



Short Papers

of the

9th Conference on

Cloud Computing,

Big Data

&

Emerging Topics

[HTTPS://JCC.INFO.UNLP.EDU.AR](https://jcc.info.unlp.edu.ar)

 @CONF_CC_BD_ET
 JCC@LIDI.INFO.UNLP.EDU.AR

Short Papers of the 9th Conference on Cloud Computing, Big Data & Emerging Topics

La Plata, Buenos Aires, Argentina.

June 22–25, 2021

Short papers of the 9th Conference on Cloud Computing Conference,
Big Data & Emerging Topics / editado por Armando De Giusti...[et al.].
-1a ed. - La Plata: Universidad Nacional de La Plata. Facultad de
Informática, 2021.

Libro digital, PDF

Archivo Digital: descarga y
online ISBN 978-950-34-2016-4



1. Computación. 2. Actas de Congresos. 3. Conferencias. I. De Giusti,
Armando, ed.
CDD 004.071

Preface

Welcome to the short paper proceedings of the 9th Conference on Cloud Computing, Big Data & Emerging Topics (JCC-BD&ET 2021), held in an interactive, live online setting due to COVID-19 situation. JCC-BD&ET 2021 was organized by the III-LIDI and the Postgraduate Office, both from School of Computer Science of the National University of La Plata.

Since 2013, this event has been an annual meeting where ideas, projects, scientific results and applications in the cloud computing, big data and other related areas are exchanged and disseminated. The conference focuses on the topics that allow interaction between academia, industry, and other interested parties.

JCC-BD&ET 2021 covered the following topics: cloud, edge, fog, accelerator, green and mobile computing; big data; data analytics, data intelligence, and data visualization; machine and deep learning, and special topics related to emerging technologies. In addition, special activities were also carried out, including 1 plenary lecture and 1 discussion panel.

In this edition, 20 short papers were accepted after peer-review process. These short papers correspond to initial research with preliminary results, on-going R+D projects, or postgraduate thesis proposals. The authors of these submissions came from the following 6 countries: Argentina, Austria, Brazil, Chile, Ecuador, and Spain. We hope readers will find these contributions useful and inspiring for their future research.

Special thanks to all the people who contributed to the conference's success: program and organizing committees, authors, reviewers, speakers, and all conference attendees.

June 2021

Armando De Gisuti
Marcelo Naiouf
Laura De Giusti
Enzo Rucci
Franco Chichizola

Organization

General Chair

Armando De Giusti Universidad Nacional de La Plata and CONICET,
Argentina.

Program Committee Chairs

Marcelo Naiouf	Universidad Nacional de La Plata, Argentina.
De Giusti Laura	Universidad Nacional de La Plata and CIC, Argentina.
Enzo Rucci	Universidad Nacional de La Plata and CIC, Argentina.
Franco Chichizola	Universidad Nacional de La Plata, Argentina.

Program Committee

María José Abásolo	Universidad Nacional de La Plata and CIC, Argentina
José Aguilar	Universidad de Los Andes, Venezuela
Jorge Ardenghi	Universidad Nacional del Sur, Argentina
Javier Balladini	Universidad Nacional del Comahue, Argentina
Oscar Bria	Universidad Nacional de La Plata and INVAP, Argentina
Silvia Castro	Universidad Nacional del Sur, Argentina
Laura De Giusti	Universidad Nacional de La Plata and CIC, Argentina
Mónica Denham	Universidad Nacional de Río Negro and CONICET, Argentina
Javier Diaz	Universidad Nacional de La Plata, Argentina
Ramón Doallo	Universidade da Coruña, Spain
Marcelo Errecalde	Universidad Nacional de San Luis, Argentina
Elsa Estevez	Universidad Nacional del Sur and CONICET, Argentina
Aurelio Fernandez Bariviera	Universitat Rovira i Virgili, Spain
Héctor Florez Fernández	Universidad Distrital Francisco José de Caldas, Colombia

Fernando Emmanuel Frati	Universidad Nacional de Chilicito, Argentina
Carlos Garcia Garino	Universidad Nacional de Cuyo, Argentina
Carlos García Sánchez	Universidad Complutense de Madrid, Spain
Adriana Angélica Gaudiani	Universidad Nacional de General Sarmiento, Argentina
Graciela Verónica Gil Costa	Universidad Nacional de San Luis and CONICET, Argentina
Roberto Guerrero	Universidad Nacional de San Luis, Argentina
Waldo Hasperué	Universidad Nacional de La Plata and CIC, Argentina
Francisco Daniel Igual Peña	Universidad Complutense de Madrid, Spain
Tomasz Janowski	Gdansk University of Technolgy, Poland
Laura Lanzarini	Universidad Nacional de La Plata, Argentina
Guillermo Leguizamón	Universidad Nacional de San Luis, Argentina
Edimara Luciano	Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Emilio Luque Fadón	Universidad Autónoma de Barcelona, Spain
Mauricio Marín	Universidad de Santiago de Chile, Chile
Luis Marrone	Universidad Nacional de La Plata, Argentina
Katzalin Olcoz Herrero	Universidad Complutense de Madrid, Spain
José Angel Olivas Varela	Universidad de Castilla-La Mancha, Spain
Xoan Pardo	Universidade da Coruña, Spain
María Fabiana Piccoli	Universidad Nacional de San Luis, Argentina
Luis Piñuel	Universidad Complutense de Madrid, Spain
Adrian Pousa	Universidad Nacional de La Plata, Argentina
Marcela Printista	Universidad Nacional de San Luis, Argentina
Dolores Isabel Rexachs del Rosario	Universidad Autónoma de Barcelona, Spain
Nelson Rodríguez	Universidad Nacional de San Juan, Argentina
Juan Carlos Saez Alcaide	Universidad Complutense de Madrid, Spain
Aurora Sánchez	Universidad Católica del Norte, Chile
Victoria Sanz	Universidad Nacional de La Plata, Argentina
Remo Suppi	Universidad Autónoma de Barcelona, Spain
Francisco Tirado Fernández	Universidad Complutense de Madrid, Spain
Juan Touriño Dominguez	Universidade da Coruña, Spain
Gabriela Viale Pereira	Danube University Krems, Austria
Gonzalo Zarza	Globant, Argentina

Sponsors

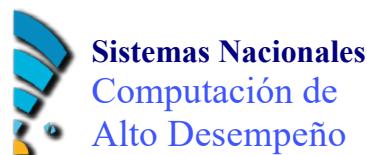


Table of Contents

Cloud, Fog, and High-Performance Computing	1
An approach to residential energy savings using IoT and Cloud Computing to provide real-time feedback. <i>Guillermo Friedrich, Guillermo Reggiani.</i>	2
A contribution to security in IOT system. Reconfigurable Logic Device Technology and Fog Computing. <i>Osvaldo Marianetti, Pablo Godoy, Ernesto Chediak, Carlos García Garino.</i>	6
Forest Fire Simulation in High Performance Computing. <i>Mónica Denham, Sigfrido Waidelich, Viviana Zimmerman, Karina Laneri.</i>	10
Early Experiences Migrating CUDA codes to oneAPI. <i>Manuel Costanzo, Enzo Rucci, Carlos García-Sánchez, Marcelo Naouf.</i>	14
Artificial and Computational Intelligence	19
Classic and recent (neural) approaches to automatic text classification: a comparative study with e-mails in the Spanish language. <i>Juan M. Fernandez, Nicolás Cavasin, Marcelo Errecalde.</i>	20
MbedML: A Machine Learning Project for Embedded Systems. <i>César A. Estrebou, Martín Fleming, Marcos Saavedra, Federico Adra.</i>	25
The current role of machine learning and explainability in actuarial science. <i>Catalina Lozano, Francisco P. Romero, Jesus Serrano-Guerrero, Jose A. Olivas.</i>	29
Speech emotion representation: A method to convert dis-crete to dimensional emotional models for emotional in-ference multimodal frameworks. <i>Fernando Elkfury, Jorge Ierache.</i>	33
Querying on Google Sheets. Designing a Sentiments Analysis Alternative for rating tweets regarding the Ecuadorian 2021 Presidential Campaigns. <i>Sariah López-Fierro, Carlos Chiriboga, Rubén Pacheco.</i>	37
An approach for the analysis of news during COVID-19 in the Chubut province. <i>Pablo Toledo Margalef, Emanuel Balcazar, Leo Ordinez, Claudio Delrieux, Lucila Allende.</i>	42
Intelligent Anomaly Detection System for IoT. <i>Diego Angelo Bolatti, Marcelo Karanik, Carolina Todt, Reinaldo Scappini, Sergio Gramajo.</i>	47

Distributed Cybersecurity Strategy, applying Intelligence Operation concept through data collection and analysis. *Ignacio Martín Gallardo Urbini, Patricia Bazán, Paula Venosa, Nicolás Del Río.* 51

Evaluation of a heuristic search algorithm based on sampling and clustering. *Maria Harita, Alvaro Wong, Dolores Rexachs, Emilio Luque.* 55

Modelling and Simulation of the COPD Patient and Clinical Staff in the Emergency Department (ED). *Mohsen Hallaj Asghar, Alex Vicente-Villalba, Alvaro Wong, Dolores Rexachs, Emilio Luque.* 59

Big Data 63

Big Data Technology for monitoring ICT service data. *Marcelo Dante Caiafa, Ariel Aurelio, Adrian Marcelo Bustos.* 64

Proposal of a Data Warehouse for Scholarly Institutions built on Institutional Repositories. *Pablo C. de Albuquerque, Gonzalo L. Villarreal, Marisa R. De Giusti.* 69

Semantic Web 74

Semantic Web for interoperable food safety legislation data: A case study. *Carlos Enrique Pintor, Carlos Francisco Ragout, Diego Torres, Alejandro Fernandez.* 75

Towards Ubiquitous and Actionable Augmented Reality Browsers by using Semantic Web Technologies. *Martín Becerra, Jorge Ierache, María José Abásolo.* 80

Smart Cities and Emerging Topics 84

Thermodynamic Dissipative Systems and Information Theory to Study the Social Component of a Smart City. *Gabriele De Luca, Thomas J. Lampoltshammer, Felipe Vogas.* 85

Videogames and virtual assets exchange. *Flavio A. Garrido, Hernán D. Merlino.* 89

Cloud, Fog, and High-Performance Computing

An approach to residential energy savings using IoT and Cloud Computing to provide real-time feedback

Guillermo Friedrich¹ and Guillermo Reggiani¹

¹ Universidad Tecnológica Nacional, Bahía Blanca, 8000, Argentina
`{gfried, ghreggiani}@frbb.utn.edu.ar`

Abstract. In recent years there has been a growing development of applications oriented to energy saving, based on the Internet of Things and cloud computing. These developments have not only economic motivations, but also environmental ones, related to the reduction of greenhouse gas emissions. The energy sector is perhaps the main global contributor to the emissions of these gases. In the present work, the development of a system based on IoT and CC for the monitoring of energy consumption at the residential level is described. It is organized according to the three-tier model: Edge, Platform and Enterprise. At the Edge level, some innovations are proposed, such as indirect energy sensing and the connection of sensors using the electrical network for data communication. Both would enable an agile deployment of the sensor network. The objective of the system is to provide the user with feedback about their energy consumption and certain environmental variables, in such a way that they can manage their energy consumption, while still achieving an adequate level of comfort.

Keywords: Internet of Things, Cloud Computing, Energy Monitoring.

1 Introduction

The emission of greenhouse gases (GHG) is one of the great current problems, due to its consequences on climate change. As part of the actions agreed by almost all countries, inventories of these emissions are made periodically. 53% of GHG emissions in 2016 in Argentina corresponded to the energy sector: 5.1 points due to residential electricity consumption and 7.4 by residential fuel burning [1]. With some variations, this situation is common to different countries. Around one third of the energy in the world is consumed in large public buildings, and in some countries they contribute about 40% of the total consumption [2]. Saving energy not only has an economic impact, but also an environmental one, although the economic saving could be, probably, the main motivation to make more rational consumptions. In this sense, the feedback that users receive about their consumption is a key factor.

Periodic bills for electricity and gas consumption also serve as feedback, but are not as effective in causing changes in consumption patterns. [3] presents a review of several works about the effectiveness that feedback schemes have on consumption patterns. It could be observed that thanks to the feedback, consumption savings of between 5% and 20% were produced. Also, feedback works well if the following conditions are met: delivered regularly; presented plainly and engagingly; tailored to

the householder; interactive and digital; capable of providing information by appliance; accompanied by advice for reducing electricity use; associated with a challenging goal for energy conservation. In [4] a review of different schemes for energy consumption monitoring is presented, classifying the measurements as direct and indirect. Its effectiveness in producing energy savings is highlighted and it is observed that when the cost of energy is important, the savings tend to be greater.

In recent years there has been a growing development of applications oriented to energy saving, based on the Internet of Things (IoT) and Cloud Computing (CC). Some proposals deal with the automation of energy management, according to models that try to save energy consumption at the same time as providing well-being to the inhabitants of so-called smart buildings [5] and smart houses [6], while others are oriented towards provide feedback to users, so they can manage smarter their consumption.

IoT aims to connect real world devices, sensors and actuators, in order to gather the generated information and process it to produce actions on it. IoT devices generally have little processing power. CC has virtually unlimited storage and processing capacity. The conjunction of IoT and CC allows the development of complex solutions, with ease of access by users through web interfaces and / or mobile applications. The main companies that offer cloud services are oriented to the world of IoT [7], such as Amazon, Google and Microsoft, although it is also worth mentioning others such as Mathworks [8], which allows combining their analytical tools with IoT.

This paper presents a system for monitoring residential energy consumption, based on a network of IoT sensors with different communication alternatives with the cloud, through a local gateway, which also implements some basic functions. In the cloud, data is processed and information and analysis are generated for the user. An outstanding feature of this project is to take advantage of the power lines to connect the sensors to the local gateway, which implies greater flexibility for their installation. The communication protocol used between the local gateway and the cloud is MQTT [9], while in the proximity network can be used MQTT with those devices that support it or an ad-hoc protocol for the most basic devices.

2 Variables of Interest. Measurement Strategies

In the case of electrical energy, the variables to be measured for a direct measurement [4][9] are voltage and current. The advantage of the direct strategy is that it enables accurate measurement and is suitable for variable power consumption. It requires taking into account electrical safety aspects.

Another possibility is to carry out indirect measurements, for example using acoustic, piezoelectric or luminic sensors, among others. These types of sensors allow knowing the on or off status of certain devices from noise or vibrations. In these cases a standard power or consumption should be considered and then the operating time should be measured. This strategy would also be suitable for non-invasively monitoring boilers, generators, engines, etc.

In addition to energy, it is also necessary to sense other variables, such as exterior and interior temperature, ambient humidity, light intensity, presence of people, etc., which allow knowing the context in which energy is being consumed, so that, as a result of an analysis, the system can provide suggestions to the user, as well as so that the user can decide based on greater quantity and quality of information. As an example: that the heating is not on when the maximum comfort temperature has been exceeded, or when the outside temperature is above a certain minimum.

3 Network Architecture

Industrial Internet Consortium defines a three-tier model for the IoT architecture [10]: Edge, Platform, and Enterprise. Edge collects data from devices through the proximity network. Platform receives, processes and forwards data from Edge to Enterprise and control commands from Enterprise to Edge; it can also offer some non-domain-specific services, such as queries and analytics. Enterprise implements domain-specific applications, decision support systems, and provides interfaces for end users; it receives data from Edge and Platform and sends control commands to Platform and Edge. Fig. 1 shows a simplified schema of the system over the three-tier model.

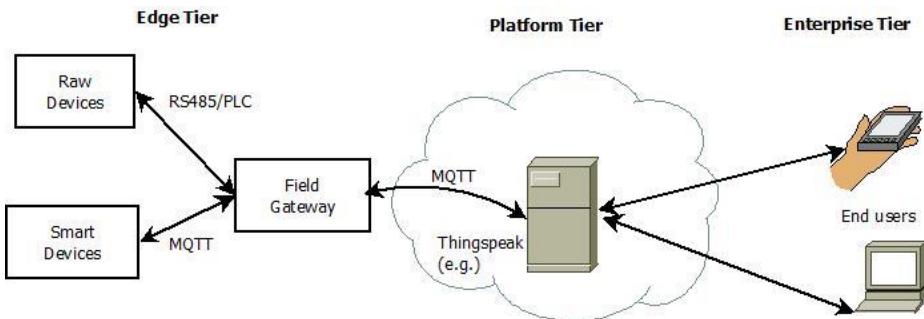


Fig. 1. Structure of the IoT + CC system according to the three-tier architecture.

For this stage, ThingSpeak™ [8] has been adopted to implement cloud processing. It provides an IoT analysis platform service that allows to add, visualize and analyze live data streams in the cloud. Additionally, it allows to write and execute Matlab™ code to perform preprocessing, visualization and analysis. ThingSpeak uses channels to store data sent from applications or devices, and data can be read from these channels using HTTP calls and the REST API. You can also use the MQTT subscription method to receive messages every time the channel is updated. However, it would be possible in the future to replace it with another service, even one developed ad-hoc.

Connecting the sensors to the field gateway allows at least two options. In the case of sensors and smart devices, with sufficient memory, processing and networking capacity, they can be connected directly via the local wired or wireless network available on site. On the other hand, for sensors and devices with limited processing and communication capacity, it would be possible to implement an RS-485 network at

9600 or 19200 bit/s, using the electrical cabling as the physical medium. This has the advantage that it does not require additional cabling, simplifying the deployment of the devices.

4 Conclusions and future works

The conjunction of IoT and CC is a valuable tool for energy monitoring and conservation. Based on a certain variety of data collected by the sensors, and the possibility to process a large quantity of them in the cloud, it is possible to obtain online information and analysis, to support decision-making aimed at saving energy.

Future work will continue with the analysis and testing of different indirect measurement strategies, aimed at facilitating the deployment of the sensors, as well as progress in the design of an ad-hoc platform for the storage, processing, analysis and visualization of information.

References

1. República Argentina. Ministerio de Ambiente y Desarrollo Sostenible (2019). *Inventario Nacional de Gases de Efecto Invernadero de Argentina*, ISBN 978-987-47482-4-9, <https://inventariogei.ambiente.gob.ar/files/inventario-nacional-gei-argentina.pdf>, last accessed 2021/04/08.
2. J. Mastelic, L. Emery, D. Previdoli, L. Papilloud, F. Cimmino and S. Genoud, Energy management in a public building: A case study co-designing the building energy management system, 2017 International Conference on Engineering Technology and Innovation (ICE/ITMC), pp. 1517-1523, 2017.
3. Desley Vine, Laurie Buys, Peter Morris, The Effectiveness of Energy Feedback for Conservation and Peak Demand: A Literature Review, Open Journal of Energy Efficiency (OJEE), 2013, 2, pp. 7-15.
4. Fadi Al-Turjman, Chadi Altrjman, Sadia Din and Anand Paul, "Energy monitoring in IoT-based ad hoc networks: An overview". Elsevier, Computers & Electrical Engineering, Vol. 76, June 2019, pp. 133-142.
5. Emna Taktak and Ismael Bouassida Rodriguez, "Energy consumption adaptation approach for Smart Buildings". 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications, Hammamet, Tunisia, October 30 - Nov. 3, 2017, pp. 1670-1377
6. A.R. Al-Ali, I.A. Zualkernan, M. Rashid, R. Gupta and M. Alikarar, "A smart home energy management system using IoT and big data analytics approach", IEEE Transactions on Consumer Electronics, vol. 63, no. 4, pp. 426-434, 2017
7. Paola Pierleoni, Roberto Concetti, Alberto Belli and Lorenzo Palma, "Amazon, Google and Microsoft Solutions for IoT: Architectures and a Performance Comparison". IEEE Access, vol. 8, 2020, pp. 5455-5470.
8. ThingSpeak™, <https://thingspeak.com>, last accessed 2021/04/08.
9. S. Ereno- Quincozes, E. R. Reginaldo-Tubino and J. Kazienk, "MQTT protocol: fundamentals tools and future directions", IEEE Latin America Transactions., vol. 17, no. 9, pp. 1439-1448, Sep. 2019.
10. Industrial Internet Consortium Reference Architecture, https://www.iiconsortium.org/IIC_PUB_G1_V1.80_2017-01-31.pdf, last accessed 2021/04/08.

