17 eran artículos de revistas, conferencias y congresos, 14 responden a trabajos de tesis y 1 a un libro. La Tabla 3 muestra la distribución de los trabajos.

Los 9 trabajos resultantes son [9], [14], [16], [20], [22], [23], [25], [27] y [29]. De éstos el trabajo [20] usa la API de YouTube, mientras que los restantes la API de Twitter.

Tabla 3: Trabajos identificados por tipo

Artículo	[5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22]
Tesis	[23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36]
Libro	[37]

De acuerdo con el relevamiento bibliográfico, solo las redes sociales de Twitter y YouTube muestran casos de acceso al contenido generado por los usuarios. Como se mencionó anteriormente, las redes sociales más difundidas son YouTube, Facebook, Instagram, Pinterest, LinkedIn, Snapchat, Twitter, WhatsApp, TikTok, Reddit y Nextdoor. Por lo anterior, es que se analizó para cada una de estas redes si tenía disponible una API para acceder al contenido de los usuarios más allá de que no hayan aparecido en la búsqueda bibliográfica. La Tabla 4 muestra para cada red social, su tiene API disponible a la fecha para acceder a contenido generado por el usuario y la url de acceso a la API general de la red.

Tabla 4: Análisis de redes sociales

Red social	API Contenido	Url desde donde se la accede
YouTube	Si	https://developers.google.com/youtube/v3
Facebook	No	https://developers.facebook.com/
Instagram	No	https://developers.facebook.com/docs/instagram
Pinterest	No	https://developers.pinterest.com
Linkedin	No	https://developer.linkedin.com/
Snapchat	No	https://developers.snap.com/
Twitter	Si	https://developer.twitter.com/en
WhatsApp	No	https://developers.facebook.com/products/whatsapp/
TikTok	No	https://developers.tiktok.com/
Reddit	Si	https://www.reddit.com/dev/api/
Nextdoor	No	No posee API

Luego, y de acuerdo con lo relevado y analizado, sólo es posible acceder al contenido generado por los usuarios de las redes sociales Twitter, YouTube y Reddit

3 Líneas de Investigación Futura

A partir del relevamiento bibliográfico y análisis de APIs de redes sociales, se espera implementar en AlfaDatizando el acceso y visualización del contenido generado por los usuarios en las redes sociales YouTube y Reddit, considerando que la plataforma ya considera la visualización de contenido generado por los usuarios de Twitter.

Referencias

1. Tindall, D., McLevey, J., Koop-Monteiro, Y., & Graham, A. (2022). Big data, computational social science, and other recent innovations in social network analysis. Canadian Review of Sociology/Revue canadienne de sociologie
2. Yu, J., & Muñoz-Justicia, J. (2022). Free and low-cost twitter research software tools for social science. Social Science Computer Review, 40(1), 124-149.
3. Lliteras A., Artopoulos A., Fernandez A. & Huarte J. AlfaDatizando: a Data Visualization Platform to work Computational Thinking in Digital Humanities. Laclo 2022. In press
4. Pew Research. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/
5. Hajikhani, A., & Porras, J. (2018). Advanced Methods: Operationalizing Social Network Services Data—Deep Content Analysis to Comprehend Brand Presence. In Innovation Discovery: Network Analysis of Research and Invention Activity for Technology Management (pp. 471-502).
6. Babvey, P., Lipizzi, C., & Ramirez-Marquez, J. E. (2019, December). Dissecting Twitter discussion threads with topic-aware network visualization. In 2019 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 1359-1364). IEEE.
7. Juric, T. (2022). Ukrainian refugee integration and flows analysis with an approach of Big Data: Social media insights. medRxiv.
8. Uyheng, J., & Carley, K. M. (2021, May). Computational Analysis of Bot Activity in the Asia-Pacific: A Comparative Study of Four National Elections. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 727-738)
9. Abu-Salih, B., Qudah, D. A., Al-Hassan, M., Ghafari, S. M., Issa, T., Aljarah, I., ... & Alqahtan, S. (2022). An Intelligent System for Multi-topic Social Spam Detection in Microblogging. arXiv preprint arXiv:2201.05203.
10. Rodríguez-Ibáñez, M., Gimeno-Blanes, F. J., Cuenca-Jiménez, P. M., Soguero-Ruiz, C., & Rojo-Álvarez, J. L. (2021). Sentiment Analysis of Political Tweets from the 2019 Spanish Elections. IEEE Access, 9, 101847-101862.
11. Chen, S., Chen, S., Wang, Z., Liang, J., Wu, Y., & Yuan, X. (2018). D-map+ interactive visual analysis and exploration of ego-centric and event-centric information diffusion patterns in social media. ACM Transactions on Intelligent Systems and Technology (TIST), 10(1), 1-26.
12. Hussain, J., Satti, F. A., Afzal, M., Khan, W. A., Bilal, H. S. M., Ansaar, M. Z., ... & Lee, S. (2020). Exploring the dominant features of social media for depression detection. Journal of Information Science, 46(6), 739-759.

13. Masood, F., Almogren, A., Abbas, A., Khattak, H. A., Din, I. U., Guizani, M., & Zuair, M. (2019). Spammer detection and fake user identification on social networks. IEEE Access, 7, 68140-68152.
14. Tago, K., & Jin, Q. (2018). Influence analysis of emotional behaviors and user relationships based on twitter data. Tsinghua Science and Technology, 23(1), 104-113.
15. Burini, F., Cortesi, N., Gotti, K., & Psaila, G. (2018). The urban nexus approach for analyzing mobility in the smart city: towards the identification of city users networking. Mobile Information Systems, 2018.
16. Aggrawal, N., & Arora, A. (2019). Behaviour of viewers: YouTube videos viewership analysis. International Journal of Business Innovation and Research, 20(1), 106-128.
17. Saeed, M. U., & Hassan, T. U. (2020). Relationship Among the Attributes of World Countries and Their Coverage in Tweets of International News Agencies: 2010–2016. Indian Journal of Science and Technology, 13(08), 966-982.
18. Madanian, S., Airehrour, D., Samsuri, N. A., & Cherrington, M. (2021, October). Twitter Sentiment Analysis in Covid-19 Pandemic. In 2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 0399-0405). IEEE.
19. Burini, F., Cortesi, N., Gotti, K., & Psaila, G. (2018). Research Article The Urban Nexus Approach for Analyzing Mobility in the Smart City: Towards the Identification of City Users Networking.
20. Ashraf, N., Zubiaga, A., & Gelbukh, A. (2021). Abusive language detection in youtube comments leveraging replies as conversational context. PeerJ Computer Science, 7, e742.
21. Wu, B., Cheng, W. H., Zhang, Y., Cao, J., Li, J., & Mei, T. (2018). Unlocking author power: On the exploitation of auxiliary author-retweeter relations for predicting key retweeters. IEEE Transactions on Knowledge and Data Engineering, 32(3), 547-559.
22. Shugars, S., Gitomer, A., McCabe, S., Gallagher, R. J., Joseph, K., Grinberg, N., ... & Lazer, D. (2021). Pandemics, protests, and publics: Demographic activity and engagement on Twitter in 2020. Journal of Quantitative Description: Digital Media, 1.
23. Haselwood, S. M. (2018). Evolution of educator professional development in the age of social media: A case study of the# Oklaed community of practice on Twitter (Doctoral dissertation, Oklahoma State University).
24. Krochmal, G. (2020). Sentiment of tweets and socio-economic characteristics as the determinants of voting behavior at the regional level. Case study of 2019 Polish parliamentary election. arXiv preprint arXiv:2010.03493.
25. Kouvela, M. (2020). Bot detective: explainable bot detection in twitter. A thesis submitted in fulfillment of the requirements for the degree of Master of Data & Web Science.
26. Binzagr, F. A. (2022). Social Intelligence Approach for Service-Oriented Software (Doctoral dissertation).
27. Vitali, L. (2020). Analysis and detection of social bots on Twitter mimicking human interests in people or contents.
28. Alduaiji, N. (2019). Edge Attribute-enhanced Community Discovery in Social Networks.
29. Rai, A. (2018). Inferring landscape preferences from social media using data science techniques (Doctoral dissertation, University of Illinois at Urbana-Champaign).
30. Vayansky, I. (2018). An evaluation of geotagged Twitter data during Hurricane Irma using sentiment analysis and topic modeling for disaster resilience (Doctoral dissertation, Coastal Carolina University).
31. Pitenis, Z. (2019). Detecting offensive posts in greek social media.
32. Rehman, F. U. (2018). Towards a Framework for Multiscale Social Event Extraction and Visualization (Doctoral dissertation, Université Grenoble Alpes).

33. Du, J. (2019). VaxInsight: an artificial intelligence system to access large-scale public perceptions of vaccination from social media.
34. Matamoros-Fernandez, A. (2018). Platformed racism: The Adam Goodes war dance and booing controversy on twitter, YouTube, and Facebook (Doctoral dissertation, Queensland University of Technology).
35. Apong, R. A. A. H. M. (2018). Mining negation and uncertainty in social healthcare networks. The University of Manchester (United Kingdom).
36. Saleem, H. M. (2022). Abusive language through the lens of online communities.
37. Nahili, W., Rezeg, K., & Miloudi, L. (2018). Towards better decision-making with twitter sentiment analysis. Business Intelligence & Big Data

Eficiencia energética en el hogar, una propuesta tecnológica basada en simulación

Javier Marchesini[1], Pablo Santibáñez Acuña[1], Ariel Pessotano[1], Leandro Sosa[1], Pablo Salani[1], Julián Abregú[1], Leopoldo Nahuel[1], Agustín Álvarez Ferrando[1]

[1] Grupo de Investigación & Desarrollo Aplicado a Sistemas informáticos y computacionales
Facultad Regional La Plata, Universidad Tecnológica Nacional, Av.60 esq. 124 s/n, La Plata
Buenos Aires, Argentina
{jmarchesini, psatibanes lnahuel, aaferrando}@frlp.utn.edu.ar
{arielpessotano, lhsosa, psalani, jabregu}@alu.frlp.utn.edu.ar

Abstract. El presente trabajo, tiene como finalidad difundir alcance, objetivos y avances sobre desarrollo de tecnologías informáticas para educación y concientización en eficiencia energética. El consumo energético mundial se encuentra en aumento exponencial, conduciendo a un posible desequilibrio energético y un mayor impacto ambiental. Por consecuencia, la eficiencia energética se convirtió en una de las estrategias más importantes para reducir el consumo energético, pudiendo aportar un conjunto de acciones como educación y concientización en ahorro y uso responsable de la energía. Por ello, mediante las actividades de I&D, buscamos brindar Tecnologías de Información y Comunicación (TIC) basadas en simulación, destinadas a educar y concientizar en temáticas como sustentabilidad, eficiencia y gestión energética, a efectos de asegurar buenos hábitos del uso de energía en el hogar. Consideramos que las TIC suministran medios que acompañen los procesos de educación y aprendizaje en temas de eficiencia energética en ámbitos hogareños, despertando interés y motivación a partir de su utilización.

Keywords: Eficiencia Energética, Simulación, TIC

1 Introducción

Históricamente, la situación energética mundial, ha experimentado cambios constantemente y hoy en día continua, pero a ritmo acelerado. La energía se convirtió en un recurso esencial para el desarrollo de la vida humana y en un factor fundamental para el crecimiento de los países, siendo empleada para infraestructura, producción, transporte y necesidades de las poblaciones modernas. El crecimiento socioeconómico de las poblaciones hace que día a día, se consuman grandes cantidades de energía conduciendo a incrementos significativos de demanda energética. Esto posiciona al mundo en una difícil situación principalmente por dos razones, en primer lugar, porque las principales fuentes energéticas son de carácter no renovables, y en segundo

lugar, al ser provenientes de la quema de combustibles fósiles (petróleo, gas natural y carbón mineral) genera efectos indeseables en el medio ambiente.

La Argentina, no queda al margen de las problemáticas mencionadas. El consumo energético, en los últimos años, se incrementado significativamente tanto en el sector industrial como en el residencial conduciendo a problemas de demanda y déficit energético, que generalmente tienden a resolverse mediante la oferta, generando energía o recurriendo a importaciones por no poder satisfacer la demanda, impactando directamente en aspectos económicos. La energía que se consume proviene de diferentes fuentes que conforman la matriz energética, caracterizada por tener una alta dependencia de los fósiles/hidrocarburos, principalmente del gas natural, resultando el 51,8% en la matriz energética según el Balance Energético Nacional del año 2021.

Por estos motivos, debemos ser conscientes de la importancia de hacer un uso responsable de energía, acompañando al ahorro energético a fin de asegurar la disponibilidad energética asegurando el abastecimiento para el desarrollo sostenible y contribuir a la conservación del medio ambiente.

Motivado por este contexto, la eficiencia energética se convirtió en una de las estrategias más importantes, destinadas a la reducción de los consumos energéticos. La International Energy Agency (IEA) la considera como "El principal combustible para el desarrollo sostenible". Es una forma de gestionar el crecimiento de la energía, obteniendo un resultado igual con menor consumo o un resultado mayor consumiendo lo mismo [2]. Nos permite adoptar un conjunto de acciones y medidas como la educación y concientización en ahorro y uso responsable de la energía.

Consideramos que el uso de uso de Tecnologías de Información y Comunicación (TIC) pueden desempeñar un papel sumamente importante como motor para la eficiencia energética. Contribuyen al desarrollo ágil, económico y masivo de herramientas de simulación dando apoyo a la eficiencia energética, promoviendo el compromiso de la población en el consumo responsable de energía. Esta hipótesis es parte de las investigaciones abordadas en un Proyecto de Investigación y Desarrollo (PID), homologado por la Secretaría de Ciencia, Tecnología de la del Rectorado de la Universidad Tecnológica Nacional (UTN).

El objetivo general, es desarrollar herramientas software basadas en simulación que permitan estimar consumos energéticos de energías facturables como la energía eléctrica y el gas natural, representando diferentes escenarios de consumo partiendo del uso hogareño. Se espera que con la herramienta se puede educar y generar conciencia en el ahorro y uso racional de la energía en el hogar.

2 Marco Teórico

Las políticas eficiencia energética son un factor esencial para la reducción de los consumos energéticos. Estas políticas, con el apoyo de las de TIC basadas en simulación, ayudan a la concientización y aprendizaje sobre las temáticas de ahorro y uso racional de la energía.

La simulación es una herramienta que nos permite establecer un marco experimental para resolver problemas, describir el comportamiento y predecir comportamiento

ante determinados cambios en el sistema. Resulta un recurso muy valioso para la enseñanza en los temas de eficiencia energética, dado que los conocimientos se adquieren a partir de la acción e interacción con eventos que se generan ante diferentes escenarios. Es útil para alcanzar un aprendizaje significativo, permitiendo recrear experiencias sobre la realidad.

2.1 Por qué la simulación como técnica para generar conciencia

Las técnicas de simulación son útiles para alcanzar un aprendizaje significativo, permitiendo recrear experiencias sobre la realidad. Esto da lugar a que los usuarios aprendan a partir de la acción e interacción con eventos que se generan sobre escenarios simulados. Lo hace partícipe de una vivencia que le permitirá desarrollar hábitos, destrezas, esquemas mentales, entre otras características que le sirvan como punto de apoyo para la mayor comprensión de una disciplina.

2.2 Características principales de un entorno de simulación

Las características de un entorno de simulación van a depender del área de conocimiento a aplicarlo. Por tal razón, es difícil realizar una clasificación general a todas las herramientas. Según plantea por J.M. Ruiz Gutiérrez [3] las características comunes más importantes de un software de simulación para considerarla una herramienta de aprendizaje son:

- Interfaz de Usuario
- Posibilidad de Conexión con el exterior.
- Incorporación de módulos de planificación del aprendizaje.
- Posibilidad de conexión con otros programas.
- Lenguaje de programación gráfica
- Posibilidad de ampliación de biblioteca de objetos.
- Interfaces Hombre Máquina.
- Instrumentación Virtual.

2.3 Importancia del interfaz de usuario

Las simulaciones son eficaces a la hora de generar resultados precisos, pero hay limitaciones como pueden ser la interfaz y experiencia de usuario, que impide que muchas personas no comprenden plenamente su propósito o el significado transmitido de los resultados.

Por tal motivo, HIX, D. y Hartson, H. R, en su investigación [4] definen […] "Para los usuarios, la interfaz es el sistema" […]. Se considera que una interfaz es la parte de un sistema con la que los usuarios interactúan, vinculando procesos perceptivos y cognitivos, convirtiéndose en un factor fundamental para el éxito de cualquier sistema, incluidas las simulaciones. Por este motivo, una interfaz defectuosa, como mencionan los investigadores James H. Gerlach y Feng-Yang Kuo [5] puede atrapar al

usuario en situaciones no deseadas, afectando así a la actitud de los usuarios hacia la aplicación. La eficacia de un sistema puede verse obstaculizada muy rápidamente si hay defectos en la navegación, el diseño de la interfaz y la disposición.

3 Una propuesta tecnológica

Actualmente, trabajamos en la definición y especificación funcional de una herramienta basada en simulación, que a partir de artefactos hogareños permita contabilizar consumos y costos, representar diferentes escenarios y proporcionar una comunicación a los usuarios de buenos hábitos y recomendaciones destinados a generar conciencia en el uso de los recursos energéticos.

Nuestra propuesta nace de la necesidad de brindar, a las personas que utilicen la herramienta de software, un panorama más general sobre el consumo real que posee en su inmueble, teniendo en cuenta tanto la energía eléctrica como el gas y que pueda interactuar de manera dinámica con la aplicación a través de los distintos simuladores.

Como una primera aproximación al desarrollo nos encontramos prototipando la interfaz de usuario de un simulador de consumo eléctrico (Fig. 1), basándonos en las premisas que hemos considerado dentro del marco teórico. Con la ayuda de este prototipo podremos probar alternativas de flujo que nos permitan, a la hora del desarrollo mismo de la aplicación, elegir la opción que brinde la mejor experiencia al usuario

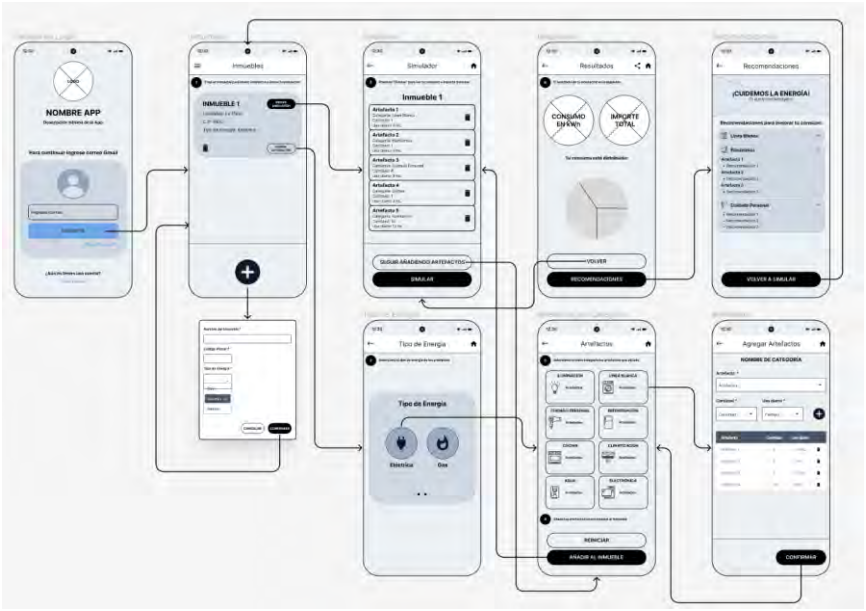


Fig. 1. Propuesta – Wireframes & Screenflow

Para conseguir la flexibilidad de simular distintos escenarios de consumo, la aplicación estará dividida en un conjunto de aplicaciones especializadas en un tipo de

fuente de energía que confluirán en recomendaciones de posibles alternativas y mejoras para el usuario.

