



SHORT PAPERS

OF THE 8th

CONFERENCE ON CLOUD COMPUTING, BIG DATA & EMERGING TOPICS



Short Papers of the 8th Conference on Cloud Computing Conference, Big Data & Emerging Topics (JCC-BD&ET 2020)

La Plata, Buenos Aires, Argentina.
September 8–10, 2020

Short Papers of the 8th Conference on Cloud Computing, Big Data & Emerging Topics: JCC-BD&ET 2020 / editado por Armando De Giusti, Marcelo Naiouf, Franco Chichizola, Enzo Rucci, Laura De Giusti - 1a ed. La Plata: Universidad Nacional de La Plata. Facultad de Informática, 2020.

Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-950-34-1927-4



1. Conferencias. 2. Publicaciones Científicas. 3. Computación. I. De Giusti Armando, ed. II. Título.
CDD 004.071

Scientific Committee

Coordinator: Armando De Giusti (*UNLP, Argentina*)

Abásolo, María José (UNLP, Argentina)
Aguilar, José (Universidad de Los Andes, Venezuela)
Ardenghi, Jorge (UNS, Argentina)
Balladini, Javier (UNCOMA, Argentina)
Bria, Oscar (UNLP, Argentina)
Castro, Silvia (UNS, Argentina)
Chichizola, Franco (UNLP, Argentina)
De Giusti, Laura (UNLP, Argentina)
Denham, Mónica (UNRN, Argentina)
Diaz, Javier (UNLP, Argentina)
Doallo, Ramón (Universidad da Coruña, España)
Errecalde, Marcelo (UNSL, Argentina)
Estevez, Elsa (UNS, Argentina)
Fernandez Bariviera, Aurelio (Universitat Rovira i Virgili, España)
Fрати, Fernando Emmanuel (UNDeC, Argentina)
Garcia Garino, Carlos (UNCuyo, Argentina)
Gaudiani, Adriana Angélica (UNGS, Argentina)
Gil Costa, Graciela Verónica (UNSL, Argentina)
Guerrero, Roberto (UNSL, Argentina)
Hasperué, Waldo (UNLP, Argentina)
Igual Peña, Francisco Daniel (Universidad Complutense de Madrid, España)
Janowski, Tomasz (Gdansk University of Technology, Polonia)
Lanzarini, Laura (UNLP, Argentina)
Leguizamón, Guillermo (UNSL, Argentina)
Luciano, Edimara (Pontificia Universidade Católica do Rio Grande do Sul, Brasil)
Luque Fadón, Emilio (Universidad Autónoma de Barcelona, España)
Marín, Mauricio (Universidad de Santiago de Chile, Chile)
Marrone, Luis (UNLP, Argentina)
Naiouf, Marcelo (UNLP, Argentina)
Olcoz Herrero, Katzalin (Universidad Complutense de Madrid, España)
Olivas Varela, José Angel (Universidad de Castilla-La Mancha, España)
Pardo, Xoan (Universidad da Coruña, España)
Piccoli, María Fabiana (UNSL, Argentina)
Piñuel, Luis (Universidad Complutense de Madrid, España)
Pousa, Adrian (UNLP, Argentina)
Printista, Marcela (UNSL, Argentina)
Rexachs del Rosario, Dolores Isabel (Universidad Autónoma de Barcelona, España)
Rodríguez, Nelson (UNSJ, Argentina)
Rucci, Enzo (UNLP, Argentina)
Saez Alcaide, Juan Carlos (Universidad Complutense de Madrid, España)
Sánchez, Aurora (Universidad Católica del Norte, Chile)

Sanz, Victoria (UNLP, Argentina)
Suppi, Remo (Universidad Autónoma de Barcelona, España)
Tirado, Francisco (Universidad Complutense de Madrid, España)
Tourinho Dominguez, Juan (Universidad da Coruña, España)
Viale Pereira, Gabriela (Danube University Krems, Austria)
Zarza, Gonzalo (Globant, Argentina)

Contents

Cloud and High-Performance Computing	1
Integration of Sensor Networks with Cloud Computing. <i>Santiago Medina, Fernando Romero, Fernando G. Tinetti</i>	2
Cloud TACs: OpenStack and Learning and Knowledge Technologies for teaching - learning of IT Infrastructures using and manipulating technologies. <i>Guillermo Baldino, Damian Ferrara, Ivan Añasco, Luciano Heredia, Nahuel Baez, Leopoldo Nahuel, Javier Marchesini</i>	6
Collaborative, distributed and scalable platform based on mobile, cloud, micro services and containers for intensive computing tasks. <i>David Petrocelli, Armando De Giusti, Marcelo Naiouf</i>	10
Finger-vein individuals identification on massive databases. <i>Sebastián Guidet, Ricardo J. Barrientos, Fernando Emmanuel Frati, Ruber Hernández-García</i>	14
Performance Analysis and Optimizations Techniques for Legacy Code Numerical Simulations. <i>Federico J. Diaz, Fernando G. Tinetti</i>	18
Artificial and Computational Intelligence	22
AI for Hate Speech Detection in Social Media. <i>Andrés Montoro, José A. Olivas, Adan Nieto</i>	23
Are statistics and machine learning enough to make predictions and forecasts? <i>Antonio Lorenzo, José A. Olivas</i>	27
Dynamic Data Driven approach to improve the performance of a river simulation. <i>Adriana Gaudiani, Emilio Luque</i>	31
Evaluation of the quality of the "Montecarlo plus K-means" heuristics using benchmark functions. <i>Maria Harita, Alvaro Wong, Dolores Rexachs, Emilio Luque</i>	36
From Fuzzy Deformable Prototypes to Elastic Patterns: Preliminary proposal. <i>Ruben Rodriguez-Cardos, José A. Olivas</i>	40
Towards Smart Data Technologies for Big Data Analytics. <i>Maria José Basgall, Marcelo Naiouf, Francisco Herrera, Alberto Fernández</i>	44
E-Government and Data Quality	48
A framework for linking open environmental data. <i>Juan Santiago Preisegger, Ariel Pasini, Patricia Pesado</i>	49
Framework for Data Quality Evaluation Based on ISO/IEC 25012 and ISO/IEC 25024. <i>Julieta Calabrese, Silvia Esponda, Patricia Pesado</i>	53

Cloud and High-Performance Computing

Integration of Sensor Networks with Cloud Computing

Santiago Medina, Fernando Romero, Fernando G. Tinetti¹

III-LIDI, Facultad de Informática Universidad Nacional de La Plata, La Plata, Argentina

¹ CIC – Comisión de Investigaciones Científicas de la Pcia. de Buenos Aires

{smedina, fromero, fernando}@lidi.info.unlp.edu.ar

Abstract. This article presents a possible approach for the integration of wireless sensor networks with applied IoT (Internet of Things) platforms related to Fog Computing concepts. Several ideas are presented related to: a) definition of the main characteristics to be considered in the deployment of sensor networks, b) evaluation of alternatives for combining wireless communication technologies and, c) experimentation with different IoT platforms for local data pre-processing that will then optimize the data flow of information to the cloud.

Keywords: Sensor Network, LoRa, WiFi, WSN.

1 Introduction

Wireless Sensor Networks (WSN) are one of the fastest growing and most widely applied data processing systems in recent years. WSN are an alternative concept of the MANET (Mobile ad hoc network), focused directly on the interaction with the environment where they are deployed rather than with people [1] [2].

One of the first development areas has been in military applications aimed at surveillance in conflict zones. Among some of the first projects with similarities to the current characteristics of sensor networks we could mention: a) The Chain Home Project, a ring of early warning radars that could detect and track aircraft during World War II, b) The SOSUS project, used in the cold war by the United States Army to track Soviet submarines, and c) The NORAD (North American Aerospace Defense Command) Project, also developed during the Cold War, for control and air defense in the United States.

The growth of Sensor Networks is directly related to the evolution in the development of microcontrollers and communication modules. Sensor network nodes are based on the combination of these two types of hardware. Microcontrollers usually centralize and convert the data generated by their connected sensors and, then, send them through a wireless communication module. Besides, communication devices are the basis of the network architecture as they provide the channels for the data flow. There are many wireless communication technologies used in the WSN, such as Bluetooth, WiFi, Zigbee, LoRa (Long Range), etc. Depending on the type of network to be deployed, the distances to be covered and the type of power available, a detailed analysis must be carried out to select the correct technology.

The application of WSN are varied and constantly growing, being military applications, control systems in agriculture, home automation, and climate sensor networks. Also, the WSN are directly related to the concept of IoT, interconnecting objects (*things*) that are part of our daily life to the Internet, integrating services that facilitate daily tasks of people. Connecting devices to the internet, specifically to Cloud Computing environments, generates a massive flow of data that can bring latency problems and increase several costs involved. In this context, Fog Computing introduces the idea of generating an intermediate instance of processing and control between the devices and the cloud in order to locally compute as much as possible with limited power and storage so as to reduce the data to be sent to the cloud [3] [4].

2 Main Tasks

The main aspects to consider in the deployment of a wireless sensor network connected to the cloud have to be specifically defined. Initially, it is necessary to test wireless communication technologies and analyze the platforms that allow the connection to the cloud with an intermediate level of processing and storage. We are currently working on two specific lines of work: low-power with long-range wireless technologies and IoT platforms.

Wireless technologies have evolved to the so-called LPWANs (Low Power Wide Area Networks). In LPWANs there are several options that allow including nodes in distances of the order of kilometers, while requiring low power consumption. We are analyzing LoRa, specifically in terms of communication integrity and reliability. LoRa, is a radio modulation technology developed by the company Semtech, which defines and owns the physical layer. The data communication layers (called LoRaWAN) are openly developed by a non-profit organization called the Lora Alliance. Semtech is responsible for marketing the devices [5] [6]. It enables long-range, low-power communication very well-suited for sensor networks.

There are several options for sensor data processing. Sending the collected data directly to the cloud imposes an eventually high network bandwidth as well as a proportional cloud data processing (and sometimes storage) facilities. In many applications, it is necessary to act according to the processing results, with its corresponding response time. More sensors usually imply more data, increasing costs for the application deploying and maintenance. We are approaching this problem by means of IoT platforms that run locally and are responsible for being part of the data processing. Thus, we expect to reduce the amount of data sent to the cloud as well as to reduce their value. In this way, the system is including tasks and processes related to the concept of Fog Computing. There are several platforms enabling to build the complete application including Fog Computing such as thinger.io, Node-RED, AWS Greengrass, as well as others [7] [8] [9]. Those platforms have similar operating concepts, they a) allow the creation and control of virtual devices that represent real nodes of the network, b) provide a context for managing data flow, and storage and handling of operations to be executed in the nodes.

3 Methodology

We are testing a basic network topology integrating two wireless communication technologies: WiFi and LoRa, as shown in Fig. 1. While long-range data transfers are provided by LoRa, small WiFi subnets are set up for relatively short-range data transfers (about 50 meters). WiFi subnets may include a relatively large number of Sensor Nodes as well as a single LoRa transceiver acting as a centralized subnet Router Node which also takes the role of network gateway role. Each sensor node is handled by a NodeMCU development board based on the ESP8266 [10]. Heltec WiFi LoRa 32 development board [11] was used for the Router Node, which is based on the ESP32 [12] and includes a LoRa SX1276 transceiver module. The whole network in Fig. 1 also includes a Master Node, which will have access to the cloud or to a standard serve-like computer with access to the cloud.

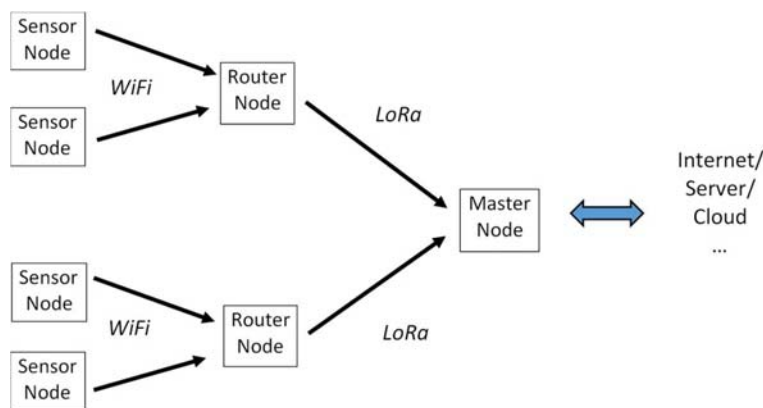


Fig. 1. Network Architecture.

The experiments were focused on identifying three data/network characteristics: transmission distance, integrity of the communication and energy consumption of the nodes [13]. A local instance of Thingier.io was used as an IoT platform directly integrated to the cloud. Thingier.io handles virtual devices as nodes and generates data flows for those nodes. It was tested specifically with a node Wi-Fi interconnection network. This platform allows a simple data management context on the devices as well as a specific communication API (Application-Programming Interface) [14]. For the completion of the tests, it is planned to integrate the Node-RED platform [15] for easier communication between the LoRa gateway and the Thingier.io server.

4 Main Tasks

We have presented a general approach for integrating WSN to the cloud. Even when the approach is not conceptually original, our focus is on defining and experimenting

the main characteristics of fog computing-like deployments. We have started our work with a state-of-the-art proof-of-concept installation. Many of the technologies, hardware modules and APIs are relatively new and we are currently gaining experience on them, in our first experiments. The short-term immediate performance experiments will be oriented to characterize wireless communications details such as modules distances, rural and urban environments, average power consumption, and amount of data handled per node.

Once we effectively integrate an IoT platform to some cloud service/s we will start the performance analysis via experimentation work. Initially, some performance analysis will be focused on IoT, cloud computing, and general integration functionalities. Once defined the most important deployment functionalities, the experimentation focus will be shifted to timing parameters like communication latency and bandwidth and hardware computing and storage requirements.

References

1. Sohraby, K., Minoli, D., Znati T.: Wireless sensor networks: technology, protocols, and applications. John Wiley & Sons (2007).
2. Akyildiz, I., Vuran, M.: Wireless sensor networks (Vol. 4). John Wiley & Sons (2010).
3. Bonomi, Flavio, et al. "Fog computing and its role in the internet of things." Proceedings of the first edition of the MCC workshop on Mobile cloud computing (2012).
4. Asemani, Malihe, et al. "A Comprehensive Fog-Enabled Architecture for IoT Platforms." International Congress on High-Performance Computing and Big Data Analysis. Springer, Cham, (2019).
5. LoRa Documentation, www.semtech.com 11.
6. LoRaWAN Documentation, www.lora-alliance.org
7. Thinger.io Documentation, www.thinger.io
8. Node-RED Documentation, www.nodered.org
9. AWS GreenGrass, www.aws.amazon.com/es/greengrass/
10. ESP8266 Technical Reference Version 1.3, Espressif Systems, 2018.
https://www.espressif.com/sites/default/files/documentation/esp8266-technical_reference_en.pdf
11. WIFI LoRa 32 Documentation, <http://www.heltec.cn>
12. ESP32 Datasheet Version 2.3, Espressif Systems, 2018.
https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf
13. Medina, S., Romero, F., De Giusti, A. E., & Tinetti, F. G., "Experiencias de Análisis de Consumo Energético en Redes de Sensores". XXV Congreso Argentino de Ciencias de la Computación (La Plata, 2019).
14. Aghenta, Lawrence Oriaghe, and Mohammad Tariq Iqbal. "Low-Cost, Open Source IoT-Based SCADA System Design Using Thinger. IO and ESP32 Thing." Electronics 8.8 (2019).
15. Selvaperumal, Sathish Kumar, et al. "Integrated Wireless Monitoring System Using LoRa and Node-Red for University Building." Journal of Computational and Theoretical Nanoscience 16.8 (2019).