A contribution to security in IOT system. Reconfigurable Logic Device Technology and Fog Computing.

Osvaldo Marianetti¹, Pablo Godoy¹, Ernesto Chediak¹ and Carlos García Garino¹

¹ Universidad Nacional de Cuyo. Facultad de Ingeniería, Mendoza, Argentina
olmarianetti@gmail.com, pablodgodoy@gmail.com,
ernestochediack@gmail.com, cgarcia@itu.uncu.edu.ar

Abstract. Internet of Things (IoT) presents a scenario in which billions of devices are interconnected and distributed almost anywhere, from the human being bodies to the most remote areas of the planet. In general, computer attacks, can steal or modify important data, bring down critical online services or obtain money illegally. On the other hand, in an IoT context, in addition to all these actions, there are possibilities of doing physical harm to people at a distance or manipulating critical infrastructures. This work proposes a FPGAs (Field Programmable Logic Array), with reconfiguration capabilities and great computational power, as a development alternative to the problems presented by the secure implementation of IoT systems.

Keywords: IoT, security, FPGA, reconfigurable.

1 IoT data security and integrity issues.

IoT augurs a very promising and interesting future. However, there are several security issues to be addressed: a) Technology heterogeneity: Protocol conversions are necessary to make compatible the security mechanisms implemented by different manufacturers; b) IoT computing capacity devices currently do not satisfy the by the security requirements available on other platforms; c) IoT communications are based mostly on wireless technologies.

This technology can suffer many different types of attacks, because the information exchange in IoT devices is quite predictable. Wireless sensor network (WSN) applications are a part of the IoT paradigm. WSN and the IoT share the same application scenarios. Sensor nodes within a WSN network can monitor and interact with each other just as physical and virtual objects do in IoT [1].

2 Wireless sensor network nodes. Design alternatives.

Wireless Sensor Networks (WSN) are based on groups of wireless connection embedded device nodes (sensors, microcontroller or processor plus a transmitter / receiver

module, and so on). One of the main problems to be solved in practice is processing resources optimizing. WSNs uses nodes with general-purpose processors or microcontrollers in their deployment. General-purpose processors are designed to support virtually all types of applications. However, these tools are high priced and their power consumption is not optimized. In the literature the called *soft-core* processors (configurable architecture ones) have been proposed in order to circumvent the above cited drawbacks [2]. This choice optimizes the processor architecture so that it can be tailored to the needs of sensor network applications. There are FPGAs completely optimized on low-power. In the case of Xilinx and Altera, someone boards look promising since both are coupled with powerful specialized blocks and have a static power consumption between 41 mW and 197 mW. However, for applications with less advanced calculus at high-speed, the IGLOO platform [3] provides a limited static power consumption within the μ W and the mW range. The costs of these devices have also dropped considerably.

3 FPGA technology at Fog Computing and Edge layer nodes.

Fog computing is a cloud technology than can be potentially useful in order to improve IoT deployment. The devices generated data are not uploaded directly to the cloud. Instead, the information is preprocessed first in smaller decentralized data centers. This concept encompasses a network that extends from its own limits, where the terminals or sensors generate the data, to the central destination of the data in the public or private cloud or in a proper data center.

The goal of fog computing is to shorten the communication paths between the cloud and the devices in order to reduce the throughput of data on external networks. The nodes fulfill the role of intermediate layer in the network in which it is decided which data are processed locally or remotely. The three layers of a Fog computing infrastructure are:

- a) Edge layer: comprises all the smart devices, place at the edge of the network, of an IoT architecture. The data that generated in this layer is processed in the same node or is to send to a server in the fog layer.
- b) Fog layer: it is based on a proper quantity of high-performance servers that receive the data from the first layer, prepare and send it to the cloud if necessary.
- b) Cloud layer: the cloud layer is the upper layer of a fog computing architecture.

In fog computing the resources for data storage and preparation are distributed in the intermediate layer in the network by means of fog nodes or pre-processing units. Security issues are usually approached considering traditional solutions. A wider security vision is required from the design, where threats are addressed proactively. Reconfigurable logic technology can enable efficient, scalable, and sustainable solutions in this case. The great computational capacity of FPGAs together the possibility to process different types of information, offer a response to address the following requirements:

- a) Capacity and dynamic load management: FPGAs allow the resources available for a given task to be adapted at runtime without complex infrastructures.

- b) Security: Due to the nature of reconfigurable hardware, the system is more resistant to attacks. In addition, more powerful encryption hardware systems based can be added without affecting the operation of the application [4].
- c) Software infrastructure simplification: All the functionality of the Fog IoT node can be included on a FPGA. Then the maintainability and operating cost of the platform are improved.

4 FPGA based WSN and edge layer nodes prototypes.

The design of an architecture of an embedded system based on soft_processors is discussed in this section. The approach for an architecture of a WSN and/or Edge node based on traditional components can be seen in Figure 1. The main drawback of this approach is that components are not reconfigurable and cannot be adapted for flexible working conditions.

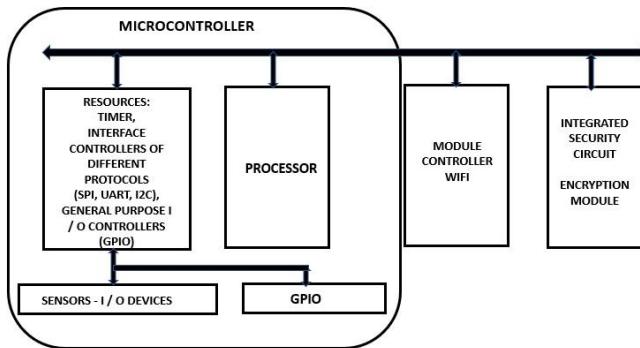


Figure 1. Architecture of a WSN and/or Edge node based on traditional components.

FPGAs is a valuable alternative in order to improve the operability of WSN nodes as has been previously considered in sections 2 and 3. On the other hand, the implementation of FPGA based nodes on the Edge layer can improve the security of the overall system. The proposed architecture has the same functionality of commercial systems including security components in its design, while the characteristics of the nature of reconfigurable hardware are implicitly considered. Then the system is more resistant to attacks. In FPGA-based architecture, its functional units (memories, ports, controllers, timers, etc.) are reconfigurable and adaptable to new requirements, even remotely. A FPGA based architecture [5] scheme is presented in Figure 2. For the development of the embedded system the Quartus II development environment (versions 13.1 web edition and Quartus Prime Lite Edition 17.0) has been used. The QSYS tool of these environments for the generation of the SOPC (system programmable on chip) and the NIOS II software build tool for Eclipse environment have been used for programming the NIOS II / e soft_processor. In a previous work [6] the authors have designed a FPGA based WSN processor node.

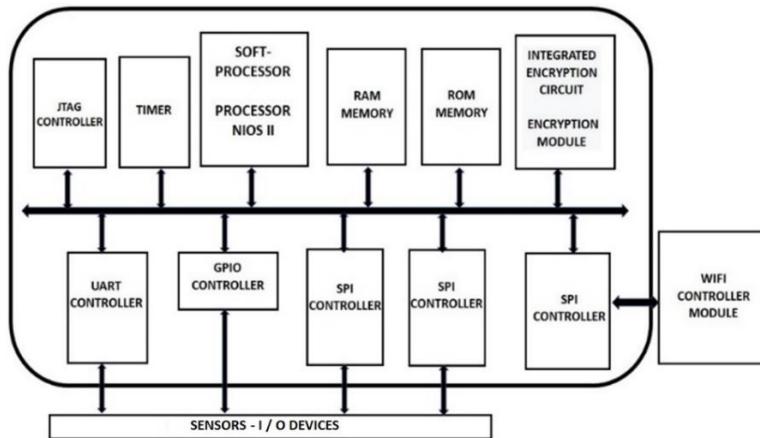


Figure 2. System embedded in reconfigurable hardware.

5. Conclusion.

The proposed prototype can be considered as a proof of concept that allows to research architectures of programmable systems on chip (SOPC) based on FPGAs. This prototype can be optimized in order to operate as a node of a WSN and also in applications of IoT systems. For instance, the proposed tool can be used in order to implement gateways or nodes of the Edge layer [7].

References

1. Baktyan, A. A., & Zahary, A. T. A Review on Cloud and Fog Computing Integration for IoT: Platforms Perspective. *EAI Endorsed Transactions on Internet of Things*. (2018).
 2. Lei Zhou, Qingxiang Liu, Bangji Wang, Peixin Yang, Xiangqiang Li and Jianqiong Zhang. Remote System Update for System on Programmable Chip Based on Controller Area Network. *School of Physical Science and Technology, Southwest Jiaotong Universit*. (2017).
 3. IGLOO platform. <https://www.mdpi.com/1424-8220/12/9/12235/pdf>
 4. A Lattice Semiconductor White Paper IoT Sensor Connectivity and Processing with Ultra-Low Power, Small Form-Factor FPGAs. (2018).
 5. O. Marianetti, P. Godoy, E. Chediak, D. Fontana. La tecnología de lógica reconfigurable como alternativa en la solución a los problemas de seguridad en IoT. XXVI Jornadas de investigación: “Avances y desafíos de la ciencia en pandemia”. UNCUYO. (2020).
 6. O. Marianetti, A. Iglesias, L. Arce. Diseño de un prototipo de procesador soft-core para aplicaciones en nodos de WSN. <https://doi.org/10.18682/cyt.v1i17>. Online ISSN 2344-9217 | Print ISSN 1850-0870. Universidad de Palermo. Facultad de Ingeniería (2017)
 7. Jhansi Naga Sai Surekha, Archana, Hannah Priyanka, Munavvar Hussain. Raju Institute of Technology, Narsapur, India. “An FPGA Implementation of Health Monitoring System using IOT.” http://ijcrt.org/papers/IJCRT_185534.pdf

Forest Fire Simulation in High Performance Computing

Mónica Denham^{1,2[0000-0001-9132-1018]}, Sigfrido Waidelich^{1[0000-0002-2434-580X]}, Viviana Zimmerman^{3[0000-0001-5737-348X]}, and Karina Laneri^{2,4[0000-0001-8536-4695]}

¹ Universidad Nacional de Río Negro. Sede Andina. Centro Interdisciplinario de Telecomunicaciones, Electrónica, Computación y Ciencia Aplicada (CITECCA). Río Negro, Argentina.

² Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, Argentina.

³ Centro Regional Universitario Bariloche, Universidad Nacional del Comahue, Argentina.

⁴ Grupo de Física Estadística e Interdisciplinaria. Gerencia de Física - Comisión Nacional de Energía Atómica. Río Negro, Argentina.

mdenham@unrn.edu.ar

Abstract. This work presents a research line related to the simulation of wild fires. It is being developed in San Carlos de Bariloche, Río Negro, Patagonia Argentina. Our multidisciplinary team is composed by computer scientists, physicists, atmospheric and biological scientists and electronic engineers, coming from different institutions: CONICET, Universidad Nacional de Río Negro, Centro Atómico Bariloche, Universidad Nacional del Comahue. As a result of this interaction we've developed a forest fire simulator with a visual interface that allows to test different scenarios for fire propagation. The application is tailored according to the needs of local firefighters with whom we define the main features needed to eventually use our simulator for management purposes.

Keywords: Forest fire behaviour · Simulation · High Performance Computing.

1 Introduction

Each year forest fires affect hundreds of thousands, or even millions of hectares of natural vegetation across the world. They are a real hazard in Patagonia (Argentina), as well as in the whole country and all over the world. Immediate consequences are the lost of native and exotic vegetation, death of animals, reduction of natural and forested areas, CO₂ liberation into the atmosphere, etc. Some of the most important indirect consequences are the acceleration of global warming, land desertification, inundations and lost of buildings through interface zones.

Frequency and severity of wild fires had increased during last decades. A real feedback exists between big forest fires and current global warming. Climate

warming causes rising temperatures, drier conditions and longer fire seasons. In turn, huge forest fires release big amount of CO₂ to the atmosphere, which increases the green house phenomenon, rising global climate warming [5].

Even though great effort and work is done in this area in our country, the development of computational tools for forest fire simulation are not mature enough. Moreover, fuel types (based on typical vegetation in a region) are not developed for our landscapes.

Our team is in a constant communication with firefighters of the Departamento de Incendios, Comunicaciones y Emergencias of the Parque Nacional Nahuel Huapi (ICE-PNNH). ICE-PNNH staff contribute to the guidance on new functionalities for the simulator and requirements to be able to use it to simulate different scenarios of interest.

In this work we expose the main lines and developments of our research work.

2 Current Developments

The main goal of our work is to develop a high performance application to simulate and visualize wild fire propagation. This application can be useful in the decision-making process, for mitigation, control and eventually prevention of forest fires.

Our forest fire simulator is based on a cellular automaton (CA) often used to model the spread of a given phenomena in 2 dimensions. Within that CA the space is discretised into a grid of cells, while time is discretised into equal time steps. Then, a mathematical model for forest fire spread runs over all cells of the grid for each simulation time step.

On a previous step, a fitting procedure to real fire scars was implemented. The model parameters were estimated using genetic algorithms over millions of simulations. The best set of parameters found were used to simulate and visualize the fire propagation [3].

Besides the forest fire simulator, our research group had developed an application for calculating the Fire Weather Index (FWI) [1]. The FWI deals with the state of fine and coarse vegetation (fire fuel) in combination with weather conditions. This development was requested by the ICE-PNNH and then, the FWI assessment was included as a board that can be deployed on the simulator menu [2] [6].

At the same time, we are working in the design, implementation and testing of a new mathematical model for forest fire behaviour. This model is based on differential equations for Reaction-Diffusion-Convection processes for 2D propagation (RDC model from now on). This dynamical model aims to understand the key mechanisms driving fire propagation in the Patagonian region [2].

2.1 High Performance simulator implementation details

The core of our simulator is the CA. It was developed in parallel for NVIDIA GP-GPUs (General Purpose Graphic Processing Unit), matching High Perfor-

mance Computing paradigm requirements. CUDA (Compute Unified Device Architecture) was used as programming model as well as C and C++ extension. The simulator user interface is also executed on the graphic card and execution times match real time requirements.

During simulation time, a fire behaviour model is executed within each CA cell. Previously, the fire behaviour model was implemented using a statistical model (according to [4]). This model was created by the study of different wild fires occurred in the NW Andean Patagonia.

Given that fuel models are still not available for our region and vegetation acts as fire fuel, we used land vegetation as input for our forest fire simulator. Required simulator inputs are: vegetation cover (3 kinds of vegetation are taken into account: shrubland, forest and non fuel), wind speed, wind direction, terrain altitude, slope and aspect. Simulator's inputs and outputs are raster files.

The simulator shows the area of interest, where each cell is coloured by the vegetation type. Then, the user can start fire spreading by setting one (or more) ignition points. The simulator runs and shows 10 instances of progress of the fire. When a cell is reached by the fire, it is coloured red indicating its new state. Furthermore, when user passes the mouse pointer over each cell, a label indicates how many times the fire had reached that cell (due to the 10 simulation instances). This is an indicator of cell ignition probability.

In addition, the user can set firebreaks all over the area of interest. Then, when fire reaches a cell with a firebreak, it can't propagate through that cell. Fire progress can go forward or backward. Setting firebreaks and the possibility of rewinding the fire progress, allow the user to decide where a treatment of the fuel can mitigate or halt fire spread.

As mentioned, the FWI computation was included in the simulator. Once FWI value is calculated for a specific date (using weather data as input), it is possible to calculate the fire risk for each cell (fire risk is based on the FWI value and local cell vegetation).

Different data layers describe the fire scenery (vegetation, altitude, slope, aspect, wind velocity and direction, contour lines). These data layers can be turned on and off in order to display the specific data during the simulation. Moreover, these layers can be overlapped using their transparencies.

Fig. 1 shows the execution of 10 different simulations (red, yellow and orange cells) after setting 2 ignition points. New ignition points can be set by the user during simulation. The map was zoomed and rotated in order to focus on a particular area of the map.

3 Next steps

Different goals are proposed within the scope of this project. Short term goals deal with the iterative steps needed for implementation, testing and fitting of the new RDC model. Meticulous analysis of input and output simulations of each of the RDC equations are ongoing. After model testing, using synthetic maps that will be compared with simulations, real maps will be used to validate the

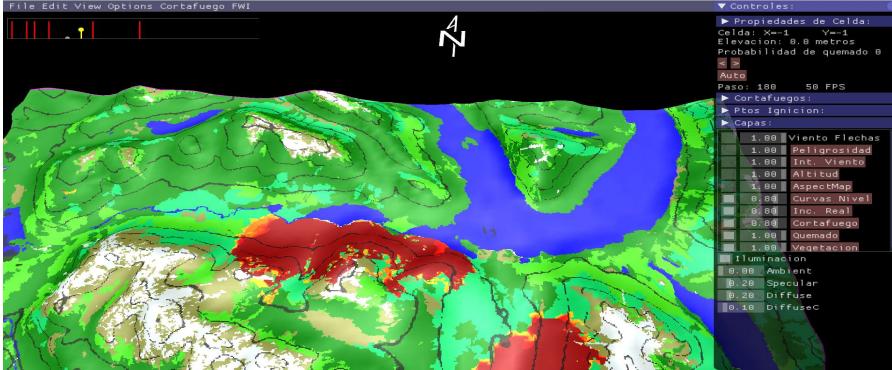


Fig. 1. Simulator's main view. Each vegetation kind is coloured with different colors. Cells in red, yellow and orange are burned. This scenery is located at the South of Lago Mascardi, Patagonia Argentina

mathematical model. This validation phase will be accomplished by fitting the model to real data, finding the best set of parameters that will be used to show simulations on the visual interface.

In a near future, internet repositories will be used for hosting the source codes and the user manual documentation of our applications.

We hope to increase the communication with different fire management institutions, municipalities and research groups, in order to work together on the requirements needed for our forest fire simulator.