4 Conclusiones y Trabajo Futuro

En este trabajo, hemos indicado las problemáticas a las cuales se enfrenta el mundo en temas energéticos. Así mismo, exponemos la importancia de las TIC integradas con técnicas de simulación, y cómo estas apoyan los procesos a la educación en sustentabilidad, eficiencia y gestión energética a efectos de asegurar buenos hábitos de uso de la energía. En las etapas iniciales del proyecto, las actividades se focalizaron conocer y descubrir el estado del arte de la simulación, la relación con los métodos de enseñanza y las TIC, con el objetivo de establecer las características con las que debe contar la herramienta propuesta.

En conclusión, observamos que la simulación resulta una técnica importante para la herramienta software propuesta, siendo un gran aporte para la eficiencia energética. Consideramos que los usuarios que vayan a utilizar la herramienta se educarán y generará conciencia a partir de la interacción con los diferentes procesos provistos.

En futuras etapas, estudiaremos el Programa Nacional de Etiquetado de Viviendas, con el objetivo de integrar mecanismos que permitan realizar una simulación del etiquetado de vivienda, partiendo de datos hogareños y así conocer el Índice de Prestaciones Energéticas (IPE) de una vivienda. Se pretende ofrecer una evaluación energética de viviendas, sin necesidad de tener conocimientos avanzados sobre calculo energético.

References

1. Balance Energético Nacional de la República Argentina - Año 2021 https://www.argentina.gob.ar/economia/energia/hidrocarburos/balances-energeticos, ultimo acceso 2022/08/19.
2. International Energy Agency – Energy Efficiency, https://www.iea.org/topics/energy-efficiency, ultimo acceso 2022/08/19.
3. José M. Ruiz Gutiérrez, “La Simulación como Instrumento de Aprendizaje (Evaluación de Herramientas y estrategias de aplicación en el aula)”, https://docplayer.es/8550830-La-simulacion-como-instrumento-de-aprendizaje-evaluacion-de-herramientas-y-estrategias-de-aplicacion-en-el-aula.html, ultimo acceso 2022/08/19.
4. HIX, D. y Hartson, H. R.; Developing user interfaces: ensuring usability through product and process. New York: John Wiley & Sons; 1993.
5. Gerlach, J.H., Kuo, F-Y, “Understanding Human-Computer Interaction for Information Systems Design, MIS Quarterly”, 1991.

Identificación de Personas en Sistemas de Videovigilancia sin uso de Reconocimiento Facial

Tomas Cannatella, Miguel Méndez-Garabetti y Pablo Javier Sández

Laboratorio de Investigación en Ciencia y Tecnología, Facultad de Ciencias Sociales y Adminsitrativas, Universidad del Aconcagua, Mendoza, Argentina.
tomas742011@gmail.com

Abstract. En este artículo se plantea la idea de un sistema biométrico de reconocimiento de personas. Con el objetivo de poder reconocer a una persona por su forma de andar utilizando Inteligencia Artificial. Analizando distintas maneras de implementación en base a desarrollos realizados por otras personas.

Keywords: Cycle Gait · Gait recognition · Soft Bio-metrics · Joint Learning · Network visualization

1 Motivación

El desarrollo de este trabajo surgió principalmente por el interés en desarrollar un sistema biométrico utilizando la inteligencia artificial con la diferencia de proponer algo distinto a lo que estamos acostumbrados a ver. En este caso, la idea principal es poder implementarlo en el área de seguridad para una vivienda proponiendo un sistema biométrico que no sea tan invasivo y tedioso de activar o desactivar cuando una persona entra o sale de la casa. Este trabajo consistio en realizar una revisión de literatura para luego poder realizar una evaluacion sobre si es viable o no realizar una implementación sobre la idea.

2 Introducción

Antes de la existencia de los sistemas biométricos, las maneras de distinguir a las personas eran únicamente por documentos de identificación o contraseñas. Pero gracias a los grandes avances de la tecnología, las identificaciones de las personas han sido más complejas y seguras. En la actualidad hay muchas maneras de controlar el acceso a un lugar, ya sea por una persona o distintos sistemas biométricos. Existen distintos sistemas biométricos que permiten verificar la identidad de una persona ya sea por voz, huella dactilar, rostro, etc. Cada uno con sus ventajas y desventajas. Por ejemplo algunos pueden ser más invasivos que otros, en este artículo se hará hincapié especialmente en los sistemas biométricos que permitan identificar a una persona por su forma de moverse o desplazarse, sin necesidad de recurrir a técnicas de reconocimiento facial.

3 Material y Método de Búsqueda

Como método de elaboración de este desarrollo de revisión se realizó una búsqueda sobre el reconocimiento de personas con inteligencia artificial. Pero la mayoría de artículos trataban sobre la identificación de personas con reconocimiento facial. Luego se hizo una búsqueda un poco más exhaustiva donde se encontraron algunos artículos sobre la identificación de personas mediante su forma de andar. La cantidad de artículos que se encontraron fueron muy pocos, por lo tanto lo que se hizo fue armar una lista con palabras claves que estén relacionado específicamente en el tema. Con esta técnica se obtuvieron un mejor resultado de búsqueda. Específicamente en el motor de búsqueda de Google Scholar y Scopus. De esta manera se pudo recolectar los artículos que hablan sobre la identificación de una persona por su forma de andar utilizando Inteligencia Artificial (IA). Luego para buscar más artículos se obtuvieron algunos artículos en las referencias que mencionaban los artículos.

4 Sistemas Biométricos

Como se mencionó al comienzo, existe gran variedad de sistemas biométricos para verificar la identidad de una persona, pero algo en común que tienen éstos es que están conformados por dos pasos fundamentales para la verificación de la identidad de una persona: 1) reclutamiento, y 2) utilización. En el primer paso se obtienen los datos de la persona para luego almacenarlo en el sistema. Luego en el segundo paso, la persona utiliza el sistema y el mismo compara con los datos almacenados analizando de esta manera el porcentaje de acierto o error [2].

Estos dos pasos están conformados de las siguientes fases:

- **Captura:** Se recogen datos físicos, biológicos o de comportamiento del usuario.
- **Preprocesado:** Adapta los datos para posteriormente poder realizar una extracción de los mismos.
- **Extracción de características:** Adapta los datos para posteriormente poder realizar una extracción de los datos.
- **Comparación**: Las características de las muestras se comparan con el patrón ya almacenado.

5 Sistema biometrico por su forma de caminar

La forma de caminar de una persona es una característica biométrica que define como es la forma que una persona se mueve, está dada por un comportamiento periódico que se compone por varias fases donde se identifican diferentes conductas naturales del sujeto. Estas fases se la conocen como fase de estancia cuando el pie esta en contacto con el suelo y balanceo donde el pie no tiene ningún tipo de contacto. Y se conoce como un ciclo al intervalo de tiempo entre los dos estados [5].

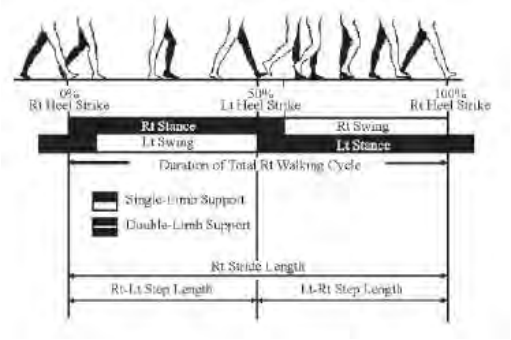


Fig. 1. Fases del proceso de caminar.

A diferencia de otros sistemas biométricos como la utilización de la voz, huella dactilar o el rostro, para reconocer a una persona por su forma no necesitamos de una gran calidad de video para poder identificar al mismo. Además que la misma es menos invasiva que los otros sistemas biométricos ya que no necesitamos cooperación de la misma.

5.1 Aplicaciones

Con la alta cantidad de cámaras se volvió mucho más común la utilización de este tipo de sistemas para la identificación de personas. Se puede utilizar tanto en lugares abiertos como cerrados. Este tipo de sistemas biométricos se puede aplicar en sistemas de video vigilancias, sistemas particulares de seguridad o controles de acceso [5].

5.2 Prototipo de implementación

Durante la investigación sobre el tema se encontraron bastantes propuestas para implementar este sistema biométrico. Como por ejemplo el uso de sensores inerciales que en las pruebas de entrenamiento de la red neuronal tuvo un resultado entre el 94% y 98% de eficiencia. También había otras propuestas interesantes como la utilización del sensor de Kinect de la consola de videojuegos de Microsoft que facilitaba la implementación, ya que generaba modelos 3D de la persona. Pero la implementación que mas me llamo la atención fue la que se menciona en el artículo desarrollado por ***Prasad Pai*** que habla sobre las firmas digitales utilizando la forma de caminar de las personas, ya que se utiliza la librería Tensor Flow para el desarrollo y es algo mucho más reciente lo que

se propone. En ese artículo se muestra un pequeño prototipo mostrando que no hace falta comenzar un desarrollo desde cero sino que se puede hacer utilizando como base un trabajo ya desarrollado con las librerías de Tensor Flow. Se utiliza una grabación entre seis a diez segundos la cual es de poca duración para lo que queremos hacer. Entonces en esa secuencia lo que se hace es realizar un estudio de todos los patrones posible de la persona teniendo en cuenta que la persona siempre va a estar caminando de manera diferente para aumentar la cantidad de datos. Por ejemplo algunas veces pude caminar rápido, otras veces puede caminar lento, también podría caminar de izquierda a derecha o viceversa y así infinitas posibilidades [1][7][6][8][4][9][3].

Fig. 2. Izquierda: Secuencia de entrada,Centro: Detección de persona, Derecha: Resultado de detección.

6 Conclusiones y Trabajo Futuro

Este tipo de sistema biométrico es bastante interesante a comparación de los otros y se podrían realizar grandes aplicaciones con el mismo. A pesar de la complejidad que lleva implementarlo podría ser un gran avance para la seguridad en la videovigilancia y así poder evitar actos delictivos. Como también poder aplicarlo en otros campos que no estén relacionado con la videovigilancia ya sea por ejemplo en la medicina para detectar enfermedades a los pacientes. Seguir con la investigación del sistema biométrico por la forma de andar de la persona y realizar una comparativa para ver cuál de los distintas implementación es mejor para llevar a cabo este sistema biométrico.

Referencias

1. Generating digital signatures with gait — Towards Data Science, https://towardsdatascience.com/generating-digital-signatures-with-the-gait-of-people-3a66f0c44b7b
2. Sistemas de aprendizaje automático para reconocimiento de personas mediante gait
3. Balazia, M., Plataniotis, K.N.: Human gait recognition from motion capture data in signature poses. IET Biometrics **6**(2), 129–137 (mar 2017). https://doi.org/10.1049/IET-BMT.2015.0072, https://onlinelibrary.wiley.com/doi/full/10.1049/iet-bmt.2015.0072 https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2015.0072 https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/iet-bmt.2015.0072
4. Bashir, K., Xiang, T., Gong, S.: Gait recognition without subject cooperation. Pattern Recognition Letters **31**(13), 2052–2060 (oct 2010). https://doi.org/10.1016/J.PATREC.2010.05.027
5. Comparación De Sistemas De Reconocimiento Biométrico De Personas Usando Características De La Forma De Andar, D.Y., Gabriel Sanz, S.: Universidad Autónoma de Madrid Escuela politécnica superior Proyecto fin de carrera (2012)
6. Delgado, R., Tutorizado, E., Ramos Cózar, J., Secretario Del Tribunal, E.:
7. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A.: Gait recognition for person re-identification. Journal of Supercomputing **77**(4), 3653–3672 (apr 2021). https://doi.org/10.1007/S11227-020-03409-5/FIGURES/10, https://link.springer.com/article/10.1007/s11227-020-03409-5
8. Sánchez, A., Pantrigo, J.J., Rubio, A., Virseda, J., Rey, U., Carlos, J., Tulipán, C..: Un Estudio sobre la Identificación de Personas basada en su Movimiento al Caminar (Gait)
9. Zhao, H., Wang, Z., Qiu, S., Wang, J., Xu, F., Wang, Z., Shen, Y.: Adaptive gait detection based on foot-mounted inertial sensors and multi-sensor fusion. Information Fusion **52**, 157–166 (dec 2019). https://doi.org/10.1016/J.INFFUS.2019.03.002

Predicción de la Respuesta en un Sistema de Búsqueda de Respuesta Semántico.

Matías Oyarzun and Sandra Roger

Grupo de Investigación en Lenguajes e Inteligencia Artificial (GILIA),
Facultad de Informática. Universidad Nacional del Comahue
Buenos Aires 1400, (8300) Neuquén
matias.oyarzun@est.fi.uncoma.edu.ar
roger@fi.uncoma.edu.ar

Resumen En este artículo se describe un primer prototipo que se ha desarrollado para la tarea de Predicción de la Categoría planteada en el desafío SMART[1]. Este problema se puede plantear como una tarea de clasificación multiclase, pues toma preguntas en lenguaje natural y devuelve la categoría (resource, literal, boolean) a la que pertenecen. Para el entrenamiento, se utilizaron los datasets de DBPedia y Wikidata de los SMART 2020 y 2021. En este prototipo, se entrenaron 4 modelos de aprendizaje automático con distintas combinaciones de los datasets para hallar el más preciso. El mejor modelo, obtuvo una precisión del 97, 2 % y 96,8 % para los datasets de DBpedia y Wikidata, respectivamente. En ambos casos, se utilizó el clasificador Support-Vector Machines (SVM). Posteriormente, se busca también la construcción de un modelo para la tarea de Predicción del Tipo de Respuesta. Ésto permitirá, finalmente, la implementación de un Sistema de Búsqueda de Respuestas eficiente.

Keywords: Clasificación de Preguntas, Aprendizaje Automático, SMART, Question Answering.

Contexto Este trabajo está parcialmente financiado por la UNCo, en el marco del nuevo proyecto de investigación *Tecnologías Semánticas para el desarrollo de Agentes Inteligentes*. Como así también, lo financia parcialmente el Consejo Interuniversitario Nacional (CIN) con una Beca de Estímulo a las Vocaciones Científicas 2021. Este trabajo formará parte de la propuesta de tesis final de carrera.

1. Introducción

El proceso de QA (*Question Answering* - QA) consta de una etapa de análisis de la pregunta, recuperación de los datos relevante de fuentes de conocimiento y la extracción de la información concreta y correcta como respuesta.

El análisis de la pregunta es fundamental. En este sentido, continuando con [4] nos concentramos, en una primera etapa, en la predicción de la respuesta esperada a partir de la pregunta de entrada. En este sentido, se ha realizado un análisis

[1] https://smart-task.github.io/

de las diferentes metodologías y estudio de herramientas disponibles para realizar una implementación eficiente.

Dentro de la tarea de QA, focalizada en la predicción del tipo de respuesta, existen distintas competencias, una de ellas es el desafío denominado SMART *SeMantic Answer Type and Relation Prediction Task*, de la cual se han realizado hasta el momento dos instancias de tales competencias en los años 2020 [2] y 2021, y una última que se encuentra en proceso de este año.

Para el desafío, es posible realizar una clasificación del tipo de respuesta granular con ontologías de Web Semántica populares como DBpedia (∼760 clases) y Wikidata (∼50K clases). En esta competencia se cuenta con dos tareas principales e independientes: 1) predicción del tipo de respuesta y 2) predicción de un conjunto de relaciones usadas para la identificación de la respuesta correcta.

La primer tarea, predicción del tipo de respuesta, consiste en la predicción de la "categoría" (*resource, literal* y *boolean*) y la predicción del "tipo de la respuesta". En el caso de que la respuesta sea *resource*, el tipo de la respuesta son clases de ontologías. Si es literal, entonces el tipo de la respuesta puede ser un número, una fecha o una cadena. Finalmente, si es *boolean*, el tipo de la respuesta es siempre *boolean*.

La tarea de predicción de relaciones para la pregunta es una tarea difícil: algunas relaciones están alejadas semánticamente, a veces los tokens que deciden las relaciones están distribuidas a lo largo de la pregunta, algunas relaciones están implícitas en el texto, entre otras.

Se implementó un módulo para la clasificación del tipo de respuesta utilizando aprendizaje automático. Para lograr ésto, la competencia dispone de varios corpus que se emplearon para el entrenamiento y testeo de la clasificación de la categoría y el tipo de respuesta para cada una de las diferentes ontologías que proponen. Se llevó a cabo el entrenamiento y testeo de diversos modelos de aprendizaje automático, entre ellos se encuentran *Support-Vector Machine, Logistic Regression, Naive Bayes y Decision Tree*. A partir de la precisión de cada uno de ellos, se efectuó una comparación para hallar el mejor modelo.

Asimismo, se pretende diseñar y desarrollar un módulo para la segunda tarea de predicción de relaciones usando tanto la ontología de DBpedia como la de Wikidata. Al igual que en la primera tarea, se provee de corpus para trabajar.

2. Nuestra Propuesta

2.1. Preprocesamiento de los Datos

Antes de la construcción de los modelos de aprendizaje automático, se llevó a cabo un preprocesamiento de los datos, donde se realizó una limpieza y transformación de los mismos para un formato deseado.

Tanto *datasets* de entrenamiento como los de testeo son provistos por la competencia SMART, ambos en formato JSON. Ésto permitió cargar los mismos utilizando una representación tabular, lo que nos proporciona una fácil manipulación de los datos.

Para lograr una construcción robusta de los modelos de aprendizaje, se eliminaron aquellas filas que no presentaban valores o que contenían valores inválidos. A su vez, se eliminaron de las preguntas aquellas palabras que no eran alfanuméricas. Posteriormente, se realizó una tokenización de las preguntas, donde se dividió cada pregunta en partes más pequeñas llamadas tokens (cada palabra de la misma). Estos tokens se utilizaron para realizar una normalización del texto, en la cual se efectuó el *stemming* y lematización. El *stemming* es el proceso de reducir la inflexión de las palabras a sus formas raíz, incluso si la propia raíz no es una palabra válida en la lengua. Mientras que la lematización, reduce las palabras inflexionadas asegurándose de que la palabra raíz pertenece a la lengua, esta raíz se llama lema y es la forma canónica de un conjunto de palabras.

Inicialmente, en este estudio se buscaba darle un trato especial tanto a las partículas interrogativas de las preguntas como a las stopwords, por lo que se decidió adaptar los *datasets* para lograr ello. Sin embargo, estos nuevos *datasets* funcionaban con menos eficacia. Por lo tanto, se optó por dejar las partículas interrogativas y las *stopwords* tal cual provienen del *dataset* original.

2.2. Selección de Características

Efectuar una buena selección de características para los modelos de aprendizaje, aporta ciertos beneficios al proceso de aprendizaje, pues reduce la dimensionalidad, eliminación de ruido, entre otros. Todo esto permite mejorar la velocidad de cómputo a la hora de entrenar y evaluar el clasificador, y hasta evitar el *overfitting.*

Para lograr independizar el modelo a utilizar con respecto a los recursos lingüísticos, se analizó la frecuencia de términos (TF) y la frecuencia de términos inversa (TF-IDF) en la extracción de características utilizando CountVectorizer y TFIDFVectorizer. La evidencia empírica concluye que aplicar TFIDFVectorizer junto con unigramas y bigramas es la opción más adecuada para esta tarea.