Cloud TAC: OpenStack and Technology Learning and Knowledge for teaching IT Infrastructure

Guillermo Baldino, Damian Ferrara, Ivan Añasco, Luciano Heredia, Nahuel Baez, Leopoldo Nahuel, Javier Marchesini

Group R & D applied to computing and computer systems -GIDAS
Technology National University (UTN) - Regional Faculty of La Plata (FRLP)
Av. 60 esq. 124 s / n, La Plata, Buenos Aires, Argentina
{gbaldino, dferrara, ianasco, lheredia, nbaez, lnahuel,
jmarchesini}@frlp.utn.edu.ar

Abstract. In today's university environment, most students are digital natives. Therefore, it is difficult to imagine their academic life without relating it to the various cloud tools for communication and collaborative work. In this context, university professors work in new scenarios of communication and collaborative work in the classroom. This represents a transformation in the teaching-learning process assisted by new ICTs in the cloud. Working in the cloud offers the opportunity to transmit new knowledge when using pedagogical strategies supported by computer technologies. With the combination of ICTs and modern teaching-learning processes, the concept of Learning and Knowledge Technologies (TAC) is valuable. This work exposes the academic experience of researching and developing Cloud Computing using an OpenStack configuration so that students can empower themselves with the knowledge and use of cloud technologies. Thus, to be able to teach concepts and practices on IT Infrastructure including activities such as: design, configuration, implementation and administration of a private cloud for academic uses.

Key words: TAC, OpenStack, IT Infrastructure, private clouds.

1 Introduction

Systems and technological tools currently going to the Cloud Computing paradigm, which is constantly growing. More and more companies and research groups are working together to take advantage of the opportunities offered by Cloud tools [1]. These they consume resources as services: SaaS (Software as a Service), PaaS (Platform as a Service) or IaaS (Infrastructure as a service).

Students generally need computing infrastructure resources to carry out academic software development professorships or workshops. To satisfy this need, virtual machines with different operating systems are offered, specific tools that have a network (VPN), among others.

In this way, we support learning processes on IT infrastructure management [2], giving space for teachers to guide students to manage and configure IaaS requests. These resources could be obtained simply, quickly and easily by means of Cloud Computing

through the IaaS services. Virtual machines, computing, storage, among others, can be ordered and managed by the same student upon request from the private cloud itself. The management and use of these technologies allow the student to be trained with the necessary technical-technological knowledge such as the analysis, design and development of software systems [3].

The main objective of this work is to facilitate the academic community teaching-learning process for these since students are active participants in the private cloud. Also show the theory in a practical way, so that the student can get closer to new technologies through practical contact.

2 Technologies used: OpenStack for academic cloud

OpenStack is a free and scalable software platform designed to offer public or private clouds, enabling IaaS. It has different components with specific functions and can be installed separately or together, depending on the distribution. This integration is through API, which each service offers and consumes. Thanks to these APIs, the services can communicate with each other and allow one service to be replaced by another with similar characteristics as long as the way of communication is respected. In other words, OpenStack is extensible and meets the needs of those who want to implement it [4]. The main components are [5]:

Compute (Nova): OpenStack core, designed to manage and automate groups of equipment resources, being able to work with virtualization technologies.

Object Storage (Swift) - Module responsible for redundant, scalable, and fault-tolerant storage of objects and files.

Networks (Neutron): In charge of network management.

Block Storage (Cinder) - Provides persistent block-level storage devices, allowing search and recovery of virtual machines.

Identity service (Keystone): Service that offers user authentication and security policies.

Image Service (Glance): Provides virtual machine creation, search and recovery service. Manages all images of operating systems.

Dashboard (Horizon) - Provides administrators and users with a graphical interface to access, provision, and automate cloud resources.

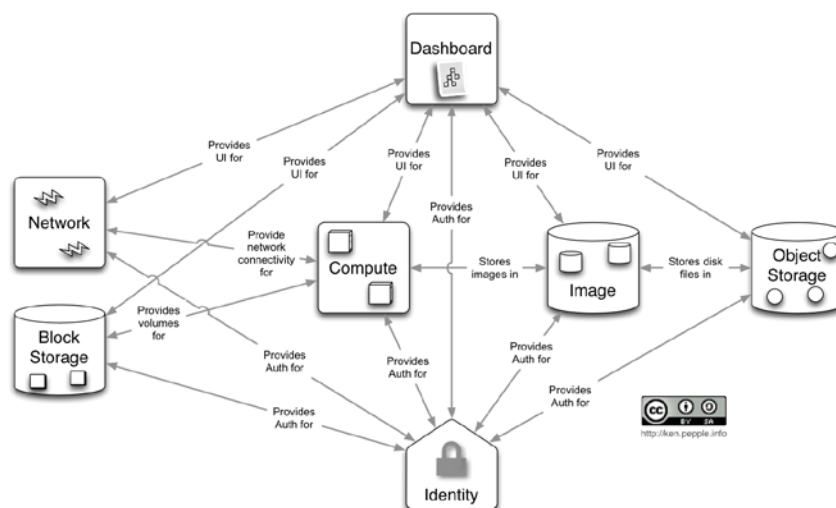


Fig. 1. Representative image of the intercommunication of modules in OpenStack [5]

3 Proposal

This work proposes to implement an IaaS platform based on OpenStack, which allows us to offer Cloud Computing resources to professors and students of the academic community of the UTN-FRLP. This promotes the use of free and open source software, as services such as Amazon Web Service, Microsoft Azure and Google Cloud are difficult to access due to their high licensing costs.

It seeks to implement and provide services through a private cloud, as part of the initiatives of the research area "Computer Science in Education and the 21st Century Classroom" of GIDAS. This technological experimentation is part of the Project approved by the UTN Secretariat for Science and Technology called "Computer innovation in learning and knowledge technologies applied to the improvement of educational processes". This work allows us to bring the use of OpenStack closer, not only providing the possibility of using virtual instances to the rest of the academic community, but also promoting its use and seeking to generate knowledge in students about the advantages of using this type of technology.

On the other hand, we consider that we are bringing a tool to centralize the deployment of software projects and execute them on different platforms. We simplify this task by giving all students the possibility of implementing their own virtual servers as part of a private cloud within the Faculty. During the 2019 academic year, the first experimentation experience with OpenStack was carried out in the subject of Resource

Management (hardware-software) of the 4th year in the Information Systems Engineering degree at UTN-FRLP. For this activity in the classroom, a workshop the OpenStack was designed with an installation and configuration guide so that students in the computing cabinet can deploy a Cloud and experience its use and administration. What allowed training in a practical way on the operation and resources that can be obtained from an IaaS. From this experience, it was possible to obtain feedback from the students for the use of these resources in various subjects and to improve the teaching mechanisms, being able to have virtual machines, computing, storage, among others, in a centralized way and that could be ordered and managed by the same student on demand from the private cloud itself. In addition, he provided us with initial information on the types of resources that will be consumed and those necessary to deploy Cloud Computing.

4 Conclusions and Future Work

From the design of these cloud technologies in the scope of the Faculty, it is expected to achieve an effective use of the use of TIC, putting not only knowledge but also tools from the perspective of usability, implementation and maintenance. This Project originates from the need to train researchers and thematic areas related to the management, deployment and manipulation of hardware technologies, within the scope of their use in the different professorships of the Career. In this aspect, the use of Openstack has been chosen as a starting point since its versatility and use allow activities to be carried out in such a way as to achieve an adequate transfer of knowledge in the research group, and its opportunity to transfer knowledge to different Chairs. . As future work, this project hopes to continue advancing in the implementation of a Private Cloud using its own technological resources. At the same time, incorporate new teaching-learning strategies through interaction with students and specific needs.

References

1. Murazzo, M.: Analysis of an open source Cloud Computing infrastructure. At: <http://sedici.unlp.edu.ar/handle/10915/53514> (2015).
2. Bustichi, G., Mosconi, E.: Applied methodologies for the development of autonomous learning. ISBN: 978-987-34-1796-6. II° Conference on Teaching Practices at the Public University (La Plata, 2018).
3. López, M.: From ICT to TAC: the importance of creating digital educational content. Didactic, Innovation and Multimedia Magazine (DIM). At: <http://www.pangea.org/dim/revista.htm>
4. Galarza, B., Zaccardi, G., Belizán, M., Duarte, D., Morales, M., Encinas, D. : Performance of the Cloud Computing for HPC: Deployment and Security. ISBN: 978-987-42-5143-5. At: <http://sedici.unlp.edu.ar/handle/10915/62576> (2017)
5. Openstack Components and Services. Last access 2020/11/06. <https://www.openstack.org/software/project-navigator/openstack-components>

Collaborative, distributed and scalable platform based on mobile, cloud, micro services and containers for intensive computing tasks

David Petrocelli^{1,2}, Armando De Giusti^{3,4} and Marcelo Naiouf³

¹ PhD Student at Computer Science School, La Plata National University, 50 and 120, La Plata, Argentina

² Professor and Researcher at Lujan National University, 5 and 7 routes, Luján, Argentina

³ Instituto de Investigación en Informática LIDI (III-LIDI), Computer Science School, La Plata National University - CIC-PBA, 50 and 120, Argentina

⁴ CONICET - National Council of Scientific and Technical Research, Argentina
dmpetrocelli@gmail.com, degiusti@lidi.info.unlp.edu.ar,
mnaiouf@lidi.info.unlp.edu.ar

Abstract. Compute-heavy workloads are traditionally run on x86-based HPC platforms and Intel, AMD or Nvidia GPUs; these require a high initial capital expense and ongoing maintenance costs. ARM-based mobile devices offer a radically different paradigm with substantially lower capital and maintenance costs and higher gains in performance and efficiency in recent years. When compared to their x-86 brethren, they have become ubiquitous in consumer markets and are making steady gains in the server market. Given this shifting computer paradigm, it is conceivable that a cost- and power-efficient solution for our world's data processing would include those very same ARM-based mobile devices while they are idling. Given that context, we developed and deployed an auto-scalable, distributed and redundant platform on the basis of a cloud-based service managed via container orchestration and microservices that are in charge of recycling and optimizing these idle resources. We tested the platform performing distributed video compression. We concluded the system allows for improvements in terms of scalability, flexibility, stability, efficiency, and cost for compute-heavy workloads.

Keywords: Kubernetes & Containers, Pipelines, Microservices, Cloud Computing & Storage, Mobile Computing, Distributed & Collaborative Computing

1 Introduction

Developing and deploying a high-quality, distributed, collaborative, and scalable software platform to process intensive tasks requires the adoption of the newest techniques, technologies, tools, and infrastructure patterns that allow taking advantage of all the available benefits on today's computing resources (On-Premise, Cloud and Mobile). We have implemented the following set of features: a) Build lightweight and more scalable applications (Microservices), b) Integrate auto-scalable infrastructure to guarantee

an cost and power efficient usage of resources (Container and Container Orchestration) and c) Reuse processing cycles from idle devices (Mobile devices)

Microservices allow for building smaller, lighter, reusable, and self-deployable software components which communicate through a simple and lightweight protocol [1]. Implementing a containerized [2] infrastructure allows microservices to be deployed and run as a naturally distributed application. This provides developers and operations engineers great benefits such as, nimble deployment of software changes simplifying the backups, replication and moving of applications and their dependencies. Operation engineers also use an orchestration layer to track which containers are running and to control, monitor, and scale (shrink or enlarge) applications [3].

In terms of computing power, mobile devices based on ARM chips have long idle periods while they are charging [4]; if properly managed them, they could become massively distributed data centers, consuming only a fraction of the energy [5] for the same computing power [6] as their traditional counterparts.

2 Collaborative, distributed and scalable platform for HPC

Based on the features we described earlier and considering our previous study [7], we developed and deployed the platform based on a model composed by a) Kubernetes container orchestration service (Cloud); b) Tasks management dockerized Microservices; c) Dockerized Queue Middleware and Database System; d) HTTP storage system; e) x86 and ARM-based mobile workers. (see Fig. 1).

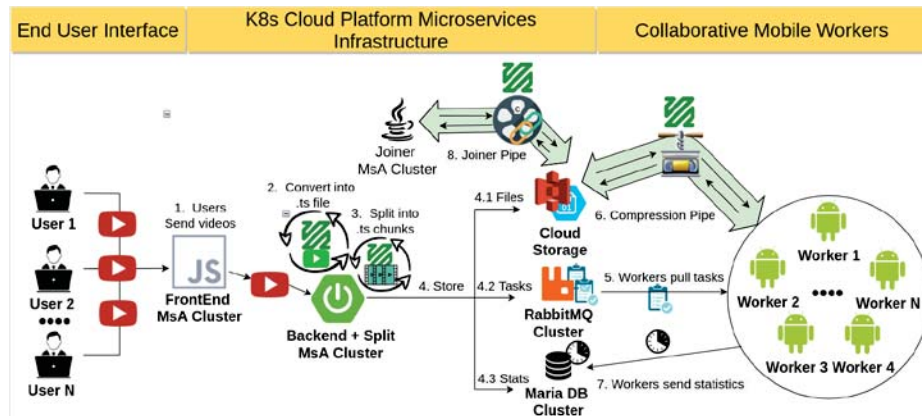


Fig. 1. Functional Diagram of the Collaborating Computing Network.

We used Kubernetes (K8s) AWS (EKS) and Azure (AKS) K8s SaaS clusters to host, orchestrate, heal and monitor our distributed Dockerized services [8]. We configured an auto-scalability mechanism based on CPU and memory container resources utilization thresholds, guaranteeing services (pods) high availability, efficient distribution and scaling (shrink/enlarge) across the cluster nodes. Finally, we configured DNS to interconnect containers, provide traffic routing and load balancing.

The front-end service allows clients to upload source video files and profile coding parameters. It also allows users to visualize information about their in-progress tasks and the result. Once tasks are received in backend microservices, source files are converted into a Video Transport Stream File (ts) due to is the recommended format for video streaming [9] and is tested impacts positively on latency, playback compatibility and viewing experience [10]. Once converted, files are split in smaller chunks [11] and stored in low-cost blob cloud storage [12]. We both use Azure Storage and Amazon S3 [12], to store data reliably and cheaply, upload and download stream files and distribute content via delivery networks (CDN) for lower latency and content caching.

Using official Bitnami RabbitMQ and MariaDB Kubernetes helm charts [13], we built an auto-scalable and fault-tolerant queue and database system where jobs and statistics are published respectively. RabbitMQ is used to securely and asynchronously store, publish and distribute backend service jobs to worker processing nodes. High Availability is guaranteed by RabbitMQ Policies and non-losing tasks are guaranteed by implementing a manual ACK mode model where should a server error, client-side issue, execution timeout happen, RabbitMQ thread moves the task back to queue. MariaDB is used to register job information (parameters, chunks, completed tasks and storage endpoint) the user interface periodically uses to be refreshed. Furthermore, it stores information about executed tasks (task, worker node and executed time). We use these statistics to evaluate the platform and worker efficiency in different scenarios. High Availability is configured via Galera active-active multi-master topology, guaranteeing scalability, smaller latencies, no slave service and no lost transactions.