References

1. Canadian Forest Fire Weather Index (FWI) System. Available at: <https://cwfis.cfs.nrcan.gc.ca/background/summary/fwsi>, accessed on 2019-07
2. Denham, M., Laneri, K., Zimmerman, V., Waidelich, S.: First steps towards a dynamical model for forest fire behaviour in argentinian landscapes. Journal of Computer Science and Technology **20**(2), e09 (Oct 2020)
3. Denham, M., Laneri, K.: Using efficient parallelization in graphic processing units to parameterize stochastic fire propagation models. Journal of Computational Science **25**, 76 – 88 (2018)
4. Morales, J.M., Mermoz, M., Gowda, J.H., Kitzberger, T.: A stochastic fire spread model for north patagonia based on fire occurrence maps. Ecological Modelling **300**(0), 73 – 80 (2015)
5. Tiribelli, F.: Phd Thesis: Cambios en la inflamabilidad con la edad post-fuego en bosques y matorrales del NO de la Patagonia: composición, estructura y combustibles finos. Universidad Nacional del Comahue. Centro Regional Universitario Bariloche. (2019)
6. Waidelich, S., Zimmerman, V., Laneri, K., Denham, M.: Fire weather index assessment and visualization. In: de Universidades con Carreras en Informática, R. (ed.) Actas XXV Congreso Argentino de Ciencias de la Computación (CACIC 2019). Red de Universidades con Carreras en Informática, SEDICI (2019)

Early Experiences Migrating CUDA codes to oneAPI

Manuel Costanzo¹ , Enzo Rucci¹ *, Carlos García-Sánchez² , and Marcelo Naiouf¹ 

¹ III-LIDI, Facultad de Informática, UNLP – CIC.

La Plata (1900), Bs As, Argentina

{mcostanzo,erucci,mnaiouf}@lidi.info.unlp.edu.ar

² Dpto. Arquitectura de Computadores y Automática, Universidad Complutense de Madrid. Madrid (28040), España

garsanca@dacya.ucm.es

Abstract. The heterogeneous computing paradigm represents a real programming challenge due to the proliferation of devices with different hardware characteristics. Recently Intel introduced oneAPI, a new programming environment that allows code developed in DPC++ to be run on different devices such as CPUs, GPUs, FPGAs, among others. This paper presents our first experiences in porting two CUDA applications to DPC++ using the oneAPI dpct tool. From the experimental work, it was possible to verify that dpct does not achieve 100% of the migration task; however, it performs most of the work, reporting the programmer of possible pending adaptations. Additionally, it was possible to verify the functional portability of the DPC++ code obtained, having successfully executed it on different CPU and GPU architectures.

Keywords: oneAPI · SYCL · GPU · CUDA· Code portability

1 Introduction

In the last decade, the quest to improve the energy efficiency of computing systems has fueled the trend toward heterogeneous computing and massively parallel architectures [1]. One effort to face some of the programming issues related to heterogeneous computing is SYCL³, a new open standard from Khronos Group. SYCL is a domain-specific embedded language that allows the programmer to write single-source C++ host code including accelerated code expressed as functors. In addition, SYCL features asynchronous task graphs, buffers defining location-independent storage, automatic overlapping kernels and communications, interoperability with OpenCL, among other characteristics [2].

Recently, Intel announced the *oneAPI* programming ecosystem that provides a unified programming model for a wide range of hardware architectures. At the core of the oneAPI environment is the Data Parallel C++ (DPC++) programming language, which can be summarized as C++ with SYCL. Additionally,

* Corresponding author.

³ <https://www.khronos.org/registry/SYCL/specs/sycl-2020/pdf/sycl-2020.pdf>

DPC++ also features some vendor-provided extensions that might be integrated into these standards in the future [3].

Today, GPUs can be considered the dominant accelerator and CUDA is the most popular programming language for them [4]. To tackle CUDA-based legacy codes, oneAPI provides a compatibility tool (`dpct`) that facilitates the migration to the SYCL-based DPC++ programming language. In this paper, we present our experiences from porting two original CUDA apps to DPC++ using `dpct`. Our contributions are: (1) the analysis of the `dpct` effectiveness for CUDA code migration, and (2) the analysis of the DPC++ code's portability, considering different target platforms (CPU and GPUs).

2 The oneAPI Programming Ecosystem

oneAPI⁴ is an industry proposal based on standard and open specifications, that includes the DPC++ language and a set of domain libraries. Each hardware vendor provides its own compatible implementations targeting different hardware platforms, like CPUs and accelerators. The Intel oneAPI implementation consists of the Intel DPC++ compiler, the Intel `dpct` tool, multiple optimized libraries, and advanced analysis and debugging tools [5].

3 Experimental Work and Results

3.1 Migrating CUDA Codes to oneAPI

`dpct` assists developers in porting CUDA code to DPC++, generating human readable code wherever possible. Typically, `dpct` migrates 80-90% of code in automatic manner. In addition, inline comments are provided to help developers finish migrating the application. In this work, we have selected two CUDA applications from the CUDA Demo Suite (CDS)⁵. Both codes were translated from CUDA to DPC++ using the `dpct` tool.

Matrix Multiplication (MM) This app computes a MM using shared memory through tiled approach. Fig. 1 shows an example of the memory transference translations. Because `checkCudaErrors` is a utility function (it is not part of the CUDA core), `dpct` inserts a comment to report this situation. Then, the programmer must decide whether to remove the function or redefine it.

Fig. 2 shows the kernel invocations. At the top, the original CUDA kernel's call and, at the bottom, the migrated DPC++ code (only a portion is included due to the lack of space). On the one hand, `dpct` adds comments informing the programmer that it is possible that the size of the *work-group* exceeds the maximum of the device, being his responsibility to prevent this from happening. On the other hand, the resulting code is longer and more complex than the

⁴ <https://www.oneapi.com/>

⁵ <https://docs.nvidia.com/cuda/demo-suite/index.html>

```
// copy host memory to device
checkCudaErrors(cudaMemcpy(d_A, h_A, mem_size_A, cudaMemcpyHostToDevice));

// copy host memory to device
/*
DFTI003: Migrated API does not return error code. (t, 0) is inserted. You may need to rewrite this code.
*/
checkCudaErrors((q_ctl.memcpy(d_A, h_A, mem_size_A).wait(), 0));
```

Fig. 1: MM memory transference. Up: Original CUDA code. Down: Resultant DPC++ code.

```
for (int j = 0; j < nIter; j++) {
    if (block_size == 16) {
        MatrixMulCUDA<16> <<< grid, threads >>>(d_C, d_A, d_B, dimsA.x, dimsB.x);
    } else {
        MatrixMulCUDA<32> <<< grid, threads >>>(d_C, d_A, d_B, dimsA.x, dimsB.x);
    }
}

for (int j = 0; j < nIter; j++) {
    if (block_size == 16) {
        /* DFTI048:11: The workgroup size passed to the SYCL kernel may exceed the limit. To get the device limit,
        query info::device::max work group size. Adjust the workgroup size if needed. */
        q_ctl.submit([j](sycl::handler & cgh) {
            sycl::range<2> As_range_ctl(16, 16); sycl::range<2> Bs_range_ctl(16, 16);

            sycl::accessor<float, 2, sycl::access::mode::read_write, sycl::access::target::local> As_acc_ctl(As_range_ctl, cgh);
            sycl::accessor<float, 2, sycl::access::mode::read_write, sycl::access::target::local> Bs_acc_ctl(Bs_range_ctl, cgh);

            cgh.parallel_for<class kernel3>(sycl::nd_range<3>(grid * threads, threads),
                [<] (sycl::nd_item<3> item_ctl) {
                    MatrixMulCUDA<16>(d_C, d_A, d_B, dimsA[2], dimsB[2], item_ctl,
                        dpct::accessor<float, dpct::local, 2>(As_acc_ctl, As_range_ctl),
                        dpct::accessor<float, dpct::local, 2>(Bs_acc_ctl, Bs_range_ctl));
                });
        });
    } else {
```

Fig. 2: MM kernel call. Up: Original CUDA code. Down: Resultant DPC++ code (portion).

CUDA original code. However, it is important to remark that this code is the result of an automatic translation. By following the DPC++ conventions, it could be significantly simplified.

Finally, Fig. 3 shows part of the kernel bodies, resulting in very similar codes. `dpct` manages to correctly translate the local memory usage, although it defines the arrays outside the loop as opposed to the CUDA case. In addition, it can be noted that `dpct` effectively translates the `unroll` directive and the synchronization barriers.

Reduction (RED) This app computes a parallel sum reduction of large arrays of values. The CUDA code includes several important optimization strategies like reduction using shared memory, `_shfl_down_sync`, `_reduce_add_sync` and `cooperative_groups reduce`.

In this case, `dpct` is not able to translate advanced functionalities such as CUDA *Cooperative Groups*. Fig. 4 presents the comment inserted by `dpct` to inform the programmer about this issue. Even so, the tool manages to translate most of the original CUDA code, leaving little work to the programmer.

```

// Loop over all the sub-matrices of A and B required
// to compute the block sub-matrix
for (int a = aBegin, b = bBegin; a <= aEnd; a += aStep, b += bStep) {
    // Declaration of the shared memory array As/Bs used
    // to store the sub-matrix of A/B
    shared_float As[BLOCK_SIZE][BLOCK_SIZE];
    shared_float Bs[BLOCK_SIZE][BLOCK_SIZE];
    // Load the matrices from device memory to shared memory;
    // each thread loads one element of each matrix
    As[ty][tx] = A[a + wA * ty + tx];
    Bs[ty][tx] = B[b + wB * ty + tx];
    // Synchronize to make sure the matrices are loaded
    __syncthreads();
    // Multiply the two matrices together;
    // each thread computes one element of the block sub-matrix
    #pragma unroll
    for (int k = 0; k < BLOCK_SIZE; ++k) {
        Csub += As[ty][k] * Bs[k][tx];
    }
    // Synchronize to make sure that the preceding computation
    // is done before loading two new
    // sub-matrices of A and B in the next iteration
    __syncthreads();
}

// Loop over all the sub-matrices of A and B required
// to compute the block sub-matrix
for (int a = aBegin, b = bBegin; a <= aEnd; a += aStep, b += bStep) {
    // Declaration of the shared memory array As/Bs used
    // to store the sub-matrix of A/B
    shared_float As[BLOCK_SIZE][BLOCK_SIZE];
    shared_float Bs[BLOCK_SIZE][BLOCK_SIZE];
    // Load the matrices from device memory to shared memory;
    // each thread loads one element of each matrix
    As[ty][tx] = A[a + wA * ty + tx];
    Bs[ty][tx] = B[b + wB * ty + tx];
    // Synchronize to make sure the matrices are loaded
    item_ctl.barrier();
    // Multiply the two matrices together;
    // each thread computes one element of the block sub-matrix
    #pragma unroll
    for (int k = 0; k < BLOCK_SIZE; ++k) {
        Csub += As[ty][k] * Bs[k][tx];
    }
    // Synchronize to make sure that the preceding computation
    // is done before loading two new
    // sub-matrices of A and B in the next iteration
    item_ctl.barrier();
}

```

Fig. 3: MM kernel. Left: Original CUDA code. Right: Resultant DPC++ code.

```

cg::thread_block_tile<32> tile32 = cg::tiled_partition<32>(cta);

/* DPC10070: Migration of this CUDA API is not supported by the Intel(R) DPC++ Compatibility Tool. */
cg::thread_block_tile<32> tile32 = cg::tiled_partition<32>(cta);

```

Fig. 4: RED kernel. Up: Original CUDA code. Down: Resultant DPC++ code.

3.2 Experimental Results

Two hardware platforms were used for the experimental work. The first comprises an Intel Core i3-4160 3.60GHz processor, 16GB main memory and a NVIDIA GeForce RTX 2070 GPU. The second has an Intel Core i9-10920X 3.50GHz processor, 32GB main memory, and an Intel Iris Xe MAX Graphics GPU, from the Intel DevCloud⁶. oneAPI and CUDA versions are 2021.2 and 10.1, respectively. In addition, different workloads were configured for MM ($nIter = 10$; $wA, wB, hA, hB = \{4096, 8192, 16384\}$). Finally, to run DPC++ code on NVIDIA GPUs, several modifications had to be made to the build, as it is not supported by default⁷.

Table 1 shows the execution times of MM (CUDA and DPC++ versions) on the different experimental platforms. Before analyzing the execution times, it is important to remark that the DPC++ code was successfully executed on all the selected platforms and that the results were correct in all cases.

On the RTX 2070, the DPC++ code presents some overhead compared to the original code. However, it should be noted that these results are not final since the oneAPI support for NVIDIA GPUs is still experimental⁸. In fact, currently the code generation does not consider any particular optimization passes.

The DPC++ code was compiled and successfully executed on two different Intel devices: a CPU and a GPU. In this way, we verified its functional portability

⁶ <https://software.intel.com/content/www/us/en/develop/tools/devcloud.html>

⁷ <https://intel.github.io/llvm-docs/GetStartedGuide.html>

⁸ <https://www.codeplay.com/portal/news/2020/02/03/codeplay-contribution-to-dpcpp-brings-sycl-support-for-nvidia-gpus.html>

Table 1: MM execution times on the target platforms

Size	NVIDIA RTX 2070 (CUDA)	NVIDIA RTX 2070 (oneAPI)	Intel Core i9-10920X	Intel Iris Xe MAX
4096	1.3	1.4	9.2	6.3
8192	11.1	15.3	102.8	50.4
16384	89.3	122.9	919.5	401.1

on different architectures. Little can be said about its performance due to the absence of an optimized version for both Intel devices. However, there is probably significant room for improvement considering that the ported code was compiled and executed with minimal programmer intervention.

4 Conclusions and Future Work

In this paper, we present our first experience migrating CUDA code to DPC++ using the Intel oneAPI environment. First, we were able to test the effectiveness of `dpct` for the selected test cases. Despite not translating 100% of the code, the tool does most of the work, reporting the programmer of possible pending adaptations. Second, it was possible to verify the functional portability of the obtained DPC++ code, by successfully executing it on different CPU and GPU architectures.

As future work, we are interested in deepening the experimental work. In particular, we want to include other test cases, hardware architectures, and metrics (like performance portability).

References

- [1] H. Giefers et al. “Analyzing the energy-efficiency of sparse matrix multiplication on heterogeneous systems: A comparative study of GPU, Xeon Phi and FPGA”. In: *2016 IEEE ISPASS*. 2016, pp. 46–56.
- [2] Ronan Keryell and Lin-Ya Yu. “Early Experiments Using SYCL Single-Source Modern C++ on Xilinx FPGA”. In: *Proceedings of the IWOCL ’18*. Oxford, UK: ACM, 2018. doi: 10.1145/3204919.3204937.
- [3] S. Christgau and T. Steinke. “Porting a Legacy CUDA Stencil Code to oneAPI”. In: *2020 IEEE IPDPSW*. May 2020, pp. 359–367. doi: 10.1109/IPDPSW50202.2020.00070.
- [4] Manuel Costanzo et al. “Comparison of HPC Architectures for Computing All-Pairs Shortest Paths. Intel Xeon Phi KNL vs NVIDIA Pascal”. In: *Computer Science – CACIC 2020*. Vol. 1409. 2021, pp. 37–48. doi: 10.1007/978-3-030-75836-3_3.
- [5] Nikita Hariharan et al. “Heterogeneous Programming using OneAPI”. In: *Parallel Universe* 39 (2020), pp. 5–18.

Artificial and Computational Intelligence

Classic and recent (neural) approaches to automatic text classification: a comparative study with e-mails in the Spanish language

Juan M. Fernandez^{1,2}, Nicolás Cavasin³, and Marcelo Errecalde⁴

¹ Master Student at Computer Science School, La Plata National University

² Professor and Researcher at Luján National University

³ Luján National University

⁴ Professor and Researcher at LIDIC, San Luis National University

{jmfernandez, ncavasin}@unlu.edu.ar, merreca@uns1.edu.ar

Abstract. Currently, millions of data are generated daily and its exploitation and interpretation has become essential at every scope. However, most of this information is in textual format, lacking the structure and organisation of traditional databases, which represents an enormous challenge to overcome.

Over the course of time, different approaches have been proposed for text representation attempting to better capture the semantic of documents. They included classic information retrieval approaches (like Bag of Words) to new approaches based on neural networks such as basic word embeddings, deep learning architectures (LSTMs and CNNs), and contextualized embeddings based on attention mechanisms (Transformers). Unfortunately, most of the available resources supporting those technologies are English-centered.

In this work, using an e-mail-based study case, we measure the performance of the three most important machine learning approaches applied to the text classification, in order to verify if new arrivals enhance the results from the Spanish language classification models.

Keywords: Text Classification · SVM · Word2Vec · LSTM · BERT

1 Introduction

As a result of the massive access to the internet, millions and millions of data are daily generated and its exploitation and interpretation has become essential at every scope. Information retrieval and text mining became, along the years, the most popular investigation fields, specially in the text classification field [5]. Following this direction, papers about text classification can be found since 1957, where the research work only proposed text classification using the words frequency method [9]. Since then, diverse approaches have been developed for text representation and the knowledge creation using them as a data source. Nevertheless, most of the resources available are English-centered, leaving a reduced group of alternatives for the remaining languages. At the same

time, there are not many works with empirical comparisons measuring those new approaches' performance in languages like Spanish, where reliable resources are not frequently available for the implementation of those new text classification approaches. This work presents experiments comparing the performance of the three most relevant approaches of machine learning applied to text classification in order to measure how beneficial their contributions to non-English languages are.

2 Related work

As stated before, even though in the last 60 years several approaches for automatic text classification proliferated, there are not enough researches about the performance of those different strategies on non-English languages. Below is a brief review of the three strategies used in the frame of this research for the study comparison.

#1: BoW+SVM One of the simplest methods for document representation, and also one of the oldest, is called Bag of Words (*BoW*) or vector space model [11]. This technique generates a vector that represents a document using the frequency count of each term inside the document [6] and is called that way because words are taken as features and documents like collections of unordered words [8]. This representation strategy has simplicity as advantage and also the possibility of applying any classification technique to the final representation. One of the most frequently used is the Support Vector Machine (SVM), created in the mid 1990s, which won popularity due to some attractive features and its empirical performance. SVM is based on the statistical learning theory principle Structural Risk Minimization (SRM), which consists in finding the optimal hyperplane that guarantees the smallest real error [7]. For the distances calculus and hyperplanes pursuit, SVM uses functions called kernels [12].

#2: Word2Vec+LSTM A fairly current line of research includes the usage of contextual information in conjunction with simple neural-network models to obtain words and phrases representations in the vectorial space [14]. One of the most popular models is Word2Vec, which has two different architectures namely CBow and Skip-gram [10]. These models of word embeddings are usually complemented with recurrent neural-networks as Long Short-Term Memory (*LSTM*). These neural-networks provide two new features that drastically improve the performance against conventional neural-networks for text processing: they are able to identify the order of text sequences in documents and process different length documents. [1].

#3: BERT As an evolution of the previous strategy, in 2017, a new neural-network architecture, called *Transformer* [13], arose simpler and parallelizable and is only based on attention mechanisms that completely avoid recurrences and convolutions. They can be described as the assignment of a query and a set of key-value pairs to an output where the query, the keys and the values are all vectors. From this logic rises what in nowadays literature is known as the text representation models' actual state-of-art, called Bidirectional Encoder

Representations from Transformers (**BERT**) [4]. Briefly, this framework has two steps: initial pre-training and posterior fine-tuning. During the pre-training step, the model is trained with unlabelled data in different tasks. Then, during the fine-tuning step, the BERT model is first initialized with the pre-trained model's parameters and are finally adjusted using labelled data from posterior tasks.