El *target* de esta tarea, la categoría, se encuentra en forma de cadena en los *datasets* originales. Sin embargo, se determinó transformar en una etiqueta numérica de tal manera que permita a los modelos trabajar correctamente.

2.3. Diseño de los Modelos de Aprendizaje

Para comenzar con los experimentos iniciales, se optó por seguir un enfoque en dos fases. En la primer fase, se realizó la clasificación de las categorías de las preguntas. Posteriormente, para la segunda fase, se pretende diseñar un modulo para la predicción de tipos de las preguntas para aquellas para las que se predijo que la categoría era recurso o literal.

Para la clasificación de categorías se realizaron pruebas sobre 4 tipos de clasificadores: *Naive Bayes* (NB), *Support-Vector Machine* (SVM), *Decision Tree* (DT) y *Logistic Regression* (LR), pues eran los más utilizados por los participantes de la competencia [1] [5] [3]. Para los modelos se utilizó el *dataset* provisto por SMART tanto para el año 2020 como el 2021, efectuándose pruebas con los mismos por separado y combinándolos. En cada prueba se hacían cambios en la

Tabla 1: Dataset provisto por SMART para el año 2020

Dataset	**Preguntas**		**Categorías**		
	Train	**Test**	**Boolean**	**Literal**	**Resource**
DBpedia	17,571	6,883	2,799	5,188	9,584
Wikidata	18,251	4,571	2,139	4,429	11,683

Tabla 2: Dataset provisto por SMART para el año 2021

Dataset	**Preguntas**		**Categorías**		
	Train	**Test**	**Boolean**	**Literal**	**Resource**
DBpedia	29,336	7,334	1,794	3,363	24,179
Wikidata	34,843	8,711	1,693	3,663	29,487

Tabla 3: Dataset combinado 2020-2021

Dataset	**Preguntas**		**Categorías**		
	Train	**Test**	**Boolean**	**Literal**	**Resource**
DBpedia	39,512	4,381	2,525	5,147	31,840
Wikidata	44,688	4,571	2,556	5,303	38,855

parametrización de cada uno de los modelos, tanto de las características utilizadas para el entrenamiento como de los parámetros recibidos. A través de estas pruebas, se logró obtener un mejor modelo, en el cual se utilizaron unigramas y bigramas de palabras ponderados por TFIDF como características para entrenar un clasificador SVM con un kernel lineal y sus parámetros por defecto.

3. Resultados Experimentales

3.1. *Datasets*

La Tabla 1 y la Tabla 2 presentan las estadísticas descriptivas para los dos conjuntos de datos, DBpedia y Wikidata, provistos por la competencia SMART para los años 2020 y 2021 respectivamente. Wikidata tiene un poco más de recursos que tipos de respuestas literales, en comparación con DBpedia. A su vez, la Tabla 3 nos presenta el resultado de combinar los *datasets* de DBpedia y Wikidata para los años 2020 y 2021.

A continuación se expondrán los resultados obtenidos en distintas pruebas realizadas. Particularmente, la tarea de clasificación de la categoría fue evaluada en términos de la precisión de la clasificación.

La Tabla 4 presenta los resultados obtenidos a partir de la experimentación realizada al *dataset* por separado correspondiente a los años 2020 y 2021 respectivamente. Como se puede apreciar la mejor precisión es obtenida para ambos años es con el método de SVM tanto para del *dataset* de DBPedia como para Wikidata. Si analizamos los resultados de la del *dataset* combinado de ambos años la misma tendencia se mantiene para el caso de DBPedia, pero no ocurre lo mismo para el caso de Wikidata donde la mejor precisión se obtuvo para el método DT.

Tabla 4: Resultados *dataset* 2020, 2021 y combinando ambos años

2020			**2021**			**2020-2021**		
Dataset	**Método**	**Train**	**Dataset**	**Método**	**Train**	**Dataset**	**Método**	**Train**
DBpedia	LR	0.931	DBpedia	LR	0.958	DBpedia	LR	0.906
	SVM	**0.945**		**SVM**	**0.972**		**SVM**	**0.939**
	NB	0.889		NB	0.938		NB	0.850
	DT	0.913		DT	0.961		DT	0.911
Wikidata	LR	0.922	Wikidata	LR	0.956	Wikidata	LR	0.925
	SVM	**0.936**		**SVM**	**0.968**		SVM	0.963
	NB	0.885		NB	0.943		NB	0.902
	DT	0.914		DT	0.960		**DT**	**0.972**

4. Conclusiones

A partir de los datos de DBpedia y Wikidata provistos por SMART, se llevaron a cabo varias pruebas sobre la normalización del texto, desde darle un trato especial a las partículas interrogativas hasta remover las stopwords de las preguntas. De la misma manera, se realizaron pruebas sobre distintos modelos para encontrar el más adecuado para el problema de clasificación.

En otros estudios [5] [3] realizados por participantes del desafío SMART, se utilizaron los modelos BERT ajustados. Sin embargo, debido a las limitaciones en los recursos informáticos, se utilizaron los modelos mencionados para tener un coste computacional barato durante el desarrollo. Esto abre la puerta a futuras mejoras, en la cual se podría utilizar redes neuronales o enfoques híbridos para lograr una mayor eficacia, así como intentar aumentar los *datasets* de entrenamiento para lograr un balanceo en la cantidad de categorías existentes de cada tipo.

Asimismo, se pretende diseñar y desarrollar un prototipo para la tarea de predicción de relaciones usando ambas ontologías, y con ello finalizar las principales tareas del desafío.

Referencias

1. C. Kim and E. Jimenez-Ruiz. CitySAT: a System for the Semantic Answer Type Prediction Task. In *CEUR Workshop Proceedings*, volume 3119, pages 77–88. CEUR, 2022.
2. N. Mihindukulasooriya, M. Dubey, A. Gliozzo, J. Lehmann, A.-C. N. Ngomo, and R. Usbeck. SeMantic AnsweR Type prediction task (SMART) at ISWC 2020 Semantic Web Challenge. *CoRR/arXiv*, abs/2012.00555, 2020.
3. C. Nikas, P. Fafalios, and Y. Tzitzikas. Two-stage Semantic Answer Type Prediction for Question Answering using BERT and Class-Specificity Rewarding. In *SMART@ ISWC*, pages 19–28, 2020.
4. M. Oyarzun and S. Roger. Tecnologías de sistemas de QA aplicadas a la Web Semántica. In *XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021, Chilecito, La Rioja)*, 2021.
5. V. Setty and K. Balog. Semantic Answer Type Prediction using BERT: IAI at the ISWC SMART Task 2020. *arXiv preprint arXiv:2109.06714*, 2021.

Videojuegos: del Ocio a Herramientas de Enseñanza

José Ernesto Bocci Cordón, Miguel Mendez Garabetti y Pablo Javier Sández

Laboratorio de Investigación en Ciencia y Tecnología, Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua, Mendoza, Argentina
joseebocci@gmail.com

Resumen Estudiar es algo que conlleva un esfuerzo mental y por eso muchos niños lo encuentran demandante y tienen poca motivación para hacerlo. Del mismo modo, hacer ejercicio tiene la carga de un esfuerzo físico, por lo que también ciertas personas lo encuentran poco disfrutable. Debido a esto surgió la idea de integrar dos áreas que parecen no tener relación pero que día a día crecen más hermanadas: los videojuegos y la educación/ejercitación. Desde los años 2000 se plantean teorías e inclusos estudios donde se demuestra que bajo condiciones de enseñanza o ejercitación estimuladas por un videojuego se puede llegar a que los participantes de la actividad tengan más disfrute en la misma y además prefieran seguir o repetir dicha actividad en el tiempo. Hoy en día hay muchos ejemplos de videojuegos aplicados a la enseñanza debido al auge que ha cobrado la programación. Expondré casos de videojuegos representativos de la enseñanza.

Keywords: Videojuegos · Enseñanza · Aprendizaje · Método de estímulo · Programación.

1. Motivación

Mi motivación para el desarrollo de este trabajo proviene de que considero que por mucho tiempo se ha idealizado a los videojuegos como algo ajeno a la enseñanza y dedicado exclusivamente a niños. Presento distintos trabajos de investigación para demostrar que hoy en día esto no es así y que los videojuegos pueden ser y son aprovechados con beneficios, por ejemplo en el área de la educación.
Este estudio busca demostrar como se debería integrar en la enseñanza los videojuegos y lo que ello conlleva.
Este trabajo es la base para desarrollar mi tesis de último año.

2. Introducción

La tecnología sigue avanzando sin esperar a nadie que se acostumbre, día a día hay mayores avances en todas las áreas, sin embargo, el área que más

importante es para la sociedad lleva tiempo sin reinventarse.
La educación, en los distintos niveles, sigue siendo igual que hace muchos años. Hoy los alumnos tienen textos más actuales y cuadernos más modernos pero el método de un profesor dando la clase y ellos escuchando sigue siendo la norma en la mayoría de instituciones.
En el año 2005 se puede apreciar los avances que harán que la enseñanza sea más disfrutable para los alumnos. Y en el año 2011 se forma la idea de utilizarlo como herramienta evaluativa [6].
Hay ventajas en la forma en que un juego recompensa al jugador para que el mismo se mantenga jugando, este estímulo hace que la persona no sienta una carga al realizar la actividad y que quiera seguir para mejorar su puntuación y seguir recibiendo dicho estímulo.
No solo a niños les serviría este estímulo, sino también a adultos mayores [4].

3. Videojuegos

Hoy el mercado de los videojuegos ha logrado ganar un lugar fuerte en la industria. Y esto se debe a como ha ido aumentando su alcance.
Los videojuegos ya no son algo exclusivo de niños, hay adultos e incluso adultos mayores que juegan videojuegos por la diversión o satisfacción que estos les producen [12].
Los smartphones también se han vuelto algo normal en nuestras vidas, esto facilita que aquellos que no tengan el tiempo para pensar en jugar algo más demandante busquen jugar algo rápido en su celular [3].
Todo esto ha creado una sociedad que convive con los videojuegos, quizás un padre/madre no lo haga directamente, pero ve a su hijo/a jugar y termina conociendo y hasta participando de la actividad.
Si ya están tan presentes en nuestras vidas, ¿por qué no se ven tanto en el ámbtio de la enseñanza?
Los videojuegos enseñan distintas habilidades como la toma de decisiones, el pensamiento crítico, la coordinación espacial, coordinación mano-ojo [9]. Facilitan el entregar un mensaje ya que se puede hacer una historia con personajes y así contarla de forma más amena [7].
Logran desviar la atención de la tarea como un esfuerzo mental o físico y obtienen los mismos resultados, sino mejores, que las tareas realizadas de forma no "gamificada"[1] [8]. Con estos beneficios parece obvio que se debería integrar en todas las áreas que puedan hacer uso de estas habilidades.
La realidad es que lograr que un juego sea práctico para la enseñanza es complicado, debe abordar temas a explicar, no debe ser demasiado complejo ni demasiado sencillo, debe ser agradable de jugar sino pierde todo el estímulo que se busca.
Entonces no siempre se puede considerar que el videojuego se puede usar en la enseñanza.
A continuación, muestro algunos ejemplos de juegos que si cumplen con lo antes dicho gracias al enfoque de este trabajo:
Wii Sports. Es un juego de deportes enfocado a la movilidad con distintos mi-

nijuegos en los cuales se puede jugar en grupo o solitario. Permite desarrollar coordinación espacial, además de proveer un refuerzo positivo al hacer ejercicio, desviando la atención de la actividad física y convirtiéndola en un desafío contra un oponente.
Nintendo Switch Sports. Similar al juego anterior pero más actual.
Scratch. No es un videojuego, es un programa enfocado a enseñar programación jugando. En el mismo se tiene que hay bloques de programación con diferentes acciones. Los bloques se conectan entre sí para dar una secuencia de órdenes que luego son ejecutadas y se muestra el resultado final en una ventana donde el personaje del programa (un gato) realiza la secuencia.
MOBI. Este juego sale mencionado en 'Entertainment Computing' [10], el mismo se basa en un modelo llamado Lean UX. Lamentablemente hoy no se encuentra disponible, pero lo que proponía era que MOBI, personaje principal, debía ir resolviendo diferentes problemas para ayudar a un amigo. La resolución de estos problemas lleva a la creación de un videojuego más serio. Utiliza la propuesta de "gamificación"para poder enseñar programación a niños.
En "Las posibilidades educativas de los videojuegos. Una revisión de los estudios más significativos"[9] se mencionan y describen: PC Futbol, The Machine Incredible, Los Lemmings, Carmen Sandiego, Simon the Sorcerer, La Pantera Rosa, Indiana Jones y el Destino de la Atlántida, Civilization II. Todos enfocados a la enseñanza, en el informe además se detalla como otras áreas pueden aprovechar los videojuegos.
También se utilizan para demostrar fenómenos físicos y poder sustituir a una práctica con elementos reales pudiendo tener un ambiente simulado y controlado, como pasa en los artículos "Potencialidad de los videojuegos en el aprendizaje de Física"[5] y 'El uso de videojuegos en un laboratorio de Física' [11]. Tal es la importancia que han llegado a obtener los videojuegos que en el 2012 la autora Laura Vadillo planteaba como se deben moldear las futuras redes para la aceptación de la inserción de los videojuegos en la sociedad [2].

4. Aplicación y enseñanza

En un estudio realizado en 2018 [13] se observa como la línea de pensamiento no debe ser siempre negativa hacia los videojuegos, y como su uso tiene cualidades positivas que impactan en los jugadores que los practican.
Para llegar a obtener estos beneficios se puede usar un modelo "gamificado"sin que sea adictivo. Esto es debido a la naturaleza propia de los videojuegos que realizan una interacción estímulo-recompensa que puede llegar a ser perjuidicial.
Otra opción es utilizar los juegos preexistentes antes nombrados y los otros tantos sin nombrar.
También es posible hacer uso de una combinación de ambos, organizar partidas de algún juego y proveer una retroalimentación por fuera del mismo. Un ejemplo de esto sería utilizar una partida del juego Çivilization IV.[el] cual consiste en elegir uno de los distintos imperios que se ofrecen y crearlo desde los comienzos, se avanza por las distintas edades completando misiones y así se va obteniendo

mejor equipamiento en el juego. Luego de la partida preguntar el orden de ciertos acontecimientos, edades aproximadas de cambio de épocas, que puntos fuertes traía cada época para la civilización elegida, etc.
De este modo se tiene que los alumnos al momento de jugar tienen la posibilidad de aprender de manera más distendida el cómo se fue produciendo el avance de las distintas civilizaciones.
No es necesario que sea un juego complejo, el juego "Kahoot.[es] un ejemplo de esto, en el se utiliza una planilla de preguntas sobre un tema específico. Al momento de jugar aparecen en pantalla 1 pregunta y entre 2 a 4 respuestas posibles. Hay un tiempo para responder cada pregunta, mientras más rápido se conteste correctamente más puntos se ganan. Esto crea un ambiente de competitividad en el que se pueden beneficiar los alumnos ya que intentan superarse entre sí.
Además de la enseñanza clásica en un aula o curso también son útiles para una persona autodidacta.
"Flexbox Froggy.[es] un juego online el cual consiste en ubicar una rana sobre un nenúfar. Esta colocación se realiza mediante código CSS. Si una persona está aprendiendo diseño web y desea una forma más sencilla que solo leyendo documentación, de esta manera puede jugar y obtener los conocimientos de qué hará su código cuando lo implemente de manera real en un proyecto.

5. Desventajas

Hasta ahora he hablado de cómo los videojuegos pueden ayudar y enriquecer la educación, pero no he mencionado los aspectos negativos que pueden llegar a tomar.
Uno de los principales es el mismo Método de Estímulo-Recompensa. Puede suceder que si el niño es demasiado inmaduro se acostumbre a recibir una recompensa por cada "tarea"bien realizada, y luego cuando crezca se dará cuenta de que no todo lleva una recompensa directa o material. Esto puede producir problemas en su concepto preestablecido de qué es una recompensa y cómo la obtiene.
Otro problema posible es la generación de .ªdicción.ª los videojuegos. En el afán de seguir recibiendo la estimulación proveniente de la enseñanza "gamificada"puede suceder que el jugador entienda que todo se puede aprender jugando y busque jugar constantemente sin importar si el juego le aporta algo beneficioso en si.
Todas estas desventajas están llevadas al extremo para explicar el punto, quizás no sucedan, pero la posibilidad de que lo hagan no es nula.

6. Conclusión

En definitiva opino que los videojuegos pueden ser una poderosa herramienta para la enseñanza de un abanico de materias. Esta enseñanza está principalmente orientada a niveles primarios, aunque no excluye niveles secundarios ni universitarios. La mayoría de situaciones de motivación para estudiar pueden resolverse

planteando un sistema que recompense al alumno por hacer algo que le gusta y también mostrar que no es necesario quedarse con una sola impresión. Ciertos juegos pueden ser rejugados infinidad de veces y cada vez enseñar o aprender algo distinto del mismo.
Creo firmemente que con un modelo organizado y divido con los programas de diferentes materias se puede enseñar varios tópicos exclusivamente mediante videojuegos y sus referencias.
Todos los días surgen nuevas ideas y más proyectos orientados a enseñar a futuras generaciones de diversas formas, estas deben ser apoyadas y revisadas para poder lograr que aumente el nivel educativo que se puede alcanzar mediante los videojuegos.