Meanwhile, workers are continuously pulling from the RabbitMQ queue to obtain tasks and compress using the FFmpeg library. When chunks are completely processed, the Joiner backend microservice unifies parts and uploads it to the cloud storage endpoint. Both worker and joiner process tasks parallelly using Linux pipelines. While the source is stream downloaded via HTTP GET curl request, is also processed by FFmpeg and parallelly streamed to the cloud storage via HTTP PUT curl request. As a result, disk and memory operations are reduced, improving system performance.

3 Platform test and obtained results

So far, we have evaluated K8s microservices behaviour, scalability (defining auto-scale sets based on CPU and memory pod usage) and reliability. First, we forced crashing instances and verified K8s pod recreation and data integrity. Later, we stressed K8s services instances and checked horizontal scaling and load-balancing.

We have experimented with video compression tasks, following streaming best-practices [10][11], selecting representative source videos (see Table 1) and defining a set of h.264 compression profiles (see Table 2) to be executed on ARM-based and x86-based devices. Thus, we obtained performance and power usage metrics.

Table 1. Most relevant source videos features (codecs and bitrate) used for compression tests

Source Video File	Duration	Size	Size Screen	V. Codec	V. Bitrate	V. Prof.	V. Level	V. fps	A. Codec	A. Bitrate	A. Sample	A. Channel
3dmark_4k_120fps.mkv	2m 35 segs	487 MB	3840x2160	AVC x264	27545 Kbps	high	@L6	120	Vorbis	160	48000	2
bbb_4k_60fps.mp4	10 m 34 segs	642 MB	3840x2160	AVC x264	8000 Kbps	high	@L5.1	60	AC-3	320	48000	6
L.G_4k_30fps.mp4	1 m 6 segs	266 MB	3840x2160	AVC x264	34000 Kbps	high	@L5.1	30	aac	192	44100	2

Table 2. Most relevant compression profiles properties (size, codecs, resolution and bitrate)

Compression profile	Size Screen	V. Codec	V. Bitrate	V. Prof.	V. Level	V. Preset	A. Codec	A. Bitrate	A. Sample	A. Channel
4K Encoding	4096x2160	AVC x264	15600 Kbps	High	L@5.1	very slow	ac3	512 Kbps	48000	6
Full HD Encoding	1920x1080	AVC x264	3900 Kbps	High	L@4.1	slow	ac3	320 Kbps	48000	6
HD Encoding	1280x720	AVC x264	2000 Kbps	Main	L@4.1	medium	aac	320 Kbps	44100	2
480p Encoding	852x480	AVC x264	900 Kbps	Main	L@3.1	fast	aac	256 Kbps	44100	2





4 Preliminaries conclusions

K8s architecture running platform microservices, based on the experiments we made, might be considered as an interesting infrastructure for HPC tasks. The results showed stability, scalability, good response time and efficiency for different scenarios. Regarding mobile workers, we tested ARM devices are capable of encoding video with a competitive power and cost advantage over traditional x86 architecture. We have recently improved, via pipeline implementation, the worker stability, performance and efficiency.

References

1. Fritsch J., Bogner J. et al From Monolith to Microservices: A Classification of Refactoring Approaches. First International Workshop, DevOps 2018, France. (2018)
2. Matthias K., Kane S. Docker: Up & Running - 2nd Edition, Shipping Reliable Containers in Production. O'Reilly Media. (2018)
3. Dobies J., Wood J. Kubernetes Operators - Automation the container Orchestration Platform. O'Reilly Media. (2020)
4. Chethan K, Chiranthan H and D'Silva K. A Distributed Computing Infrastructure Using Smartphones. International Advanced Research Journal in Science, Engineering and Technology JSS Academy of Technical Education Vol. 4, Special Issue 8 (2017)
5. Pramanik P., Sinhababu N et al. Power Consumption Analysis, Measurement, Management, and Issues: A State-of-the-Art Review of Smartphone Battery and Energy Usage. IEEE Access 7. DO 10.1109/ACCESS.2019.2958684. (2019)
6. Andatech - The Snapdragon 865 Performance Preview: Setting the Stage for Flagship Android 2020, <https://bit.ly/2ZzFdd6>, last accessed 2020/05/20
7. Petrocelli, D., De Giusti, A. E. & Naiouf, M. Hybrid Elastic ARM&Cloud HPC Collaborative Platform for Generic Tasks. Springer. Communications in Computer and Information Science. Cloud Computing and Big Data (2019)
8. Hausenblas M., Schimanski S. Programming Kubernetes: Developing Cloud-Native Applications. O'Reilly Media. (2019).
9. ETSI - Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream, <https://bit.ly/36segJI>, last accessed 2020/05/20
10. 2019 Global Media Format Report, <https://bit.ly/2ZBmZli>, last accessed 2020/20/05
11. Dash Industry Forum - Guidelines for Implementation: DASH-IF Interoperability Points, <https://bit.ly/3ghFxTT>, last accessed 2020/05/20
12. Daher Z., Hajjdiab H. Cloud Storage Comparative Analysis Amazon Simple Storage vs. Microsoft Azure Blob Storage. Int. Journal of Machine Learning, Vol. 8, No. 1. (2018)
13. Bitnami Helm Charts, <https://bit.ly/3gsNci7>, last accessed 2020/20/05.

Finger-vein individuals identification on massive databases

Sebastián Guidet¹ , Ricardo J. Barrientos^{2,3} , Fernando Emmanuel Frati¹ , and Ruber Hernández-García² 

¹ Department of Basic and Technological Sciences,
Universidad Nacional de Chilecito, Chilecito, La Rioja, Argentina
sguidet@undec.edu.ar, fefrati@undec.edu.ar

² Laboratory of Technological Research in Pattern Recognition (LITRP),
Faculty of Engineering Sciences, Universidad Católica del Maule, Talca, Chile
rbarrientos@ucm.cl, rhernandez@ucm.cl

³ Department of Computer Science and Industries, Faculty of Engineering Science,
Universidad Católica del Maule, Talca, Chile

Abstract. In massive biometric identification, response times highly depend on the searching algorithms. Traditional systems operate with databases of up to 10,000 records. In large databases, with an increasing number of simultaneous queries, the system response time is a critical factor. This work proposes a GPU-based implementation for the matching process of finger-vein massive identification. Experimental results show that our approach solves up to 256 simultaneous queries on large databases achieving up to 136x.

Keywords: High Performance Computing, identification of individuals, local linear binary pattern, finger veins, GPU.

1 Introduction

Massive identification of individuals by using biometric techniques is a difficult problem in modern society. Particularly, vein-based biometric provides universality, distinctiveness, permanence, and acceptability. In the literature, different approaches based on finger-vein recognition [4] report several advantages of the finger-vein biometrics, such as high accuracy, high resistance to criminal manipulation (very difficult to copy or forge), authentication speed, compact size, liveness detection, and does not suffer damage or change over time.

On the other hand, the searching process on a biometric database consists of an exhaustive searching by calculating similarity metrics between the stored elements and the sample to be identified. As the size of the database increases, the identification accuracy decreases, while the recognition process execution time increases significantly. Besides, a high rate of queries per unit time is to be expected in this type of system. Thus, it is essential to ensure a reasonable response time [1].

This paper proposes a searching method based on GPGPU (general-purpose computing on graphics processing units) for finger-vein individuals identification

on massive databases. Our approach guarantees an adequate response time by using the Vertical Linear Binary Pattern (LLBP_v) descriptor and the Hamming distance as a similarity function. As far as we know, no real application has been proposed for finger-vein massive identification on GPU platforms, which is the main novelty of our approach.

2 Finger-vein identification system

A finger-vein identification system comprises of four main processes. Initially, a near-infrared (NIR) imaging device (700-1000 nm) captures an image of the finger-vein patterns. Later, the pre-processing stage obtains the region of interest (ROI) and enhances image quality. For ROI segmentation, we adopt the method proposed in [5], which is robust to finger movement and rotation. Also, the limited adaptive histogram equalization technique (CLAHE) is applied to adjust differences in illumination and contrast [6]. During the feature extraction process, the final enhanced image is represented by using the Vertical Local Line Binary Pattern (*LLBP_v*) descriptor. This descriptor decreases the computation-time and its straight-line shape extracts robust features from images with unclear veins [3]. Finally, the searching process is performed on the database by using a similarity function. The following section describes this last procedure in more detail.

3 Searching process on a massive database

Individuals identification consists of a 1:N exhaustive searching on the database, which means comparing the individual's sample against each record of the database. Each comparison computes a similarity score between the extracted binary code (i.e LLBP_v descriptor) and the stored codes. The Hamming's distance similarity function is adopted for this calculation due to its effectiveness for comparing binary codes [3]. When two codes correspond to the same finger, the similarity score tends to be 0. Instead, if the codes are from different fingers, the value is closer to 1.

The searching procedure returns a list of 32 records sorted by the similarity score in ascending order. We only obtain the first 32 results because it is the lowest perfect recognition range for *LLBP_v* with the best accuracy performance [2].

This process must be performed for every query received by the system. Thus, the workload of the system increases with a high rate of queries per unit time, and the data volume to be processed increases significantly, therefore the response time should be reduced and to use a GPU is a suitable solution.

3.1 GPU-based searching algorithm

Aiming to speed up the computation time of the searching process, our proposed solution divides the similarity calculation tasks among all GPU threads. Through coalescing access, consecutive threads access to consecutive memory addresses, facilitating I/O operations. The entire database is copied to the global memory (DRAM), and for decreasing the read operations latency, the records to

be processed are moved to the shared memory (Flash L1 type) of each CUDA Block.

Each GPU thread makes comparison calculations between the query to be identified and the records on the database. Our proposed algorithm uses the shared memory to store the query and also a set of *heap* data structure, and this implies that it is only possible to execute 128 threads per CUDA Block in our kernel. Each GPU thread keep the lowest distances found in its heap stored in shared memory. We choose a heap data structure because its efficiency in the insertion and extraction operations. As a result of this process, each CUDA Block reduces the database to 128 heaps of 32 records each. Then, taking account that a warp (32 threads) is the minimum execution unit in GPU, the threads of the first warp of each CUDA Block access the data previously found, and reduce them to elements stored in 32 heaps in shared memory. Finally, the first thread of each CUDA Block reduce the elements of the 32 previous heaps to just one heap with the 32 lower elements, also stored in shared memory. Each CUDA Block transfers its 32 elements to CPU, where all of them are merged, and a sequential quicksort algorithm is performed over these elements to select the lowest final distance.

When the system replies to multiple simultaneous queries, it repeats the whole process described for each query request.

4 Experimental work

The experimental environment consists of a GPU NVIDIA GeForce GTX 1080 Ti with 3584 CUDA cores and 11GB GDDR5X memory, and the host computer is composed of $2 \times$ Intel Xeon Gold 6140 CPU @ 2.30GHz, in total 36 physical cores, 24.75MB L3 cache and 128GB RAM.

To evaluate the performance of the proposed algorithm responding to a high query traffic, we use the BigFVDB dataset, which was generated in our previous work [2]. For these experiments we used 339,968 samples, due the capacity limits on the GPU global memory. It should be noted that a query in BigFVDB can only have one possible result, but it performs 1:N matching comparisons.

The experiments were performed by increasing the number of available queries on the system, starting with 4 up to 512 by increasing in power of two. To obtain an unbiased result and to guarantee the stability of the results, the time measurements were averaged by repeating each test 100 times. Besides, it was checked that in all experiments the same results were obtained for the same comparisons. In all cases, the maximum expected response time was 10 seconds (time threshold), which was defined based on previous evaluations [2]. However, it could vary according to the system requirements and the size of the database to be processed. Figure 1 summarizes the obtained results. It is worth highlighting that the proposed method solves up to 512 queries while keeping the response time less than the time threshold, keeping a very similar time in solving each different query.

It should be clarified that for the calculation of speed - up ($\text{Time(Sequential)} / \text{Time(Parallel)}$) in the Figure 1. (b) the time of execution of the sequential version in CPU was taken as reference.

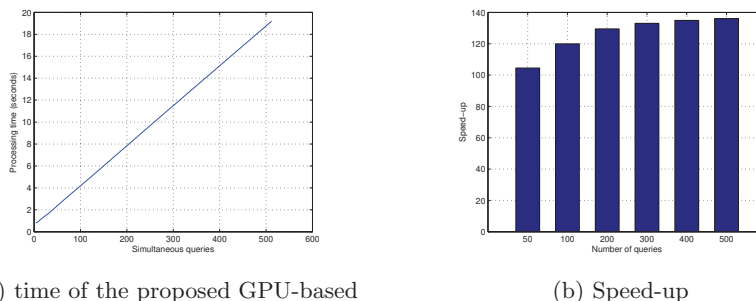


Fig. 1: Running time of our proposed GPU-based algorithm and its speed-up solving different quantity of queries.

5 Conclusions

This paper presents a GPU-based implementation for massive finger-vein individuals identification. The proposed method aims to reduce the computation time of the searching process over a high query traffic. We used a set of heaps as auxiliary structures to keep the lower elements found, and we also propose an algorithm in GPU to reduce the elements of the heaps to obtain the final query result.

The experimental validation shows that the proposed approach obtains a linear behavior as the workload increases. The proposed method keeps response times lower than 10 seconds with an increasing number of simultaneous queries, without requiring to increase the involved hardware resources, achieving up to 136x of speed-up solving 500 queries.

Future work plans to evaluate the proposed approach by increasing the number of individuals in the database, to reach 16 million records. Using a database of this size faces the issue of the overall memory capacity of the GPU.

References

1. Cappelli, R., Ferrara, M., Maltoni, D.: Large-scale fingerprint identification on gpu. *Information Sciences* 306, 1–20 (2015)
2. Hernández-García, R., Guidet, S., Barrientos, R.J., Frati, F.E.: Massive finger-vein identification based on local line binary pattern under parallel and distributed systems. In: 2019 38th International Conference of the Chilean Computer Science Society (SCCC). pp. 1–7. IEEE (2019)
3. Rosdi, B.A., W.Shing, C., Suandi, S.A.: Finger vein recognition using local line binary pattern. *Sensors* 11, 11357–11371 (2011)
4. Shaheed, K., Liu, H., Yang, G., Qureshi, I., Gou, J., Yin, Y.: A systematic review of finger vein recognition techniques. *Information* 9(9), 213 (2018)
5. Yang, L., Yang, G., Yin, Y., Xiao, R.: Sliding window-based region of interest extraction for finger vein images. *Sensors* 13(3), 3799–3815 (2013)
6. Zuiderveld, K.: Contrast Limited Adaptive Histogram Equalization. In: Heckbert, P.S. (ed.) Chapter VIII.5, *Graphics Gems IV*, pp. 474–485. Academic Press Professional, Inc. (1994)

Performance Analysis and Optimizations Techniques for Legacy Code Numerical Simulations

Federico J. Díaz and Fernando G. Tinetti¹

III-LIDI, Fac. de Informática, Universidad Nacional de La Plata, Argentina

¹Also with Comisión de Inv. Científicas de la Provincia de Buenos Aires

fernando@info.unlp.edu.ar

Abstract. Numerical simulations used today by scientists in various disciplines, are frequently based on implementations created when the predominant computing hardware was sequential by design. In this simulations, new features are added or updated, when new discoveries are made, but the computational implementation remains unchanged, not taking advantage of modern hardware architectures. This “legacy code” study cases, presents the opportunity to create a set of techniques and tools, oriented to perform optimizations from a computational and software engineering points of view. As an example, in conjunction with an astrophysics research group, a real-world case numerical integrator optimization is presented, were these techniques were applied, showing the results obtained.