3 Research methodology

This research work uses a dataset generated from academic questions made by e-mail by students of the National University of Luján to the administrative staff on topics related to academic activities. From a total of 24700 e-mails, 1000 interactions have been selected, labelled by a domain expert and assigned based on four classes (public transport discount ticket, admission to the university, admission requirements, other topics). For the experiments, the original e-mails were used without any human supervision on semantic nor syntactic mistakes. For the approach based on **BoW+SVM** the text was normalized removing stop-words, adding static attributes (such as question's length and punctuation marks usage) and using variations of *n-grams* and characters. On the other hand, for the approaches based on **Word2Vec+LSTM** and **BERT**, the text sequences related to the selected interactions were just normalized. Only for **Word2Vec+LSTM** the stop-words were removed due to the fact that **BERT** was experiencing a decrease in its performance when they were not there. Additionally, Spanish pre-trained word embeddings[2] were used for **Word2Vec+LSTM**. Regarding **BERT**, two pre-trained models were used. One of them is Spanish-native [3] and the other, called *Multilingual* [4], was developed for several languages. For the evaluations, a cross validation approach was adopted with a *5-fold* on the 80% of the training instances while the models were *tested* with the remaining 20% of the instances using *accuracy*, *precision*, *recall* and *f1-score*.

4 Experimental results

In every approach a search for the best hyper-parameters was applied to each strategy, getting the following results⁵:

Table 1: Results of the different learning strategies.

Strategy	Accuracy	Precision	Recall	F1-score
BoW+SVM	0.870	0.862	0.830	0.840
Word2Vec+LSTM	0.835	0.814	0.841	0.820
BERT (Multilingual)	0.860	0.838	0.842	0.840
BERT (BETO)	0.890	0.870	0.885	0.876

⁵ Experiments available at github.com/jumafernandez/clasificacion_correos

Results show **BERT** as the most effective approach for classifying this dataset, using the previously mentioned Spanish pre-trained model. Nevertheless, the difference with the resulting metrics regarding **BOW+SVM** were 0.03 or smaller.

5 Conclusions and future work

Based on the previous results, and considering the 30 years of evolution that this discipline has experienced since **BOW+SVM** first appearance and **BERT**'s presentation, we have verified that the traditional representation and classification methods are still a very competitive option.

However, it is important to keep in mind that e-mails in general, and this dataset in particular, have some features that do not help these cutting-edge models due to its informal manners and syntactic mistakes which are frequently seen in this type of communication model. That is why the precision gap between cutting-edge models and traditional ones is expected to maximize when datasets with cleaner texts are used. At the same time, and for our collection, this could be solved using spell-checkers to purge the documents during the pre-processing step.

References

1. Aggarwal, C.C., et al.: Neural networks and deep learning. Springer **10**, 978–3 (2018)
2. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (August 2019), <https://crscardellino.github.io/SBWCE/>
3. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Fanny, F., Muliono, Y., Tanzil, F.: A comparison of text classification methods k-nn, naïve bayes, and support vector machine for news classification. Jurnal Informatika: Jurnal Pengembangan IT **3**(2), 157–160 (2018)
6. Harish, B.S., Guru, D.S., Manjunath, S.: Representation and classification of text documents: A brief review. IJCA, Special Issue on RTIPPR (2) pp. 110–119 (2010)
7. Islam, M.R., Chowdhury, M.U., Zhou, W.: An innovative spam filtering model based on support vector machine. In: CIMCA-IAWTIC'06. vol. 2, pp. 348–353. IEEE (2005)
8. Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K.: Fast text categorization using concise semantic analysis. Pattern Recognition Letters **32**(3), 441–448 (2011)
9. Luhn, H.P.: A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development **1**(4), 309–317 (1957)
10. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
11. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)
12. Skiena, S.S.: The data science design manual. Springer (2017)

13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
14. Wu, L., Yen, I.E., Xu, K., Xu, F., Balakrishnan, A., Chen, P.Y., Ravikumar, P., Witbrock, M.J.: Word mover's embedding: From word2vec to document embedding. arXiv preprint arXiv:1811.01713 (2018)

Acknowledgement

Authors are grateful to the Center of Research, Teaching and TIC Extension from the National University of Luján (CIDEATIC) for providing the computational resources that allowed this project's experiments to be executed.

MbedML: A Machine Learning Project for Embedded Systems

César A. Estrebou^{1[0000-0001-5926-8827]}, Martín Fleming², Marcos Saavedra², and Federico Adra²

¹ Instituto de Investigación en Informática LIDI, Facultad de Informática,
Universidad Nacional de La Plata
{cesarest}@lidi.info.unlp.edu.ar

² Facultad de Informática, Universidad Nacional de La Plata

Abstract. This article describes the tasks being carried out within the framework of a research and development project on machine learning techniques and algorithms applied to small devices. It includes a brief review of the available technologies, online development platforms, work methodology and open source software for the implementation of solutions. Experiments carried out on a proper implementation of an inference algorithm for convolutional neural networks are also presented with interesting preliminary results regarding existing implementations.

Keywords: Machine learning · Embedded Systems · Internet of Things · Convolutional Neural Networks

1 Introduction

As reported by Cisco [1] and Statista[2], it is estimated that the number of IoT devices connected to the Internet by 2021 will be between 10,600 and 13,800 million. Many of these devices, limited in both hardware resources and processing capacity, upload information to the cloud to be processed, which leads to problems [3, 4] related to bandwidth, response delays, high computational and storage costs, higher energy consumption, among others. To address these problems, the popular concept of *Edge Computing* arises, which refers to the transfer of total or partial computing from the cloud to the devices located at the edge of the network. In this way you can take advantage of the computing power of these low-power devices that can execute millions of instructions per second despite their limitations. In general, machine learning and in particular deep learning have the potential to make important contributions since they can provide robust solutions [5] as long as they can be adapted to the capacity of the edge computing devices.

Both the area related to *edge devices* and machine learning have received a lot of attention from large hardware and software companies. This drive that combines machine learning techniques with different platforms for embedded systems, brings together and facilitates the development of applications that run on small microcontrollers, something that until recently seemed unthinkable.

With all of the above as motivation comes this research and development project aimed at documenting, studying and implementing traditional machine learning techniques adapted to small devices.

2 Software and Hardware Tools

The main objectives of this project are to document, analyze, test and develop both machine learning and deep learning tools and techniques adapted to devices with limited computing capacity and resources.

Part of the work carried out in the project includes the study of different software tools that develop solutions based on machine learning techniques on microcontrollers. The following sections briefly discuss the work done so far.

2.1 Online Development Platforms, Frameworks and Libraries

There are several online platforms that automate the entire development process, or at least a good part of it. Platforms such as *AlwaysAI*, *Edge Impulse*, *Qeexo*, and *Cartesiam.AI*, with just data loading and microcontroller model selection, automatically scan a wide variety of algorithms with different configurations and perform the deployment of the final application in the microcontroller. The *OctoML* platform optimizes models previously built by the user using machine learning techniques for efficient execution in the microcontrollers it supports. Although in general, most of the platforms support the Arm Cortex-M MCUs, the low number of supported devices is objectionable.

The open source libraries and frameworks for developing machine learning applications on microcontrollers are few. The software with the greatest potential that has been surveyed and analyzed so far is briefly described below:

- MicroML: Implements Scikit-learn (Python) algorithms in C code. Supports Decision Trees, Random Forest, XGBoost, Gaussian NB, SVM, SEFR.
- eMLearn: Supports Decision Trees, Random Forest, XGBoost, Gaussian NB, Keras fully connected neural networks and audio features extraction.
- TensorFlow Lite: Support for neural networks generated with TensorFlow. Requires 32-bit MCU architectures (ARM and ESP32). Provides tools to adapt your models to microcontrollers.
- TinyML: Supports TensorFlow Lite neural network models in MCU ARM.

2.2 Project Microcontrollers

At this time the project has several development boards to perform tests on the different implementations of machine learning algorithms. In the future, several more are planned to be incorporated. The MCUs currently being experimented on are *Arm Cortex-M3*, *Tensilica L106* and *Xtensa LX6* and the technical characteristics can be seen in the table 1. The decision of the embedded models is based on aspects such as local availability, low cost, computing capacity (medium

to low) and availability of open source software. Regarding connectivity, it was decided to incorporate both IoT and non-IoT devices, since from the point of view of machine learning there are many popular devices without this feature.

Table 1: Relevant technical characteristics of the MCUs used in the project.

Board	MCU	Cores	Clock	Data	Prog.	Connectivity	US\$
Stm32f103c8t6	Arm Cortex-M3	1	72MHz	20KiB	64KiB	No	3,50
NodeMCU ESP8266	Tensilica L106	1	80MHz	80KiB	32KiB	Wi-Fi	3,50
Esp32-Wroom	Xtensa LX6	2	160MHz	520KiB	448KiB	Wi-Fi+BT	8,00

2.3 Machine Learning for Microcontrollers

Due to hardware limitations in terms of data and program memory, it is impossible to perform the generation of the machine learning models (training) on the microcontroller (MCU). For this reason, the model is built in a traditional way on a computer and the corresponding verification tests are performed.

Once the model is generated, a transformation tool is used to reduce its size and thus to fit the MCU's limitations. These types of tools export the model data, adapting from the data types to including the necessary code to execute the model. Finally, tests are performed to verify the effectiveness of the adaptation.

3 Experiments and Results in Convolutional Networks

At the time of writing this article, the team is developing tests on convolutional neural network models on the 3 MCUs used in the project. To perform the tests, it was decided to build a convolutional model with the ability to distinguish a digit in an image. As a data source, the UCI repository database [7] was selected, which is a reduced version of the MNIST database [8] widely used to evaluate image classification algorithms in different areas such as computer vision, machine learning and neural networks. This dataset comprises some 5,620 grayscale images with handwritten digits centered in an 8x8 pixel area.

A convolutional neural network [6] (CNN or ConvNet) was used as a model. This type of network has in its architecture a series of convolutional layers with a nonlinear activation function in its output such as ReLU or tanh. Each layer of the network transforms the representation by applying filters that give a higher level of abstraction. For example, the first layer detects edges, then the second layer detects contours from those edges, then the third layer detects structures from contours and so on until an object is detected.

The objective of the 2 experiments performed was to measure the performance of a convolutional model on different MCUs through the implementations of 2 libraries. For this, a convolutional network was trained to classify 8x8 digit images from the UCI database. In the first experiment the Eloquent TinyML library was used in 2 of the MCUs, leaving out of the test the ARM Cortex-M3 MCU because it was no longer supported. In the second experiment we used a

proprietary implementation of the convolutional neural network inference algorithm that works on the different architectures of the 3 MCUs. Table 2 shows the test results of each experiment. In this it can be seen that the proposed algorithm occupies much less program memory and data than the Eloquent TinyML implementation. Even the significant difference in average inference runtime can also be observed.

Table 2: Experiment results

Exper.	Library	Data Mem.	Prog. Mem.	MCU	Time	Accuracy
1	Eloquent TinyML	Xtensa LX6	23.12 KiB	470.6 KiB	2270 us	97,8 %
2	Own	Xtensa LX6	14.21 KiB	277.1 KiB	606 us	97,8 %
1	Eloquent TinyML	Tensilica L106	51.20 KiB	391.5 KiB	11167 us	97,8 %
2	Own	Tensilica L106	31.01 KiB	274.0 KiB	7568 us	97,8 %
2	Own	Stm32103	2.14 KiB	27.3 KiB	8624 us	97,8 %

4 Final Comments

This article has presented a research and development project of machine learning applied to microcontrollers with low computational power and limited data and program memory. A brief review of both the technologies and the available software has been included and a proper implementation of convolutional neural networks has been presented with satisfactory preliminary results. In the future, we will continue to experiment with algorithms and machine learning techniques in small devices, both our own and those of third parties.

References

1. Cisco, Cisco Annual Internet Report (2018–2023) White Paper <https://www.cisco.com/> Last accessed 30 March 2021
2. Statista, Internet of Things (IoT) and non-IoT active device connections worldwide from 2010 to 2025 <https://www.statista.com/> Last accessed 30 March 2021
3. Farhan, L., Kharel, R., Kaiwartya, O., Quiroz-Castellanos M., Alissa, A. and Abdulsalam M., A Concise Review on Internet of Things (IoT) - Problems, Challenges and Opportunities, 11th International Symposium on Communication Systems, Networks & Digital Signal Processing, pp. 1-6. Budapest, Hungary, 2018
4. Shekhar, S. and Gokhale A., Dynamic Resource Management Across Cloud-Edge Resources for Performance-Sensitive Applications, 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 707-710, NJ, USA, 2017
5. Sharma, K., and Nandal, R., A Literature Study On Machine Learning Fusion With IOT”, 3rd International Conference on Trends in Electronics and Informatics, 2019
6. Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. 1st edn. MIT Press, 2016
7. Blake, C.L. and Merz, C.J., Optical Recognition of Handwritten Digits Dataset, 1998. <https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits> Last accessed 30 March 2021
8. LeCun, Y. and Cortes, C., MNIST handwritten digit database, 2010 <http://yann.lecun.com/exdb/mnist/>. Last accessed 30 March 2021

The current role of machine learning and explainability in actuarial science

Catalina Lozano, Francisco P. Romero, Jesus Serrano-Guerrero, Jose A. Olivas

¹ Universidad de Castilla La Mancha, 13071 Ciudad Real, España

Catalina.lozano@alu.uclm.es,

{FranciscoP.Romero, Jesus.Serrano, JoseA.Olivas}@uclm.es

Abstract. Actuarial science seeks to evaluate, predict and manage the impact of future events. Nowadays, the actuary faces the challenge of predicting and managing risks efficiently, with a universe of information growing exponentially in real-time and with a business dynamic that demands constant competitiveness and innovation. The techniques associated with data engineering and data science open a window of tools that seek, through technology, to improve the processes of product design, pricing, reserves and establishment of market niches practically and realistically, considering the pros and cons that brings the availability and constant updating of information, as well as the computational times that this implies. Therefore, this article aims to review the application of Explainable Machine Learning techniques as an alternative to the development of more efficient and practical actuarial models.

Keywords: Machine Learning, Actuarial Models, Explainability

1 Introduction

Actuarial science, seeking risk modelling through mathematical and statistical techniques, faces new challenges every day, both in the volume of existing information to improve its modelling capacity and the nature of the different problems. The techniques associated with Artificial Intelligence and Machine Learning provide a series of tools whose purpose is to improve the processes of product design, pricing, reservations, and establishment of market niches practically and realistically [1]. However, there is an essential limitation in the practical application of complex models that are difficult to interpret and audit, which is related to the strict regulations that regulate the financial sector, as it is a systemic risk business and of vital importance for the world economy, so that for specific processes this type of model is usually ruled out. In this sense, the use of explainable AI technique - Local Interpretable Model-Agnostic Explanations (LIME) [2], The Partial Dependence Plots (PDPs) [3] - would make it possible to understand and evaluate the capacity of the results of the proposed models and investigate the relationships between variables, thus facilitating the understanding and monitoring of the adequacy of these models.

The analysis of the main issues related to the article could be divided two main lines are identified and presented below.

2.1 Artificial Intelligence in actuarial science

In the actuarial field, standard models and different combinations of techniques that are already well established in the field continue to be applied; however, in recent years, different works have appeared related to the application of techniques more closely linked to Artificial Intelligence, such as the following: Genetic Algorithms [4] [5], Artificial Neural Networks, Regression Trees [6], Random Forests and Fuzzy Logic [1]. Regarding the specific application of machine learning techniques to the area of insurance, it is usual to find applications from ANOVA applications to classification models by trees and random forests [6]; Markov Decision Process (MDP) [7], fuzzy generalized probabilistic OWA (FGPOWA) [8] and SMuRF [9] in order to improve precision and computational performance.

2.2 Explainable in actuarial science

Data science has evolved rapidly as computational capacity has advanced, and so has the complexity of the models and the difficulty of understanding the relationship of variables. This is the reason why in the last two decades, there has been an increase in the analysis and application of explainability techniques as a final step in the development of Machine Learning models. This set of techniques includes the application of graphical analysis such as LIME [2], X-Shap [10] or Partial Dependence Plots (PDPs) [3]. These techniques provide a graphical explanation of the influence of the variables on the predictions made by the model. Other approaches are those based on game theory, in which the interaction effects between characteristics and the understanding of the structure of the global model based on the combination of many local explanations of each prediction are explored. [11]. It is also worth mentioning the copulas analysis, where it is possible to intuitively construct the relationships between them using the statistical properties of the variables [12]. Any machine learning technique that intends to be applied to real problems in the finance and insurance area must offer transparent and replicable modelling that allows for review and audibility by control agencies and stakeholders. Since this has been a permanent limitation in applying sophisticated or black box models, a further challenge lies in the definition of a framework for the development of explainability analysis within this context [13]. The alternatives based on model agnostic methods that allow, in an aggregated manner, the evaluation of relationships between variables, facilitating the global understanding of the models, should be highlighted. [14]. [15] [16].

The case study involves using different models to estimate the probability of credit default (PD), in line with the requirements of IFRS 9 for an Expected Loss model, for the U.S. Trade and Industry debt portfolio. For this purpose, use was made of public and freely available information published by the World Bank, where there are quarterly default rates for different types of portfolios and information on macroeconomic variables with the same updating periodicity. In the model adjustment process, the classic models based on Autoregressive Vectors are those that offer the best results, in addition to presenting by definition a reasonable degree of interpretability, while in terms of predictive power, the neural network models (NARX and ANN) are those that offer the best results. However, their results cannot be directly explained. An approach based on locally over-trained decision trees was used, making it possible to establish the relationships between the variable to be explained and the independent variables. Specifically, it is possible to identify that both models preponderate the influence of *Expenses at real prices for personal consumption in Durable goods*, this may reflect the direct relationship between the production of goods and services, and the investment and consumption in the medium term. In the final segmentation, *net savings* are included as a decisive factor. However, despite the favorable results, associated with the predictive capacity of the evaluated models, it is not common to see their application, as it is not possible to clearly identify the type of relationships established. Applying explanatory models, it is possible to reinforce the relationship of relevant variables to compare the logic established by the models, facilitating their understanding and revision.

4 Conclusions and Future Work

Currently, a window of possibilities is open for the application of explainability models to actuarial science problems. This is mainly due to the possibility to develop a standardization of the review using explainability techniques in the review and audit processes of artificial intelligence models, which to date have demonstrated better predictive capacity, and their limitation is associated with their understanding and replicability. These kinds of techniques help the companies to create more risk models in the machine learning way.

References

- [1] A. F. Shapiro, "Fuzzy logic in insurance," *Insurance: Mathematics and Economics*, pp. 119-132, 2000.
- [2] M. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," O'reilly, 12 Agosto 2016. [Online]. Available: <https://arxiv.org/abs/1602.04938>.