Referencias

1. New study recommends using active videogaming ("exergaming") to improve children's health, https://www.elsevier.com/about/press-releases/archive/research-and-journals/new-study-recommends-using-active-videogaming-exergaming-to-improve-childrens-health
2. Videojuegos - las infraestructuras de telecomunicaciones en el futuro de los videojuegos — coit — colegio oficial ingenieros de telecomunicación, https://www.coit.es/archivo-bit/septiembre-2012/videojuegos-las-infraestructuras-de-telecomunicaciones-en-el-futuro-de
3. Videojuegos y aprendizaje - begoña gros salvat, alejandro català bolós, carles feixa pampols, javier jaén martínez, pilar lacasa díaz, m. luisa lamazán Álvarez, rut martínez borda, laura méndez zaballos, jose antonio mocholí agües, isidro moreno sánchez, xavier vilella i miró, antònia bernat cuello, manel camas magri, juan josé cárdenas ballestero - google libros
4. Bock, B.C., Dunsiger, S.I., Ciccolo, J.T., Serber, E.R., Wu, W.C., Tilkemeier, P., Walaska, K.A., Marcus, B.H.: Exercise videogames, physical activity, and health: Wii heart fitness: A randomized clinical trial. American Journal of Preventive Medicine **56**, 501–511 (4 2019). https://doi.org/10.1016/J.AMEPRE.2018.11.026
5. Bouciguez, M.J., Santos, G., Guerrero, M.J.A.: Potencialidad de los videojuegos en el aprendizaje de física (2013), http://sedici.unlp.edu.ar/handle/10915/74437
6. Esnaola, G., Yuste, R., de Ansó, M.B., Borrero, R.: Videojuegos en aula: una herramienta de evaluación educativa (5 2011), http://sedici.unlp.edu.ar/handle/10915/26540
7. Irigaray, M.V., Luna, M.D.R.: Cine y video en el aula: La enseñanza de la historia a través de videojuegos de estrategia. dos experiencias áulicas en la escuela secundaria. Clío Asociados. La historia enseñada pp. 411–437 (5 2015). https://doi.org/10.14409/CYA.V0I18/19.4758
8. Núñez-Barriopedro, E., Sanz-Gómez, Y., Ravina-Ripoll, R., Núñez-Barriopedro, E., Sanz-Gómez, Y., Ravina-Ripoll, R.: Los videojuegos en la educación: Beneficios y perjuicios. Revista Electrónica Educare **24**, 240–257 (8 2020). https://doi.org/10.15359/REE.24-2.12
9. Pindado, J.: Las posiblidades educativas de los videojuegos. una revisión de los estudios más significativos pp. 55–67 (2005), https://idus.us.es/handle/11441/45601
10. Ramos-Vega, M.C., Palma-Morales, V.M., Pérez-Marín, D., Moguerza, J.M.: Stimulating children's engagement with an educational serious videogame

using lean ux co-design. Entertainment Computing **38**, 100405 (5 2021). https://doi.org/10.1016/J.ENTCOM.2021.100405
11. Sagastume, J.I.G., Devece, E., Torroba, P.L., Videla, F.A.: El uso de videojuegos en un laboratorio de física (2013), http://sedici.unlp.edu.ar/handle/10915/37778
12. Salvat, B.G.: Certezas e interrogantes acerca del uso de los videojuegos para el aprendizaje. Nº **7**, 251–264 (2009), https://idus.us.es/handle/11441/58304
13. Vaamonde, A.G.N., Toribio, M.J., Molero, B.T., Suárez, A.: Cognitive, psychological, and personal benefits of the use of video games and e-sports: a review **3**, 1–14 (2018), www.revistapsicologiaaplicadadeporteyejercicio.org

Generación de comentarios a partir de código fuente utilizando Transformers

Cristian Vincenzini and Sandra Roger

Grupo de Investigación en Lenguajes e Inteligencia Artificial (GILIA), Facultad de Informática. Universidad Nacional del Comahue. Neuquén
cristian.vincenzini@est.fi.uncoma.edu.ar, roger@fi.uncoma.edu.ar

Resumen En este trabajo se utilizará un modelo de generación de lenguaje natural (GLN) para crear comentarios a partir de código fuente. Para esto, se aprovechará la técnica de transferencia de conocimiento en un modelo basado en mecanismo de atención, utilizando pequeños conjuntos de datos de entrenamiento sobre grandes modelos ya entrenados y posteriormente se analizarán los resultados obtenidos.

Keywords: Generación del Lenguaje Natural, Aprendizaje Automático, Código, Comentarios

Contexto Este trabajo se desarrolla en el marco de la tesis final de la carrera Licenciatura en Ciencias de la Computación.

1. Introducción

Los programadores utilizan los comentarios con diversos fines: para especificar los requerimientos del software que desarrollan, comunicarse con otros desarrolladores, informar tareas que faltan realizar, pero fundamentalmente se utilizan para describir la funcionalidad de algún fragmento de código. Este tipo de comentarios en particular, contienen una gran cantidad de información que puede aprovecharse para mejorar el mantenimiento, la fiabilidad del software y para facilitar la comprensión del programa [1,8]. La generación de lenguaje natural (GLN) a partir de código atañe a proveer de manera automatizada una descripción resumida de un fragmento de código. Los modelos de GLN basados en el mecanismo de atención -conocidos como Transformers- se presentan aún como un campo poco explorado. Hasta la fecha, distintas arquitecturas entrenadas con millones de datos para diferentes tareas se hicieron disponibles para uso público. Una de las características de estas arquitecturas es que permiten refinar su aprendizaje para realizar una tarea específica, mecanismo que se conoce como *transferencia de aprendizaje*. Mediante esta técnica es posible utilizar estos grandes modelos ya entrenados y especializarlos, obteniendo resultados aceptables en un tiempo considerablemente inferior al de entrenar un modelo desde cero.

Para este trabajo utilizaremos un transformer conocido como T5 (2020)[7]. Este modelo permite procesar múltiples tareas relacionadas a la GLN. La tarea a desarrollar para este trabajo será la generación de comentarios a partir de código fuente. Para ello tomaremos un modelo ya entrenado que será refinado sobre un conjunto de datos propio y finalmente analizaremos los resultados obtenidos.

2. Trabajos Relacionados

En los últimos años han surgido varias arquitecturas basadas en el mecanismo de atención, y utilizadas para diversas tareas en el campo de la ingeniería del software. Tenemos por ejemplo CodeBERT[3], un modelo de propósito general que permite, entre otras tareas, buscar código utilizando lenguaje natural, generar documentación de código fuente, etc. Otro ejemplo es TransCoder[4], un transpilador que permite la traducción de un código fuente a otro entre diversos lenguajes de programación. Estos trabajos han alcanzado el estado del arte.

3. Nuestra Propuesta

El modelo de lenguaje utilizado se basa en la arquitectura de transformadores de codificación-decodificación de CodeTrans [2]. En dicho trabajo se utilizaron tres tamaños del modelo T5 para realizar los entrenamientos: *small*, *base* y *large* (60, 220 y 770 millones de parámetros respectivamente). Estos modelos fueron entrenados para diferentes tareas en varios lenguajes de programación utilizando transferencia de aprendizaje. La transferencia de aprendizaje consiste en dos etapas. Una etapa inicial de entrenamiento auto-supervisado donde se utilizan datos sin etiquetar y una segunda etapa, conocida como *fine-tuning* donde el modelo se entrena para una tarea específica utilizando datos etiquetados. Otra técnica utilizada en CodeTrans es el entrenamiento multi-tarea. Esta estrategia consiste en entrenar un modelo en mútiples tareas, utilizando datos etiquetados y sin etiquetar. Esta metodología permite, además, realizar un *fine-tuning* posterior a través de la transferencia de aprendizaje.

En este trabajo tomamos la arquitectura T5 ya entrenada con los datos de CodeTrans y realizamos *fine-tuning* sobre cada uno de ellos, utilizando sets de datos propios para realizar diversos experimentos. En particular, tomamos los modelos realizados por transferencia de aprendizaje (TF) y multi-tarea (MT) para los tres tamaños considerados: small, base y large.

4. Configuración de la Experimentación

Utilizamos dos conjuntos de datos para desarrollar los experimentos. Todos contienen 120 líneas de entre las cuales se seleccionaron de forma aleatoria 20 líneas para formar un conjunto de datos de test, el resto se utilizó para el entrenamiento del nuevo modelo. Los datos consisten en tuplas que referencian pares de funciones -en algún lenguaje de programación- y el comentario en inglés que describe la funcionalidad de la misma.

Seleccionamos a GO como primer *set* de datos. GO es un lenguaje de programación imperativo[1], desarrollado por Google en el año 2009. Las funciones y sus respectivos comentarios fueron extraídos en su mayor parte utilizando GitHub

[1] `https://go.dev/`

Tabla 1: Resultados de la evaluación de nuestro *dataset test* en todas las tareas para el lenguaje GO. Se utilizó BLEU-4 tanto para evaluar los modelos de CodeTrans[2] como los propios.

	Nuestra salida	Code-Trans
go-tf-s	**3,53**	2,75
go-tf-base	**13,96**	11,34
go-tf-large	**20**	12,55
go-mt-tf-s	12,77	16,44
go-mt-tf-base	11,19	11,69
go-mt-tf-large	**17,56**	9,95

Copilot [2]. El siguiente *set* de datos seleccionado es PROLOG. PROLOG es un lenguaje lógico que utiliza predicados como elementos de ejecución. Los datos fueron tomados de diversas fuentes. Durante la selección, se evitó utilizar predicados provenientes de módulos o paquetes de terceros, también se evitó usar sintaxis de gramáticas de cláusulas definidas.

5. Resultados y discusión

Se llevaron a cabo experimentos para evaluar las tareas de *Transfer Learning* (TF) y de *Multi-Task Learning* con *Fine-Tune* (MT-TF), debido a que en ambas tareas se emplea la utilización de un *Fine-Tune*. En cada tarea, se realizaron experimentos para los modelos de tres tamaños distintos: pequeño, base y grande (*S*, *B* y *L*, respectivamente). Para evaluar estas tareas se utilizó la medida BLEU-4 [6]. Los experimentos relacionados al lenguaje GO se realizaron para comparar cómo se comportaba nuestro *finetune*, construido con el corpus descripto en 4 con respecto a los modelos de CodeTrans de [2] utilizando nuestro *dataset-test*. La Tabla 1 muestra los resultados. El desarrollo del *Fine-tune* sobre nuestro corpus brindó buenos resultados en la tarea de TF y solo en el mt-tf-large se logró una mejora.

Por otro lado, se hicieron evaluaciones de estas mismas tareas en los modelos pequeños y bases, como así también los modelos grandes en el lenguaje PROLOG. Cabe destacar que CodeTrans no cuenta con este lenguaje. Al igual que con el lenguaje GO, se construyó un corpus tanto para realizar el *fine-tune* como el *dataset* utilizado en el testeo. El objetivo fue probar cómo se comportaban los modelos para un lenguaje no contemplado previamente. Además, elegimos PROLOG por ser un lenguaje de programación declarativo, a diferencia de los otros lenguajes utilizados. En esta oportunidad se utilizaron dos métricas: el BLEU y el ROUGE [6]. Dentro de las variantes de la medida ROUGE nos concentramos en la ROUGE-L, con la ventaja de que puede capturar la estructura del nivel de oración de una manera natural.[5]

La Tabla 2 muestra la evaluación realizada para el lenguaje de programación PROLOG en las dos tareas y modelos (*S, B, L*) en los cuales se hicieron, también, las evaluaciones de GO. Como se puede apreciar los valores del BLEU-4 son de un dígito en su totalidad. Siendo el mejor resultado para el *Transfer Learning* con un modelo *Large*. Por el contrario, las medidas ROUGE-L arrojan valores

[2] `https://github.com/features/copilot`

Tabla 2: Resultados de la evaluación de todas las tareas para el lenguaje de programación PROLOG.

MEDIDA	PRO-TF-S	PRO-TF-B	PRO-TF-L
ROUGE-L	24.4537	28.2623	**31.4393**
BLEU-1	18.79	20.56	20.56
BLEU-2	10.71	12.21	13.43
BLEU-3	5.74	6.27	9.07
BLEU-4	0	0	**6.05**
	PRO-MT-TF-S	PRO-MT-TF-B	PRO-MT-TF-L
ROUGE-L	27.5264	29.0601	28.5726
BLEU-1	20.56	19.85	17.02
BLEU-2	13.43	11.6	9.53
BLEU-3	9.07	7.33	5.72
BLEU-4	5.1	4.34	0

Tabla 3: Ejemplo de una pregunta del test. La referencia humana fue tomada del recurso donde se extrajo el programa.

Programa	`circle(X, Y, R):- number(X), number(Y), number(R), R >0.`
Referencia Humana	succeeds if the item represents a valid circle x,y represents the centrepoint of a circle on an x,y plane r represents the radius of the circle.
PRO-TF-BASE	computes a circle from the coordinates.
BLEU-4 y BLEU-3	0.0
BLEU-2	6.79
BLEU-1	11.54

Tabla 4: Ejemplo de una pregunta del test. La referencia humana fue tomada del recurso donde se extrajo el programa.

Programa	`reverse([], Z, Z).` `reverse([H|T], Z, Acc) :- reverse(T, Z, [H|Acc]).`				
Referencia Humana	reverses a list of any length.	BLEU-1	BLEU-2	BLEU-3	BLEU-4
pro-tf-s	reverses the array of elements in the array z.	20.22	0.0	0.0	0.0
pro-tf-b	reverses the order of the elements in the list.	38.46	17.9	0.0	0.0
pro-tf-l	reverses an array.	16.67	0.0	0.0	0.0
pro-mt-tf-s	reverses the reverse of the given array.	28.22	0.0	0.0	0.0
pro-mt-tf-b	reverses a list of numbers.	66.67	63.25	58.48	50.81
pro-mt-tf-l	reverses the order of the elements in the array.	20.22	0.0	0.0	0.0

más interesante. Igualmente el mejor resultado sigue siendo para el *TF* de tamaño *Large*. Estos valores podrían deberse a las características propias de los comentarios para este tipo de lenguajes. Como lenguaje declarativo, podría pensarse que los comentarios tienen cierta semi-estructuración donde se describen, muchas veces, las características de los parámetros que intervienen haciendo uso de los nombres de las variables y de los predicados (Tabla 3). Además, la longitud de los comentarios son relativamente más cortos. De igual manera, como se puede apreciar en dicho ejemplo, las métricas podrían no ser del todo adecuadas. La Tabla 4 muestra la salida de las dos tareas y de los modelos utilizando

distintos tamaños. También se observa su puntuación BLEU-n en relación a su referencia humana y su programa asociado.

6. Conclusiones y trabajos futuros

Se estudiaron diversas arquitecturas en la tarea de generación de comentarios a partir de código fuente. Estudiamos más a fondo la arquitectura de codificación-decodificación de CodeTrans. Se construyeron dos corpus, uno para el lenguaje GO y otro para el lenguaje PROLOG. La elección del primero radica en la selección de un lenguaje ya presente en la arquitectura antes mencionada. En cuanto al segundo, se consideró utilizar un lenguaje basado en otro paradigma de programación y que no estuviera presente. La experimentación relacionada al lenguaje GO fue satisfactoria para los modelos estudiados, obteniendo resultados significativos. En cuanto al lenguaje PROLOG, los resultados considerando la medida BLEU no fueron del todo satisfactorios. Como mencionamos en la Sección 5 hay muchos factores a considerar. Por ejemplo, las características de los comentarios, o bien la elección adecuada de la medida de evaluación para este tipo de lenguaje, entre otros. Sobre estos temas se seguirá trabajando.

Referencias

1. K. K. Aggarwal, Y. Singh, and J. K. Chhabra. An integrated measure of software maintainability. In *Annual Reliability and Maintainability Symposium. 2002 Proceedings (Cat. No. 02CH37318)*, pages 235–241. IEEE, 2002.
2. A. Elnaggar, W. Ding, L. Jones, T. Gibbs, T. Feher, C. Angerer, S. Severini, F. Matthes, and B. Rost. Codetrans: Towards cracking the language of silicon's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2104.02443*, 2021.
3. Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, M. Gong, L. Shou, B. Qin, T. Liu, D. Jiang, et al. Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155*, 2020.
4. M.-A. Lachaux, B. Roziere, L. Chanussot, and G. Lample. Unsupervised translation of programming languages. *arXiv preprint arXiv:2006.03511*, 2020.
5. C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
6. C.-Y. Lin and F. J. Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, 2004.
7. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
8. L. Tan, D. Yuan, G. Krishna, and Y. Zhou. icomment: Bugs or bad comments? In *Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles*, pages 145–158, 2007.

Caracterización de Variables para el Análisis del Índice de Vegetación*

Carolina Villegas[1], Agustina Buccella[1][0000−0002−8516−7453], Alejandra Cechich[1][0000−0003−4804−6270], and Ayelén Montenegro[2]

[1] GIISCO Research Group
Departamento de Ingeniería de Sistemas - Facultad de Informática
Universidad Nacional del Comahue
Neuquen, Argentina
{carolina.villegas,agustina.buccella,alejandra.cechich}@fi.uncoma.edu.ar
[2] Instituto Nacional de Tecnología Agropecuaria (INTA)
Alto Valle de Río Negro y Neuquén
montenegro.ayelen@inta.gob.ar

Abstract. Los procesos de análisis de datos nos permiten encontrar información o patrones que no son simplemente visibles. Esta información descubierta es sumamente importante para la toma de decisiones de las organizaciones. En particular, en el dominio de la fruticultura es importante conocer factores que produzcan variaciones en los índices de vegetación. Así, en este trabajo caracterizamos estos factores o variables, para luego comprobar algunas de ellas en un caso particular provisto por el INTA Alto Valle. Este es un trabajo preliminar desarrollado en el marco del trabajo de Tesis de la carrera de Licenciatura en Sistemas de Información.

Keywords: Análisis de Datos · Taxonomía · Índice NDVI

1 Introducción

Los sistemas para análisis de datos resultan interesantes hoy en día debido a la cantidad de información recolectada por las diferentes organizaciones y su necesidad de explotación. Para la creación de este tipo de sistemas se requiere de la aplicación de un proceso bien definido con actividades o fases que deben incluir la selección de las fuentes de datos, el procesamiento de información relevante, el almacenamiento en un repositorio específico y el análisis de los datos almacenados.

En este trabajo nos centramos en el dominio de la fruticultura que posee un especial interés en la región del Alto Valle de Río Negro y Neuquén, ya que es una área productiva destinada a la producción de peras y manzanas. En particular, nos avocamos al análisis de índices de vegetación de forma tal de dar los

* Este trabajo esta parcialmente soportado por el proyecto UNCOMA 04/F0012 "Variedad en Big Data" 2022-2025

primeros pasos para determinar causas y factores que los condicionan. En esta ocasión, analizamos el índice de Vegetación de Diferencia Normalizada (NDVI) que describe la salud de la vegetación midiendo la diferencia entre el infrarrojo cercano (lo que refleja la vegetación) y la luz roja visible (lo que absorbe la vegetación)[3] [11]. Este índice tiene alta relevancia en el campo de la fruticultura y es también altamente estudiado por el Instituto Nacional de Tecnología Agropecuaria (INTA) en su estación experimental agropecuaria Alto Valle[4]. Para el instituto, es importante conocer los factores que puedan determinar las variaciones del índice en las diferentes regiones. Sin embargo, dichos factores no son pocos ni tampoco fáciles de medir o analizar y es aquí donde los procesos de análisis de datos toman importancia.

Considerando este contexto, en este trabajo realizamos un proceso de análisis preliminar que posee dos objetivos: (1) caracterizar los factores que influyen en el índice NDVI según la evidencia en la literatura y (2) comprobar algunos de estos factores en datos provistos por el INTA.

El artículo se organiza de la siguiente manera. La sección siguiente describe los trabajos relacionados que han aplicado algún proceso de análisis de datos y que involucren a su vez el estudio del índice NDVI y los factores influyentes. Luego describimos nuestra caracterización y presentamos un caso de estudio que en base a datos provistos por el INTA intenta comprobar la incidencia de algunos de los factores caracterizados en el NDVI de la región del Alto Valle.

2 Trabajos Relacionados y Caracterización de Variables para el índice NDVI

En la literatura hay varios trabajos que realizan análisis del índice NDVI considerando diferentes factores y obteniendo resultados diferentes e interesantes. Por razones de espacio, hemos resumido los trabajos analizados en la Tabla 1. En la misma destacamos, de cada trabajo, el objetivo, fuentes de información, período analizado, técnicas de análisis aplicadas y los resultados obtenidos.

Para la caracterización de variables o factores que pueden incidir en el valor del índice NDVI hemos seguido los pasos propuestos por [2] para la construcción de taxonomías. Estos pasos involucran la identificación de requerimientos y el análisis de la información existente a fin de elaborar las taxonomías candidatas. En nuestro caso, los factores significativos para el índice NDVI se determinaron a partir de los trabajos relacionados y del análisis con usuarios expertos. En la Figura 1 podemos observar la taxonomía realizada que direcciona nuestro primer objetivo planteado en este trabajo.