Keywords: High Performance Computing, Optimization, Numerical integrators Legacy Numerical Software.

1 Introduction

Scientific disciplines often require complex numeric simulations to compute the models that describe real world/physical processes. The natural complexity of the real world requires that the simulations process large volumes of data, and perform computational-heavy calculations, in order to obtain the desired results.

The software used to compute these numerical simulations was often created in a time where the predominant computing technologies were sequential in nature. This software, today referred as “Legacy code” [1], is widely used amongst the scientific community.

Scientists update the implementation of the models, introducing new features as new discoveries are made in their respective disciplines. These new features frequently do not include modern optimization techniques, thus maintaining the sequential nature of the original implementation.

The shift of paradigm in current hardware design, moving away from sequential single-core processors to parallel and distributed computing creates a new opportunity to potential performance improvement of legacy code. But performing these optimiza-

tions requires a good knowledge on these mentioned parallel and distributed architectures, which is usually agnostic to the actual model implemented.

1.1 Performance Optimizations

We can mention 2 types of performance-oriented optimization categories, as follows:

- 1) **Automatic optimizations.** These optimizations are implemented through compiler options. A good example of them are the well-known optimizations flags `-O[1..3]`, which provide a very easy way of obtaining good performance results [2]. Other examples are the automatic inlining of functions, using the flag `-inline` on the ifort intel Fortran compiler [3], or the shared-memory parallelization performed by the OpenMP library [4], using compiler directives to generate automated threads in for-cycles.
- 2) **Manual optimizations.** These optimizations, require some understanding of the underlying code structure, in order to successfully obtain good performance results, that don't affect the execution numerical results. These are non-trivial, and require a good amount of profiling and research in order to be completely done.

1.2 Software engineering improvements

One of the characteristics of a “legacy code” implementation, is defined by software that is still being used today, but that have not been updated to modern software paradigms. The most predominant feature of Fortran code, is the usage of the GOTO statement.

While some of the GOTO usage can be removed automatically (like, for example, when it is used to create a FOR-like structure) [5], there are other cases where it requires manual interaction. Removing GOTO statements from legacy code, should be considered a must, before performing optimizations, if these statements are related to the portion of the program that needs to be optimized.

2 Performance Analysis of a Real-World Case

An example of a numerical integrator that has “legacy code” embedded into its core functionality, is the Mercury [6] N-Body integrator, and widely used by planetary astronomers around the world. Mercury is completely developed in Fortran, integrating the SWIFT [7] libraries for numerical simulation, and performs calculations of close-encounters in bodies. The Mercury integrator, has the ability to perform computations with “small bodies” and “large bodies”. The main difference between them, is that the small bodies don't produce interactions between them, only with a central star, and other large bodies, while the large bodies, do interact between each other, adding complexity to the simulation. The more “large bodies” used in the context, the more compute-intensive the simulation becomes. Hence, common simulations involve a mixture of large and small bodies, with hundreds of small bodies, and tens of large ones. For the performance analysis, the research group from the *Facultad de Ciencias*

Astronomicas y Geofisicas of the *Universidad Nacional de La Plata*, provided 2 real case scenarios, that used the simulator with 2 distinct execution paths. The first case was all small bodies, one large body, and the central star, and the second case, was a collection of large bodies, all interacting between each other.

2.1 Profiling Mercury

Initially, the GNU profiler gprof [8] was used to analyze the compute-intensive parts of the code, so that the key areas of Mercury to optimize were detected. There are some subprograms appearing in the top-ten most time consuming ones, for experiments with predominant “small bodies” and “large bodies”. The rest of the top-ten most time-consuming subprograms depend on the kind of bodies being simulated. We also implemented a wall-clock like time metric. This allows to measure the real-world time execution experienced by a human observer, as the gprof output only measures processor timing, not taking into account external interferences. Comparing both approaches indicated a significant difference: it was discovered that the input/output frequency and volume should be optimized as well.

3 Applying Optimizations

After the performance metrics were obtained, a number of optimizations have been applied. For each optimization applied, the numerical result was carefully controlled, so that consistency was maintained. When we found a numerical difference, the results were sent back to the scientists’ research group for approval. As a general method, it is always best to perform sequential optimizations first. Then, with the optimized code, move to implement parallel optimizations.

We applied several sequential optimizations (in critical subprograms) such as the automatic ones (-O2), I/O removal, removal of GoTo statements, and intrinsic operations replacement. We found that subprogram inlining along with operations reordering and redundant operations removal were the most successful in terms of providing performance enhancement.

The small bodies case was almost discarded for including parallel computing, because the small bodies do not interact, they only require a small amount of processing power to update their velocities and positions in each cycle. The large bodies case is particularly well suited for including parallel computing, as they have to interact between every other large body in the simulation to update their properties. Thus, the parallelization of the execution is a must to reduce the simulation time. In this case, a specific effort was made to eliminate data dependency computations for aiding the inclusion of parallel computing (e.g. via OpenMP further implementation).

4 Results and Future Works

Given the unoptimized initially legacy code, the sequential optimizations provided great performance gains. For the small bodies case, the performance gain was about 50% reduction in runtime. The large bodies case, was also parallelized, and the performance gain provided by the OpenMP implementation in 8 cores was about a 40% runtime reduction. We initially take all the optimization techniques as a guideline, given its application on a real-world numerical integrator. The potential of automatically apply some of them, has to be investigated further.

While the inline optimization is provided by some compilers (e.g. the Intel compiler), it cannot always be properly implemented. GoTo statements usually prevents the usage of many of the compilers optimizations, including inlining. As of now, there is no automated way of removing GoTo statements in general, but some of them do have a structure (e.g. GoTo statements used to replace For-like iteration structures).

The proper usage of cache memory, should be a topic of future research, with an increased number of bodies in the integrators. Also, the future line of work, of parallelizing further, using SIMD processors, like GP-GPU, could potentially increase the size of the input values, to use hundreds of thousands of elements.

References

1. Fernando G. Tinetti, Mariano Méndez, Armando De Giusti.: Restructuring Fortran legacy applications for parallel computing in multiprocessors. *The Journal of Supercomputing*, Volume 64, Issue 2, pp. 638-659.126 (2013).
2. GNU GCC Compiler Homepage, <https://gcc.gnu.org/onlinedocs/gcc/Optimize-Options.html>, last accessed 2020/03/30.
3. Intel Fortran Compiler Homepage, <https://software.intel.com/en-us/fortran-compiler-developer-guide-and-reference-inline-forceinline-and-noinline>, last accessed 2020/03/30.
4. OpenMP Architecture Review Board., “*OpenMP Application Programming Interface*”, *Version 5.0*, (2018).
5. Mariano Méndez, Fernando G. Tinetti.: Change-driven development for scientific software, *The Journal of Supercomputing*, Springer, Volume 73, Issue 5, pp. 2229-2257, (2017).
6. Chambers, J.E; Migliorini, F., “*Mercury - A New Software Package for Orbital Integrations*”, *Bull. American Astron. Soc.*, 29, 1024. (1997).
7. *SWIFT* Homepage, <http://www.boulder.swri.edu/~hal/swift.html> last accessed 2020/03/30.
8. Jay Fenlason, “*GNU gprof manual*” Homepage, <http://sourceware.org/binutils/docs/gprof/index.html>, last accessed 2020/03/30

Artificial and Computational Intelligence

AI for Hate Speech Detection in Social Media *

Andres Montoro¹, Jose A. Olivas¹[0000-0003-4172-4729] and Adan Nieto²[0000-0002-7899-4725]

¹ Department of Information Technologies and Systems, University of Castilla-La Mancha,
13071 Ciudad Real, Spain

`andres.montoro@alu.uclm.es`, `joseangel.olivas@uclm.es`

² Department of Public and Corporate Law, University of Castilla-La Mancha, 13071 Ciudad
Real, Spain

`adan.nieto@uclm.es`

Abstract. The main goal of this work focuses on solving the problem of analyzing the data coming from Social Media and exploring the mechanisms for the extraction and representation of knowledge from all the different disciplines outside the world of Information Technologies. Soft Computing and Big Data techniques are used to deal with the challenges mentioned. This paper shows a mechanism to detect hate speech in Social Media using Soft Computing and Sentiment Analysis, and it also establishes the base of a doctoral thesis.

Keywords: Soft Computing, Fuzzy Logic, Computing with words, Social Media Mining, Big Data, Text Mining.

1 Introduction.

In the age of Big Data millions of people are generating data in social media. This kind of data is unstructured, noisy, non-formatted and has variable length. Inside the universe of social media, the relations between entities (Social networks) becoming an extraordinary vehicle for the bulk dissemination of messages.

Effective social media analysis requires collecting information about individuals (users) and entities (social networks, sites, etc), analyzing the interactions between them and discovering patterns to understand human behavior [10].

The analysis of social media presents many challenges that basic techniques of Natural Language Processing (NLP) or Text Mining [1] cannot resolve. Some of these challenges are the Big Data Paradox [10], in other words, the volume of data from social media analysis is clearly Big, which also implies the problem of analyzing data in real time. Noise Removal Fallacy [10], social media data has a lot of noise and a blind removal of this can cause loss of knowledge and the definition of noise could change

* This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grants MERINET and SAFER: TIN2016-76843-C4-2-R and PID2019-104735RB-C42 (AEI/FEDER, UE).

according to the problem to be solved. Cross media data [9] i.e. how to exploit diverse data coming from social media (text, links, multilingual data, slang text, and so on).

2 Motivation.

Words play the main role in social media analysis and overall in human information processing. When we work with words we struggle with imprecision. The concept of computing with words was developed by Zadeh in [7]. In short, it's a field closely related with Fuzzy logic and Soft Computing in which the items to be computed are words, phrases and propositions drawn from a natural language [8].

Soft Computing is born as a set of techniques that groups the use of fuzzy methodologies. It was defined in 1994 by Zadeh [6] as a mixture of different methods that in one way or another cooperate from their foundations. The main components of Soft Computing are: Fuzzy Logic, Probabilistic Reasoning systems, Neural Networks, and either Evolutionary computing [2] or Metaheuristics [5].

Our investigation is focused in this field, using soft computing to deal with the problems found in social media analysis and extending the application of this field to other human disciplines like law or criminology.

3 Case study: Sentiment Analysis for the prevention of hate speech in social media.

The Internet has changed the conditions of communication in society and has become a new area of criminal opportunity different from that of the physical world [3] due to, among other things, its characteristics of neutrality, absence of censure and its constant development. This has led to a wider dissemination of hate crimes and therefore a greater effect.

In this case, it is developed a computational mechanism capable of identifying and classifying according to their intensity, hate messages in social media using techniques of Sentiment Analysis, Natural Language Processing and Fuzzy Logic. The starting point is a taxonomy designed from the current legality and the knowledge of an expert allows to determine the intensity of the hate speech and the particularities that compose it to inform of the pertinent decisions to be taken considering the prevalent legality and the corporate social responsibility of each company.

In literature there exist many approaches to identifying hate speech, in [4] resumes many of that using Natural Language Processing and Machine Learning. Some of its mentioned approaches are:

- Based in message characteristics.
- Using corpora to detect hate terms.
- Meta-information to encourage the model.
- Classification methods.
- Sentiment analysis.

3.1 Model.

All the previous approaches show the diversity of hate speech detection methodologies on the web.

The study not only identifies hate speech, it is also able to classify each message according to its intensity using knowledge engineering and soft computing. This process was developed following these phases:

1. Knowledge acquisition establishes the basis for the development of the taxonomy for the identification of Violent and Hateful Comments.
2. Extraction of ontology from the domain assists in the extraction of hate terms resulting from a message gathering experiment.
3. Taxonomy of violent and hateful communication. It is the result of modelling all the knowledge extracted in the form of a knowledge map.
4. Detection of violent and hateful communication using natural language processing techniques and the ontology.
5. Construction of fuzzy models based on the designed taxonomy, making use of linguistic labels extracted from the dataset using sentiment analysis techniques.

The result is a computer system capable of identifying and classifying hate messages in social media.

3.2 Prototype

Despite the knowledge extracted from the expert, the main source of knowledge is the article 510 of the Spanish Criminal Code. The interpretative framework is very wide, and the legally protected right refers to multiple groups. For the development of a functional prototype, the target group of hatred has been established to be Arabs and/or Muslims.

The prototype has been developed to detect hate messages in Spanish given the criminal awareness component of the model based on Spanish legislation. Its main design is based on the architecture of a fuzzy rule-based system adapted to the domain of the problem. This system is composed by the following steps:

- Input:
 - Gathering potential hate message.
- Fuzzification interface:
 - Obtaining values from ontology using natural language processing to extract relevant terms in the message.
- Inference mechanism:
 - Using taxonomy to label the extracted atomic expressions.
 - Establishing the linguistic labels of the proper fuzzy model, which is built with the knowledge base obtained with the rules derived from the knowledge extraction mechanism.
- Defuzzification interface:
 - Grouping values and extract membership grade.

- Output:
 - Reporting the intensity of hate speech and the reasoning flow.

Scaling to other languages would not be limiting. Thanks to the fuzzy system the taxonomy becomes a reasonably accurate metric for measuring the intensity of hate speech.

4 Future Work.

The main purpose of this paper is to show our investigation line and draw attention to Soft Computing and Big Data analytics like emergency fields in the coming years.

The immediate work focuses on improving the treatment of hate speech in social media exposed in the case study, detecting fake news from the perspective of soft computing and introducing the world of law using computational intelligence to model compliance system from organization analysis through risk assessment and prediction to recommendation systems. This is the starting point for the preparation of a doctoral thesis focused in applying Soft Computing and Big Data Analytics techniques to solve problems in different fields. One of them is law thanks to our collaboration with the Institute of European and International Criminal Law¹.

References

1. Aggarwal C.C., Zhai C. An Introduction to Text Mining. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. (2012).
2. Bonissone, P. P. Soft computing: the convergence of emerging reasoning technologies. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 1(1), 6–18. (1997).
3. Miró Llinares, F. *El cibercrimen. Fenomenología y criminología de la delincuencia en el ciberespacio* (1st ed.). Madrid, España: Marcial Pons (2012).
4. Schmidt, A., & Wiegand, M. A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. (2017).
5. Verdegay, J. L., Yager, R. R., & Bonissone, P. P. On heuristics as a fundamental constituent of soft computing. *Fuzzy Sets and Systems*, 159(7), 846–855. (2008).
6. Zadeh, L. A. Fuzzy logic and soft computing: issues, contentions and perspectives (pp. 1–2). *Proc. IIZUKA'94: 3rd Int. Conf. on Fuzzy Logic, Neural Nets and Soft Computing*, Iizuka, Japan (1994).
7. Zadeh, L. A. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2), 103–111. (1996).
8. Zadeh, L. A. *Computing with Words: Principal Concepts and Ideas (Studies in Fuzziness and Soft Computing)* (1st ed.). Heidelberg, Berlin: Springer (2012).
9. Zafarani, R., & Liu, H. Connecting Corresponding Identities across Communities. *International AAAI Conference on Web and Social Media*, 41–49. (2009).
10. Zafarani, Reza, Abbasi, M. A., & Liu, H. *Social Media Mining*. Cambridge University Press, (2009).