- [3] B. Greewell, "pdp: An R Package for Constructing," *The R Journal*, vol. 9, pp. 421-436, 2017.
- [4] J. B. Gray and G. Fan, "Classification tree analysis using TARGET," *Computational Statistics and Data Analysis*, pp. 1362-1372, 2008.
- [5] C. M. Farrelly, S. Namuduri and U. Chukwu, "Quantum Generalized Linear Models," 2019. [Online]. Available: <https://arxiv.org/abs/1905.00365v1>.
- [6] R. Henckaerts, M.-P. Côte and K. A. Roel Verbelen, "Boosting insights in insurance tariff plans," 2019. [Online]. Available: <https://arxiv.org/abs/1904.10890v1>.
- [7] E. Krasheninnikova, J. García, R. Maestre and F. Fernández, "Reinforcement learning for pricing strategy optimization in the insurance industry," *Engineering Applications of Artificial Intelligence*, p. 2019, 8-19.
- [8] M. Casanovas, A. Torres-Martínez, Merigó and J. M., "Fuzzy Logic Tools for Pricing Strategy in the Insurance Sector," *International Journal of Machine Learning and Cybernetics*, 2019.
- [9] S. Devriendt, K. Antonio, T. Reynkens and R. Verbelen, "Sparse Regression with Multi-type Regularized Feature Modeling," *Insurance: Mathematics and Economics*, vol. 96, pp. 248-261, 2021.
- [10] L. Bouneder, Y. Léo and A. Lachapelle, "X-SHAP: towards multiplicative explainability of Machine Learning," 2020. [Online]. Available: <https://arxiv.org/abs/2006.04574>.
- [11] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, "Explainable AI for Trees: From Local Explanations to Global Understanding," 2019. [Online]. Available: <https://arxiv.org/abs/1905.04610>.
- [12] B. Barr, K. Xu, C. Silva, E. Bertini, R. Reilly, C. B. Bruss and J. D. Wittenbach, "Towards Ground Truth Explainability on Tabular Data," 2020. [Online]. Available: <https://arxiv.org/abs/2007.10532>.
- [13] K. Kuo and D. Lupton, "Towards Explainability of Machine Learning Models in Insurance Pricing," 2020. [Online]. Available: <https://arxiv.org/abs/2003.10674>.
- [14] C. Lorentzen and M. Mayer, "Peeking into the Black Box: An Actuarial Case Study for Interpretable Machine Learning," Swiss Association of Actuaries SAV, 2020.
- [15] R. Kshirsagar, L.-Y. Hsu, V. Chaturvedi, C. H. Greenberg, M. McClelland, A. Mohan, W. Shende, N. P. Tilmans, R. Frigato, M. Guo, A. Chheda, M. Trotter, S. Ray and A. Lee, "Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans," 2020.
- [16] M. Ariza-Garzón, J. J. Arroyo, A. Caparrini and S.-V. Maria-Jesus, "Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending," *IEEE Access*, vol. 8, pp. 64873- 64890, 2020.

Speech emotion representation: A method to convert discrete to dimensional emotional models for emotional inference multimodal frameworks

Fernando Elkfury¹[0000-0003-2131-604X], Jorge Ierache^{1,2} [0000-0002-1772-9186]

¹ Instituto de Sistemas Inteligentes y Enseñanza Experimental de la Robótica, Universidad de Morón (1708) Morón Argentina.

² Laboratorio de Sistemas Información Avanzados Universidad de Buenos Aires (C1063) Ciudad Autónoma de Buenos Aires, Argentina.
`{felkfury, jierache}@unimoron.edu.ar`

Abstract. Computer-Human interaction is more frequent now than ever before, thus the main goal of this research area is to improve communication with computers, so it becomes as natural as possible. A key aspect to achieve such interaction is the affective component often missing from last decade developments. To improve computer human interaction in this paper we present a method to convert discrete or categorical data from a CNN emotion classifier trained with Mel scale spectrograms to a two-dimensional model, pursuing integration of the human voice as a feature for emotional inference multimodal frameworks. Lastly, we discuss preliminary results obtained from presenting audiovisual stimuli to different subject and comparing dimensional arousal-valence results and it's SAM surveys

Keywords: Emotions, Multimodal Framework, Affective computing.

1 Introduction

Even though speech is the most traditional way of human communication, as a feature for emotion recognition it is not as expressive as one may think. According to Albert Mehrabian [1] voice tone can only transmit a 38% of the emotions a person might feel at a given time. Despite being a feature with a low percentage of expressiveness, in a multimodal environment of emotion analysis it is meaningful for correctly inferring the emotional state of a person by comparing and correcting data from other sensors or methods of emotion assessment. Human-computer interaction is now more and more frequent due to accelerated technological development, although, they are often lacking an affective component. Thus, one of the main goals of the recent computer-human communication development is to improve user experience through making interaction between computers and humans as natural as it is between persons [2]. Current works in this field, such as [3] [4] [5] [6], infer emotion in a categorical manner, usually partially matching Ekman's model [7]. This work, additionally to the categorical approach, provides a preliminary architecture to obtain valence and arousal from a voice sample in the context of a multimodal emotional dimensional approach for emotion elicitation and representation, based on the use of a CNN [8] classifier for determining valence,

and a method to calculate arousal based on the measurement of the voice source dB. The CNN classifier from recent associated projects [9][10] is capable of up to 92% accuracy for a Spanish dataset. In the following section we present and develop our classifiers and the proposed conversion method. In the third section we discuss our preliminary tests. And lastly in the fourth section we discuss partial results and conclusions.

2 Solution development

To improve and facilitate integration of the human voice as a component in an emotional inference multimodal framework we develop a convolutional neuronal network (CNN) classifier that is trained with Mel scale [11] spectrograms from the audio samples in the ELRA [12] emotional data set. We work with the seven emotional labels proposed by Ekman, joy, fear, sadness, anger, disgust, surprise plus a neutral emotion considered by most data sets and tools available. To represent emotions dimensionally we use a Russell's circumflex of emotion [13] updated in the work of Sherer [14] so we can keep working with eight emotions. (See Fig. 1.)

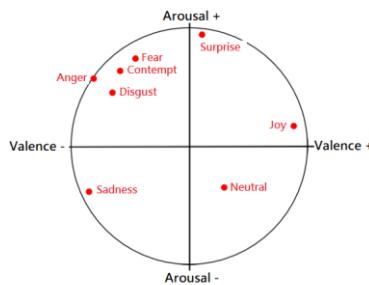


Fig. 1. Circumflex based on Russel and Sherer's work.

To convert emotional models, we start with the subtraction of the probability obtained for the "joy" label of the classifier and the probability of the most negative emotion as is proposed by Leanne in [15]. For example, for a given prediction of the classifier where "joy" equals to 0.8 anger 0.2 "fear" 0.1 and "saddens" 0.3, valence value would be in this case 0.5 from the subtraction $0.8 - 0.3$. According to Leanne the "surprise" emotion is not taken in consideration for the calculation, so we add up to that statement saying that "neutral" also should not be considered. We now must find an associative feature to arousal values. We propose using the difference between the mean dB values of subsequent samples taken during a subject's testing session and the mean dB values of the sample tested. In formula 1 we see a brief description of what is proposed, "x" is the average dB value of the previous samples, "y" is the dB value of the current sample from which we want to obtain the value of arousal and "n" is the number of samples taken in the session so far including the current one. Then the difference between the current sample and the previous average is calculated and rescaled to place it where it corresponds in the circumflex.

$$E = \frac{(\sum_{i=1}^n x)}{n} - y \quad (1)$$

The figure 3 below demonstrates the proposed architecture's workflow to obtain Arousal/valence values from a speech voice sample.

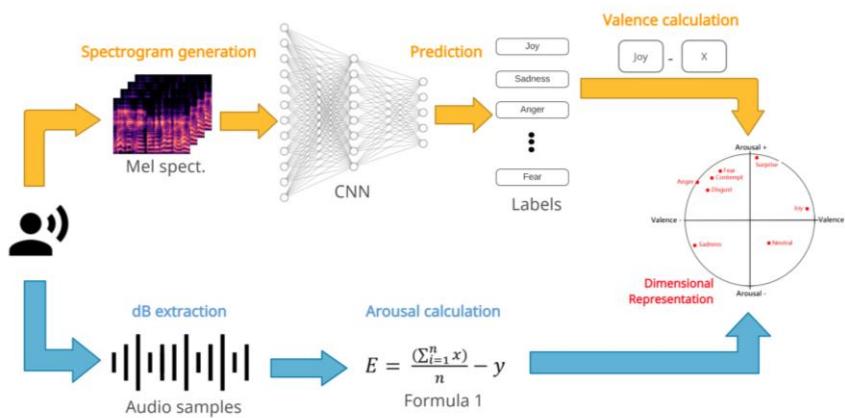


Fig. 2. Example of our architecture's workflow.

3 Test and Results

To evaluate the representation quality of the proposed transformation method, preliminary tests were carried out. In Fig. 4 below, we compare Arousal/Valence values from our CNN classifier and the proposed transformation method (blue mark) with SAM surveys [16] classifications from various subjects (black marks) for a given audio stimuli. Achieving quadrant matching between results.

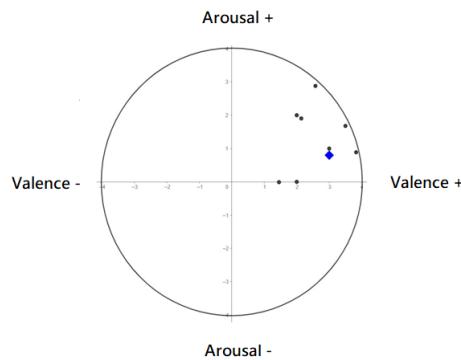


Fig. 3. Example of a test result

4 Conclusions

First of all, dB are a relative measurement unit, in this case they are used for the simple reason of showing the differences in the sample's time series amplitude in a more intuitive and easy-to-work way. The scale may require modifications, which could be determined under empirical tests. Summarizing this arousal values provide us with the visualization of the relationship that might exist between the voice volume and the changes from one emotional state to another. Also, this method relies on the history of a series of samples taken, so the measurement becomes more reliable once a definite trend of the average is established.

References

1. Mehrabian, A.: Communication Without Words. *Communication theory*, 193-200 (2017).
2. Planet, S.: Reconocimiento afectivo automático mediante el análisis de parámetros acústicos y lingüísticos del habla espontánea. (2013).
3. Sánchez-Gutiérrez, M.E., Albornoz, E.M., Martínez-Licona, F., Rufiner, H.L., Goddard, J.: Deep Learning for Emotional Speech Recognition. *Lecture Notes in Computer Science*. 311–320 (2014).
4. Kerkeni, L., Serrestou, Y., Mbarki, M., Raouf, K., Ali Mahjoub: Automatic Speech Emotion Recognition Using Machine Learning. *Social Media and Machine Learning*. (2020).
5. Mustaqeem, Kwon, S.: A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors (Basel)*. 20, 183 (2019).
6. Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y., Kwon, S., Baik, S.W.: Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*. 78, 5571–5589 (2017).
7. Ekman, P.: Basic Emotions. In *Handbook of Cognition and Emotion*, pp. 45–60. John Wiley & Sons, Ltd (2005)
8. K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 193–202 (1980).
9. Elkfury, F., Ierache, J.: Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educación emocional. XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021). Chilecito, 2021. In press.
10. Elkfury, F., Ierache, J.: Clasificación y representación de emociones en el discurso hablado en español empleando Deep Learning. *RISTI - Revista Ibérica de Sistemas e Tecnologías de Información*, versión impresa ISSN 1646-9895 n°42. In press.
11. Volkmann J., Stevens S. S., Newman E. B.: A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 185-190 (1937).
12. ELRA catalog page, <http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/>, last accessed 2021/4/8.
13. Russell, J. A.: A circumplex model of affect. *Journal of Personality and Social Psychology*, vol. 39, 1161–1178 (1980).
14. Scherer, K. R.: What are emotions? And how can they be measured? *Social Science Information*, vol. 44, 695–729, (2005).
15. Loijens L., Krips O.: FaceReader Methodology Note. <https://www.noldus.com/facereader/resources>, last accessed 2021/4/8.
16. Lang, P. J.: The cognitive psychophysiology of emotion: Fear and anxiety. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders*, 131–170 (1985).

Querying on Google Sheets

Designing a Sentiments Analysis Alternative for rating tweets regarding the Ecuadorian 2021 Presidential Campaigns

Sariah López-Fierro¹ , Carlos Chiriboga², and Rubén Pacheco³

¹ Universidad T. F. Santa María, Chile

sariah.lopez@sansano.usm.cl

² Soluciones Wandarina S. A., Ecuador

carlos@wandarina.com

<http://www.wandarina.com>

³ Universidad de E. Espíritu Santo, Ecuador

rpachecov@uees.edu.ec

Abstract. This document contains an approach for the implementation of a sentiment analysis alternative after using Google Sheets for the rating of tweets retrieved with respect to the Ecuadorian Presidential 2021 Campaign.

Keywords: Google Clouds · Google Sheets · Querying · Data Analysis · Sentiment Analysis.

1 Introduction

Since 2004, when “Web 2.0” was loosely expressed and defined at a Silicon Valley Conference [1], software technologies worked on applications to improve the interaction and collaboration between the users, via the web [2].

During those years, Google not only managed to become the most popular search engine within the web [4], but it also evolved after introducing solutions such as Google News (2002), Gmail (2004), Google Maps (2005), among others; and expanding their business after buying popular web platforms such as Blogger (2003), Youtube (2006), Upstartle (2006), among others. Thus becoming in the “symbol of online innovation” [2].

In 2006, when Google Sheets was recently launched, some publications compared it to Microsoft’s office Excel. While its innovative online collaboration was highlighted, limitations in presentation, features, and formats were also noted [2] [5].

About 15 years after its release, Google Sheets is not only a collaborative sheet. Along the mathematical functions common in other spreadsheets, it also allows users to “program” or write custom functions through the Script Editor, which it is a JavaScript platform; and to “query” similar to SQL environments, which makes it capable of processing petabytes of data [3].

Being Google Sheets part of Google Cloud, thus being backed up by the power of High Performance Computers, this article aims to describe an alternative platform

for analysing data available for any user, while describing its implementation to process and interpret sentiment analysis from tweets retrieved during the Ecuadorian 2021 Campaigns.

2 Methodology: Querying on Google Sheets

2.1 Retrieving Tweets

To retrieve the tweets, we used Martin Hawksey's TAGS¹. An available online script that allows you to easily set up a Twitter account and, through the Twitter API, collect the tweets. The only change that we made to his work was the addition of more fields in the "Archive" sheet that would allow us to obtain more details for a deeper analysis.

M. Hawksey's fields are: *id str, from user, text, created at, time, geo coordinates, user lang, in reply to user id str, in reply to screen name, from user id str, in reply to status id str, source, profile image url, user followers count, user friends count, user location, status url, entities str*.

Our fields were: *id str, text, status url, retweet count, favorite count, created at, user created at, in reply to screen name, user geo enabled, user location, place, in reply to user id str, in reply to status id str, user id str, user name, from user, user profile image url, entities str, lang, user description, user followers count, user friends count, user favourites count, user statuses count, source, extended tweet*.

2.2 Organizing Data

Google Sheets limits the use of its cells per document up to 5'000.000. Therefore, we were unable to work on the original TAGS file alone. Hence we established that each time we reached approximately the 100,000 row (100.000 tweets with details of the aforementioned fields), we would copy the data into a new spreadsheet.

Thus our step to organize the data involved to separate the tweets up to 100.000, in different files (from now on we will refer to them as **1-Archive**), also put them in different folders, consequently being able to work with them independently and then combine the results into a final step. Theoretically using the Divide-and-Conquer technique approach.

2.3 Cleaning Data

Our data was a mix of original tweets, retweets, and duplicate tweets that were separated into different files. As a result, in each one, we started by removing the duplicates through the *unique* command and then calling them to a new file (from now on we will refer to these new files as **2-Filter**).

Even though Google's "Sheets data connector for BigQuery" [3] would allow us to deal with large datasets "at once"; another way around to stick only in Google Sheets, was to use *importrange* command. Since this command works always with

¹ <https://tags.hawksey.info/get-tags/>

calls of 10.000 rows, we used it multiple times to import the 100.000+ rows from each **1-Archive** to fill the **2-Filter**.

Then, we separated the original tweets on a new sheet of the same document. From this new sheet, we used the *query* command.

2.4 Designing a Sentiment Analysis Alternative on Google Sheets

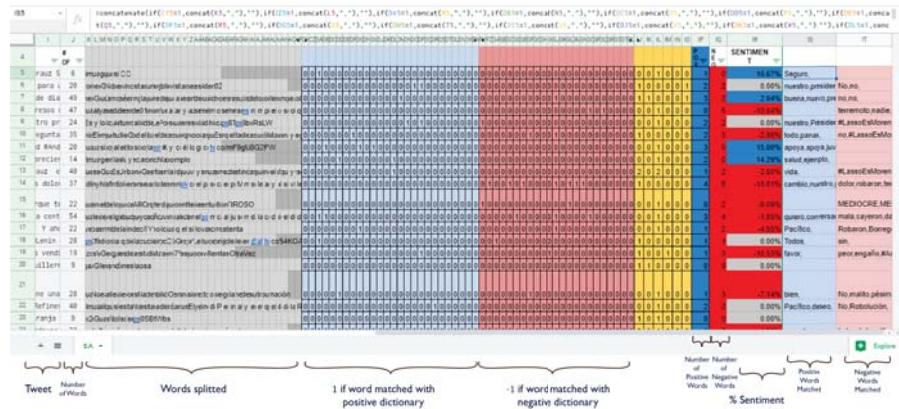


Fig. 1. Google Sheets - Screenshot of the Distribution of cells for analyzing tweets

Our approach for analyzing tweets in spreadsheets was to compare each word from each tweet against a dictionary of positive words and another of negative words. Add a positive value to each positive word found, or a negative value to each matching negative word, then calculate them. Besides, in order to get more accuracy in the rating, we decided to vary the value of the words from the dictionaries.

The final result determined the sentiment of each tweet. The higher the percentage of positive words, assures a positive tweet; and likewise, the lower the percentage of negative words identified a negative sentiment.

Besides, when a tweet had a result of "0", we marked it as "Neutral" and it meant that a same number of positive and negative words were found in the tweet. Or that none of its words matched with any vocabulary included in the dictionaries.

Setting up the dictionaries We created a new file (from now on we will refer to it as **Dictionaries**) with two sheets. One sheet contained all the positive words, and the other the negative words. We only used one column from each sheet and filled them with all the respective words. As these dictionaries were made in Spanish, and we were about to match exact words, our dictionaries sought to include possible conjugations and variations with accent and without accent.

Matching Words In a new spreadsheet (from now on we will refer to it as **3-SA**), we located the original tweets. The procedure described in this section was implemented in cells that were part of the same text row to be analyzed. To make clearer our explanation, figure 1 shows the distribution of the cells within the file.

First, we started by counting the words included in each tweet.

Second, we split the words of the tweets in different cells. In figure 1, these cells are grey colored.