[3] El NDVI se calcula a partir de imágenes de satélite y puede tomar el valor de -1 a 1 siendo 1 lo más saludable posible

[4] https://inta.gob.ar/altovalle/sobre-812000

Propuesta	Objetivo	Fuentes/Periodo	Análisis	Resultados
Ovando et.al 2019 [7] (Córdoba)	evaluar tendencias en el cambio de la profundidad de la napa freática	sensores MODIS, pluviómetro y freatímetro. Periodo: 2001-2018	regresión lineal y coeficiente R^2	profundidad con mayor tasa de precipitación y alto NDVI
Ortiz et.al 2012 [6] (Chile)	evaluar los cambios de P. tamarugo con el nivel freático	21 pozos con freatímetros y 3 imágenes satelitales. Periodo: 1997-2005	modelo de regresión lineal	+ de 10mts de profundidad napa, tamarungos con niveles bajos de NDVI
Jin et.al 2014 [4] (China)	analizar la vegetación y profundidad del nivel freático	<400 mediciones de napas, relieve, humedad, lluvias. Periodo: variable	estadísticas básicas, análisis de correlación, etc.	relación NVDI con relieve, humedad, lluvias, altura de la napa (5mts)
Nallan et.al 2015 [5] (India)	vegetación (NDVI) y el impacto de estructuras captación de agua	imágenes satelitales (MODIS 250m). Periodo: 2001-2012	análisis espacial de puntos calientes	la construcción de estructuras captación de agua generó aumento en NDVI
Pascoa et.al 2020 [8] (Península Ibérica)	vegetación potencial con las precipitaciones y evapotranspiración	imágenes satelitales (MODIS). Periodo:1971-2000	clustering (k-means)	relación entre la vegetación, precipitaciones e índice de aridez con aguas subterráneas
Zerda & Tiedemann et.al 2010 [13] (Santiago del Estero)	dinámica interanual y mensual del NDVI, analizando bosque y pastizal natural	imágenes satelitales (SPOT 4-Vegetation). Periodo:1999-2002	análisis de varianza	precipitaciones, evapotranspiración y estación del año con NDVI y pastizales naturales
Gaitan et.al 2015 [3] (Argentina)	tendencia del NDVI como indicador de la degradación de tierras	2898 imágenes satelitales (MODIS). Periodo: 2000-2014	regresión lineal entre el tiempo y NDVI	tendencias negativas y positivas de NDVI en el territorio
Poudel et.al 2021 [10] (California)	productividad del sistema de riego y rendimiento con respecto a la evapotranspiración, NDVI, uso del agua, etc.	imagenes satelidades (EEFlux, NDVI), reportes no espaciales. Periodo: 2018-2019	coeficientes de variación, random forest, correlaciones	relaciones entre la evapotranspiración NDVI, uso del agua por riego y rendimiento de los cultivos
Yujie et.al 2019 [12] (Mundo)	los factores con mayor influencia en el índice NDVI	imágenes satelitales de GIMMS NDVI y CRU. Periodo: 1982-2015	análisis Theil-Sen, test Mann-Kendall, Pearson, BRT	NDVI con lluvias, temperatura, suelo, población, elevación, luz nocturna

Table 1. Resumen de propuestas utilizando el índice NDVI

3 Instanciación de la Taxonomía: Análisis del índice NDVI en datos del INTA Alto Valle

Para realizar la instanciación de la taxonomía seguimos un proceso de desarrollo de sistemas de análisis de datos definido en trabajos previos [1] que incluyen, entre otras actividades, la limpieza y/o preprocesamiento, almacenamiento y la analítica de datos.

Considerando nuestro segundo objetivo planteado de analizar algunos de los factores caracterizados previamente (Figura 1) con el índice NDVI, hemos aplicado un proceso de análisis de datos con datos provistos por el INTA Alto Valle. Dichos datos cubren un periodo de 2015-2022 y contienen información sobre la zona de Villa Regina, Río Negro. En particular, los conjuntos de datos que hemos tenido que preprocesar contenían información de: (1) mediciones realizadas en freatímetros[5], (2) el índice NDVI asociado a la zona de cada freatímetro en píxeles de 10x10 metros, y (3) datos meteorológicos de la estación que se encuentra ubicada en la región analizada.

Para el procesamiento de los datos, hemos primero diseñado las transformaciones necesarias incluyendo eliminación de nulos, normalización y agregación de los valores considerando la medición de cada freatímetro por cada mes y año

[5] Los freatímetros son dispositivos instalados en puntos particulares que miden la altura de la napa freática

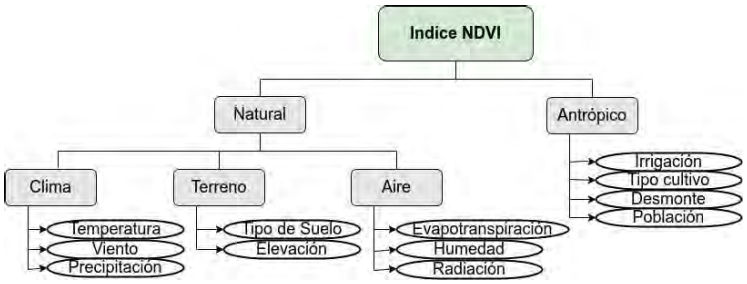


Fig. 1. Factores que influyen en el índice NDVI según la literatura y expertos

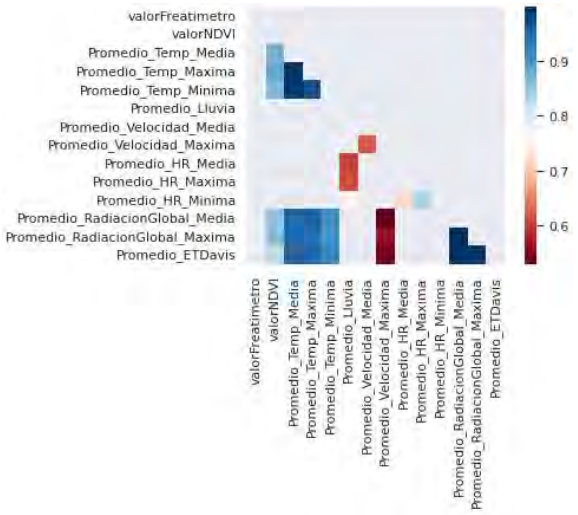


Fig. 2. Relación del índice NDVI con otras variables cuando la correlación es mayor a 0,5

con su cota, valor NDVI, temperatura (min, max y media), lluvia, velocidad del viento (media, max. y dirección), humedad (min, max y media), radiación (media, max.) y evapotranspiración (ET Davis). Luego, con estos datos normalizados hemos analizado sus distribuciones para, en algunos casos, realizar nuevos preocesamientos o limpiezas. Por último, realizamos análisis de correlación [9], buscando identificar relaciones en los datos, y a partir de los estudios previos y nuestra taxonomía (Figura 1), determinar cuáles serían variables predictoras para el resultado del índice NDVI.

Podemos observar en la Figura 2 que las variables con más relación con el índice NDVI son las temperaturas, la radiación y la evapotranspiración. Sin embargo, por ejemplo otros factores como la napa freática (valorFreatímetro) no obtuvieron buenas correlaciones, con lo que se requerirá en un futuro analizar las causas.

4 Conclusiones y Trabajo Futuro

En este artículo hemos descripto un trabajo preliminar para cumplir con los objetivos de caracterizar en base a la literatura y los usuarios expertos los factores que influyen en el índice de vegetación NDVI e instanciar esas características sobre datos provistos por el INTA. Los análisis realizados mostraron buenas correlaciones entre el índice y algunos de los factores caracterizados. Como trabajos futuros, se plantea la recolección de más datos que posean otros de los factores caracterizados para que junto a este trabajo, se continúen analizando relaciones del índice NDVI y otros factores.

References

1. Buccella, A., Manrique, D., Troncoso, D., Cechich, A.: Experiences from a data analysis of crimes against humanity. Journal of Computer Science and Technology **21**, e3 (04 2021). https://doi.org/10.24215/16666038.21.e3
2. Choksy, C.: 8 steps to develop a taxonomy. Information Management Journal pp. 30–41 (2006)
3. Gaitan, J., Bran, D., Azcona, C.: Tendencia del ndvi en el período 2000-2014 como indicador de la degradación de tierras en argentina: ventajas y limitaciones. AgriScientia **32** (12 2015). https://doi.org/10.31047/1668.298x.v32.n2.16559
4. Jin, X., Guo, R., Zhang, Q., Zhou, Y., Zhang, D., Yang, Z.: Response of vegetation pattern to different landform and water-table depth in hailiutu river basin, northwestern china. Environmental Earth Sciences **71** (06 2014)
5. Nallan, S., Armstrong, L., Tripathy, A., Teluguntla, P.: Hot spot analysis using ndvi data for impact assessment of watershed development. In: ICTSD 2015 (04 2015). https://doi.org/10.1109/ICTSD.2015.7095869
6. Ortiz, M., Morales-Salinas, L., Candia, P., Acevedo, E.: Estimation of water table depth from landsat ndvi in the pampa del tamarugal (chile). Revista de Teledeteccion pp. 42–50 (01 2012)
7. Ovando, G., Bocco, M., Bollatti, P., Sayago, S., Andreucci, A., Collino, D.: Análisis de la tendencia del nivel de napa freática y su relación con las precipitaciones, evapotranspiración potencial y ndvi en marcos juárez (córdoba). In: XI CAI - JAIIO 48. pp. 73–82. SAIIO (2019)
8. Pascoa, P., Gouveia, C., Kurz-Besson, C.: A simple method to identify potential groundwater-dependent vegetation using ndvi modis. Forests **11**(2) (2020). https://doi.org/10.3390/f11020147
9. Peck, R., Olsen, C., Devore, J.: Introduction to Statistics and Data Analysis - Third Edition. Thomson, USA (2008)
10. Usha, P., Haroon, S., Sajjad, A.: Evaluating irrigation performance and water productivity using eeflux et and ndvi. Sustainability **13**(14) (2021). https://doi.org/10.3390/su13147967
11. Weier, J., Herring, D.: Measuring vegetation (ndvi & evi). NASA Earth Observatory, Washington DC (2000)
12. Yujie, Y., Shijie, W., Xiaoyong, B., Qin, L., Luhua, W., Shiqi, T., Zeyin, H., Chaojun, L., Yuanhong, D.: Factors affecting long-term trends in global ndvi. Forests **10**(5) (2019). https://doi.org/10.3390/f10050372
13. Zerda, H., Tiedemann, J.: Dinámica temporal del ndvi del bosque y pastizal natural en el chaco seco de la provincia de santiago del estero, argentina. Ambiencia **6**, 13–24 (02 2010)

Inteligencia artificial: herramienta diagnóstica para el cáncer de mama

Gianfranco Jesús Curci Robledo, Miguel Mendez Garabeti, Pablo Javier Sandez

Laboratorio de Investigación en Ciencia y Tecnología, Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua, Mendoza, Argentina.
gianuniversidad@gmail.com

Resumen. La medicina es uno de los campos del conocimiento que más podrían beneficiarse de una interacción cercana con la computación y las matemáticas, mediante la cual se optimizarían procesos complejos e imperfectos como el diagnóstico diferencial. De esto se ocupa el aprendizaje automático, rama de la inteligencia artificial que construye y estudia sistemas capaces de aprender a partir de un conjunto de datos de adiestramiento y de mejorar procesos de clasificación y predicción. En este artículo, se dispone trabajar con datos de pacientes de un padecimiento muy frecuente en la actualidad, que es el cáncer de mama. Para mejorar la precisión diagnóstica precoz de dicha con el uso de redes neuronales.

Keywords: inteligencia artificial · diagnóstico clínico · cáncer de mama · minería de datos.

1 Introducción

La detección temprana de cáncer de mama contribuye a una reducción eficaz de la mortalidad producida por esta enfermedad. Dicha detección se hace en una primera instancia a través de la palpación personal de la mama, por parte del paciente, seguida de estudios imagenológicos y de laboratorio. Entre los estudios de laboratorio, encontramos aquellos dirigidos a la expresión de marcadores tumorales aumentados en sangre y estudios relacionados con la expresión de ciertos genes. Teniendo todos estos estudios, ya el diagnóstico certero depende de la habilidad del oncólogo a cargo del caso. Muchas veces de las cuales, este diagnóstico por falta de análisis del caso de los estudios genéticos anteriormente nombrados, se asume que el paciente, a pesar de tener alterados genes como: BRCA1 y el BRCA2, está sano o bien no es un posible candidato a desarrollar algún tipo de cáncer de mama en el futuro. Para evitar este sesgo diagnóstico, este trabajo plantea de manera teórica la posibilidad de utilizar diferentes herramientas de la inteligencia artificial, de manera de poder plantear algún modelo experimental de IA que permita en base a la expresión de los mencionados los genes alterados y otros estudios positivos. Brindar un porcentaje que refleje la posibilidad de desarrollar o no la enfermedad a futuro, es decir, usar una predicción. Este modelo se utilizaría como una "segunda opinión" que puede utilizar el medico a cargo, para justificar la realización de un tratamiento preventivo o no.

2 Descripción del problema

Cada año el cáncer de mama va en incremento, aunque los criterios de los especialistas realicen el trabajo, hay una imposibilidad en el recurso humano de atender estos volúmenes, así como el grado de certeza con respectos a los errores en los diagnósticos y las consecuencias en la salud mental de aquellos pacientes incurridos. Mientras que los avances en la Inteligencia Artificial (IA) son rápidos, abarcantes y cada vez más efectivos, existe una controversia en el sentido de la aceptación para la detección computarizada de cáncer de mama, la confianza del paciente, por un lado, la competitividad del algoritmo por parte de la comunidad médica. Diseñar un modelo en I.A. para lograr ser efectivo en la detección de cáncer de mama, permite no solo acercarse a un procedimiento válido, sino de bajo coste, preciso o confiable, traducido en una mayor capacidad en atención de pacientes.
La implementación para el procesamiento y almacenamiento de la base de datos, son posibles por costos bajos que implica el clouding, muy contrario a la infraestructura tecnológica compleja y costosa. Asumir una efectividad muy alta para la detección de cáncer de mama (benignos como malignos), es aceptable en la medida técnica, económica y administrativa, otros algoritmos de esta naturaleza, se enfocan en el aspecto académico o limitan sus funciones en lo tecnológico, y no tanto en una respuesta en el campo oncológico. Diseñar y poner a prueba un algoritmo para la detección del cáncer en el ámbito social, tecnológico y administrativo, basado en Redes Neuronales Artificiales (RNA), son tan solo algunos aspectos que contribuyen y dinamizan al sector de la salud. El problema a resolver está dirigido al grado de efectividad en el modelo por pronósticos para la detección del cáncer de mama.

3 Análisis del problema

Aunque una RNA tiene a efectos imitar el sistema neuronal del humano, en realidad no funciona idéntico, ni siquiera por el uso de millones de neuronas para realizar las funciones que modela, sin embargo, la interconexión, los pesos y el resultado, en función y logro, es muy similar. Para comprender un tanto sobre el paralelismo idealizado entre una RNA y una Red Neuronal Humana (RNH). En resumen, el problema a resolver está dirigido al grado de efectividad en el modelo por pronósticos para la detección del cáncer de mama.

4 Propuesta de modelo

La iniciativa del desarrollo de modelos de inteligencia artificial empleando Redes Neuronales Artificiales para el pronóstico del cáncer de mama, crean una esperanza tanto para los afectados directamente, como para la comunidad científica, abordar temas complejos y con un constante incremento, sirven como coadyuvantes el criterio y valoración del cuerpo médico ante los protocolos existentes.

El modelo propone una solución complementaria para la labor de los especialistas en el protocolo para la detección del cáncer de mama. El algoritmo estará basado en Redes Neuronales Artificiales (RNA) en el campo de la Inteligencia Artificial, como una propuesta para promover la resolución ante la problemática de salud pública y los índices muy altos y progresivos del cáncer de mama a nivel mundial. Se utilizará un perceptrón multicapa con un aprendizaje de tipo supervisado y una función de activación de tipo sigmoidal. La primera fase del modelo, parte de la carga y verificación de los datos sobre el estado de pacientes con cáncer de mama, el proceso se hace de forma semiautomática derivando en un dataset. La información de susodicho, estaría recopilada de pacientes que tengan aquellos genes específicos alterados y sus derivados, tales como BRCA1 o BRCA2. Esta información se usará entrenamiento del perceptrón y seria provista por la IMBECU.
La segunda fase consiste en la depuración de datos o Data Cleaning, a través de técnicas del Data Mining, se categorizan cada uno de los componentes para su exploración, manipulación y análisis. Entre las propuestas analizadas, se utilizará Weka como software elegido para esta tarea, debido a su optimización y aval científico por la Universidad de Waikato. Se conformará, además, el grupo de entrenamiento y el conjunto de validación de datos.
La tercera fase es la Exploración de análisis de los datos o Exploratory Data Analysis (EDA), se evalúan los datos, coherencia y consistencia que inciden sobre los resultados. La cuarta fase consiste en el entrenamiento, un algoritmo para establecer pautas que acondicionan el escenario para aplicar las RNA, definir entradas, la capa oculta, salida y, el resultado, tal como lo realizaría un perceptrón.
La cuarta fase consiste en el entrenamiento, un algoritmo para establecer pautas que acondicionan el escenario para aplicar las RNA, definir entradas, la capa oculta, salida y, el resultado.
La quinta fase constaría de evaluar la mejor arquitectura posible para el perceptrón, teniendo en cuenta la tasa de aprendizaje, número de capas y neuronas requeridas para el proceso, como bien se sabe este proceso, aunque sea poco ortodoxo, es más de prueba y error; el resultado obtenido se evaluará con el conjunto de validación y se ajustará la arquitectura, tasa de aprendizaje, numero de capas y función de activación.

5 Resultado esperado

Se espera que dicho modelo pueda predecir y clasificar de manera correcta los pacientes de manera correcta. En aquellos que si tienen alta probabilidad de tener cáncer de mama de los que no. Ayudando así a la decisión del especialista, de la posibilidad de sugerir un tratamiento preventivo, un seguimiento exhaustivo o bien confirmar con otra herramienta diagnostica, que efectivamente dicho paciente no desarrollara la enfermedad. Para lograr dicha hazaña, se calcula que al menos la muestra de pacientes debe ser de al menos 100, a mayor cantidad de muestras, diversificación y redundancia de algunos datos, mayor será la credi-

bilidad de los resultados arrojados por la IA. Podemos decir que dicha es eficaz, cuando su porcentaje de precisión supere el 95 por ciento.

6 Conclusiones

La iniciativa del desarrollo de modelos de inteligencia artificial empleando Redes Neuronales Artificiales para el pronóstico del cáncer de mama, crean una esperanza tanto para los afectados directamente, como para la comunidad científica, abordar temas complejos y con un constante incremento, sirven como coadyuvantes el criterio y valoración del cuerpo médico ante los protocolos existentes.

Este proyecto busca lograr los siguientes objetivos:

- Crear otras alternativas complementarias para abordar pacientes con cáncer como un complemento a la comunidad científica.
- Coadyuvar al criterio y valoración del cuerpo médico ante los protocolos existentes.
- Optimizar el número de atributos de entrada (función extracción) en el conjunto de datos para eliminar la limitación de los conjuntos de datos disponibles relacionados con el cáncer.