¹ <https://blog.uclm.es/idp/>

Are statistics and machine learning enough to make predictions and forecasts?*

Antonio Lorenzo^{1,2} [0000-0003-0752-6980] and José A. Olivas² [0000-0003-4172-4729]

¹ Coordinator of the Department of Business Intelligence, Castilla La Mancha Government, Toledo, Spain.
alorenzo@jccm.es

² SMILe (Soft Management of Internet and Learning). Information Technologies and Systems Institute, University of Castilla La Mancha, Ciudad Real, Spain.
JoseAngel.Olivas@uclm.es

Abstract. Currently the techniques used to predict the future are statistical and machine learning techniques. The first continues the trend of historical data. The second learns from previous cases training. Both use historical information but do not take in mind key factors that can make the final result change. A knowledge-based framework is presented that allows predictions of some kind of events to be made using artificial intelligence techniques. This requires an expert to enter the key factors that can change the trend of historical data into the system. The current framework has been applied prior to happening to two use cases, obtaining good preliminary results, in the framework of the developing of a PhD Thesis.

Keywords: Forecast, Trend, Prediction, Statistics, Machine Learning, Expert Knowledge.

1 Introduction.

Throughout the history of humanity, knowing what will happen in the future has been a constant. Knowing what the weather will be like tomorrow, how the stock market will behave or if your football team will win the next game, are questions that are asked daily. Statistical techniques have traditionally been used to predict the future. Lately machine learning techniques are being used. But these are not always enough because they do not take in mind some key factors that influence the final result. This doctoral thesis is born with the motivation to establish a framework to predict some types of events.

* This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET: TIN2016-76843-C4-2-R (AEI/FEDER, UE).

2 Prediction: Breaking the trend...

Reviewing the literature to know what terms are used to know future events, we have observed that two terms are commonly used to know the future: forecast and prediction. In these papers, both terms are used interchangeably, but they are not exactly the same. After this revision, the conclusion obtained is that both terms are used ambiguously, but in general, “forecast” refers to the analysis of data from a time series following the trend, and “prediction”, as the forecast plus other factors that can change the trend (Selvin et al. [1], Minh et al [2], Sezer et al. [3], Stuparu et al. [4]). In general, all forecasts are predictions, but not all predictions are forecasts.

The next step was to search the literature to find what techniques were used in forecasting and/or prediction. Traditional techniques use statistics with analysis of times series and regression Atsalakis et al. [5]. The newest techniques use artificial intelligence like machine learning (Garcés Ruiz et al. [6], Vaidehi, V. et al [7], Atsalakis et al. [8], Nojek, S. et al. [9]). After analyzing some papers, it is concluded that statistical techniques work well when the trend of historical data is maintained, and machine learning algorithms work correctly when the model has been previously trained with the type of case to be predicted. When the above conditions are not met, current prediction techniques are not enough.

We try to define a methodology and apply it to two case studies, based on knowledge, which allows predictions to be made in some cases because we want to improve the results of current techniques. The methodology starts from the analysis with the data of the current forecasting and/or prediction techniques, adding expert knowledge that indicates which elements, ideas or aspects may have a determining role in the result. For this, we are looking for an expert in the field been able to identify the key elements that have influenced the result. It is about looking for previous cases that are similar to the future event, establishing an analogy between both events. If previous events, some premises produced some results, in future events, we can establish that if they are part of the premises, they will also be part of the results.

For this, artificial intelligence techniques such as heuristics, rule-based systems, learning by analogy and case-based reasoning (CBR) are used. Analysis of historical and current data can only determine “what has happened” and “why it has happened”. If you want to determine “what will happen”, additional descriptive knowledge based on heuristics should be applied to the descriptive analysis of the data.

The methodology is not applicable in all scenarios. There are four scenarios to determine the future: the first is certain (practically 100% of the information is available). The second is forecasting (there is a linear relationship between the historical data and the results establishing a projection), the third is random (the results do not depend only on the historical data). The fourth scenario is prediction (much of the information is unknown and there is no linear relationship between the historical data and the outcome). The methodology is developed in this last scenario and is defined to predict the result of the event, not when the event will happen

The prediction is complex because the variables that form it are unknown, as well as the relationships between them. Making a prediction is difficult. The work done to date has consisted of defining a ten-step prediction methodology and it will be successful if

the methodology improves the results of current forecasting and/or prediction techniques. To simplify complexity and to be able to work with the problem of "prediction", knowledge must be represented and uncertainty must be managed. It is necessary to represent knowledge to identify what concepts and strategies have been used successfully in previous use cases to be formulated at a higher level of abstraction and can be used in other analogue use cases. Knowledge has an apparent simplicity for humans, but it is very complex to manage it artificially. All representation is an imperfect approximation of reality. There is no way of representing knowledge as rich as natural language. Knowledge in humans is not structured; instead the representation of knowledge in machines needs to be structured. To represent knowledge, logic, rules or semantic networks can be used [10]. None of them is complete. The representation of knowledge must allow identifying, model, representing and using that knowledge. Selecting the way of representing conditions, focusing on some aspects of reality and forgetting others. In the management of uncertainty, imprecision is something innate to the human being, both in his way of thinking and in his way of speaking. In the real world there are numerous sources of uncertainty. The information may be imprecise, incomplete and erroneous. Statements like "Luis is much older than Ana" are difficult to represent with predicates of bivaluated logic. Fuzzy logic manages imprecise quantifiers "quite", "often", "sometimes" ... Sometimes information is true but the defined model is imprecise. To manage uncertainty, certainty factors and fuzzy logic are used. The certainty factors expresses the reliability with which we can accept the hypothesis in the case of having the evidence. Fuzzy logic affirms that statements are more or less true in certain contexts and more or less false in a different one. To do this, it manages imprecision by indicating the degree of membership of its members to a set.

3 Cases studies.

The methodology has been applied to two case studies Lorenzo, A. et al. [11]. The aim in both cases was to predict the number of Deputies that each political party will obtain. The artificial intelligence technique of "Rules Based System" is applied. The methodology was applied prior to the celebration of both events and had different results. In April 2019 Spanish General Elections, the expert did not fully appreciate the keys factors. Surveys got a best result. In the General Elections of November 2019, the methodology improved the results predicted by the surveys. The expert fully agreed with the key factors that influenced the results.

4 Conclusions.

We begin with reviewing the literature to find out what techniques are used to make forecasts and predictions. The usual are statistical techniques and machine learning ones. The first case works well when the trend continues. The second case works well when the model has been previously trained with the type of case to predict. The main problem of prediction is complexity because, a priori, there are many variables and the relationship between them is unknown. To reduce complexity, artificial intelligence

techniques are used: heuristics, case-based reasoning, learning by analogy, and rule-based systems. The objective of the thesis is to propose a framework, based on knowledge, which allows predicting some events to improve the effectiveness of current techniques. We have applied it to two use cases before they occurred. In the first case, predicting the electoral results for the general elections in Spain in April 2019, did not work well because the expert did not correctly determine the key factors. In the second case, predicting electoral results in Spain in the general elections of November 2019, worked better than traditional methods because the expert determined the key factors well. The framework is not yet complete because to systematically work on the prediction problem, we must be able to generalize the problem and predict outcomes for new events. Therefore, the next steps we are working on are representing knowledge and managing uncertainty.

5 References.

1. Selvin, S., Vinayakumar, R., Gopalakrishnan, E. A., Menon, V. K., & Soman, K. P. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. In 2017 international conference on advances in computing, communications and informatics (icacci) pp. 1643-1647. IEEE.
2. Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, 6, pp. 55392-55404.
3. Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, p. 106181.
4. Stuparu, D., Bachmann, D., Bogaard, T., Twigt, D., Verkade, J., de Bruijn, K., & de Leeuw, A. (2017). Case studies of extended model-based flood forecasting: prediction of dike strength and flood impacts. In *EGU General Assembly Conference Abstracts Vol. 19*, p. 17173.
5. Atsalakis, G., & Valavanis, K. P. (2010). Surveying stock market forecasting techniques—Part I: Conventional methods. *Journal of Computational Optimization in Economics and Finance*, 2(1), pp. 45-92.
6. Garcés Ruiz, A., Molina Cabrera, A., Ocampo, T., & Mirledy, E. (2006). Stock market forecasting using intelligent techniques. *Tecnura* 9 (18), pp. 57-66.
7. Vaidehi, V., Monica, S., Mohamed Sheik Safeer, S., Deepika, M., & Sangeetha, S. (2008). A prediction system based on fuzzy logic. *Proc. of the World Congress on Engineering and Computer Science, WCECS 2008*.
8. Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying stock market forecasting techniques—Part II: Soft computing methods. *Expert Systems with Applications*, 36 (3), pp.5932-5941.
9. Nojek, S., Britos, P., Rossi, B., & García Martínez, R. (2003). Sales Forecast: Comparison of Neural Network Based Forecast versus Statistical Method. *Technical Reports in Software Engineering*, 5(1), pp.1-12. (In Spanish).
10. Brachman, R. J., & Levesque, H. J. (1985). *Readings in knowledge representation*. Morgan Kaufmann Publishers Inc.
11. Lorenzo, A., & Olivás, J. A. (2019). A Case Study of Forecasting Elections Results: Beyond Prediction based on Business Intelligence. *Journal of Computer Science & Technology*, 19 (2) pp. 143-152.

Dynamic Data Driven approach to improve the performance of a river simulation

Adriana Gaudiani¹ and Emilio Luque²

¹ Universidad Nacional de General Sarmiento, Instituto de Ciencias, Buenos Aires, Argentina
agaudiani@campus.ungs.edu.ar,

² Universitat Autònoma de Barcelona, Dept. de Arquitectura de Computadores y Sistemas Operativos, 08193, Bellaterra(Barcelona), España

Abstract. In this research we incorporate the contributions of the dynamic data driven systems development that is based on the possibility of incorporating data obtained in real time into an executing application, in particular a simulation. This paper reports on the first phase of our research in which we have used this idea to enhance the simulation quality of a river flow simulator by dynamic data inputs during the computational execution. We had presented an optimization methodology of this simulator model in previous works but in this opportunity, we could handle those time periods when a sudden level change takes place in the river and we could improve the forecasting prediction. These results are the path towards the development of an automatic calibration framework fed with real-time data.

Keywords: simulator optimization, dynamic data driven, model calibration, real-time data

1 Introduction

A simulation system is continually influenced by real time data for better analysis and prediction of the system under study but uncertainties is inherent in modeling studies. It is vital that these models should be equipped with robust calibration and uncertainty analysis techniques, as explained in [4]. The accuracy of computer simulation depends largely on having reliable data.

In previous researches, we dealt with uncertainty in the values of the input parameters of a river basin model and the impact of this uncertainty in estimating daily water height and forecasting reliability. We proposed an optimization via simulation methodology for enhancing the simulation quality looking for the best set of parameter values to calibrate the computational model [2]. To carry out this research work we used an hydrodynamic model, Ezeiza, for the Paraná River, which is located in Argentina. This system domain is characterized by a large number of parameters and finding the optimal set is a computationally intractable problem. To deal with this issue we implemented a heuristic in two-phases in order to reduce the search space of the parameter values, as we explain in detail in [3]. In order to make the optimization method clearer, we show an outline with its main idea in Fig. 1. We achieved a reduction of 13-19% in the

simulation errors compared to the classical simulation. In more recent work we apply this optimization scheme tacking advantage of the inherent continuity to the simulated system as we explain in [6].

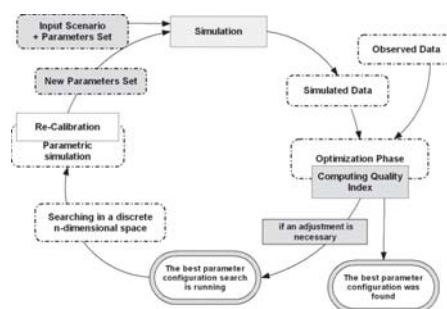


Fig. 1: Calibration process of the hydrodynamic model

In this paper we focus on enhancing our previous results and achieving better simulation quality. The improvement is based on the fact that the simulated system is a dynamic system and the adjusted parameters values of the model may change over time during the simulation. On the other hand, we try to extend this profit to most of the stations located in the domain of the Paraná River. Our main objective is to detect when the simulator needs a new calibration in a decision-making process. This process is carried out in a dynamic way and what we mean is that this detection occurs while simulation is running.

To address this problem we present a Dynamic Data Driven (DDD) approach to improve the Paraná River basin simulation based on our previous two-phases enhancing scheme. The simulation shows that the new method can give a 50% reduction in prediction error in Rosario city compared with our previous approach. One obstacle to implement successfully this new scheme is the rapid growth of computing time required to complete the simulation tuning. High performance computing provides the infrastructure and tools to handle this problem.

2 Dynamic Data Driven technique for a better prediction

In general, traditional simulations separate from real systems by using a few real-time data as say [1]. DDD can be considered as a new simulation paradigm aimed at achieving a better prediction of the behaviour of the system under study. As Xialun Hu says in [7] DDD Simulation connects Real-Time data with simulation and incorporating real-time data into a running simulation model has the potential to significantly improve simulation results. Many authors present their jobs in this line [5]

We propose to consider the assimilation of the real data, that is, the heights of the water in the river bed measured at each monitoring station. Without assimilating these observed data into the real system, the difference between the simulated data and the

real data is likely to grow on undesirable values. This is the main idea that motivates directing our research in this direction.

Incorporating real-time data from the system under study we are able to incorporate dynamic data into a running simulation model in order to greatly improve simulation results by computing the simulation error at each time step. This study prompted us to define a conceptual framework that dynamically detects the simulation error rate and makes the decision to start a new calibration. This approach requires predefining a threshold that will determine when a new calibration process is necessary. This threshold value was determined based on the knowledge we acquired of the system during the optimization work prior to this research.

The results obtained with the experimentation carried out to verify the effectiveness of the presented technique encourages us to begin the development of a conceptual environment that dynamically adjusts the simulation. Below we present the proposed methodology.

3 Methodology

We proposed a methodology guided by a continuous assimilation of the real data during the simulation. This research is based on our previous work and we need to deal with several challenges.

A dynamic assimilation function run at the same time with the river simulation which process the real time data and determining in a dynamic way the quality of the simulation, in order to forecast the water wave propagation as accurately as possible. It is therefore necessary that simulated and real data be fed into the DDD function and this process is responsible for stopping the simulation each time a new calibration is needed.

Figure 2 we can see the steps of the DDD assimilation module running together with the river simulator. In this graph, we show the successive calibration steps which will perform the search of a new set of adjusted parameters.

Each time that a search of the adjusted set of parameters is required it is performed as an optimization problem which minimize the quality index (QI) of each simulation scenario by using a parametric simulation. This optimization methodology take place in two phases which combines an optimization heuristics and a simulation analysis. We had to reduce the search space to make the optimization problem a computational treatable problem, as we described in [3].