Third, we added new cells for comparing each word of the tweet (up to this step, the words from the tweets were already spread across different cells) with the positive dictionary and each with the negative. Thus, for example in a tweet with 70 words, there will be at least 70 new more cells for comparing each word with the positive dictionary, and 70 new more cells for comparing them to the negative dictionary. Each time that a word matched with the positive or negative words, the number “1” was set in the respective cell, in figure 1 these cells are colored with blue for the positive and red for the negative columns.

2.5 Results

As we mentioned earlier, positivity and negativity were initially calculated as a percentage, increasing the probability of assertiveness when these values were further from 0.

These ratings may allow us to have a general glance of the sentiment shared through Twitter, regarding the Ecuadorian 2021 Presidential Campaigns. Hence, as a matter of example, and in order to attempt to get a more meaningful result. After standardizing the locations shared by the users, we were also able to obtain the sentiment through places, see figure 2 which displays the results from February 1st to the 7th. In this figure it is possible to observe that in Pichincha, for instance, there are more positive tweets towards Lasso (374) than against him (305). While, there are more negative content rated towards Arauz (437), than positive tweets (336).

Fig. 2. Rating tweets according to location

2.6 Limitations in the analysis

These are some of the limitations that we were able to find in our approach.

- To keep the document as small as possible (the fewer cells the better), we limited the analysis up to the 70th word of a tweet. If a tweet had more words than that, we ignored them.
- Ironies expressed with text or emojis were not well rated.
- If there was a tweet that replied an user, but mentioning a candidate, it was marked the sentiment towards the candidate.

3 Conclusion and Future Work

Analyzing data is mostly related to technologies that involved advanced technical support and constant rent in available public clouds. Our work included a different alternative that may respond to unattended parties who are more familiar with Google Suites.

Furthermore false-positives results allowed us to notice the limitations of measuring the sentiment with our approach. Thus for increasing the accuracy of analysing the tweets we have considered to measure the sentiment not only by words, but also in some cases by “phrases”; to determine sentiment according to context, to attempt to discriminate trolls from real accounts.

Acknowledgement

This research has been supported and funded by the Technological Research Department of the company Soluciones Wandarina S. A. of Ecuador.

References

1. Yu, C., and Du, H. (2007). Welcome to the World of Web 2.0. *The CPA Journal*, 77(5), 6.
2. Rienzo, T., and Han, B. (2009). Teaching Tip: Microsoft or Google Web 2.0 Tools for Course Management. *Journal of Information Systems Education*, 20(2), 123.
3. Gundrum, D. (2019, January 15). Connecting BigQuery and Google sheets to help with hefty data analysis. Retrieved March 15, 2021, from <https://cloud.google.com/blog/products/g-suite/connecting-bigquery-and-google-sheets-to-help-with-hefty-data-analysis>
4. Evans, M. P (2007). Analysing Google rankings through search engine optimization data. *Internet research*.
5. Firth, M., and Mesureur, G. (2010). Innovative uses for Google Docs in a university. *Jalt call journal*, 6(1), 3-16

An approach for the analysis of news during COVID-19 in the Chubut province

Pablo Toledo Margalef¹, Emanuel Balcazar³, Leo Ordinez², Claudio Delrieux², and Lucila Allende³

¹ Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH)- CCT-CENPAT
-CONICET, Puerto Madryn, Argentina
ptoledo@cenpat-conicet.gob.ar

² Laboratorio de Investigación en Informática (LINVI), Facultad de Ingeniería,
UNPSJB, Puerto Madryn, Argentina

³ Facultad de Ingeniería, UNPSJB

Abstract. The present work exposes preliminary results on utilizing web scraping and data mining techniques to analyze news articles published during the COVID-19 pandemic in the Chubut province. Analysis of extracted articles was made using Latent Dirichlet Allocation obtaining promising results.

Keywords: LDA · COVID-19 · media · news · NLP · scraping

1 Introduction

This work is framed in a project which aims at constructing knowledge in order to evaluate the current and predict the future socio-economic situation of the Chubut province. In particular, focusing on vulnerable population in the context of pandemic generated by the COVID-19. The target region is limited by the geographical limits of the Chubut province, adjusting the territorial scale to cities, towns and rural communes. The knowledge part of interest for this article will be constructed by extracting, processing, and automatically analyzing news articles published in local press with provincial scope. The goal is to evaluate the evolution of different topics that affects the community. Results will be obtained through web scraping and the application of Natural Language Processing Techniques (NLP).

The main objective of the current work is the construction of knowledge about the current situation of the Chubut province regarding the COVID-19 pandemic through extraction and analysis of news around topics affected by the sanitary context. The use of an unsupervised technique, like Latent Dirichlet Allocation (LDA), leads us to a discovery of such topics instead of a confirmation about preestablished subjects.

2 Materials and Methods

For the analysis of news LDA is used. This technique is part of Natural Language Processing (NLP) [8], and is considered an Unsupervised Learning method [4].

In the case of NLP, if the observations are words collected into documents, the model posits that each document is a mixture of topics which can be attributable to the presence of certain terms and that each word's presence can be attributable to one or more of the document's topics [2].

The media chosen for extraction were only those from the Chubut province. This was to limit the number of results and focus particularly in that region. The selection was also influenced by the popularity of their web portal, the amount of news posted and the territorial representativeness such that the complete provincial territory is included.

The extraction process was divided into steps to facilitate development of different components, each with its own responsibility within the extraction and analysis of articles. The steps are:

1. **Google Search:** as a first step, the search equations for each site are retrieved from the database. An equation indicates the date and URL to be used for the search. Then a personalized Google search engine performs the actual search and obtains the URLs for the links meeting the criteria.
2. **Extraction:** the links obtained in the previous step are received by a component whose task is to extract the articles in HTML format.
3. **Cleaning:** The HTML is processed extracting the relevant sections, such as the title, subtitle, body, date and original link. This way we obtain a clean version of the article.
4. **Normalization:** The articles are normalized in a common format to store them in the database, this eases the creation of a unified dataset across all sites used.
5. **NLP:** Natural Language Processing is applied once the articles are stored in the database. A series of modifications are made to the articles so they are useful in future processing, each term is taken to its root and unnecessary words (*i.e.*, stop words) are removed.

In Fig. 1 the scraping process is sketched. Note that step 3. implicitly establishes a *plugin* architecture since each website has a different HTML structure.

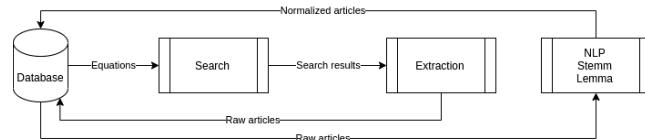


Fig. 1: Web scraping process of news media.

The tools used in the previous steps were developed in NodeJS, using AdorniJS for one of the components. Python as NLP module, and PostgreSQL to store extraction a post processing results. Rabbit MQ was used for communication between each component.

Extracted articles have the following structure:

- **title**: title of the article.
- **snippet**: brief article summary obtained through Google.
- **link**: original article link.
- **displayLink**: base URL of the site where the article was extracted from.
- **body**: body of the article, most important field in the dataset.
- **published**: date the article was published on.
- **expected_date**: expected publication date. It is used to control whether the article corresponds to the date of the search equation used.
- **is_useful**: Indicates whether the article is useful for processing. It is not field in use.
- **analyzed**: Indicates if the article is already processed by NLP.

From March 2020 to March 2021, more than 62,000 items were obtained, from which almost 85,000 unique words were found. Fig. 2 shows the dashboard of the application developed for the extraction of news. The system keeps on extracting data, so this numbers are expected to increase over time. The finishing date of this process is yet to be decided. In the figure, the top 5 of recurrent words and the different news sites extracted along with their amount of articles, are depicted. The most recurring words are: *provincia*, *case*, *work*, *do /make* (in Spanish, *hacer*) and *power* (it can also be, “can”, since the word was *poder*).



Fig. 2: Screenshot of the application developed for the scraping.

To process data and generate the corresponding models, an implementation in Python was performed. The pre-processing of the documents was made using *Spacy*⁴, an NLP library providing functionalities to generate the root of the terms, being it lemmatization or stemming. To generate the models we used *Gensim*, this library allows the construction of topic models through various methods, including LDA[9]. Visualizations were possible through *pyLDAvis*[10]. Lastly the retrieving of the data from storage mediums and the raw data processing was possible using data science tools such as *pandas*[7] and *numpy*[3].

⁴ <https://nightly.spacy.io/>

Applying these methods to news is an insightful technique for knowledge construction. Previous applications can be found on finance articles [5], detecting risk of a pandemic [1], or even analyzing patterns within media coverage of health communications on early stages of the COVID-19 outbreak[6].

3 Results

Using a data set of 1,103 articles, a series of models were constructed. At first lemmatization and stemming was not used, and 15, 30 and 60 topics were extracted. It was observed that the resulting topics were widely spaced from each other or overlapped. There was no uniformity between the relevance of each topic, resulting in topics excessively larger than others. After lemmatization and stemming were applied, keeping the quantities of extracted topics, no improvement was observed. Figures 3a 3b respectively show the previous situations for the case of 30 topics.

One last experiment was conducted using only articles from a single day, but this time only 6 topics were extracted, applying lemmatization and stemming. A clear separation between topics and uniformity among the topic sizes was observed (see Fig. 3c).

A visualization of the described models can be found at the following link⁵.

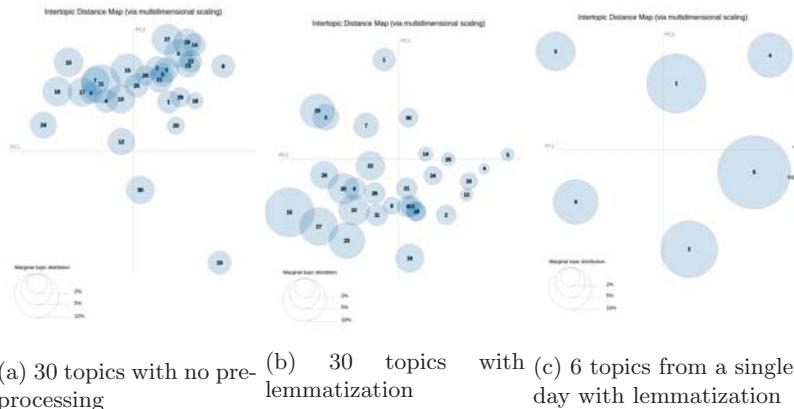


Fig. 3: LDA experiments.

4 Conclusions and Future Work

This ongoing research showed the potential in analyzing news articles for the understanding of a complex phenomenon such as that experienced by the COVID-

⁵ <https://papablo.gitlab.io/resultados-short-paper-analisis-noticias-chubut/>

19 pandemic. The automatic extraction of information and its analysis using techniques such as NLP showed their potential in terms of quantitative studies.

It was possible to discern a reasonable number of topics to be extracted that allowed a similar size and separation between them. However, these results were obtained using articles from a single day. Future works are to be focused on considering time as another analysis dimension in order to study the dynamic evolution of topics and terms. This would allow us to generate a hypothesis and list of terms to be followed over time.

References

1. Akrouch, M.E., Benbrahim, H., Kassou, I.: End-to-end LDA-based automatic weak signal detection in web news. *Knowledge-Based Systems* **212**, 106650 (Jan 2021). <https://doi.org/10.1016/j.knosys.2020.106650>
2. Blei, D., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022 (2003)
3. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. *Nature* **585**, 357–362 (2020). <https://doi.org/10.1038/s41586-020-2649-2>
4. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L.: Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* **78**(11), 15169–15211 (2019)
5. Kakhki, S.S.A., Kavaklıoglu, C., Bener, A.: Topic detection and document similarity on financial news. In: *Advances in Artificial Intelligence*, pp. 322–328. Springer International Publishing (2018). https://doi.org/10.1007/978-3-319-89656-4_34
6. Liu, Q., Zheng, Z., Zheng, J., Chen, Q., Liu, G., Chen, S., Chu, B., Zhu, H., Akinwunmi, B., Huang, J., Zhang, C.J.P., Ming, W.K.: Health communication through news media during the early stage of the COVID-19 outbreak in china: Digital topic modeling approach. *Journal of Medical Internet Research* **22**(4), e19118 (Apr 2020). <https://doi.org/10.2196/19118>
7. McKinney, W.: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) *Proceedings of the 9th Python in Science Conference*. pp. 51 – 56 (2010)
8. Nadkarni, P.M., Ohno-Machado, L., Chapman, W.W.: Natural language processing: an introduction. *Journal of the American Medical Informatics Association* **18**(5), 544–551 (09 2011). <https://doi.org/10.1136/amiajnl-2011-000464>
9. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. pp. 45–50. ELRA, Valletta, Malta (May 2010), <http://is.muni.cz/publication/884893/en>
10. Sievert, C., Shirley, K.E.: Ldavis: A method for visualizing and interpreting topics. pp. 63–70. Baltimor, Maryland, USA (June 2014)

Intelligent Anomaly Detection System for IoT

Diego Angelo Bolatti¹ [0000-0002-8275-4476], Marcelo Karanik¹, Carolina Todt¹
[0000-0001-8429-6141], Reinaldo Scappini¹ [0000-0001-6854-4643], Sergio Gramajo¹ [0000-0001-5091-7931]

¹ Center for Applied Research in Information and Communication Technologies at National University of Technology (UTN), Resistencia Regional Faculty (UTN-FRRe).
French St. 414, Resistencia, Province of Chaco, Argentina.
{diegobolatti, mkaranik, carolinatdt, rscappini,
sergiogramajo}@ca.frre.utn.edu.ar

Abstract. The growing use of the Internet of Things (IoT) in different areas implies a proportional growth in threats and attacks on end devices. To solve this problem, the IoT systems must be equipped with an anomaly detection system (ADS). This work introduces the design of a hybrid ADS based on the Software-Defined Network (SDN) architecture, which combines the rule-based and Machine Learning-based detection technique. Whereas the rule-based approach is used to detect known attacks with the help of rules defined by security experts. And the Machine Learning approach is used to detect unknown attacks with the help of Artificial Intelligence techniques.

Keywords: IoT, Anomaly Detection, Machine Learning, SDN.

1 Introduction

The Internet of Things (IoT) has been expanding in recent years and this is reflected in the thousands of devices connected every day, which obtain and exchange information through the web. This new paradigm is used in different sectors such as healthcare, transportation, agriculture, entertainment, and education.

The great diversity of communication, devices, technologies, and protocols makes managing the security of an IoT ecosystem a great challenge.

Designers rarely bear in mind security when it comes to IoT devices, and many of them lack essential encryption and authentication capabilities, which has led to a whole new category of attacks explicitly targeting end devices.

To address this issue, several security solutions for IoT have been proposed [1- 2]. Most of these solutions focus on the use of cryptography for preventing external attacks, such as message alteration and eavesdropping.

If some of the sensor nodes are compromised and become internal attackers, cryptographic techniques cannot detect these malicious nodes because the adversary can have a valid key to perform activities within the network.

Usually, attackers establish malicious nodes as legitimate nodes within the network to launch internal attacks, such as data alterations, selective forwarding, jamming, denial of service, and clone attacks. These attacks are destructive to IoT network operations. For this reason, the capability to detect intrusions and malicious activities within IoT networks is critical for maintaining the functionality of the IoT system.

This article introduces the design of an intelligent anomaly detection system for IoT and is organized as follows. The related work is presented in Section 2. The proposal system is introduced in Section 3 and, finally, the discussion and future works are presented in Section 4.

2 Related Work

The anomaly detection system can be the key to solving intrusions because alterations to normal behavior indicate the presence of intentional or unintentional induced attacks on the IoT network.

Implementing an Anomaly Detection System (ADS), Software-Defined Networks (SDN) architecture for instance, offers a good alternative because it provides all the benefits of virtualization, such as agility and cost-effective redundancy, and scalability.

The visibility across the network helps identify malicious actions and take appropriate action, such as quarantines.

Centralizing security control in one entity, such as the SDN controller, has the disadvantage of creating a central point of attack, but SDN can be used effectively to manage the security of the IoT environment if it is implemented securely and appropriately [3-4].

Anomaly Detection can use two detection techniques—rule-based or Machine Learning (ML)-based. The rule-based approach detects anomalies with rules defined by security experts [5] and is ideal to identify known attacks. The benefit of using this technique is that the rules are easily understood and highly accurate.

Generally, the Proposed ADS uses ML techniques [6-7] to identify unknown security attacks. To use ML effectively for cybersecurity purposes, a large amount of properly labeled training data is needed.

However, even when an algorithm has received a large amount of data, it does not ensure that it can correctly identify all new attacks. Therefore, human supervision, experience, and verification are constantly required. Without this process, even a single incorrect entry can cause a "snowball effect" and possibly undermine the solution to the point of failure. The same problem arises if the algorithm only uses its own output data as inputs for further learning. Errors are reinforced and multiplied as the same incorrect results re-enter the solution in a cycle, creating more false positives (incorrectly categorizing clean samples as malicious) and false negatives (marking malicious samples as benign).

To reduce the rate of false positives and negatives, both anomaly detection approaches can be combined, thus obtaining a hybrid system [8-9].

3 Proposal

This work proposes a hybrid anomaly detection system for IoT, which uses rule-based and ML-based with real-time data as input.

In the first part, the rule-based ADS captures the network traffic coming from the end devices through an open flow switch in the gateway, and based on some

predefined rules, classifies the incoming network traffic as normal or abnormal (attack).

After classification, the information is stored in a database allowing that data to be used in the future for training the ML-based anomaly detector.

In the second part, the learning model is trained using the labeled training data set. After training the model, we use the model to validate the classification results performed by the rule-based anomaly detector. The final prediction of the system is obtained by comparing the results of both types of detectors. If one of the systems declares a package as a failure; it will be labeled as an attack.

The anomaly detection system will be installed in the device layer of the reference architecture proposed by the ITU (International Telecommunication Union) [10].

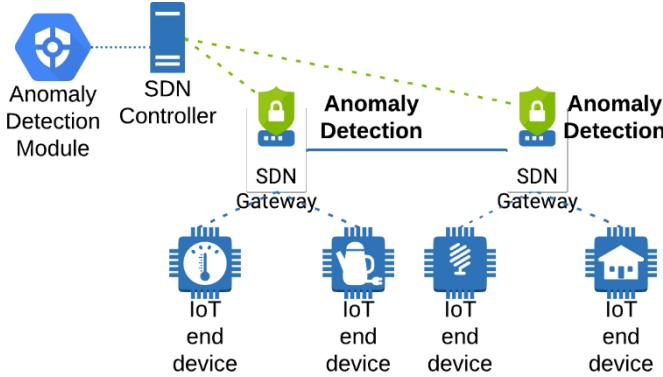


Fig. 1. Architecture of the Intelligent Anomaly Detection System for IoT.

As shown in Fig. 1, the architecture consists of the following four components:

- **IoT end device** with the mandatory capabilities of communication and optional capabilities of sensing, actuation and data capture.
- **Gateway SDN** to connect the different devices to the network through Low Power Wide Area Networks (LPWAN) such as LoRa, SigFox, NB-IoT, LTE-M, etc.
- **OpenFlow Switches** to monitor the traffic coming from the end devices.
- **SDN Controller** manages and configures the distributed network resources and provides an abstracted view of the network resources to the SDN applications via another standardized interface (i.e., application-control interface) and the relevant information and data models.
- **Anomaly Detection Module** to check the packet for anomalies and add intelligence to the SDN controller to readjust the network and maintaining the policies defined by administrators when detecting the following attacks: Denial of Service (DoS), Data type probe, Battery drain attack, Packet tampering, Jamming, Man in the Middle, and Packet delay.