Referencias

1. Jose Francisco, F., Ávila Tomás Miguel, S., Angel Mayer Pujadas, T., Victor Julio Quesada Varela: La inteligencia artificial y sus aplicaciones en medicina I: introducción antecedentes a la IA y robótica (2020)
2. Jose Francisco, F., Ávila Tomás Miguel, S., Angel Mayer Pujadas, T., Victor Julio Quesada Varela: La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas (2021)
3. Saúl Oswaldo Lugo-Reyes, F.: Inteligencia artificial para asistir el diagnóstico clínico en medicina. Revista Alergia Mexico (RAM) (2014)
4. Amrita Naika, F Lilavati Samantb, S.: Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime. Conferencia internacional de seguridad y computación (2016)
5. Pedro Isasi, F., Ávila Tomás Miguel, S., Inés M. Galván:Redes neuronales artificiales: un enfoque práctico. 1st ed. Publisher, Pearson-Prentice Hall, Madrid (2004)

Un Sistema para la Identificación de Cadáveres NN en el Contexto de Búsqueda de Personas Desaparecidas

Andrea Maldonado[1], Darío Ruano[1,2], Norma Herrera[1,2], Marcelo Martínez[3]

[1] Departamento de Informática, FCFMyN, Univ.Nacional de San Luis
[2] Laboratorio de Investigación y Desarrollo en Bases de Datos , Univ. Nacional de San Luis
[3] Jefe Interino del Cuerpo Médico Forense y Criminalístico de la Tercera Circunscripción Judicial de la Provincia de Mendoza

andreamaldonadoma@gmail.com, dmruano@unsl.edu.ar, nherrera@unsl.edu.ar, drmarcelomartinez@hotmail.com

Abstract. En este trabajo abordaremos la problemática de identificación de cadáveres en el contexto de búsqueda de personas desaparecidas, usando como base el modelo de espacio métricos. El objetivo final es el desarrollo de un sistema que que permita un manejo más ágil de la información. Para ello por cada cadáver se genera un vector que contiene la información necesaria para posteriormente realizar búsquedas por similitud.
Palabras claves: Bases de Datos, Espacios Métricos, Identificación de Personas

Este trabajo se desarrolla en el marco del Trabajo Final de la Licenciatura en Ciencias de la Computación de la alumna Andrea Maldonado, dirigido por el Lic. D. Ruano y la MCs. N. Herrera. Dada la temática involucrada se cuenta con el asesoramiento del Dr. Marcelo Martínez.

1 Introducción

En la era actual, caracterizada por la evolución de las tecnologías de la información y las comunicaciones, las ciencias de la computación son transversales a la mayoría de nuestras actividades diarias, brindando las herramientas necesarias para abordar problemas complejos y contribuyendo en la búsqueda de soluciones eficientes a problemas de interés. La medicina legal y forense no escapa a esta realidad. Existen varios temas de interés en este contexto, uno de ellos es la identificación de cadáveres NN.

Dentro de los individuos que ingresan a los distintos Institutos de Medicina Forense del país, existen casos que no poseen las condiciones adecuadas para su identificación inmediata (indocumentados, en avanzado estado de descomposición, restos óseos, etc.) o sin posibilidad de identificación (fragmentos muy pequeños, restos carbonizados, etc.). Frente a esto, las instituciones deben investigar no sólo para determinar qué fue lo que sucedió (causa de la muerte) y cuándo sucedió (data de la muerte), sino también para poder dar con la identidad del cuerpo.

La importancia de identificar a las personas cuya identidad se desconoce responde no solo al derecho fundamental de todos los seres humanos de tener una identidad sino

también a numerosas razones de tipo social que van desde la necesidad de informar a los familiares de personas desaparecidas sobre la certeza de su fallecimiento hasta el hecho de evitar que personas infractoras de la ley simulen su propia muerte.

Claramente la identificación de cadáveres está directamente relacionada con la búsqueda de personas desaparecidas. En Argentina, no existe un sistema único de procesamiento para esta problemática. Cada provincia tiene su gobierno, su sistema forense y sus protocolos. Esto dificulta el proceso de identificación de cadáveres: si una persona desaparece en Chaco y aparece un cadáver similar en Chubut, no hay una forma rápida y correcta de relacionarlos. Si se contara con una bases de datos unificada en el país, cualquier investigador podría consultar cuántos hombres de entre 20 y 30 años tienen un tatuaje en el brazo derecho evitando mirar miles de expedientes que están en diferentes jurisdicciones. El Sistema Federal de Búsqueda de Personas Desaparecidas y Extraviadas (SIFEBU) es un intento de crear esta base de datos unificada pero sin tener automatizado el proceso de búsqueda.

En [6], una de las recomendaciones dadas es estandarizar el registro de cadáveres NN por categorías, evaluando la necesidad de un registro único a nivel nacional que brinde información específica sobre las características físicas e identificadoras (como huellas y muestras de ADN, por ejemplo) de cada cadáver hallado. Recomiendan además trabajar en la posibilidad de ampliar el acceso a la información que existe en los registros civiles provinciales, en los cementerios y en otros organismos nacionales como RENAPER (Registro Nacional de las Personas)

En este trabajo abordaremos esta problemática con el fin de **dar un primer paso** a un sistema federal de identificación de cadáveres en el contexto de búsqueda de personas. Debido a la complejidad de la temática, nos centraremos en la identificación de cuerpos. La investigación forense de casos que involucran la recuperación y análisis de restos óseos es un proceso complejo en el que intervienen diferentes disciplinas científicas y que no abordaremos en este trabajo. El desarrollo de una herramienta que permita un manejo más ágil de la información en el proceso de identificación de cadáveres NN, tendrá como resultado la posibilidad real y tangible de poder colaborar en una situación tan sensible como lo es identificar el cuerpo de una persona que está siendo buscada por sus seres queridos.

Lo que resta del artículo está organizado de la siguiente manera. En la Sección 2 damos el marco teórico exponiendo una reseña sobre el modelo de espacios métricos. En la Sección 3, presentamos el desarrollo realizado hasta el momento donde hemos utilizado como base el modelo de espacio métricos.. Finalizamos en la Sección 4 dando las conclusiones y el trabajo futuro.

2 El Modelo de Espacios Métricos

Las bases de datos tradicionales son construidas basándose en el concepto de búsqueda exacta: la base de datos es dividida en registros y cada registro contiene campos completamente comparables. Las consultas a la base de datos retornan todos aquellos registros cuyos campos coinciden con los aportados en tiempo de búsqueda.

Actualmente las bases de datos han incluido la capacidad de almacenar nuevos tipos de datos tales como imágenes, sonido, video, etc. Estructurar este tipo de datos en re-

gistros para adecuarlos al concepto tradicional de búsqueda exacta es difícil en muchos casos y hasta imposible si la base de datos cambia más rápido de lo que se puede estructurar (como por ejemplo la web). Aún cuando pudiera hacerse, las consultas que se pueden satisfacer con la tecnología tradicional están limitadas en variaciones de la búsqueda exacta.

Nos interesan las búsquedas en donde se puedan recuperar objetos *similares* a uno dado. Este tipo de búsqueda se conoce con el nombre de ***búsqueda por similitud***, y surge en diversas áreas; reconocimiento de voz, reconocimiento de imágenes, compresión de texto, biología computacional, son algunas de ellas.

Todas estas aplicaciones tienen algunas características comunes. Existe un universo $\mathcal{X}$ de objetos y una función de distancia $d : \mathcal{X} \times \mathcal{X} \rightarrow R^+$ que modela la similitud entre los objetos. El par $(\mathcal{X}, d)$ es llamado ***espacio métrico*** [5]. La base de datos es un conjunto $\mathcal{U} \subseteq \mathcal{X}$, el cual se preprocesa a fin de resolver búsquedas por similitud eficientemente. Se pueden mencionar tres tipos de búsquedas que normalmente se utilizan en espacios métricos [5].

Búsqueda por rango $(q, r)_d$: dado un elemento $q \in U$ y un radio de tolerancia r, una búsqueda por rango consiste en recuperar los objetos de la base de datos que estén a distancia a lo sumo r de q, es decir: $(q, r)_d = \{u \in U : d(q, u) \leq r\}$

Búsqueda del vecino más cercano NN(q): consiste en recuperar el (o los) elemento(s) más cercano(s) a un elemento q dado. En símbolos: $NN(q) = \{u \in U : \forall v \in U, d(q, u) \leq d(q, v)\}$

Búsqueda de los k-vecinos más cercanos NNk(q): Se busca recuperar los k elementos más cercanos a q en U. Esto significa encontrar un conjunto $A \subseteq U$ tal que: $|A| = \{k \wedge \forall u \in A, v \in (U - A) : d(q, u) \leq d(q, v)\}$

Las búsquedas por similitud pueden ser resueltas trivialmente por medio de una búsqueda exhaustiva, con una complejidad $O(n)$. Para evitar esta situación, se preprocesa la base de datos por medio de un algoritmo de indexación con el objetivo de construir una estructura de datos o índice, diseñada para ahorrar cálculos en el momento de resolver una búsqueda. El tiempo total de resolución de una búsqueda puede ser calculado de la siguiente manera: $T =$ evaluaciones de $d \times$ complejidad$(d) +$ tiempo extra de $CPU +$ tiempo de I/O.

En muchas aplicaciones la evaluación de la función d es tan costosa, que las demás componentes de la fórmula anterior pueden ser despreciadas. Este es el modelo que usaremos en este trabajo.

Básicamente existen dos enfoques para el diseño de algoritmos de indexación en espacios métricos: uno está basado en Diagramas de Voronoi [5, 2, 7] y el otro está basado en pivotes [5, 3, 4].

Uno de los principales obstáculos en el diseño de buenas técnicas de indexación es lo que se conoce con el nombre de *maldición de la dimensionalidad*. El concepto de dimensionalidad está relacionado con el nivel de dificultad al buscar en un determinado espacio métrico. La dimensión de un espacio métrico se define en [5] como $\rho = \frac{\mu^2}{2\sigma^2}$, siendo μ y σ^2 la media y la varianza respectivamente de su histograma de distancias. Es decir que, a medida que la dimensionalidad intrínseca crece, la media crece y su varianza se reduce. Esto significa que el histograma de distancia se concentra más alrededor de su media, lo que influye negativamente en los algoritmos de indexación.

3 SiBDaNN: Un Sistema de Bases de Datos para la Identificación de NN's

En este trabajo abordamos la aplicación de la teoría de Espacios Métricos para la identificación de cadáveres en el contexto de búsqueda de personas desaparecidas. El objetivo es desarrollar un sistema que permita mantener una base de datos, modelizada con un espacio métrico, con información sobre cadáveres no identificados para posteriormente realizar búsquedas de personas desaparecidas. Por cada cadáver se mantendrá un vector con los datos de las características físicas del mismo. Al momento de realizar una búsqueda, se deberá ingresar el vector con los datos de las características físicas de la persona buscada para que el sistema realice una búsqueda por similitud sobre la base de datos de cadáveres. El resultado será una lista de cadáveres con características similares a la persona buscada, rankeados según el grado de similitud.

Explicamos a continuación el trabajo desarrollado hasta el momento.

3.1 Generación de Vectores Característicos

Para la elaboración del sistema, como primer paso hubo que definir un core de datos que ses adecuado a la problemática de identificación. Usar pocos datos podría provocar que las búsquedas que se realicen sean de muy baja selectividad y usar demasiados puede provocar que se descarten elementos de la base de datos que sean de interés. En este sentido, ya hemos definido un primer core de datos sobre el cual trabajar: *color de ojos, color de pelo, color de piel, existencia de tatuajes y lugar de los mismos, cicatrices y/o marcas* (lunares por ejemplo), si existen *amputaciones y en qué lugar del cuerpo*, la *contextura física (atlético, atrófico, etc.)* y finalmente la existencia de *agenesias*. Para cada dato, se transforman los valores de su dominio en números que reflejen el grado de similitud entre los valores considerados y se genera el vector correspondiente.

Claramente no todos los datos tienen el mismo grado de importancia, por ejemplo el color de pelo es menos importante que el color de ojos porque una persona puede cambiarse el color de pelo pero no el de ojos. Esto hay que tenerlo en cuenta para establecer para cada dato un peso que corresponda con el grado de importancia del mismo.

3.2 Función de Distancia

Otro punto importante es la función de distancia a utilizar en las búsquedas. En una primera etapa usaremos la función coseno [1]. Si q es la query (vector de la persona buscada) y d_j es el *j-ésimo* vector de la base de datos, entonces el grado de similitud entre los vectores $\boldsymbol{d}_j$ y $\boldsymbol{q}$ se calcula como el coseno del ángulo formado entre ambos vectores:

$$sim(d_j, q) = \frac{\boldsymbol{d}_j . \boldsymbol{q}}{|\boldsymbol{d}_j| \times |\boldsymbol{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}}$$

donde w_{iq} es el peso del *i-ésimo* dato en la consulta q y w_{ij} es el peso del *i-ésimo* dato en el vector d_j .

3.3 Indexación y Búsquedas

Con respecto al algoritmo de indexación comenzaremos usando algoritmos basados en pivotes. Cuando la base de datos se cargue con datos reales, se podrá analizar la dimensionalidad del espacio métrico sobre el cual se está trabajando y de ser necesario se cambiará el algoritmo de indexación.

Con respecto a las búsquedas, utilizaremos las búsquedas de los k vecinos más cercanos, porque es la que más se adecúa a este problema. Esto permitirá al usuario del sistema decidir cuánto elementos desea recuperar en una primera instancia y luego, de ser necesario, podrá ampliar la búsqueda; por ejemplo: puede pedir los 10 cadáveres mas parecidos a la persona buscada y posteriormente puede ampliar la búsqueda pidiendo los 10 siguientes.

4 Conclusiones y Trabajo Futuro

En este trabajo abordaremos la problemática de identificación de cadáveres en el contexto de búsqueda de personas con el fin de dar un primer paso que sirva de ayuda a una problemática tan delicada.

Para el desarrollo del sistema usamos el modelo de espacios métricos para la realización de búsquedas por similitud. Hasta el momento se ha programado el front-end del sitio web que permitirá la conexión con el sistema SiBDaNN. Como trabajo futuro nos proponemos programar el back-end e iniciar la prueba del sistema. En función de los resultados que se obtengan con la primer versión del sistema se realizarán, de ser necesario, cambios que pueden implicar: aumentar o disminuir los datos del core, cambiar la función de distancia y/o cambiar el algoritmo de indexación.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
2. S. Brin. Near neighbor search in large metric spaces. In *Proc. 21st Conference on Very Large Databases (VLDB'95)*, pages 574–584, 1995.
3. W. Burkhard and R. Keller. Some approaches to best-match file searching. *Comm. of the ACM*, 16(4):230–236, 1973.
4. E. Chávez, J. Marroquín, and G. Navarro. Fixed queries array: A fast and economical data structure for proximity searching. *Multimedia Tools and Applications (MTAP)*, 14(2):113–135, 2001.
5. E. Chávez, G. Navarro, R. Baeza-Yates, and J.L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, September 2001.
6. Procuraduría de Trata y Explotación de Personas (PROTEX) y la Colectiva de Intervención antes las Violencias (CIAV). Búsqueda de personas en democracia: Identificaciones de nn, trayectorias de vidas y cursos burocráticos. Technical report, Ministerio Público Fiscal, 2020.
7. G. Navarro. Searching in metric spaces by spatial approximation. In *Proc. String Processing and Information Retrieval (SPIRE'99)*, pages 141–148. IEEE CS Press, 1999.

Software de generación de datos de prueba para sistemas ubicuos

Mariano Andrés Acuña Ninich, Beatriz Fernández Reuter, Gabriela González

Instituto de Investigación en Informática y Sistemas de Información (IIISI)
Facultad de Ciencias Exactas y Tecnologías (FCEyT)
Universidad Nacional de Santiago del Estero
{mariano.acu.n, bfreuter, gonzalezgbr}@gmail.com

Resumen. La validación de software tiene por objetivo comprobar que un sistema cumple tanto con sus especificaciones como con las expectativas del cliente. Para el caso de los sistemas ubicuos, las técnicas de simulación de datos de prueba son de especial utilidad porque estos sistemas reciben como entrada una gran variedad de datos dinámicos lo que genera, a su vez, un alto número de escenarios posibles. El presente trabajo describe un software que busca integrar las técnicas de simulación en la generación de datos de prueba que puedan ser empleados para la validación de sistemas en ambientes ubicuos.

Palabras claves: prueba de software, sistemas ubicuos, simulación.

1 Introducción

La validación de software o, más generalmente, su verificación y validación (V & V), se crea para mostrar que un sistema cumple tanto con sus especificaciones como con las expectativas del cliente. Las pruebas del programa, donde el sistema se ejecuta a través de datos de prueba simulados, son la principal técnica de validación [1]. Las pruebas intentan demostrar que un programa hace lo que se intenta que haga, así como descubrir defectos en el programa antes de usarlo.

Como opción a las muchas técnicas de generación de datos de prueba surgen las técnicas de simulación [2], las cuales se pueden considerar aptas para generar este tipo de datos puesto que sus métodos cumplen con condiciones favorables y emplean funciones matemáticas demostradas, todo esto ayudando a reducir costos, tiempo y controlar las condiciones de experimentación.

Por otro lado, hoy en día encontramos diferentes tipos de aplicaciones, basadas en tecnologías móviles y ubicuas. El término "computación ubicua" se atribuye a Mark Weiser [3], quién decía que las tecnologías más profundas son aquellas que desaparecen y que se encuentran inmersas en la vida cotidiana, de forma tal que no se pueden distinguir de ella. Es una tendencia de las tecnologías de información y comunicación que se encuentra embebida en un gran número de pequeñas computadoras, que están equipadas con sensores y actuadores que interactúan con el medio ambiente para intercambiar datos [4]. Cuando tratamos con aplicaciones ubicuas, estas interactúan con un importante número de variables que cambian dinámicamente con el contexto. Esta característica convierte a las

aplicaciones ubicuas en grandes candidatas a ser tratadas mediante las técnicas de simulación, ya que dichas técnicas recibirán como entrada una variedad de datos dinámicos y generarán un alto número de escenarios posibles.

El trabajo desarrollado busca integrar las técnicas de simulación en la generación de datos de prueba que puedan ser empleados para la validación de un software en ambientes ubicuos.

En la sección siguiente se describe la motivación del trabajo. Luego se presentan los aportes realizados. Finalmente se detallan posibles líneas de investigación futuras.

2 Motivación

El software de generación de datos de prueba fue desarrollado en el marco de una Práctica Profesional Supervisada (PPS) de la carrera Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas y Tecnologías (FCEyT) Universidad Nacional de Santiago del Estero, en el marco del proyecto de investigación *"Métodos y Técnicas para desarrollos de Aplicaciones Ubicuas"* [5, 6, 7, 8] (23/C139) de la Secretaría de Ciencia y Tecnología de la Universidad Nacional de Santiago del Estero (SICYT UNSE).

Dicho software contribuyó directamente con uno de los objetivos del proyecto: *"Diseñar modelos y desarrollar aplicaciones ubicuas de impacto local y regional"*, dado que las aplicaciones ubicuas desarrolladas, deben ser probadas en entornos simulados antes de pasar a las pruebas en entornos reales.

Las técnicas de simulación son de especial utilidad para la generación de datos de prueba en los sistemas ubicuos porque éstos reciben como entrada una gran variedad de datos dinámicos lo que genera, a su vez, un alto número de escenarios posibles. A través de la simulación no sólo se pueden generar datos de prueba atendiendo a las particularidades de cada sistema, sino que es posible reducir costos y tiempo, que resultan considerablemente menores a los requeridos para la experimentación en el mundo real.

De esta forma se pueden realizar pruebas exhaustivas enfocadas en la funcionalidad y rendimiento utilizando datos simulados, para luego llevar a cabo pruebas específicas en entornos reales, orientadas a evaluar la satisfacción del usuario y la aplicabilidad de los sistemas.