4 Experimentation

The parameters used to calibrate the computational model are the Manning coefficient (Mn) and the leaves height (Lh). The value of these parameters must be set for each of the 75 sections necessary to discretize the domain of the system, that is, the Paraná River. We want to highlight that these values remain constant while the simulation is running since the simulation algorithm does not modify them. The DDD function is

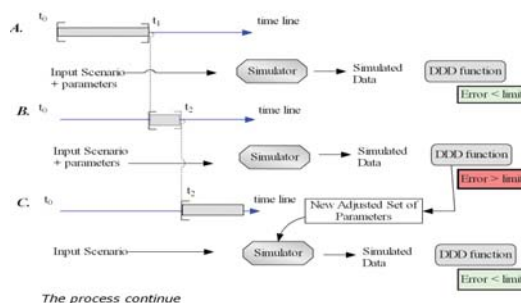


Fig. 2: DDD Optimization scheme and simulator running to calibrate the model dynamically.

continuously active testing the output values to take the decisions. Based on this knowledge, we launched a new simulation stage each time the DDD function detects a QI greater than a chosen threshold determined a priori.

For each simulation day we are receiving the measurements of the water height from 20 monitoring stations located in the river bed. We show in Figure 3 a) the difference between the simulated and real hydrograph for Rosario city and the different times that the simulation must be stopped. The simulation optimized by the new methodology is shown in Figure 3 b) and we want to remark the profit achieved. We compared the improvement achieved in Rosario Station and we obtained the following gain

- The QI in classical simulation is 0.59
- The QI in our OvS initial methodology is 0.51
- The QI in our DDD scheme is 0.29

In our previous work we obtained a gain of 13% but now we obtained a gain of 50%.

The whole process is computing time demanding due to the number of recalibration needed. This fact depends on the successive variations in the water level in the real system. We performed 5 calibrations and the whole computing job took 12 hours running on a 16 processors cluster.

5 Conclusion

We propose to handle the simulation errors arising from those events when a sudden level change takes place, which happens when water level raises or falls more rapidly than the rest of the period. This reflects the underlying model, but it is not our goal to change the simulator kernel and we have focused in a methodology that can deal with these disruptions.

The improvement achieved with our methodology reduced the simulation error about 50% compared with the best results reached in our previous research for Rosario station. Despite we improved forecasting prediction in one station as a pilot test we are encouraged to go on with the investigation. Currently, we are working on a framework to run the complete DDD optimization process and real-time data processing in order to achieve an automatic calibration.

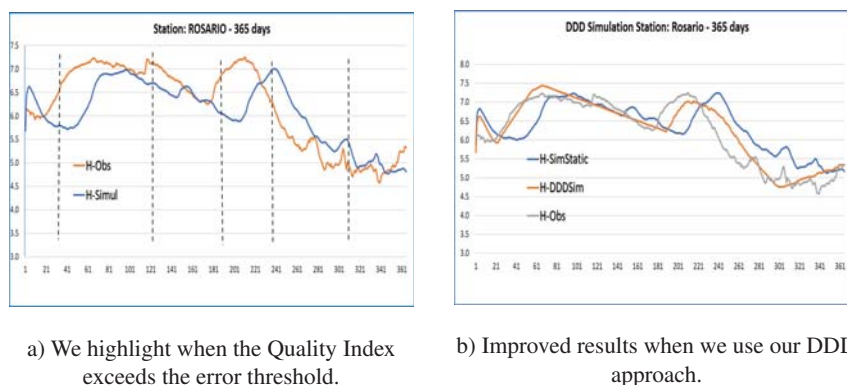


Fig. 3: Rosario forecasting improvement

We have to reduce the computing time because we need to improve as many stations as possible and the calibration should not last more than one day. That would be a reasonable time for a real-time flood forecasting taking into account the time it takes for the water wave to travel through the riverbed of the river.

References

1. Erik Blasch, Guna Seetharaman, and Frederica Darema. Dynamic data driven applications systems (dddas) modeling for automatic target recognition. In *Proceedings of SPIE - The International Society for Optical Engineering*, volume 8744, 2013.
2. A. Gaudiani, E. Luque, P. García, M. Re, M. Naiouf, and A. De Giusti. Computing, a powerful tool for improving the parameters simulation quality in flood prediction. *Procedia Computer Science*, 29:299 – 309, 2014. 2014 International Conference on Computational Science.
3. A. Gaudiani, E. Luque, P. García, M. Re, M. Naiouf, and A. De Giusti. How a computational method can help to improve the quality of river flood prediction by simulation. In M. Gomez, Sonnenschein M, and Vogel U., editors, *Advances and New Trends in Environmental and Energy Informatics.*, Progress in IS. Springer, Cham, 2016.
4. Ki Yong Lee, YoonJae Shin, YeonJeong Choe, SeonJeong Kim, Young-Kyoon Suh, Jeong Hwan Sa, and Kum Won Cho. Design and implementation of a data-driven simulation service system. In *Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory*, EDB '16, page 77–80, New York, NY, USA, 2016. Association for Computing Machinery.
5. E.E. Prudencio, P.T. Bauman, S.V. Williams, D. Faghihi, K. Ravi-Chandar, and J.T. Oden. A dynamic data driven application system for real-time monitoring of stochastic damage. *Procedia Computer Science*, 18:2056 – 2065, 2013. 2013 International Conference on Computational Science.
6. M. Trigila and Luque E. Gaudiani, A. and. Agile tuning method in successive steps for a river flow simulator. *Lecture Notes in Computer Science*, vol 10862:639–646, 2018.
7. Hu X. Dynamic data-driven simulation: Connecting real-time data with simulation. In Yilmaz L., editor, *Concepts and Methodologies for Modeling and Simulation. Simulation Foundations, Methods and Applications*. Springer, 2015.

Evaluation of the quality of the "Montecarlo plus K-means" heuristics using benchmark functions

Maria Harita  Alvaro Wong  Dolores Rexachs  and Emilio Luque 

Computer Architecture and Operating System Department,
Universitat Autònoma de Barcelona, Barcelona, Spain.
maria.haritar@gmail.com, alvaro.wong@uab.es, dolores.rexachs@uab.es,
emilio.luque@uab.es

Abstract. The evaluation in terms of quality of the results obtained from the use of a heuristic method is necessary to, first, verify the obtained results since heuristic methods do not guarantee to reach the optimum because all the possibilities are not fully explored. Secondly, it becomes interesting to validate such method, thus granting a high-quality index. Through our proposal, starting on the analysis of the literature survey on many optimization test functions, we are proposing the evaluation of a heuristic method based on Montecarlo approaches in conjunction with K-means clustering. Besides this, we aim to evaluate the results obtained through the use of some complex optimization test functions. Also, we seek to add a defined quality index to the original heuristic method relying on the consequent improvement in the results. As a side-work, we would aim to validate the heuristic mentioned above and optimize the algorithm in terms of scalability and quality.

Keywords: Heuristic method, Montecarlo, K-means, Benchmark functions.

1 Introduction

Optimization is a complex process defined as the process of finding the best solution to a given problem under certain conditions. In engineering, the objective of optimization could be to maximize the performance of a system using the minimum resources and the minimum execution time [1]. Heuristics are methods that help decide from a set of possible solutions to examine. Probabilistic methods typically consider the elements of the search space in further computations that have been selected by the heuristic. The Montecarlo-based approaches [2], usually trading the solution for a shorter runtime, which doesn't mean that the result is incorrect, just not the global optimal.

To the purpose of this work, we will analyze a problem that requires the simultaneous optimization of more than one objective, that is, a multi-objective optimization. There are extensive works dedicated to the operation of this type of methods [3] [4] [5]. Montecarlo plus K-means methodology was defined and a process was developed for improving the operation of Emergency Departments [6], based on the assumption that this is a complex system, difficult to characterize and naturally dynamic.

Our proposal develops from the idea of evaluating the Montecarlo plus K-means heuristic method using optimization test functions to evaluate the quality of the solutions provided by this heuristic. To prove the quality of optimization algorithms, optimization test functions are designed to mimic different types of workload[3]. They have many vital characteristics, such as its relevance, repeatability, scalability, transparency, and cost-effectiveness.

The heuristic method Montecarlo plus K-means[6] is explained through the next sections. Then, our proposal is delimited. In the methodology, we explain how we will proceed. Finally, in the last section, we present the conclusions and future work.

2 Related Work

The computational optimization techniques include a variety of optimization algorithms[1]. In terms of probabilistic methods, we find the heuristics which are a part of an optimization algorithm that uses information gathered by the algorithm to help decide which solution between the candidates should be tested.

The Montecarlo methods use randomness to solve problems, useful when it becomes difficult to apply other approaches. These methods are used for simulating the behaviour of many different types of systems that rely on repeated random sampling to obtain numerical results and incorporates random numbers. It is used primarily in three problem cases: optimization, numerical integration and generating draws from a probability distribution.

Emergency Departments, as described by Cabrera et al.[6], comprise a very complex system; there are a large number of possible configurations with the capacity to provide urgent medical attention and care for a certain amount of patients. Their objective is to find the best setting to minimize patient waiting times, reduce saturation and improve the use of resources. Optimization via simulation with heuristic approaches offers the best option to solve this problem, since it is difficult to experiment in real-time with the different existing parameters.

Optimization test functions[3], by its inherent properties, are convenient to evaluate and improve our proposed methodology, delimit the area of interest and perform iterations. Since we expect that the results are more accurate, the idea is to carry out the least number of iterations.

3 Proposal Methodology

The initial proposal of this work is to evaluate the quality of the Montecarlo plus K-means heuristic through the use of complex optimization functions, first delimiting a wide area of interest.

In this case, our proposed heuristic is a two-phase method: The first phase would be a coarse-grained approach consisting of a global exploration stage over the entire search space to find promising regions for optimization identified on a neighbourhood structure of the problem. This phase uses the Montecarlo heuristics plus the K-means method, returning a collection of promising regions. The second phase is a fine-grained approach, involving the search for the best solution,

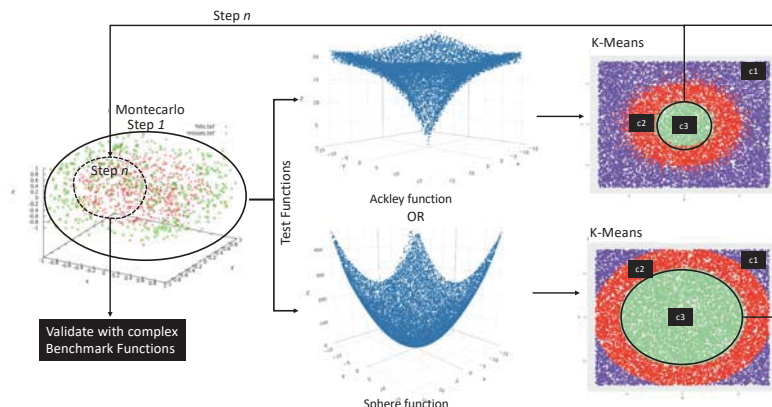


Fig. 1. Overview of the proposed methodology

be it optimal or sub-optimal, through a recursive application of our heuristics narrowing an exhaustive search within the promising regions.

To carry out the evaluation proposal, we selected the optimization test functions to assure that we find the minimum configuration value, which would be assimilated to the configuration obtained through the application of heuristic methods. This process would open up the possibility of adding a high-quality index to the heuristic method procedures carried out in addition to establishing a better value for the use of resources.

Our methodology focuses on evaluating the quality of the heuristic method. By adjusting the functions to Montecarlo methods, as illustrated in Fig.1, where we show the whole process, we apply Montecarlo to obtain a random sampling defining the domain of possible inputs. Afterwards, through the functions Ackley and Sphere, we bound an area containing a certain number of samples. These random samples represent the area where we will find the optimum value. As seen in the same Fig.1, when we apply k-means, the graphed functions seen from above show the area in which we are looking for our result. In the first case, when we use the Ackley function, we can locate a slightly smaller area in the centre containing a smaller number of values; while the second example, the Sphere function returns a broader area in the centre. These values would be similar to the one we are looking for. In this case, we can also observe that there are two different types of minimum, depending on the function.

Once we have delimited this area, we perform another iteration following the same process, and so on. After a few iterations, we will confirm that there are no variations in the returned values so we will stop iterating.

We aim to increase the quality of the results obtained and improve our heuristics using the optimization functions as evaluation benchmarks. The parameters resulting from the Montecarlo methods that we introduced within the function will allow us to locate the area of interest more precisely. By doing this, we would be on the path that leads us to an improvement of the heuristic method and also be on the road to add a quality index to the whole process.

4 Conclusions and Future Work

Concepts such as optimization have been analyzed. The proposed methodology includes the use of the Montecarlo plus K-means heuristic method and the optimization functions as evaluation benchmarks since we are seeking to improve the quality of this heuristic method taking into account that there are some scenarios that can not be solved linearly due to its characteristics and because it would require too many resources and time.

Our proposal originates from the idea of evaluating the heuristic method applied in order to find an optimal configuration. We are proposing to analyze the quality of the Montecarlo plus K-means heuristic developed to find the best possible scenario and to analyze its quality. Through the use of optimization test functions, we are trying to obtain the best solution (be it optimal or sub-optimal) in a more accurate and sophisticated way.

The heuristic method described sets us on the road to a good strategy when looking for an optimal configuration, and it is also applicable to different areas. Our idea is to finally validate the heuristic method through the use of complex optimization functions, analyze its quality and be able to add a high-quality index to the whole configuration. In the future, we will work on improving the algorithm, making it scalable, and further improving the quality index.

Acknowledgment

This publication is supported under contract TIN2017-84875-P, funded by the Agencia Estatal de Investigacion (AEI), Spain and the Fondo Europeo de Desarrollo Regional (FEDER) UE and partially funded by a research collaboration agreement with the Fundacion Escuelas Universitarias Gimbernat (EUG).

References

1. A. E. X. Li and M. Epitropakis, "Benchmark functions for cec'2013 special session and competition on niching methods for multimodal function optimization," *Technical Report, Evolutionary Computation and Machine Learning Group, Australia*, 2013.
2. C. P. Robert and G. Casella, *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2005.
3. M. Jamil and X.-S. Yang, "A literature survey of benchmark functions for global optimization problems," *Int. J. of Mathematical Modelling and Numerical Optimization*, vol. 4, 08 2013.
4. B. Qu, J. Liang, Z. Wang, Q. Chen, and P. Suganthan, "Novel benchmark functions for continuous multimodal optimization with comparative results," *Swarm and Evolutionary Computation*, vol. 26, pp. 23 – 34, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S221065021500053X>
5. M. Jamil and X.-S. Yang, "A literature survey of benchmark functions for global optimization problems," *Int. J. of Mathematical Modelling and Numerical Optimization*, vol. 4, 08 2013.
6. E. Cabrera, M. Taboada, M. L. Iglesias, F. Epelde, and E. Luque, "Optimization of healthcare emergency departments by agent-based simulation," *Procedia Computer Science*, vol. 4, pp. 1880 – 1889, 2011, proceedings of the International Conference on Computational Science, ICCS 2011. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877050911002626>

From Fuzzy Deformable Prototypes to Elastic Patterns: Preliminary proposal*

Ruben Rodriguez-Cardos (ruben.rodriguez6@alu.uclm.es)¹[0000-0002-0294-8343]
and Jose A. Olivas(joseangel.olivas@uclm.es)¹[0000-0003-4172-4729]

Department of Information Systems and Technologies, University of Castilla La Mancha, Ciudad Real, Castilla-La Mancha, Spain

Abstract. Based on previous experience in prediction systems with Fuzzy Deformable Prototypes, improvement in their functioning and deformation capacity is necessary, a new idea/concept for this is proposed. We propose a system with artificial intelligence that is capable of characterizing new situations, within the context of a PhD. thesis, capable of recognizing samples in a cognitive environment, in addition to testing its viability and performance in a non-cognitive environment a preliminary experiment is carried out, classifying handwritten digits from a reference database (MNIST) with a good success rate.