4 Discussion and Future Work

To solve the security problems of the IoT environment, this article describes an Anomaly Detection System based on two different detection approaches, rule-based

and ML-based, in different instances. The result of both techniques is compared, and if one of the systems detects an anomaly; it will be labeled as an attack. This system combines both human and machine intelligence. And the advantage of using a hybrid system is that reducing the number of false-positive and false-negative rates.

An interesting aspect to analyze is the interactions of the model (SDN Controller - SDN Gateway and SDN Gateway - SDN Gateway). In the first case the SDN Controller defines data flow control rules based on application and SDN gateway traffic. In the second case, SDN Gateways check rules and configuration updates to keep their states synchronized.

In the field of security, ML can play an important role in helping security teams make accurate decisions about security threats and incidents. But ML cannot do the job for the human engineers, developers. There is no magic solution, human experience is always necessary.

The system design is currently being finalized and the next step in this work will be to evaluate and identify the most appropriate Machine Learning technique for the system. As a final step, we intend to test the system in a real-world IoT environment to evaluate its efficiency and viability.

References

1. S. Sridhar and S. Smys, "Intelligent security framework for iot devices cryptography based end-to-end security architecture," 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, pp. 1-5 (2017).
2. Mathur, A., Newe, T., Elgenaidi, W., Rao, M., Dooly, G., & Toal, D. A secure end-to-end IoT solution. Sensors and Actuators A: Physical vol. 263, pp. 291-299 (2017).
3. Nguyen, Tam N. The challenges in SDN/ML based network security: A survey. arXiv preprint arXiv:1804.03539 (2018).
4. Tsogbaatar, Enkhtur, et al. SDN-Enabled IoT Anomaly Detection Using Ensemble Learning. IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer, Cham (2020).
5. Xie, M., Han, S., Tian, B., & Parvin, S. Anomaly detection in wireless sensor networks: A survey. Journal of Network and computer Applications, 34(4), pp.1302-1325 (2011).
6. Nguyen, T. D., Marchal, S., Miettinen, M., Fereidooni, H., Asokan, N., & Sadeghi, A. R. (2019, July). DIoT: A federated self-learning anomaly detection system for IoT. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), pp. 756-767. IEEE (2019).
7. 2-Alrashdi, I., Alqazzaz, A., Aloufi, E., Alharthi, R., Zohdy, M., & Ming, H. Ad-iot: Anomaly detection of iot cyberattacks in smart city using machine learning. In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). pp. 0305-0310. IEEE (2019).
8. Thanigaivelan, Nanda Kumar, et al. "Hybrid internal anomaly detection system for IoT: Reactive nodes with cross-layer operation." Security and Communication Networks 2018 (2018).
9. Bhatt, P., & Morais, A. HADS: hybrid anomaly detection system for iot environments. In 2018 International Conference on Internet of Things, Embedded Systems and Communications (IINTEC), pp. 191-196. IEEE (2018).
10. ITU-T Y.4000/Y.2060 (06/2012) Overview of the Internet of Things. <http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=11559&lang=en>.

Distributed Cybersecurity Strategy, applying Intelligence Operation concept through data collection and analysis

Ignacio Martín Gallardo Urbini^{1[0000-0002-1983-895X]},

Patricia Bazán^{2[0000-0001-6720-345X]},

Paula Venosa³ and Nicolás Del Río^{4[0000-0002-0889-0752]}

¹ National University of La Plata, Buenos Aires, La Plata, Argentina
ignacio.gallardou@info.unlp.edu.ar,

² LINTI. National University of La Plata, Buenos Aires, La Plata, Argentina
pbaz@info.unlp.edu.ar,

³ LINTI. National University of La Plata, Buenos Aires, La Plata, Argentina
pvenosa@info.unlp.edu.ar,

⁴ National University of La Plata, Buenos Aires, La Plata, Argentina
ndelrio@info.unlp.edu.ar ,

Abstract. This document presents a line of doctoral research that proposes a cybersecurity strategy that has not been formally standardized up to now, based on knowledge of defense intelligence operations, and applying a dynamic approach, in a context of threat risk, anticipating its effectiveness . In this way, change the current approach, leaving aside the old concept of "walled" defense, for a more innovative one, where information collectors or "spies" infiltrate "unknown terrain" or external networks to extract data and information, learn from context, analyze and detect patterns, be willing to share the knowledge, and then be able to make defensive, deterrent, or offensive decisions in real time.

Keywords: Cybersecurity, Big Data, Data Intelligence, Multivendor.

1 Introduction

Many people in the world have studied computer science, specializing in information security, cybersecurity and cyber defense; however, many of them are currently responsible for related sectors to these areas of knowledge in different parts of the world, including Government Agencies, the Army or big private organizations directly linked to society and the State. However, just few of this large population have actually devoted their time to practicing and studying intelligence strategies and tactics; perhaps it is due to this reason the rarity of bringing the security of information systems to the field of intelligence and making use of these ancient techniques. This line of research and development has the general objective of addressing a reflection on static defensive schemes and then proposing new techniques that are born in the intelligence doctrine and address the inequality between the millions of internet threats and specific objectives specially defined to combat them, applying new methods of operations. Any leader of an intelligence strategy, known to the unequal defense forces, must not be static but it should be dynamic; observing and analyzing the enemy by means of data collectors distributed in the observation field, gathering techniques, information analysis, exchanging other resources for "time", sharing the learned knowledge with other allies (interoperating with external vendors [12] for cooperation and information enrichment), and only when there is a high degree of certainty, then respond. In the specific case of an intelligence operation[1], there will be a threshold below which no further progress can be made, this line is called the "diffusion stage", and it reaches it by applying a

strategy to guarantee security called "intelligence operation" and it is what gives rise to this thesis proposal.

2 Objectives

Among the specific objectives of this research are the following:

- Plan and organize the defensive security strategy applying intelligence operations tactics and strategies[10] in order to transform the current "static" defensive attitude[2] into an innovative and "dynamic" one.
- Investigate, develop and implement computer components distributed in the network for the gathering of information, in order to efficiently maintain the picture of the threat situation both in the observation stage for learning and knowledge of the hostile context.
- Develop and implement an intelligent system applying data mining concepts and machine learning techniques that are nourished by the information obtained by the computer components named in the previous objective.
- Study and evaluate different machine learning models to check and select the most optimal and efficient one.
- Provide an Application Programming Interface and Communication Protocol for sharing the intelligence knowledge with external solutions.
- Converge in the implementation of an early and real-time detection system of anomalous patterns in the network, in order to make decisions in advance of the materialization of a possible threat.

The original contribution of this line of research is practically based on the proposal of a cybersecurity strategy not yet formally or standardized, supported by knowledge of intelligence operations for defense, and applied to a dynamic approach, in the face of the existence of a risk of threat, anticipating that it becomes effective. In this way, change the current approach, leaving aside the old concept of "walled" defense for a more innovative one, where information collectors or "spies" infiltrate "unknown terrain" or external network to extract data and information , learn from the context, analyze and detect patterns, and then early and in real time, be able to make defensive, dissuasive or offensive decisions.

3 Motivation and State of the Art

Traditional security solutions[2] focus primarily on protecting the perimeter of interest, thus focusing primarily on external threats. Yet these are constantly evolving, requiring those who wish to remain resilient in their operations to stay informed and one step ahead of attackers. For the definition of a defensive cybersecurity strategy, the same variables that are taken into account in the intelligence doctrine applied to national security can be used, where elements of aggression similar to those analyzed in a cyber attack are presented: sabotage, harassment the victim in his own land, use of irregular detachments with rapid and surprise attacks, secrecy, great mobility, temporary blockages of the basic channels of communication and supplies, and kidnapping / theft of assets. Faced with this new context of advanced cyber threats, in which criminal and hacktivist groups with political and economic interests are involved, the motivation arises to start this line of investigation in order to carry out the development of an intelligence or cyber intelligence strategy as an element key to reinforcing the information security strategy.

Currently projects that address similar objectives:

- Splunk Behavioral Analytics is a software product designed to face internal risks in organizations. It aims to cover the problem from the analysis of user behavior [3].
- FireEye Threat Analytics, a software solution that applies threat intelligence, firewall rules, and advanced security data analytics to optimize detection and response to alerts that matter [4].
- Munin, is an initiative that wants to build a low interaction honeynet, where vulnerable services that are considered critical in an organization are simulated, and that are typically published on the Internet. This project aims to collect information on botnet attacks and then study them [5].
- C1fApp is a threat feed application, which provides a single feed, both Open Source and private. Provides statistics dashboards, API open for search, useful and running for a few years. Searches are historical data [6].
- Cymon is a multi-source indicator aggregator with threats history that provides an API for searching a database along with a nice web interface [7].
- Palo Alto Artificial Intelligence and Machine Learning in the Security Operation Center - Cortex Module, provides components spread out across the enterprise and cloud, providing data to AI services that within minutes can detect new malware and identify malicious domains. The components also provide the point at which policy enforcement, based on the results of the AI services, prevent successful cyber-attacks [12].

The framework proposed in this thesis includes tactics and strategies, and procedures applied in intelligence operations included in the national intelligence doctrine itself [7], in Spanish and for public use, with a open communication protocol for sharing the learned knowledge with external vendors to be used or consuming data from them, integrating data collectors, adaptable anomaly detection modules and a frame of reference to get ahead of the enemy and thus be able to take a dissuasive, offensive or defensive action. Of these similar projects described above, none provide a comprehensive joint framework like the one proposed in this thesis.

4 Experimental Proposed Work

To validate the proposal, put the strategy into practice and test the operation, the development and implementation of a systems architecture that will be made up of different software components will be addressed. On the one hand, the network of sensors (baits) "spies" in charge of collecting information on the activity of the network will be developed. These will have the property of simulating conventional communications with each other and at the same time being able to report in real time to the expert knowledge system. On the other hand, the development of a prototype tool should be addressed to contribute to the detection of behaviors compatible with cyberattacks or cyber threats. Information processing and analysis comes into play here, invoking the different machine learning algorithms, thus converging on a system of expert knowledge and its implementation in real time. The following requirements should also be addressed:

- Observe the behavior of the tool under different cyber attack situations. To achieve this, the development of an alert system must be carried out.
- Have operational and upgradeable ease of the tool: With learning and training mechanisms as its use increases.
- Observe metric graphs, comparing the values obtained from the network flows where the monitoring system is installed.
- Identify behavior patterns: according to the observed graphs, associated with different stages of cyber attacks and classify the threats.

- Build an Application Programming Interface with a Communication Protocol to be able to share the intelligence knowledge with external similar solutions.

5 Research Methodology

As previously stated, any security solution will be tied to the strategic decision to use. However, opting for tactics and strategies applied to intelligence doctrine[9] provides dynamism and great added value at the time of defense. This research work adopts a qualitative methodology[10] for the development of the security architecture from scratch together with its component components, in order to achieve this strategy. Current security measures fall short of polymorphic threats, therefore a new line of thinking must be considered. It is reiterated that what is really critical is the total ignorance of the adversary regarding its location, magnitude, resources, behavior and capabilities, from which the first imbalance of forces arises. On the other hand, when studying defense and security activities throughout history, there are no records of any invulnerable fortress. Given these two aspects, it is proposed to analyze cybersecurity from the point of view of dynamism and intelligence, that is, leaving aside the current conception of static and centralized defense materialized in Instruction Detection Systems, Intrusion Protection Systems, or Firewalls. The intelligence doctrine[9] with its millenary experience in collecting, analyzing information and making decisions, raises a particular scenario of operations, where the reason for this research called "Cybersecurity Strategy applying the concept of intelligence operation" takes center stage. This procedure is precisely designed for contexts in which the threat is greater than the victim, there is little information about the victim, and by virtue of this imbalance is why it is planned to "Exchange resources by *time*, to be able to know the threats, anticipate to the facts and have a clear and defined overview of the context".

References

1. Andrew C.: The Secret World, a history of intelligence. (2019).
2. Endorf C., Schultz G., Mellander J.: Intrusion Detection and Prevention. 1st Edition. ISBN-10 0072229534. (2003)
3. Splunk Integrated Behavior Analytics Homepage, <https://www.splunk.com/>, last accessed 2021/03/20.
4. FireEye Overview page, <https://www.fireeye.com>, last accessed 2021/03/20.
5. Munin Homepage, <http://munin-monitoring.org/>, last accessed 2021/03/20.
6. C1fApp Homepage, <https://blog.thehive-project.org/tag/c1fapp/>, last accessed 2021/03/20.
7. Cymon Api Homepage, <https://cymon.docs.apiary.io/#>, last accessed 2021/03/20.
8. National Argentine Intelligence Law Homepage, <http://servicios.infoleg.gob.ar/infolegInternet/anexos/70000-74999/70496/norma.htm>, last accessed 2021/03/20.
9. Handel M.: Intelligence and Military Operations. (1990).
10. Sampieri R, Fenández, Collado C. and Baptista L.: Metodología de la Investigación, 5ft Edition, McGraw-Hill Interamericana. (2010).
11. Washington P.: Producción de Inteligencia Estratégica. Buenos Aires. Struhart Cia. (1983).
12. Palo Alto Artificial Intelligence Module Overview Page, https://www.paloaltonetworks.com/apps/pan/public/downloadResource?pagePath=/content/pan/en_US/resources/techbriefs/artificial-intelligence-and-machine-learning-in-the-security-operations-center, last accessed 2021/04/05.

Evaluation of a heuristic search algorithm based on sampling and clustering

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. Systems have evolved in such a way that today's parallel systems are capable of offering high capacity and better performance. The design of approaches seeking for the best set of parameters in the context of a high-performance execution is fundamental. Although complex, heuristic methods are strategies that deal with high-dimensional optimization problems. We are proposing to enhance the evaluation method of a baseline heuristic that uses sampling and clustering techniques to optimize a complex, large and dynamic system. To carry out our proposal we selected the benchmark test functions and perform a density-based analysis along with k-means to cluster into feasible regions, discarding the non-relevant areas. With this, we aim to avoid getting trapped in local minima. Ultimately, the recursive execution of our methodology will guarantee to obtain the best value, thus, getting closer to method validation without forgetting the future lines, e.g. its distributed parallel implementation. Preliminary results turned out to be satisfactory, having obtained a solution quality above 99%.

Keywords: Optimization, Heuristic methods, Clustering, Benchmark.

1 Introduction

As optimization problems become more complicated and extensive, parameterization becomes complex, resulting in a laborious, complicated task that requires a significant amount of time and resources, besides the fact that the number of possible solutions can become prohibitive in an exhaustive search. It is the reason why optimization algorithms play an important role in this transformation that usually attempt to characterize the type of search strategy through an improvement on simple local search algorithms [2]. In cases where the search space is large, metaheuristic ideas [1], which are sometimes classified global search algorithms, can often find good solutions with less computational effort. Some other approaches to achieve the optimization objectives are based on the extraction of data from probability distributions aiming for a reduction of the search space. Probabilistic distribution methods, such as Montecarlo offer flexible forms of approximation, with some advantages regarding cost. There are other approaches

that use similarity or metaheuristic algorithms to solve high-dimensional optimization problems which are validated using large-scale functions [4]. However, they are prone to fall into local optimum values. In order to solve global optimization problems, making use of global exploration there are, e.g. the naturally inspired approaches such as Genetic algorithms, Particle swarm, Grey Wolf or Ant Colony optimization algorithms [6].

Regarding the clustering methods, these are the techniques that group a series of objects, it is a pattern recognition technique which has a broad range of applications. Cluster analysis algorithms are a key element of exploratory data analysis used in e.g. data mining. Between the most widely used clustering algorithms is k-means. Here, a cluster is defined as a set of data characterized by a small distance to the cluster centers. Among the combinatorial optimization methods, for example [3], the efficiency gains regarding the application of sampling and grouping techniques are explored to solve a problem of a complex nature, because it is a dynamic and strongly human-dependent system.

In this paper, we propose an evaluation methodology of a heuristic method, based on sampling and clustering techniques comprising Montecarlo sampling, useful to obtain samples in multi-dimensional spaces, a density-based spatial analysis, and the k-means algorithm, crucial to determine and classify the data into feasible regions. For our evaluation proposal, we have selected the benchmark test functions [5] which allow testing algorithms and are useful when measuring relevant features, and can also be especially useful for understanding the algorithms applied to large-scale and continuous optimization problems. First we generate an initial sample of the benchmarks through the Montecarlo methods. Then, we apply an efficient clustering to locate feasible regions where the optimal solution might exist and can be found. The elements of the domain, which are known as candidate solutions, define such feasible regions. In our approach, we perform a density analysis to find patterns that will handle efficient grouping based on euclidian distances. Also, the recursive application of our proposal guarantees an almost optimal solution by reducing the search area, and enhancing the selection and grouping of feasible regions.

This proposal organizes as follows: the next Section presents our approach with an overview of the methodology, explaining definitions such as feasible regions, a justification of our proposal, the results obtained and the last Section discusses the open lines.

2 Proposal methodology and parameterization analysis

The design of optimization algorithms requires to make several decisions ranging from implementation details to the setting of parameter values for testing intermediate designs. Proper parameter setting can be crucial because a bad parameter setting can make an algorithm perform poorly. For our evaluation proposal, we are using the optimization benchmark functions along with an efficient grouping that performs spatial clustering by density in order to find patterns and/or variations in the landscape, in addition to k-means.

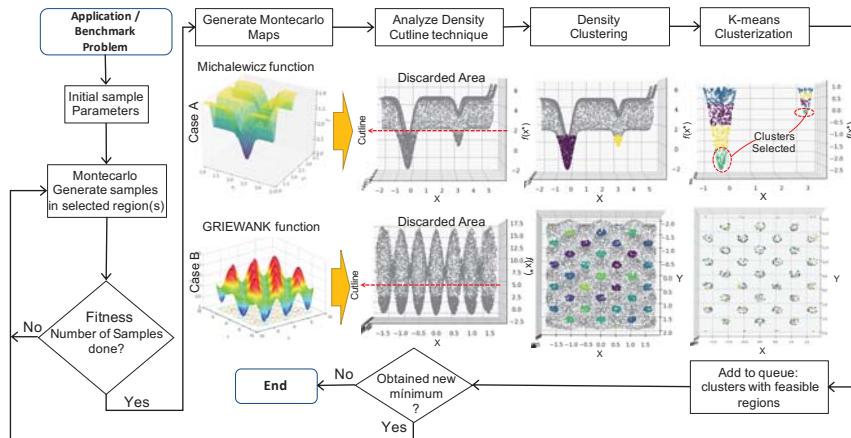


Fig. 1. Flowchart of our evaluation methodology.