3 Aporte del trabajo

La herramienta desarrollada toma diferentes datos de entrada y mediante las técnicas de simulación genera los valores de las variables identificadas en el ambiente ubicuo del sistema para poder validarlo. En base a las características de las variables para las cuales se generarán los datos de prueba, se definieron los tipos de distribuciones necesarios. Se emplearon los métodos de generación de números pseudoaleatorios congruencial mixto y uniforme, puesto que cumplen con las características que aseguran la confiabilidad de los resultados obtenidos [9].

Para desarrollar la aplicación se utilizó una arquitectura en capas. Empleando el lenguaje Python junto a las APIs Geoapify (para ubicación y puntos de interés),

Weatherapi (para datos meteorológicos) y el proyecto colaborativo OpenStreetMap (para creación de mapas).

La aplicación tiene como funcionalidad mostrar todos los puntos de interés de una categoría ubicados dentro de un radio. Mediante simulación se generan los valores de temperatura y humedad para dichos puntos. Como primera medida en la solapa "Búsqueda y Simulación" que se le presenta al usuario (Fig 1), este deberá seleccionar la forma en que desea proporcionar su ubicación o una ubicación ajena a él, estando entre las disponibles:

- Ciudad: se despliega una lista con todas las capitales de cada provincia que conforman Argentina.
- Latitud y longitud: se deben ingresar las coordenadas geográficas en grados decimales, por ej.: Latitud: 40.714, Longitud: -74.006 se corresponde a la Ciudad de Nueva York.
- Código postal: se puede ingresar como una de las siguientes opciones:
    - Primera opción: Ejemplo: G4200, Argentina o AR-G 4200, Argentina.
    - Segunda opción: Ejemplo: 4200, Argentina.
    - Tercera opción: Ejemplo: 4200.

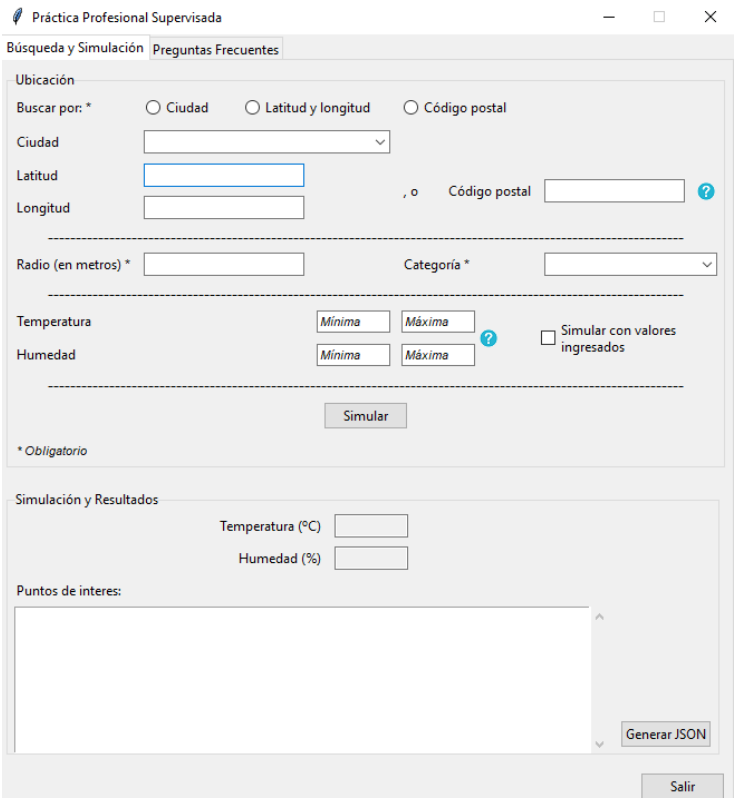


Fig. 1. Interfaz de la aplicación

Luego se debe ingresar y seleccionar un radio de búsqueda y categoría respectivamente, para acotar los puntos de interés (POIs) a mostrar. Como último apartado de la sección ubicación, se encuentra opcional el ingreso de valores mínimos y máximos de la temperatura y humedad para simular los datos de prueba finales correspondientes. De no ser ingresados, se simula con datos obtenidos automáticamente desde la API de clima.

Finalmente, se muestra en la sección simulación y resultados con sus respectivos datos característicos todos los puntos de interés encontrados (Fig. 2) y se muestran en el navegador empleando un mapa (Fig. 3). Además, la aplicación permite exportar todos los puntos generando un archivo del tipo JSON.

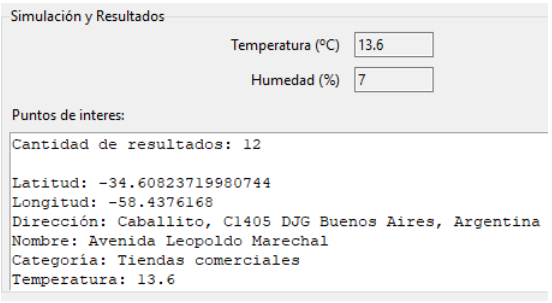


Fig. 2. Resultados de POIs obtenidos

Fig. 3. Mapa de POIs obtenidos

Para el empleo de una/s técnica/s de simulación se evaluaron los métodos teóricos para funciones continuas, siendo el más adecuado, el método de la distribución uniforme en donde la variable aleatoria se genera en un intervalo cualquiera (a, b) [9]. Para la generación del número pseudoaleatorio, también conocido como valor U o R, necesario para aplicar la distribución uniforme, se utilizó el método de generación congruencial mixto.

4 Trabajos Futuros

Se prevé ampliar la cantidad de datos contextuales a generar, incluyendo características del contexto social y de tarea del usuario, como por ejemplo, personas cercanas, tareas, actividades; características relacionadas a la salud del usuario como presión arterial, ritmo cardíaco, nivel de oxígeno en sangre, etc.

References

1. Sommerville, I. (2011). Ingeniería de Software. En I. Sommerville, & L. M. Castillo (Ed.), Ingeniería de Software (V. C. Olguín, Trad., Novena ed., págs. 28, 41, 206). Naucalpan de Juárez, México, México: Pearson.
2. Lara, C. C., "Las Técnicas de Simulación en la Validación del Software" Trabajo Final de Graduación Licenciatura en Sistemas de Información, Universidad Nacional de Santiago del Estero, Santiago del Estero, Argentina, 2011.
3. Weiser, M. (1991). La computadora del siglo XXI. The Computer for the 21st Century. Scientific American, 94-104.
4. Sakamura, K. Koshizuka, N. 2005. Ubiquitous Computing Technologies for Ubiquitous Learning. IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'05), Ieee, 11–20.
5. Unzaga, S., Durán, E. B., Álvarez, M., Salazar, N., Fernández Reuter, B., González, G., . . . Quintana Cancinos, F. (2020). Avances en métodos y técnicas para la construcción de aplicaciones basadas en computación ubicuas. Santiago del Estero, Santiago del Estero, Argentina.
6. Durán, E.B., Unzaga, S., & Álvarez, M. (2020). Instanciación del metamodelo de contexto para aplicaciones ubicuas. Santiago del Estero, Santiago del Estero, Argentina.
7. Durán, E. B., Unzaga, S., Álvarez, M. M., Salazar, N., González, G., Fernández Reuter, B., & Zachman, P. P. (2017). Métodos y técnicas para desarrollos de aplicaciones ubicuas. Santiago del Estero, Santiago del Estero, Argentina.
8. Unzaga, S., Durán, E. B., Álvarez, M., Salazar, N., Fernández Reuter, B., González, G., … Quintana Cancinos, F. (2020). Avances en métodos y técnicas para la construcción de aplicaciones basadas en computación ubicuas. Santiago del Estero, Santiago del Estero, Argentina.
9. Coss Bu, R. (2003). Simulación Un enfoque práctico. (G. Noriega, Ed.) Distrito Federal, México: Limusa.

UN MODELO DE CALIDAD PARA EL ANALISIS DE PROCESOS DE NEGOCIO DE UNA DISTRIBUIDORA DE PRODUCTOS ALIMENTICIOS

Santiago Castillo Elías, Carlos H. Salgado, Mario G. Peralta, Alberto A. Sánchez, Corina Abdelahad

Departamento de Informática, Facultad de Ciencias Físico-Matemáticas y Naturales
Universidad Nacional de San Luis
santicastilloeliasr@gmail.com, csalgado@unsl.edu.ar, mperalta@unsl.edu.ar, alfanego@unsl.edu.ar, corina.Abde@gmail.com

RESUMEN: Este trabajo se centró en la definición de un modelo de calidad basado en la ISO/IEC 25000 para la mejora de los procesos de negocio del sistema ChessERP, enfocado a distribuidoras de productos de consumo masivos de la ciudad de Rosario, donde trabajamos. Este trabajo de cátedra forma parte del trabajo por proyectos entre la materia Modelos de Calidad de Procesos del Desarrollo de Software de la Universidad Nacional de San Luis. El objetivo es poder continuar trabajándolo para que termine siendo el trabajo final de Licenciatura en Ciencias de la Computación.

Los modelos de calidad son aquellos documentos que integran la mayor parte de las mejores prácticas, proponen temas de administración en los que cada organización debe hacer énfasis, integran diferentes prácticas dirigidas a los procesos clave y permiten medir los avances en calidad.

El propósito de este trabajo está centrado en la construcción de un modelo de calidad, considerando que este es un elemento de suma importancia a la hora de determinar y definir los requisitos de la calidad del software desde el enfoque de la calidad de productos de software. El software es una de las herramientas de mayor utilidad en la optimización de procesos en las organizaciones, con el propósito de contar y ofrecer optimización, eficiencia y satisfacción de necesidades, razón por la cual el software debe contar con criterios que garanticen su calidad. Estas necesidades o requisitos explícitos y/o implícitos de la calidad se pueden especificar para el desarrollo de un software o bien para evaluar un producto de software ya construido.

Palabras Claves: Modelo de Calidad, Métricas, Norma ISO/IEC 25010, Procesos de Negocio, Software.

1. Introducción

Uno de los componentes principales de los sistemas informático lo constituye el software y la calidad de éste tendrá influencia directa en el sistema que lo contiene.

La calidad del software es presentada en la literatura a través de distintas definiciones, algunas de ellas son por ejemplo la expresada en [2], donde a la calidad de software se la define como el cumplimiento de los requisitos de funcionalidad y desempeño explícitamente establecidos, de los estándares de desarrollo explícitamente documentados, y de las características implícitas que se espera de todo software

desarrollado profesionalmente. También vemos que en ISO/IEC 25000 [1] se la define como el grado en que el producto software satisface las necesidades expresadas o implícitas, cuando es usado bajo condiciones determinadas.

Existen distintos enfoques de la calidad del software, éstos pueden ser, Calidad a nivel proceso, Calidad a nivel de producto y Calidad en uso, para cada uno de estos enfoques existen distintos tipos de modelo de calidad de software que se pueden aplicar, según se especifica en [3].

En ISO/IEC 25000 [1] se establece que un modelo de calidad es un conjunto definido de características, y de las relaciones entre ellas, que proporciona un marco de trabajo para especificar los requerimientos de la calidad y para evaluar dicha calidad.

También se explicita que el alcance de la aplicación de los modelos de calidad incluye el apoyo de la especificación y la evaluación de software y los sistemas informáticos intensivos de software desde perspectivas diferentes de los asociados con su adquisición, requisitos, desarrollo, uso, evaluación, soporte, mantenimiento, aseguramiento y control de la calidad y auditoría. Los modelos pueden, por ejemplo, ser utilizados por desarrolladores, adquirientes, personal de aseguramiento y de control de la calidad y evaluadores independientes, particularmente aquellos responsables de especificar y evaluar la calidad del producto de software.

Los Modelos de Calidad (MC), son instrumentos o artefactos específicamente diseñados y construidos para soportar la evaluación y selección de componentes de software. Permiten la definición estructurada de criterios de evaluación, la especificación de requerimientos, la descripción de componentes en relación a ellos y la identificación de desajustes de manera sistemática facilitando el proceso de evaluación y selección del software [4].

2. Modelo de Calidad ISO/IEC 25010.

En este apartado se exponen, de manera sintética, las principales características y funciones que puede cumplir el modelo de calidad establecido en la mencionada Norma ISO/IEC 25010 [5].

De manera general, el modelo de calidad que propone esta norma representa en un primer nivel las principales características de calidad que tendrá el modelo, éstas pueden subdividirse en una o varias subcaracterísticas de calidad, lo que luego permite asociarle los atributos necesarios, éstos últimos representan las cualidades o propiedades de calidad que el software debe satisfacer.

En la Norma ISO/IEC 25010 se definen dos tipos de modelos de calidad: calidad del producto y calidad en el uso. Si bien se hace una mención de ambos modelos, dada las características de este trabajo solo se desarrollará el modelo de calidad del producto.

Modelo de calidad del producto: Está compuesto por 8 características y 32 subcaracterísticas, éstas se explicitan en la Tabla 1, y se refieren a las propiedades estáticas del software y a las propiedades dinámicas del sistema informático. Este modelo es aplicable tanto a sistemas informáticos como a productos de software. A continuación en la Tabla 1 se presenta el modelo de calidad para producto de software.

Tabla 1. Modelo de Calidad del estándar ISO/IEC 25010 [1]

Características	Subcaracterísticas
Adaptación funcional	- Completitud funcional
	- Exactitud funcional
	- Adecuación funcional
Eficiencia del desempeño	- Comportamiento relativo al tiempo
	- Utilización de recursos
	- Capacidad
Compatibilidad	- Co-existencia
	- Interoperabilidad
Usabilidad	- Capacidad de reconocer la adecuación
	- Facilidad de aprendizaje
	- Operatividad
	- Protección de errores del usuario
	- Estética de la interfaz del usuario
	- Accesibilidad
Confiabilidad	- Madurez
	- Disponibilidad
	- Tolerancia a fallas
	- Capacidad de recuperación
Seguridad	- Confidencialidad
	- Integridad
	- No repudio
	- Rendición de cuentas
	- Autenticidad
Capacidad de mantenimiento	- Modularidad
	- Reutilización
	- Capacidad de ser analizado
	- Capacidad de ser modificado
	- Capacidad de ser probado
Portabilidad	- Adaptabilidad
	- Capacidad de instalación
	- Capacidad de ser reemplazado

3. Modelo Propuesto

El modelo de calidad propuesto, entre otras características y subcaracterísticas, hace foco en la: Seguridad, Satisfacción del Producto, Eficiencia y Flexibilidad. A continuación, se muestra como ejemplo la definición de características, subcaracterísticas, métricas, etc. del modelo de calidad propuesto:

Eficiencia: Recursos utilizados en relación a la precisión y la completitud con la que los usuarios alcanzan objetivos especificados.

- Tiempo de tarea: que se tarda en completar una tarea con éxito
  - Función de medición: X = T donde T = Tiempo de tarea
  - Aclaraciones: La capacidad de aprendizaje (ISO/IEC 25023) puede medirse por el tiempo que tarda un usuario normal en completar una tarea en comparación con el tiempo que tarda un experto, y cómo cambia con el uso repetido.

Satisfacción: Grado en el cual se satisfacen las necesidades del usuario cuando se utiliza un producto o un sistema en un contexto de uso especificado.

- Utilización de las características: La proporción de un conjunto identificado de usuarios del sistema que utilizan una característica particular

- Función de medición: X = A/B donde A = Número de usuarios que utilizan una característica particular y B = Número de usuarios en un conjunto identificado de usuarios del sistema.
- Aclaraciones: 1) Las características pueden ser definidas en diferentes niveles de granularidad, desde una función individual hasta un subconjunto de un sistema. 2) Un valor bajo podría indicar que la característica no es útil, o sólo es útil para un subconjunto de usuarios, o que los usuarios no entienden cómo usarla, o que no saben que existe.

Flexibilidad: Facilidad con la que un producto o elemento de software puede modificarse para cumplir con los requisitos adicionales del usuario

- Reutilización de componentes de Software: Cantidad de componentes usadas en un determinado modulo o componente, que han sido definidas en otro modulo o componente.
  - Función de medición: X = ∑Ai Ai = componente de software reutilizado.
  - Aclaraciones: Esta medición se obtiene a partir de la suma de componentes reutilizados para el desarrollo de un determinado componente de software analizando el código fuente. Componente se define de acuerdo al contexto en el que se aplique, pudiendo ser una librería, función, método, etc.

4. Caso de Estudio

Una vez que se contaba con el modelo de calidad y un conjunto de métricas e indicadores de calidad. Se procedió a realizar la instanciación y aplicación de las métricas sobre el producto ChessERP. Este es un software pensado para cubrir las operaciones específicas de una distribuidora de productos de consumo masivo, sumadas a las funciones habituales de cualquier sistema de gestión o ERP.

Permite obtener informes de gestión adecuados para la toma de decisiones e incrementar la productividad en los procesos de gestión de pedidos, picking y depósito, control de inventarios, logística y distribución, pricing, facturación, fuerza de ventas, compras, entre otros. Es una herramienta de gestión gerencial que permite calcular ventas segmentados por vendedor y supervisor, comparativas entre clientes, cobertura y rechazos, eficiencia de preventa, efectividad de visitas realizadas, seguimiento de inversión de equipos de frío, análisis financieros, entre tantas otras opciones disponibles. A partir del Modelo Propuesto previamente definido se obtuvieron los siguientes resultados sobre el producto ChessERP:

Métrica: Tiempo de tarea

- Muestra: Se midió el tiempo en realizar dos tareas: Alta de sucursal de una empresa y Alta de un talonario. Estas acciones las realizó un usuario con conocimientos del sistema, pero nunca había realizado esas acciones previamente.
- Resultados Obtenidos: Para el alta de una Sucursal nueva el tiempo de tarea fue de dos minutos y para el alta de un Talonario nuevo fue de 5 minutos. Se realizó las mismas acciones nuevamente y se tardó 1 minuto para el alta de Sucursal y 2 minutos para el alta Talonario.

Métrica: Utilización de las características

- Muestra: Se utilizó la herramienta de Google Analytics. Se ha analizado todos los módulos del sistema. El universo analizado es de 157511 usuarios.
- Resultados Obtenidos: El módulo Listado de Pedidos es el más utilizado con 9992 usuarios que utilizan la característica, por lo que se obtiene X = 0.063 o un 6.3% de usuarios utilizan dicha característica. El segundo módulo más utilizado es Visualización de un Comprobante con 5827 usuarios que utilizan la característica, por lo que se obtiene X = 0.037 o un 3.7% de usuarios utilizan dicha característica.

Métrica: Reutilización de componentes de Software

- Muestra: Se ha analizado un módulo del código fuente Front End del Módulo de Proveedores. Esta desarrollado en JavaScript, HTML y CSS utilizando el framework Angular v12. Dicho Módulo posee 5 componentes Angular.
- Resultados Obtenidos: En el módulo analizado se ha encontrado la reutilización de 7 componentes Angular dentro de las 4 componentes que conforman el módulo.

5. Conclusiones

A modo de conclusión, a partir de las métricas analizadas sobre el producto ChessERP se obtuvieron las siguientes conclusiones. En la métrica *"Tiempo de tarea"* se observa una diferencia de aprox. 50% entre las dos mediciones por lo que sugiere existe una mejora posible en cuanto a Ayuda al usuario para realizar las acciones mencionadas. En la métrica *"Utilización de las características"* se observa que el modulo más utilizado es *"Listado de Pedidos"*, esto se debe a que dicho modulo muestra los distintos tipos de pedidos presentes de las empresas como facturas, remitos, notas de crédito, etc, que es de vital importancia en el negocio. La métrica *"Reutilización de componentes de Software"* se obtuvo que aproximadamente por cada componente se reutilizaron dos componentes definidas en otros módulos, esta métrica es importante en el desarrollo ya que reutilizar software permite mejorar la productividad y la calidad del desarrollo. La métrica *"Confidencialidad por funcionalidad"* se observa que la confidencialidad no es uniforme en los módulos analizados, esto sugiere estandarizar un cierto grado de protección mínima que cumplan todos los módulos para asegurar cierto grado de confidencialidad.