Keywords: Pattern Recognition · Elastic Patterns · Enginer Strain · Springs · Intelligent Data Analysis · Deformable Prototypes.

1 Motivation/Introduction

The motivation for this work arises from different points:

1. An article by H. Bremerman[1], where he proposes the *Deformable Prototypes*.
2. Fuzzy Prototypes proposed by Lotfi A. Zadeh [8].
3. Starting from points 1 and 2, the Fuzzy Deformable Prototypes proposed by J.A. Olivas [5], these have been applied in a wide number of problems in different fields successfully. However they have a *bottleneck*, they depends on a single parameter, the *degrees of representativeness*.

This work arises especially on the basis of points 1 (with a more focused approach to this point and the use of physics) and 3 of the previous list. As entry point, a handwritten digit recognition through a classic Fuzzy Pattern system (based on the concept of Mask[6]) was done. For these reasons, the main objective of this thesis is: *Improve the application of Fuzzy Deformable Prototypes and the creation/use of a new concept to improve their deformation capacity*, we will call this new concept: **Elastic Patterns**.

* This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant MERINET: TIN2016-76843-C4-2-R (AEI/FEDER, UE).

H. Bremermann proposed that a set of equivalent classes can be represented by individual members. Furthermore, if a pattern is equivalent to a class and a serie of *affine transformations*, we can try to represent that class by an individual member of it. In that case, these individual members can be called **prototype** [1]. A prototype can be defined by a set of parameters, a *parametric representation*. Combining the work of H. Bremermann and R.Hodges a *matching function* was defined, this function is able of classifying samples from a universe U into one of the existing labels (classes) in that universe.

On the other hand, Lofti A Zadeh proposed that a fuzzy prototype is not an element, but the set of a good, poor and borderline element of a category. *Prototypicality* is a matter of degree. A fuzzy set A can be defined as the degree of membership of selected elements previously divided by the elements. So, a prototype A is a fuzzy set defined by the degree of membership of selected elements previously divided by the prototypes of the elements:

$$PT(A) = High/PT(A_{Good}) + Medium/PT(A_{Borderline}) + Low/PT(A_{Poor})$$

Fuzzy Deformable Prototypes can be described as a linear combination Fuzzy Prototypical Categories (described as tables of attributes), extending the Deformable Prototypes to the case of affinity with more than one Fuzzy Prototypical Category the definition of a real situation would be:

$$C_{real}(w_1...w_n) = | \sum \mu p_i(v_i...v_n) |$$

2 Elastic Patterns

In order to improve the bottleneck of the Fuzzy Deformable Prototypes and their capacity to deform, a new idea is proposed: *Represent a pattern* (a prototype) *by a set of springs and deform these, simulating the physical deformation that will be produced in a real spring*, this is carried out generating a deformation, by contracting or stretching, each spring individually. Measuring the deformation that the springs suffer is one of the most important aspects of this new idea. So, the *Elastic Patterns* generate a deformation on two levels in order to match perfectly with a new sample to be recognized:

- **At the level of the parameter** (spring/parameter deformation): Calculated using the concept of *Engineering Strain*. Used to measure the deformation of a spring on a single axis, obtaining the deformation suffered in function of its initial length, the deformation causes that the final length of the spring is n times the initial length[3]. Parameters that are most deformed individually should add more value to the total deformation. For example: The deformation suffered by a spring that measures 1 cm. and deforms 1 cm. more is much greater than the suffered by a spring of 20 cm. and deforms 1 cm. more.

- **At the level of the pattern** (pattern deformation): The calculation of the *Deformation Energy* corresponds to the deformation of the Elastic Pattern. To carry out this calculation a new concept is used, a *Deformation Vector*. Which is based on the concept of *Deformation Tensor*[2]. This concept normally used in mechanics of continuous media and mechanics of deformable solids, with the aim of weighting the change of shape and volume in a body. The *Deformation Energy* that a pattern undergoes to fit perfectly with a real case is the sum of each of the values of the *Deformation Vector*, in other words, the *Deformation Energy* that affects the Elastic Pattern is the sum of the deformation suffered by each parameter.

It is possible to use the following concept, inherited from the Deformable Prototypes:

A sample is classified according to the minimum Deformation Energy required for physically deforming the closest Elastic Pattern.

3 Preliminary Experiments

To test whether the proposed hypothesis makes sense and can work in a both cognitive (closer to the knowledge engineering) and non-cognitive environments, some preliminary experiments is carried out, similar to the digit recognition system described above. MNIST is, de facto, one of the most used databases in image classification and artificial intelligence tasks, due to the quantity, variety and quality of classified samples. It has been used in many projects[4] [7], so it can be considered a reference data set. For this reason, the experiment is being carried out using MNIST. The conditions under which the experiment is carried out are as follows:

- An image database is available ¹, which is a subset of MNIST, consisting of 70,000 images (matrix-coded) of handwritten digits.
- Each sample is coded as a 28 x 28 pixel matrix, whose values will be in the range 0 (a black pixel) to 255 (a white pixel). In addition, each sample has associated with which digit it represents, these digits being numbers from 0 to 9. The samples are centered in the image.
- The training set consists of 60,000 of these random samples, used to generate the different elastic patterns. The test set will be made up of the remaining 10,000 samples.

The results obtained are: the generation time of the elastic patterns is approximately **20 seconds**, success rate is approximately **80%**, and execution time is approximately **90 seconds**.

¹ <https://www.openml.org/d/554>

4 Conclusions

Once the experiment has been carried out, it is possible to draw the following conclusions:

- Conceptually the Elastic Patterns are easy to understand and use. However, the generation of Elastic Patterns depends on each problem and context in which they are applied.
- They are a good way to generate a model that represents reality in a simple way as the previous works on which they are based[1][5].
- Elastic patterns can work with a certain level of uncertainty, as they deform to perfectly match a new situation.
- The high percentage of correct results in the experiments, over 80%, shows that the proposed new concept is viable and that it is possible to use it at a not cognitive environment too.

As future work the following points are proposed: continue to study how data deformation is possible, establish a general method for the generation of Elastic Patterns, and the use of the Elastic Patterns on a more complex database, use them at a cognitive environment. Currently there is already a project in development for the detection of hereditary cancer using Elastic Patterns.

References

1. Breermann, H.: Pattern Recognition, pp. 116–159. Birkhäuser Basel (1976)
2. Chen, H.: Constructing continuum-like measures based on a nonlocal lattice particle model: Deformation gradient, strain and stress tensors. *International Journal of Solids and Structures* **169**, 177–186 (2019)
3. Chrysanidis, T.: Evaluation of out-of-plane response of r/c structural wall boundary edges detailed with maximum code-prescribed longitudinal reinforcement ratio. *International Journal of Concrete Structures and Materials* **14**(1) (2020)
4. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine* **29**(6), 141–142 (2012)
5. Olivas, J.A.: Contribución al estudio experimental de la predicción basada en categorías deformables borrosas. Ph.D. thesis, Universidad Castilla-La Mancha (2000)
6. Rodríguez-Cardos, R.: Reconocimiento óptico de caracteres en escritura manual, chap. 4,5. Universidad Castilla-La Mancha (2017), <https://ruidera.uclm.es/xmlui/handle/10578/15413>
7. Schott, L., Rauber, J., Bethge, M., Brendel, W.: Towards the first adversarially robust neural network model on mnist. arXiv preprint arXiv:1805.09190 (2018)
8. Zadeh, L.A.: A note on prototype theory and fuzzy sets. *Cognition* **12**(3), 291 – 297 (1982)

Towards Smart Data Technologies for Big Data Analytics

María José Basgall^{1,2,3}[0000-0002-7024-847X], Marcelo Naiouf²[0000-0001-9127-3212], Francisco Herrera³[0000-0002-7283-312X], and Alberto Fernández³[0000-0002-6480-8434]

¹ UNLP, CONICET, III-LIDI, La Plata, Argentina
mjbasgall@lidi.info.unlp.edu.ar

² Instituto de Investigación en Informática (III-LIDI), CIC-PBA, Facultad de Informática - Universidad Nacional de La Plata, Argentina

³ DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Granada, Spain

Abstract. Currently the publicly available datasets for Big Data Analytics are of different qualities, and obtaining the expected behavior from the Machine Learning algorithms is crucial. Furthermore, since working with a huge amount of data is usually a time-demanding task, to have high quality data is required. Smart Data refers to the process of transforming Big Data into clean and reliable data, and this can be accomplished by converting them, reducing unnecessary volume of data or applying some preprocessing techniques with the aim of improve their quality, and still to obtain trustworthy results. We present those properties that affect the quality of data. Also, the available proposals to analyze the quality of huge amount of data and to cope with low quality datasets in an scalable way, are commented. Furthermore, the need for a methodology towards Smart Data is highlighted.

Keywords: Big Data · Smart Data · Data Complexity · Data Quality.

1 Introduction

The Big Data term [13] refers to the enormous amount of data that is being generated increasingly and from several sources, with a strong relationship with both velocity and variety. However Big Data does not entail a good quality of data. In the field of Data Science, obtaining knowledge from datasets is the main task. Unfortunately, several data complexities can degrade the quality of the problems, and in turn yield to inaccurate results. Among others, class imbalance, overlapping, redundancy, outliers, missing values, can be stressed [5, 3].

Whether it is due to the nature of the problem, or due to the way in which data is obtained or generated, most of the publicly available big datasets have different qualities. This can be identified as one of the causes for being unable to replicate the good behavior of standard techniques in Big Data benchmarks [1, 2]. Because of all the above, it is crucial to identify the data complexity in order

to take action and to apply Big Data preprocessing techniques towards Smart Data [7], with the aim to learn from high quality data.

In this contribution, the data characteristics that affect the expected behavior of a knowledge extraction technique and how they are represented in the publicly available datasets for Big Data, are presented. Furthermore, we introduce the current proposals to cope with data quality, and the need for more technologies to turn Big Data into Smart Data is commented. This work is part of the ongoing doctoral thesis based on the analysis and design of preprocessing techniques for Imbalanced Big Data problems.

2 Towards Smart Data in Big Data Analytics

Identifying data complexities is very important in order to decide about which preprocessing technique or Machine Learning algorithm to apply. In data classification, class imbalance [9] refers to an uneven data distribution between the classes of a problem. The overlapping areas [5] of a problem are ambiguous regions of the data space, where there are instances belonging to different classes. In addition, if a dataset contains a subset of examples with different values in their features but representing the same concept, those instances are considered as redundant. Therefore redundancy is more than exact copies of the examples in a set of data [12, 10]. At the same time, outliers [4] relate to those instances that contain in their attributes very distant values from the common order of magnitude of the remainder. Finally, missing values [11] occurs, as its name suggests, when an instance has one or more of its features values lost.

In Academia, publicly available big datasets are of different qualities, presenting a variable type of undesirable characteristics. For instance, many Big Data classification problems are about biotechnology or related to new physics discoveries, and they usually pose a high degree of imbalance between its classes, in addition to other intrinsic complexities. Therefore, each data problem should be treated as an independent project in order to apply Smart Data technologies with the aim to improve the data quality. Therefore, if Smart Data is not taken into account, Machine Learning models could lead to misleading results.

In [1] and [2], we have studied the behaviour of the well-known oversampling technique called SMOTE [6] for Imbalanced Big Data Classification problems without following any additional Smart Data technology. The study was based using two different kind of design approaches in order to be scalable. SMOTE is one of the most widely used technique to balanced a dataset in small data scenarios due to its simplicity, but also for offering better results with respect to the standard random solutions. Those work contributions showed that the results achieved were not as good as the ones in small datasets, and one reason was pointed out as the lack of data quality.

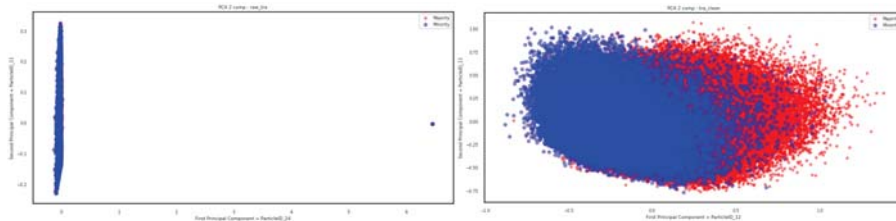
Regarding the current proposals towards Smart Data we may found two main papers. On the one hand, in [10] two novel metrics to describe the quality of a big dataset in a scalable way were proposed, jointly to another basic metrics as a Spark-package. Authors have found a redundancy of information in most of

the Big Data Classification problems. The main conclusion was that randomly decreasing the datasets up to 25% do not affect significantly the performance of the classifiers applied. On the other hand, in [8] a Smart Data based ensemble for Big Data Imbalanced Classification problems was proposed. Authors reached an improvement for data quality by combining different preprocessing techniques.

3 A case study

In this section, a brief analysis over the MiniBooNE dataset as a case study is presented. The objective is to briefly show the incidence of some of several intrinsic data characteristics in a Big Data classification task. Specifically, the most straightforward characteristics to discover in a dataset are the imbalanced degree, the replicated instances (as part of the redundancy property), missing values, and rare values of features, among others.

The MiniBooNE dataset has more than 130,000 instances and a high dimensionality (49 continuous features). It represents a binary classification problem from the physic field and it aims to distinguish signal from background, where the signal is the 36.5% of the dataset. Using a pipeline of standard techniques, we have been able to determine that MiniBooNE does not contain missing values, and a 0.36 % of the data are replicated instances (which we have removed). Furthermore, by means of a graphical analysis we have detected values far from the common order of magnitude of the ParticleID.19 feature, and the instances with those extreme values were removed from the raw dataset. The incidence of this action can be seen in Fig. 1, where the two principal components (PC) for the raw dataset (Fig. 1a) and for the new version of it (Fig. 1b) are shown.



(a) Principal Components for raw dataset (b) Principal Components for clean dataset

Fig. 1: Principal component analysis (PCA) for MiniBooNE dataset before and after cleaning stage

After the aforementioned pipeline, a Decision Tree classifier to learn from the raw and from the clean version of the dataset was applied. Table 1 shows the widespread metrics used for imbalanced classification problems. A slight improvement in detecting the minority class instances can be seen. Considering that no further preprocessing techniques have been applied to the dataset except

for these basic ones, a trend of improving classification results is evident as more Smart Data technologies are applied.

Table 1: Decision Tree classifier results for the MiniBooNE dataset

	GM	AUC	TPR	TNR
raw	0.8556	0.8590	0.7825	0.9355
clean	0.8636	0.8659	0.8025	0.9293

4 Conclusions

The data quality analysis of the Big Data sets is almost an uncharted territory, and Smart Data is also a fledgling topic. An exhaustive study of the data properties, together with the application of the proper preprocessing techniques, has become mandatory for all Data Science projects in both industry and academia. By considering a case study on Big Data classification, we have determined that even straightforward data transformation allowed at improving the modeling process, thus stressing the need towards the use and development of Smart Data technologies.

References

1. Basgall, M.J., et al.: Smote-bd: An exact and scalable oversampling method for imbalanced classification in big data. *JCS&T* **18**(03), e23 (Dec 2018)
2. Basgall, M.J., et al.: An analysis of local and global solutions to address big data imbalanced classification: A case study with smote preprocessing. In: VII JCC&BD. vol. 1050, pp. 75–85. Springer (2019)
3. Das, S., et al.: Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition* **81**, 674–693 (2018)
4. Devi, D., et al.: Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance. *Pattern Recognit. Lett.* **93**, 3–12 (Jul 2017)
5. Fernandez, A., et al.: *Learning from Imbalanced Data Sets*. Springer (2018)
6. Fernández, A., García, S., Herrera, F., Chawla, N.V.: Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**, 863–905 (2018)
7. García-Gil, D., et al.: Enabling smart data: Noise filtering in big data classification. *Inf. Sci.* **479**, 135–152 (2019)
8. García-Gil, D., et al.: Smart data based ensemble for imbalanced big data classification (2020)
9. López, V., et al.: An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **250**(20), 113–141 (2013)
10. Maillo, J., et al.: Redundancy and complexity metrics for big data classification: Towards smart data. *IEEE Access* pp. 1–1 (2020)
11. Montesdeoca, B., et al.: A first approach on big data missing values imputation. In: *IoTBDS* (2019)
12. ur Rehman, M.H., et al.: Big Data Reduction Methods: A Survey. *DSE* **1**(4), 265–284 (Dec 2016)
13. Wu, X., et al.: Data mining with big data. *IEEE TKDE* **26**(1), 97–107 (Jan 2014)

E-Government and Data Quality

A framework for linking open environmental data

Juan Santiago Preisegger¹ , Ariel Pasini , Patricia Pesado 

Computer Science Research Institute LIDI (III-LIDI)*

Facultad de Informática – Universidad Nacional de La Plata 50 y 120 La Plata Buenos Aires

* Partner Center of the Scientific Research Agency of the Province of Buenos Aires (CIC)

¹ Fellow UNLP

{jspreisegger, apasini, ppesado}@lidi.info.unlp.edu.ar

Abstract. The concept of open government has promoted several initiatives in order to progress in their implementation in different governmental agencies. Generally, implantation is carried out, in the first place, based on transparency and the opening of data, giving citizens relevant information about their community. Within that information, data about the rational use of natural resources and care of the environment can be found. In this context, and taking into account the ongoing analysis, it was discovered that, apart from the positive measures taken for the progress of the implantation, the main existing flaw is the possibility to link and relate the existing data in different data sources. The linking of data will allow for a closer analysis that can generate a wider context for information. This paper is based on a PhD thesis in process whose aim is to generate a framework that allows the linking and relationship of information related to the environment published in different open government portals.

Keywords: Open government, Open data, Data linking, Related open data, Environment.

1 Introduction

The demands of society towards their rulers and government entities, such as transparency, inclusion in decision-making processes and collaboration, are, nowadays, enormous. Therefore, a new way of government that includes citizens, allowing them to help in public politics and to participate in decision-making processes of the government, emerges [1]. This method of governing is called open government and it is defined as "...a technological and institutional platform that turns governmental data into open data to let citizens use, protect and collaborate with public decision processes, accountability and improvement of public services" [2]. It is based on three main and well-defined principles: Transparency, Collaboration and Participation [3] [4]. In papers related to this area, it could be observed that there exist standards in the dataset publication formats [5] [6], however, the main flaw is how data is loaded in an orderly manner to improve the linking and relationship to others.

Due to changes that affect the world, a significant level of awareness about the rational use of resources and the environment care was generated in the world society. The governmental agencies are not indifferent to this matter, and great progress is observed regarding data published.

As part of the PhD thesis, it was analyzed the difficulties when linking open environmental data published by different organizations; particularly, the work is focused on energy, water and air data.

Section 2 presents the state of the art on issues of open government and open data. In the third section, the main existing problem in the linking of these data is described. In section 4, the framework for linking open data, that is expected to be carried out to conclude with the PhD thesis, is presented. Finally, in section 5, conclusions and the expected results are described.

2 Open data

Currently, citizens present enormous demands towards their rulers and government entities. Among these demands, transparency and the efficient management of public assets, inclusion in decision-making processes and collaboration with different areas of society, are included. According to these requirements, and with the help of new technologies, a new way of government that includes the citizen was generated; this way allows them to generate contributions to public politics and to participate in the decision-making processes [7].

Through the implantation of open government, there was an opening of data from governmental agencies. It strengthens two activities that are part of the performance of societies: critical thinking and decision-making process, now based on more information [8]. However, the main difficulty that citizens find is that such information is spread in different data sources, such as portals and catalogues, which makes it impossible to link and relate these data in order to obtain a global vision of the topic.

3 Linking open data

Through the boom generated by this new paradigm within the governmental agencies of the world, several applications and tools were generated from different sectors to make the search, integration and data processing automatic or to better them. One of these tools is Google Dataset Search, a search engine that is specialized in finding sets of data stored in the web via keywords, as long as they use schema.org dataset tags or equivalent structures represented in Data Catalog Vocabulary (DCAT) format [9].

Various searches were carried out to obtain datasets related with energy, water and air topics published worldwide, under these schemes. Additionally, datasets published by *UN Water* [10], *U.S. Energy Information Administration* [11] and *European Statics* [12] were analyzed. They contained relevant information but, as they didn't fulfill the scheme, they weren't reached by Google Dataset Search. The searches were carried out with the following words: *Drinking water quality*, *Energy generation*, *Air quality*. From

these searches, a group of dataset of each topic was selected and the shapes, the available formats and the structures of each of them were analyzed.

Regarding air and water quality, it was possible to raise awareness of certain standard to analyze the existent magnitudes in different characteristics. It could be observed that almost the same tests on different samples were carried out to analyze different characteristics and determine if they are within the healthy margins for the human use. Nevertheless, there exist differences in the structure of dataset, such as, variation from columns to rows or the division of a field into many, which complicate the linking between different datasets of such topics.

In the case of energy area datasets, the difference between published data from different governmental agencies is even greater. It was observed that: some countries simply publish the annual percentage generated and the type of energy obtained in which different published generation stations are based on; other countries simply publish the percentages of their energy generation sources, without discerning between the stations they have; and other countries publish geo-referenced data of the location of each station, the total capacities, the types of stations and even their owners. It is possible to observe that there is a potential for the interrelationship of data in this field, with the consequent standardization and the selection of certain fields in common among different dataset.

4 Framework for linking open data

As part of the PhD thesis, it is expected to carry out the state of the art study in relation to usual problems that are detected in the publication of environmental data in different portals of open government. It is expected to identify a group of existing patterns, similitudes and differences in the different technologies used, the data publication formats and their structure, which prevents interrelation and, due to the study aforementioned, it is expected to generate a framework for linking open data.

5 Conclusion

This work suggests analyzing different data sources that governmental agencies provide and generate a framework for open data, particularly in water, air and energy data. The framework of this study allows for the selection of data available in different portals in order to change them and combine them, according to the aim the parties concerned wish to analyze.

Acknowledgments. Project co-funded by the Erasmus+ Programme of the European Union. Grant no: 598273-EPP-1-2018-1-AT-EPPKA2-CBHE-JP.

6 References

- [1] S. A. Chun, S. Shulman, R. Sandoval, and E. Hovy, "Government 2.0: Making

- connections between citizens, data and government,” *Inf. Polity*, vol. 15, no. 1–2, pp. 1–9, 2010.
- [2] J. R. Gil-García and J. I. Criado, *Las Tecnologías de Información y Comunicación en las Administraciones Públicas Contemporáneas*. 2017.
- [3] C. Calderón and S. Lorenzo, *Open Government. Gobierno Abierto*. 2010.
- [4] A. Naser and A. Ramirez, “Plan de gobierno abierto. Una hoja de ruta para los Gobiernos de la Región,” *CEPAL - Manuales*, vol. 81, p. 80, 2017.
- [5] A. Pasini, J. S. Preisegger, and P. Pesado, “Modelos de evaluación de gobiernos abiertos , aplicado a los municipios de la provincia de Buenos Aires,” *XXIV Congr. Argentino Ciencias la Comput.*, vol. XXIV, pp. 0–10, 2018.
- [6] A. Pasini, J. S. Preisegger, and P. Pesado, “Open Government Assessment Models Applied to Province’s Capital Cities in Argentina and Municipalities in the Province of Buenos Aires,” in *Communications in Computer and Information Science*, 2019, vol. 995, pp. 355–366.
- [7] A. Naser, Á. Ramírez-Alujas, and D. R. Editores, *Desde el gobierno abierto al Estado abierto en America Latina y el Caribe: Planificación para el Desarrollo*. 2017.
- [8] J. Pastor Verdú, “Conceptos y fenómenos fundamentales de nuestro tiempo,” *Unam*, p. 10, 2012.
- [9] N. Noy, M. Burgess, and D. Brickley, “Google dataset search: Building a search engine for datasets in an open web ecosystem,” in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2019, pp. 1365–1375.
- [10] <https://www.unwater.org/>
- [11] <https://www.eia.gov/>
- [12] <https://ec.europa.eu/eurostat/>

Framework for Data Quality Evaluation Based on ISO/IEC 25012 and ISO/IEC 25024

Julieta Calabrese¹ , 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, Argentina

*Partner Center of the Scientific Research Agency of the Province of Buenos Aires

¹ Fellow, UNLP

{jcalabrese, sesponda, ppesado}@lidi.info.unlp.edu.ar

Abstract. Nowadays, organizations process large volumes of data. Being able to foster and maintain data quality is one of the biggest challenges that these organizations face. To do this, there are standards, such as ISO/IEC 25012 and ISO/IEC 25024, which are intended to measure data quality based on a set of inherent and system-dependent characteristics, along with a set of associated metrics. Using standards to carry out measurements can be complex and even expensive for those with little experience in the area. In this context, we propose in this work a prototype for a tool based on ISO/IEC 25012 and ISO/IEC 25024 that, by analyzing different patterns of common errors in data, allows an organization to understand data current status.

Keywords: data quality - ISO/IEC 25012 - ISO/IEC 25024

1 Introduction

From the beginning of Software Engineering, there have always been issues related to achieving optimum quality levels in different aspects of software [1]. It is widely known that quality management is essential within any organization.

Technological advancement interferes in all sectors, from agriculture to manufacture, tourism, health care, and university; in the process, data become the most powerful organizational asset and a key aspect for decision-making. New technologies allow obtaining and storing large amounts of data (through mobile devices, sensors, and so forth) and analyzing them using different algorithms, generating an unlimited amount of processing. However, for the information obtained to be considered the most important asset for the organization, data must be of adequate quality from their

source and appropriate for the environment where they are used. The result of any decision made by the organization will be based on its information.

Currently, data have become essential in digitally transformed organizations, and data quality is essential to achieve service excellence for all stakeholders. Lack of resources to assess data quality is one of the main problems organizations currently face, as this significantly affects organizational and business effectiveness and efficiency. This context leads to focus attention on the standards defined by ISO in relation to data quality.

This work is part of a doctoral thesis for the development of a framework that simplifies data quality control for organizations.

2 ISO/IEC 25012 and ISO/IEC 25024

To organize and unify all standards related to software product quality, ISO/IEC published in 2005 the document ISO/IEC 25000:2005 - SQuaRE (System and Software Quality Requirements and Evaluation) [2], also known as the ISO 25000 family. Within this set, ISO/IEC 25012 - "Data Quality Model" [3] and ISO/IEC 25024 - "Measurement of data quality" [4] stand out for study.

ISO/IEC 25012 - "Data Quality Model" specifies a general quality model for data that are defined in a structured format within a computer system. This standard classifies quality attributes into fifteen characteristics analyzed from two points of view: inherent and system-dependent. These characteristics will be assigned different importance and priority by each evaluator based on their own specific needs. Accuracy, Completeness, Consistency, Credibility and Currentness are **inherent** characteristics; while Availability, Portability and Recoverability are **system-dependent** ones. Accessibility, Compliance, Confidentiality, Efficiency, Precision and Traceability belong to **both groups**.

On the other hand, ISO/IEC 25024 - "Measurement of data quality" provides measurements, including measurement methods and related quality measurement elements for the quality characteristics of the data quality model described above.

3 Data Quality

Due to the lack of departments specialized in data quality analysis and quality certifications, organizations face greater challenges in relation to the data they manage. As an example, a government organization published its data corresponding to Wi-Fi networks access points [5], and situations such as the following are found:

id	identificador	ubicacion	latitud	longitud
965	TUC034-02	Comuna	-26.420030	-64.775878
681	CHA002-01		0.000000	0.000000

Table 1. Wi-Fi networks access point data

The location of one of the access points is unknown and, therefore, the corresponding data about coordinates is not available. For the specific context, is it possible for this type of problem to occur? Should the missing data be mandatory? Are access points allowed that do not have location as data? Is naming fields as ID and IDENTIFIER confusing as to what they represent?

In this context, it is of interest for this work to identify a set of recurring problems with data in various organizations with different topics, generating patterns that represent common organizational failures that somehow can facilitate data analysis.

Commonly found problems are usually:

- Errors in completeness: Fields that are required but are left blank.
- Syntax errors: Some fields in upper case and others in lower case, mix of both cases in the same field, language-specific issues, such as words with tilde and others without tilde (where it should be present), problems with the "ñ", and so forth.
- Semantic errors: Errors with expected values. For example, in a field called "Country", the expected value would be the name of a country, not "María".
- Consistency errors: Contradiction errors.
- Update errors: Data are not updated with respect to the current environment.
- Traceability errors: There is no data log recording additions, modifications and/or deletions, specifying the corresponding event.
- Precision errors: Data are not accurate where they should be. This is more common in contexts where a lack of precision in a value (decimals for example) can result in a radical change.
- Understandability errors: Data are represented by symbols that cannot be understood by any type of user.
- Accessibility errors: Data cannot be accessed by users who need support due to having some type of disability.

In addition, it has been demonstrated that there are department-specific errors in data, which must be taken into account for correct evaluation.

It would be extremely useful for organizations to have a tool that can use quality standards to respond to the different patterns identified so as to understand the current status of their information through a simple process that yields an in-depth analysis of the quality of their data.

4 Proposed Development

A framework will be developed to respond to the different common failure patterns identified in current organizations from various areas. This framework will be based on the ISO standards associated with data quality, ISO/IEC 25012 - Data Quality Model and ISO/IEC 25024 - Data Quality Measurement, to define which characteristics will be evaluated, as well and the corresponding metrics. The contribution will be useful for organizations that need to understand the current state of the information they process, simplifying the application of standards.

The framework will be applied in various organizations. In particular, it will be applied in organizations that process large volumes of data and which require acceptable levels of quality.

5 Bibliography

- [1] “Guía para evaluar calidad de datos basada en ISO/IEC 25012”. Calabrese, Julieta; Esponda, Silvia; Pasini, Ariel, Boracchia, Marcos; Pesado, Patricia. Congreso Argentino de Ciencias de la Computación - CACIC 2019.
- [2] “ISO/IEC 25000:2014 Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE”.
- [3] “ISO/IEC 25012:2008. Software engineering -- Software Product Quality Requirements and Evaluation (SQuaRE) -- Data quality model”.
- [4] “ISO/IEC 25024:2015. Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Measurement of data quality”.
- [5] Data obtained from <http://datos.gob.ar>