Referencias

1. ISO/IEC 25000 Systems and software engineering-Systems and software Quality Requirements and Evaluation (SQuaRE)-System and software quality models.
2. R. Pressman, Ingenieria del Software. 6ª Ed: Mcgraw-Hill, 2005.
3. Callejas-Cuervo, M.; Alarcón-Aldana, A.C.; Álvarez-Carreño, A.M. Modelos de calidad del software, un estado del arte. En: Entramado. Enero - Junio, 2017. vol. 13, no. 1.
4. A. Villalta, J.P. Carvallo "Modelos de calidad de software: Una revisión sistemática de la literatura" en Maskana, CEDIA 2015.
5. ISO/IEC 25010: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality Models

Integración de una red de sensores con una plataforma IoT para control inteligente de aulas

Lucas Gómez D´Orazio [(1)], Santiago Medina [(1)], Diego Montezanti [(1)]

[1] Instituto de Investigación en Informática LIDI (III-LIDI),
Facultad de Informática, Universidad Nacional de La Plata – Comisión de Investigaciones Científicas de la Provincia de Buenos Aires

lucas.dorazio@alu.ing.unlp.edu.ar
{smedina,dmontezanti}@lidi.info.unlp.edu.ar

Resumen. Los recientes avances en las tecnologías de conectividad y sensado constituyen un escenario propicio para las soluciones basadas en IoT. Actualmente existe una variedad de herramientas que ayudan a explotar el potencial de estas soluciones, entre ellas las Plataformas IoT, complementadas con niveles de procesamiento cercanos a los bordes de la red. En este artículo se describe la implementación de un sistema para monitoreo y control inteligente de parámetros de interés en un edificio universitario basado en una plataforma IoT, y se brindan los resultados preliminares de la funcionalidad desplegada y las pruebas de conectividad realizadas entre los nodos sensores y el servidor.

Palabras claves: Plataforma IoT, Edge y Fog Computing, CoAP, MQTT, monitorización, control inteligente

1 Introducción y objetivos

En los últimos años, la Internet de las Cosas (IoT) ha ganado protagonismo debido a su característica de permitir la conexión de una gran cantidad de dispositivos sensores que obtienen información del entorno y enviarla a servicios en la nube para su procesamiento [1]. Esto se ha utilizado extensamente para desarrollar aplicaciones inteligentes, como gestión de tráfico, hogares inteligentes, monitoreo de fenómenos naturales y salud humana [2].

Debido a la creciente popularidad de IoT, la cantidad de dispositivos conectados a Internet se ha incrementado significativamente. Como resultado, se genera una enorme cantidad de tráfico de red, lo que conduce a cuellos de botella y puede producir limitaciones respecto a las latencias de comunicación con el *cloud* y al ancho de banda de la red [3], por lo que las infraestructuras tradicionales basadas en el *cloud* no resultan suficientes para las demandas actuales de las aplicaciones de IoT [4]. Para lidiar con estos inconvenientes, en los últimos años han surgido los paradigmas de *Fog Computing* y *Edge Computing*, que alivian estas sobrecargas trasladando parte de la potencia de cómputo cerca de los bordes de la red y lejos de los servidores centrales en la nube [1]. De esta forma, el cómputo de los datos de IoT se distribuye, permitiendo reducir las latencias de comunicación [3].

Una plataforma de IoT es un conjunto de servicios o sistemas de software que trabajan conjuntamente con nodos sensores y actuadores conectados mediante Internet. La plataforma brinda las herramientas para capturar, almacenar, procesar y presentar los datos obtenidos de los sistemas embebidos, haciendo uso de protocolos de comunicación IoT. Para ello, debe proporcionar servicios de administración de nodos finales, administración de redes y conectividad, procesamiento y análisis de datos, desarrollo de aplicaciones, seguridad, almacenamiento en bases de datos, herramientas de visualización, monitoreo y control de acceso, entre otros [4][5].

La funcionalidad de una plataforma de IoT viene dada por la interacción de sus seis bloques constitutivos: el bloque de identificación, el bloque de sensado, el bloque de comunicación (que contiene los protocolos que sirven para intercambiar datos entre los objetos conectados y el sistema de administración), el bloque computacional, el bloque de servicios y el bloque semántico [6].

También, dependiendo de la funcionalidad necesaria para una aplicación particular, se puede desplegar una solución a medida que implemente los servicios necesarios, como el almacenamiento, la visualización o la analítica de datos, prescindiendo de la utilización de una plataforma de IoT comercial.

Este trabajo se enmarca dentro de la Práctica Profesional Supervisada del alumno avanzado en la carrera Ingeniería en Computación Lucas Gómez D´Orazio. En el mismo se describen y analizan los resultados iniciales de la integración de una red de sensores con una plataforma de IoT específica (ThingsBoard [7]), y con un conjunto de servicios individuales, en la búsqueda de desplegar un sistema de monitoreo y control inteligente de aulas en un edificio universitario. La experimentación realizada apunta a generar capas de *Edge* y *Fog Computing* para la gestión de parte de la funcionalidad del proyecto.

2 Plataformas, servicios y protocolos

Thingsboard es una plataforma IoT de código abierto que permite almacenar, visualizar y analizar datos con diferentes servicios, como múltiples motores de base datos, generación de alertas, llamados a procedimientos remotos, flujos de trabajo basados en eventos y diseño de tableros dinámicos, entre otros [7]. Su bloque de comunicaciones permite la transmisión de datos entre los dispositivos conectados y el sistema de administración, a través de los protocolos estándar para IoT que se utilizan en la industria, como MQTT, CoAP y HTTP.

CoAP [8] (*Constrained Application Protocol*) es un protocolo basado en HTTP que usa comunicaciones uno a uno, y se utiliza en el hardware de IoT. Debido a esto, debe ser liviano y generar poco tráfico, por lo que utiliza UDP sobre IP. En tanto, MQTT [9] (*Message Queue Telemetry Transport*) es un protocolo de comunicaciones liviano implementado sobre TCP/IP. Usa un servidor *broker* en medio de los dispositivos que se comunican, por lo que no es comunicación M2M. Consiste en tres elementos: publicador, suscriptor y el *broker*. Los clientes publican y se suscriben a tópicos en el *broker*.

En el caso de una red WAN, MQTT se comporta mejor debido a la existencia del *broker* que se encuentra entre los dispositivos que se comunican, por lo que es útil cuando hay ancho de banda limitado. En tanto, CoAP es una buena opción para

aplicaciones basadas en servicios web. Se utiliza cuando los dispositivos necesitan transmitir y recibir a gran velocidad, ya que UDP tiene soporte para *multicast* y *broadcast* [6].

En tanto, en el bloque computacional existen diferentes plataformas de hardware diseñadas para ejecutar específicamente aplicaciones de IoT, como Intel Galileo, Raspberry PI o Arduino. Del mismo modo, hay plataformas de software que proveen las funcionalidades requeridas por las aplicaciones de IoT; entre ellas, están las plataformas a nivel de la nube, que permiten el procesamiento de los datos en tiempo real y ayudan al usuario final a obtener conocimiento extraído de grandes volúmenes de datos.

3 Desarrollo, experimentación y resultados preliminares

El objetivo del trabajo es el despliegue y configuración de una red de sensores de CO2 en el contexto de un edificio universitario cuyas aulas se van a monitorear y controlar remotamente [10]. Los sensores de CO2 en cada aula están conectados vía WiFi a una Raspberry Pi (que corresponde al nivel de Edge), que recolecta los datos de cada nodo a través de peticiones HTTP y los pre-procesa, identificando los valores provenientes de cada sensor, adaptándolos al formato adecuado (JSON) y encapsulándolos como mensajes MQTT para enviarlos al servidor ThingsBoard, tal como se muestra en la Figura 1. Este servidor centraliza el monitoreo de los niveles de CO2 de todo el edificio, evaluando la coherencia de los valores recibidos y configurando el motor de reglas de la plataforma para emitir alertas en caso de valores no deseados. Un ejemplo del panel de control y visualización (o *dashboard*) de ThingsBoard se puede observar en la Figura 2. Adicionalmente, en el nodo de nivel de Edge se desplegaron servicios individuales de base de datos de series temporales (Influx DB) y visualizador de datos (Grafana) que brindan alternativas locales a algunas de las funcionalidades provistas por la plataforma de IoT.

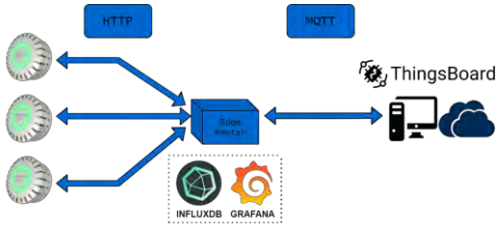


Fig 1. Arquitectura desplegada para la medición de CO2

De las pruebas realizadas con MQTT surge que el mínimo intervalo de tiempo entre envíos de mensajes que garantiza la recepción y el procesamiento de todos los mensajes es de 1.3 milisegundos. Al aumentar la frecuencia de envío, comienza a haber pérdida de mensajes. Sin embargo, las cantidades absolutas de mensajes recibidos continúan en incremento, hasta alcanzar el límite de frecuencia de envío de un mensaje cada 0.5 milisegundos. A partir de ese valor, la plataforma ya ha alcanzado la saturación en la recepción de mensajes, por lo que si se continúa con la reducción del intervalo,

comienzan a disminuir los mensajes recibidos, además del esperable aumento en las tasas de pérdidas. Las cantidades de mensajes recibidos y perdidos se observan en la Figura 3.

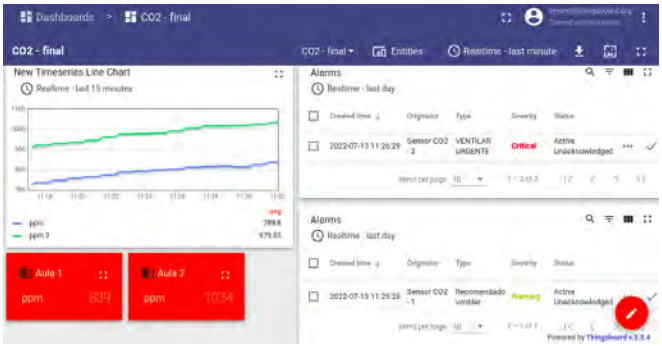


Fig. 2. Panel de monitoreo y control de ThingsBoard.

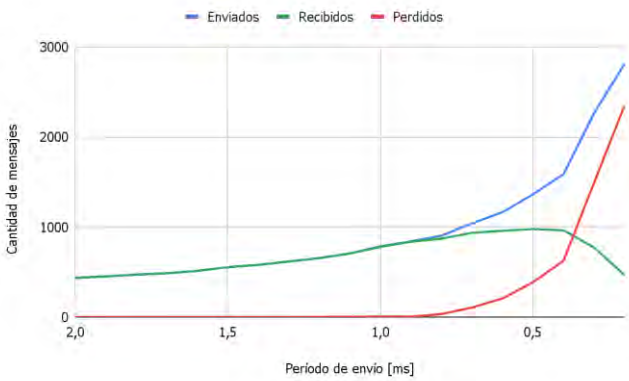


Fig. 3. Relación entre mensajes MQTT recibidos y perdidos al reducir el intervalo entre envíos.

Por último, es importante mencionar que para llegar a los máximos valores en la recepción es necesario desactivar el Algoritmo de Nagle de la comunicación TCP. Al enviarse cada mensaje en un paquete individual, se garantiza que se respeta el intervalo de tiempo que la plataforma requiere para procesar las recepciones, lo cual no ocurre cuando varios mensajes MQTT se encapsulan en un único paquete TCP.

4 Conclusiones y líneas futuras

En este trabajo, en vías de desarrollo, se han alcanzado algunos primeros resultados satisfactorios. Se ha logrado desarrollar una solución de IoT para el monitoreo de CO2 en espacios cerrados. Debido a la baja cantidad de recursos de infraestructura

necesarios para el despliegue, consideramos que la solución implementada es ligera, viable y escalable.

La plataforma de IoT seleccionada para nuestra solución (ThingsBoard) provee una variedad de servicios (integrados entre sí) que facilitan el desarrollo de proyectos. En este desarrollo se exploraron los servicios de capa de transporte, el motor de reglas y las funcionalidades de almacenamiento y visualización de datos. Además, se evaluaron las limitaciones de la plataforma para recibir mensajes en función de la cantidad de memoria asignada a la cola de mensajes, así como también el máximo valor aceptable de frecuencia de envío para los requerimientos de la solución.

En cuanto a los trabajos futuros con la plataforma, se planea investigar su integración con otros servicios externos capaces de potenciar las soluciones IoT, como también investigar los beneficios y desventajas de utilizar una versión reducida orientada al *edge computing*. Además, en el presente trabajo se utilizó la versión de ThingsBoard basada en una arquitectura monolítica, por lo que está pendiente probar la versión basada en microservicios.

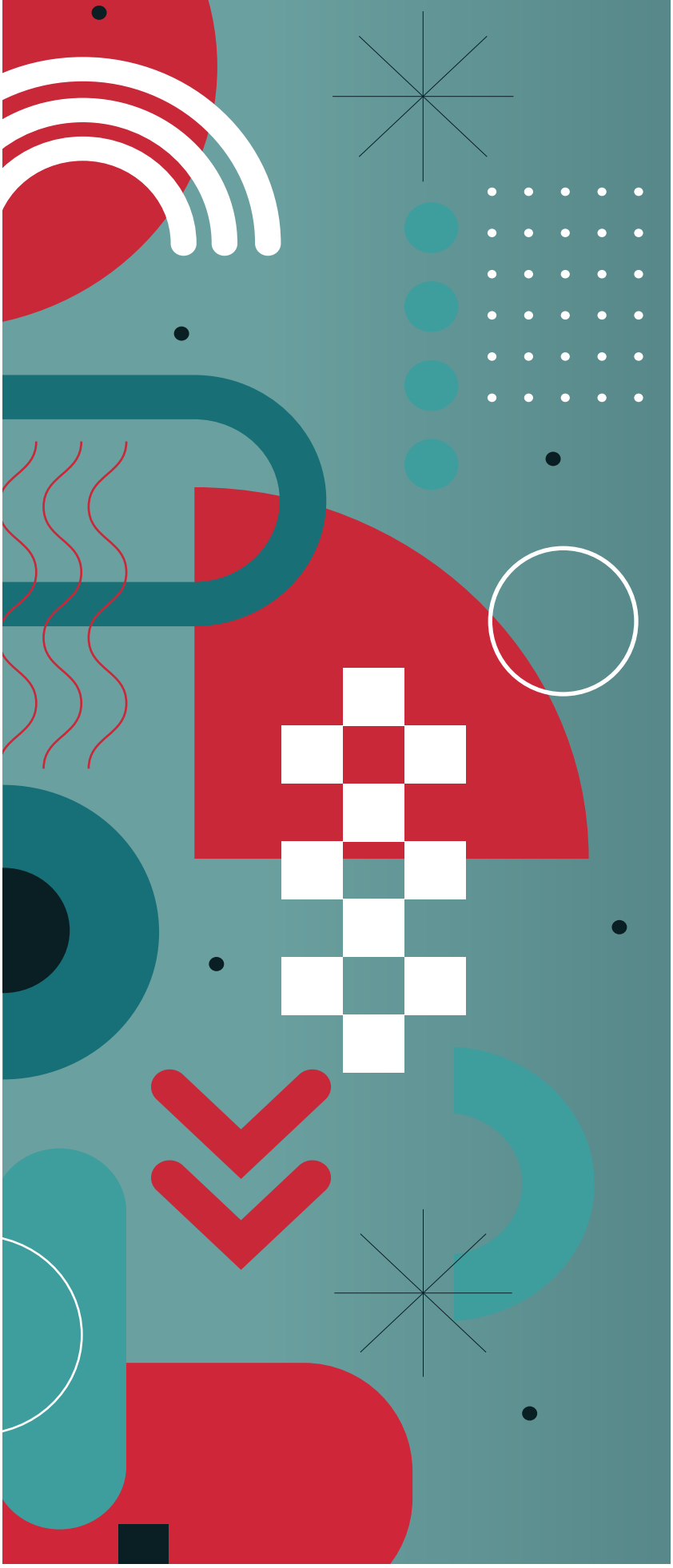
Respecto a las comunicaciones, resulta necesario realizar pruebas orientadas a hallar una relación de compromiso entre la utilización eficiente de la red (relacionada con el empaquetamiento de mensajes) y la maximización de la tasa de recepción. Además, hemos comenzado a realizar pruebas de comunicación entre los nodos sensores y la plataforma mediante CoAP, cuyos resultados nos permitirán comparar el desempeño del sistema con ambos protocolos.

Por último, y continuando con la línea de trabajo planteada en [11], se planifica ampliar la funcionalidad del sistema, agregando en las aulas nodos para la medición y control del consumo energético, conectados también al servidor ThingsBoard. De ese modo, se podrá contar con un sistema de monitoreo y control inteligente de parámetros de interés en los edificios objetivo, implementado como una solución IoT.

Referencias

1. Mohan, N., & Kangasharju, J. (2016, November). Edge-fog cloud: A distributed cloud for internet of things computations. In 2016 Cloudification of the Internet of Things (CIoT) (pp. 1-6). IEEE.
2. Tong, Y., Tian, L., Lin, L., & Wang, Z. (2020). Fault Tolerance Mechanism Combining Static Backup and Dynamic Timing Monitoring for Cluster Heads. IEEE Access, 8, 43277-43288.
3. Karagiannis, V., Desai, N., Schulte, S., & Punnekkat, S. (2020). Addressing the node discovery problem in fog computing. In 2nd Workshop on Fog Computing and the IoT (FogIoT 2020). Schloss Dagstuhl-Leibniz-Zentrum für InformatikAuthor, F.: Article title. Journal 2(5), 99–110 (2016).
4. Ullah, M., Nardelli, P. H., Wolff, A., & Smolander, K. (2020). Twenty-one key factors to choose an iot platform: Theoretical framework and its applications. *IEEE Internet of Things Journal*, *7*(10), 10111-10119.
5. M. Fahmideh and D. Zowghi, "An exploration of IoT platform development", Information Systems, vol. 87, p. 101409, 2020.
6. Hejazi, H., Rajab, H., Cinkler, T., & Lengyel, L. (2018, January). Survey of platforms for massive IoT. In *2018 IEEE international conference on future IoT technologies (future IoT)* (pp. 1-8). IEEE.
7. ThingsBoard Homepage, https://thingsboard.io, accedido el 2022/08/20.

8. CoAP Homepage, https://coap.technology/, accedido el 2022/08/20.
9. MQTT Homepage, https://mqtt.org/, accedido el 2022/08/20.
10. De Antueno, J., Medina, S., De Giusti, L., & De Giusti, A. (2020). Analysis, Deployment and Integration of Platforms for Fog Computing. Journal of Computer Science and Technology, 20(2), e12-e12.
11. Medina, S., Montezanti, D. M., Gomez D'Orazio, L., Compagnucci, E., De Giusti, A. E., & Naiouf, M. (2022). Incorporating Resilience to Platforms based on Edge and Fog Computing. In X Jornadas de Cloud Computing, Big Data & Emerging Topics (La Plata, 2022).



cacic2022