As seen in Fig. 1, we generate an initial sample using the Montecarlo methods making successive iterations of the algorithm until it converges. This parameter will be linked to the precision within the range of the search space. If the sample is satisfactory, we generate the Montecarlo map, a landscape visualization that will help to analyze variations in the parameters, such as roughness. Since we are dealing with high-dimensional data with large variations in density because of the fixed global parameters (such as distance radius), we propose to make a cutline along the $f(x^*)$ axis, which represents the value associated to each solution that is evaluated by computing the value of the objective function. It enables an efficient grouping by detecting arbitrary shapes, and automatically discovering the number of clusters through a density threshold allowing to find one cluster surrounded (although not connected) by another different cluster. It needs to have a notion of noise and be robust in detecting outliers. In the same Fig. 1 we can see that the data groups based on connected density objects, form different shapes and variations are detected. In this way, we will obtain the feasible regions, for which we will adapt the classical k-means algorithm locating the centroids along $f(x^*)$. This type of grouping will be very useful to manage the dense areas, locating the clusters in which the best values are.

When the algorithm is able to find more than one feasible region, a queue forms to analyze clusters from each region, or the ones containing the best values, until the end of the queue resulting in selecting a single cluster. In this way, we ensure that we will not be trapped in a local minimum. If the single selected cluster contains the optimal (or near-optimal) value the simulation ends, otherwise, we return to the previous step of obtaining a new sample, analyzing density and grouping. It will execute recursively until finding the optimal value or until it is not possible to find a lower value than that obtained in the last iteration. From the preliminary results we obtained when testing the Michalewicz function, the problem size is in the range of $x^* = (1.0000, 3.5000)$ and the sample size was of 69,700 which is less than 0.5%. The total search space

therefore reaches $6.2500E+08$, and its optimal value $f(x^*) = -1.8013$ located in $x^* = (2.2000, 1.57000)$. Through our proposal, the best result we obtained was $f(x^{**}) = -1.8012$ located at $x^{**} = (2.2024, 1.5709)$. The quality of such solution was of 99.9944%, which is very promising, so we believe that further testing is necessary, as well as parallel systems exploration.

3 Discussion and Open Lines

The effectiveness of heuristic methods in dealing with challenging optimization problems is a widely studied field. Understanding the limitations of existing approaches and identifying areas for improvement contributes to evaluate a system, validate the method and allow its comparison to real-world problems. For our proposal, we selected the benchmark test functions to enhance a methodology that evaluates a heuristic based on sampling and clustering.

We are proposing a calibration of the parameters involved in the density-based analysis, as well as adapting the k-means clustering to select the feasible regions, creating a model that is based on the parameters of the best solution concerning the optimal value. The preliminary results that we obtained, gave a solution quality of 99.9944%, which encourages to expand the method.

Regarding the limitations about metaheuristic methods, one issue we may find has to do with the high-dimensionality of real problems, making it difficult to characterize. Still, it is a very useful tool when it is possible to achieve high precision in the evaluation phase. Nevertheless, there are some open lines that need to be explored, such as the increase in the dimensionality, which will increase the ranges and consequently the search space. To conclude, we believe that the extrapolation of combinatorial optimization techniques along with heuristics in distributed parallel systems is a step forward in the process that allows decision-making in real-time.

References

1. Bianchi, L., Dorigo, M., Gambardella, L.M., Gutjahr, W.: A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing* **8** (06 2009) 239–287
2. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. *Siam Review* **60**(2) (2018) 223–311
3. Cabrera, E., Taboada, M., Iglesias, M.L., Epelde, F., Luque, E.: Optimization of healthcare emergency departments by agent-based simulation. *Procedia Computer Science* **4** (2011) 1880 – 1889 Proceedings of the International Conference on Computational Science, ICCS 2011.
4. Eftimov, T., Korošec, P.: A novel statistical approach for comparing meta-heuristic stochastic optimization algorithms according to the distribution of solutions in the search space. *Information Sciences* **489** (2019) 255 – 273
5. Hussain, K., Salleh, M., Cheng, S., Naseem, R.: Common benchmark functions for metaheuristic evaluation: A review. *International Journal on Informatics Visualization* **1** (11 2017) 218–223
6. Vikhar, P.A.: Evolutionary algorithms: A critical review and its future prospects. In: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC). (2016) 261–265

Modelling and Simulation of the COPD Patient and Clinical Staff in the Emergency Department (ED)

Mohsen Hallaj Asghar¹, Alex Vicente-Villalba², Alvaro Wong¹, Dolores Rexachs¹, Emilio Luque¹

¹Computer Architecture and Operating Systems Department
Universitat Autonoma de Barcelona (UAB), Campus UAB
08193 Bellaterra, Barcelona, Spain

²Escuelas Universitarias Gimbernat. Nursing School
Universitat Autonoma de Barcelona (UAB)
Barcelona, Spain

mohsenhallaj62@gmail.com, alejandro.vicente@eug.es, alvaro.wong@uab.es,
dolores.rexachs@uab.es, emilio.luque@uab.es

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is a critical and major social health problem. Comorbidities and coexisting conditions and symptoms are associated with and affected the whole body of COPD patients. Regarding the Exacerbation COPD patient, the Emergency Department (ED) and Emergency Medical Service (EMS) responsibility is managing, decision making, treating the initial response to the COPD patient. During the process, the head nurse and emergency medicine specialist should make various decisions for COPD patients. The first aims of this research is to create a new conceptual model to investigate the model of COPD patient, exacerbation COPD in EMS, Multiple COPD pathologies in ED, nurse action in the emergency box, nurse decision making, evaluation of the patient's condition, and reaction to the emergency box. The second purpose of this research is to create a computational model which will concentrate on the simulation model to use the probabilistic finite-state machines for training the nurses for professional decision with treatment and decision without treatment to evolves nurse with intervention to prevent exacerbation of COPD patients.

Keywords: Simulation, COPD, Pathologies, Emergency Department (ED), Emergency Medical Service (EMS)

Introduction:

Nowadays, in different societies, the quality of Emergency Department (ED) and Emergency Medical Services (EMS) is an essential factor for people and governments. The EMS can be a matter of life and death in different incidences such as accidents and epidemics. As we have seen in recent months with the outbreak of COVID-19, the living conditions have become much more difficult for people particularly for patients with Chronic Obstructive Pulmonary Disease (COPD) and medical staff. Emergencies are usually the first entry point for acute COPD patients that emergency personnel have to assess the patients' condition with COPD and treat, should be able to make the right and timely decisions in the shortest possible time [1]. Therefore, a Decision Support Toolkit (DST) is needed to meet the needs of patients [2]. Decision-making in the ED as a vital and substantial process depends on the medical knowledge, training of nurses, and emergency medicine specialist duties of the personnel[3]. The nurses and physicians of the ED should be able to make correct and timely decisions in emergencies applying the theoretical and practical training programs and knowledge. The most important mission of the medical schools is to train specialized and skilled physicians to provide health care services, having

sufficient knowledge to diagnose and treat diseases and also the ability to perform scientific and clinical skills. So, accurate and appropriate planning in the field of clinical education is essential in creating the capabilities of these people.

As healthcare personnel must make many decisions and also apply the results quickly, giving rise to possible errors due to lack of training in unexpected situations. If we refer to mortality data, in the 1999 report of the Institute of Medicine of the United States entitled "To Err is Human: Building a safe health system" estimated at 100,000 deaths per year due to medical errors [1]. Hence, the need to try to avoid these errors by improving the training of professionals will already arise. All increases in errors were attributed to several factors such as the lack of investment in technology and the increasing complexity of therapeutic procedures. Following this report, health educators began to add simulation components to their pedagogical activities.

Clinical simulation is a participant-centered learning technique or method offering better curves than classical learning. Thus, the main limitation for its generalized application is the highest costs derived from their training in teaching methodology, infrastructure, and the excess time spent by them and by participants themselves in each clinical activity

On the other hand, computational simulation is a genre that helps student self-evaluation, feedback in real-time, carry out simulations at any time and place without teacher on site thanks to the possibility of sending messages throughout. In short, the learning process facilitates online training for the both student and the professionals. For this reason, we think that the idea of designing a training simulator for students/professionals can further enhance the learning curve and, also, taking into account today that we live in a period of a pandemic where capacity limitations, mobility, etc, are marking academic training towards a more digitized environment given that clinical simulation is affected by the impossibility of carrying it out.

Research Objectives and Methodology:

The objective of the research proposal is to implement a training simulator that reflects, the evolutionary conceptual model behavior of COPD (First Modeling of COPD evolution patient), which is already working with great pedagogical interest, would be to know all the variables for the state of the patients and Exacerbation of COPD Patient. The Second Modeling Reasoning is to computational model for evolution behavior of COPD in the face of interventions (decision-making) by the student or professional the aim is for training/improving the nurse/student knowledge in a critical situation such as emergency box, real patient analysis feedback form simulator, improve the medical knowledge of junior student, nurse, doctor without much experience in EMS at ED. For the development of the simulator, the Iterative Spiral Development Model (IDMS) will be followed [4]. This system iterates permanently on the traditional software development cycle[5]. The objective of this process is to gradually implement the models in each cycle to define a more complex model. There are three stages of research that described below:

Stage1: Already in this research, we defined several variables which are most relevant to our conceptual model such as heart rate, blood pressure, skin color (Cyanosis), etc, which would make up the state situation of the COPD patient. Each variable is classified according to Fig 1.

Stage 2: Normally the COPD patient doesn't coverage solely one disease, for this purpose including other pathologies is mandatory for own research. In addition to the objective of design simulation is to create a methodology that helps us implement

other conceptual models of other pathologies following the same philosophy that we have been developing in the COPD model.

CYANOSIS (SKIN COLOR)	lack of Oxygen (Hypoxia) and in COPD Exacerbation the skin color in hand and lips are (purple or bluish),	Visible Variable
ACCESSORY MUSCLE	It (Intercostal retraction) includes the sternocleidomastoid, Scalene, Trapezius, which place in the diaphragm. The COPD patients need oxygen for 20-30 minutes and rubbing the muscle	Visible Variable
HEART RATE	by reference the Global Initiative Chronic Obstructive Lung Disease (GOLD) the <65 bpm and >85 bpm, 5.5 years without COPD, 9.8 years in mild (stage I, GOLD), 6.7 years in moderate (stage II, GOLD), 5.9 years in severe (stage III , GOLD)	Non- Visible Variable
OXYGEN SATURATION	OS should >90% in COPD exacerbation and the COPD patients needs 24% Or 28% Oxygen	Non- Visible Variable
PULMONARY ASCULATION	Respiratory sound provide the vital information regarding the COPD patients, COPD exacerbation has several Asculation such as (wheezing, Cracking, Stridor and Rhonchi)	Complementary Test
X-RAY	It shows Hyperinflated Lungs (the lungs appear larger than normal). X-Ray may reveal bullae (Bullae is a pocket of air that forms near the surface of the lungs)	Complementary Test
Sputum (Mucoid)	The normally is clear and white but the COPD exacerbation may darker with either a yellow or green tinge. The COPD (Needs medicine Mucolytics, such as Hypertonic saline (Nebusal), dornase alfa (Pulmozyme))	Visible Variable
Temperature	The body's temperature in COPD patients is normally high by the infection. the body temperature of COPD is >=38	Non- Visible Variable
ECG	The COPD patients in different case have, Rightward QRS Axis (90 degrees) P wave = P>2.6mm R wave = Right precordial leads (SV1, SV2, SV3) low Voltage - left side (I, aVL, V5-6)	Complementary Test
ARTERIAL BLOOD GAS	-The normal hydrogen ions (H+) in the blood are between 7.38 and 7.42 and the acidic blood (PH<7.38) -- Partial Pressure of oxygen (PAO2) should be below 75 to 10 mmHg and Acidic blood (PH<7.38) --Partial Pressure of Carbon Dioxide (PACO2) should be above (38 to 42 mmHg) and acidic (PH less than 7.38), normally in COPD patients the blood is more Acidic and PH level is low and PACO2 level is above normal	Complementary Test

Fig 1. Patient's relevant variables

Stage 3: Emergency Decision Making (EDM) is one of the critical ways of deal with any emergency in the environment and has a prominent role in loss properties, then focus on EDM is widely used in emergencies[6].

Fig 2, shows the cycles of COPD patient evaluation and decision making of the doctor/nurse to purpose of state variable of the COPD patient.

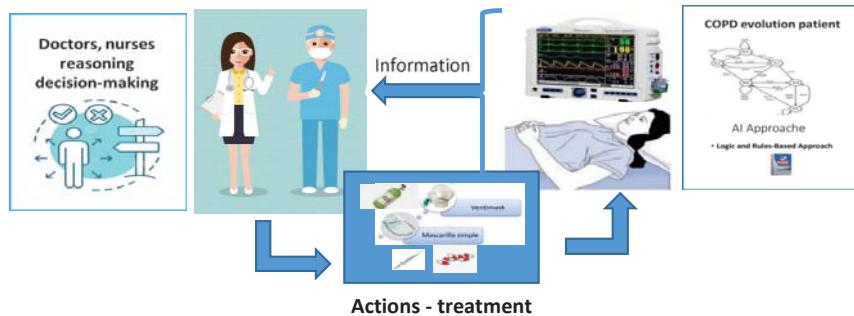


Fig 2. Patient's evaluation for decision making

Normally people make decision-making based on the potential value of losses and gains, for this reason, we propose three R's in EMS.

1. Immediately Recognition and patient Examination

The nurse activity should be immediately measured and examined these items of patients such as HR: (<59, 60-99, >100), BR: (<11, 12-19, >20)
OS: (<80, 81-89, 90-95, >95), TE: (36-37.4, 37.5-37.9, >38)

2. Immediately Right Decision

The critical idea to realize, whenever having to decide on EMS, is about the patient's status [7]. There are several aspects can help the nurse and emergency medical specialist in EMS to have immediately decision such as:

1. Is your patient going to die?
2. Your patient's stability condition and how about now
3. Is your immediate decision safe?

3. Immediately Reaction:

The nurse and EMS specialist should act immediately to save the patient's life. This ability can improve by training to enhance the capability and encounter with the real situation and simulation

Conclusions and Future Work:

This research is based on the conceptual model (Qualitative) and the computational model (Quantitative) that explores the conditions for the implementation of simulating based on COPD intervention to help/improve the quality of the medical services to aim the enhance the student/nurse knowledge. This research has high potential to gather/connect all pathologies relevant to COPD to create a highly professional conceptual and computational reference

Acknowledgments:

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

References:

- [1] A. V. Villalba, M. Antonin, D. Rexachs, E. Luque. 2019. "A Reactive "In Silico" Simulation For Theoretical learning Clinical Skills And Decision-Making". The eleventh international conference on advance in system simulation. pp.3-7. SIMUL 2019.
- [2] Elizabeth G. Bond, Lusine Abrahamyan, Mohammad K. A. Khan, Etc. 2020. " Understanding Resource Utilization And Mortality In COPD To Support Policy making: A Micro simulation Study" PLOS ONE.
- [3] A. franklin, Y.liu, Z.li, Etc. 2011. "Opportunistic decision making and complexity in emergency care". Elsevier.
- [4] D. Keyek-Franssen, Wayne. B and S. Macklin. 2006. "Learning By Doing: A Comprehensive Guide To Simulations, Computer Games, And Pedagogy In E-learning And Other Educational Experiences". Educause.
- [5] U. Yakutcan, E. Demir, John R. Hurst & Paul C. Taylor. 2020. "Patient Pathway Modeling Using Discrete Event Simulation To Improve The Management of COPD", Journal of the Operational Research Society (JORS).
- [6] Z.x. zhang, L.wang, Y.m.wang. 2018. " An Emergency Decision-Making Method Based on Prospect Theory for a Different Emergency Situation ". Springer.
- [7] J. Riancho, J. Maestre, I. del Moral and J.A. Riancho. 2012 "Highly Realistic Clinical Simulation: An Undergraduate Experience", Vol. 15, pp. 109-115. Educ Med .

Big Data

Big Data Technology for monitoring ICT service data

Marcelo Dante Caiafa¹ , Ariel Aurelio¹ , Adrian Marcelo Bustos¹ 

¹ Universidad Nacional de La Matanza, Buenos Aires, Florencio Varela 1903, 1754 Buenos Aires, Argentina
`{macaiafa,aaurelio,abusto}@unlam.edu.ar`

Abstract. Data analysis has become an important source of knowledge for organizations. An adequate treatment allows to obtain valuable information. Its massive processing is possible from Big Data technologies.

The work is based on the use of an open source platform for the processing of files generated by the communication systems of a mass service institution with three hundred branches that serves more than two million customers.

The research addresses the need to consolidate results that add value to decision-making and improve the operational efficiency of information and communication technology (ICT) services.

The objective is the development of a control panel based on measurement of key indicators. It will allow the monitoring of its operating costs and the level of quality of customer care. For this, the ELK (Elasticsearch-Logstash-Kibana) set is used, fed with the call detail records known as CDR (Call Detail Records).

Keywords: Big Data, ICT, CDR, ELK.

1 Introduction

Big data offers ICT engineers a real opportunity to capture a more comprehensive view of their operations and services [1]. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets [2]. As an example of different use cases based on CDR analysis can be mentioned, operational efficiency and improvement of the customer experience [3].

This work aims to respond to the need to analyze the operating cost of a telecommunications infrastructure of a large organization and the level of quality of the care services it provides. The result is a dashboard with consolidated information built with ELK technology, from the processing of CDRs generated by its telephony servers.

The research work focuses on the process of developing a dashboard that consolidates main indicators according to the following objectives:

1. Analysis of operating costs based on network traffic data flows according to service monitoring, to detect fraudulent behavior when it occurs.
2. Monitoring of the quality of care services, resulting from the representation of operation time for each of the interactions.

2 Work Development

The development of the work is structured in 5 fundamental stages. In the first stage, the detail of the communications system were carried out to know the data source. In the second stage, the CDR records are obtained and a data dictionary is prepared. Stage 3 deals with indexing the database. Each CDRs field with relevant information is assigned a specific type of parameter from text file to JSON document. Stage 4 runs the ETL process. A text file is build from which the Logstash module is configured for data ingestion. In the last stage, searches are carried out according to the specific objectives to be achieved, so Kibana module allows them to be consolidated into a control panel.

2.1 Data source's context and relevance

The data was provided by an organization with more than two million clients dedicated to mass consumption services. Its products are heavily regulated, that is why the competitive strategy focuses on differentiation based on the quality of customer service. It enhances the value of this work. The linking process to productive environment was necessary in order to know the details of service model and the infrastructure that supports the business services. These activities were key to understanding the technical architecture of the platform, the different models of care it supports, and the interpretation. The CDR data is used for collection, settlement, billing, network efficiency, fraud detection, value-added services, business intelligence, etc [6].

2.2 Database indexing

To build the database, you must create an index in Kibana. To do this, it is necessary to define the field type when formatting the data structure that is expected to be received. This is done through the DevTools section with PUT command.

```
PUT /cdr2020DBv2 {
  "mappings": {
    "properties": {
      "cdrRecordType": {"type": "integer"},
      "globalCallID_callId": {"type": "integer"},
      "origLegCallIdentifier": {"type": "integer"},
      "dateTimeOrigination": {"type": "integer"},
      "dateTimeOrigination_formatted": {"type": "date"},
      "dateTimeConnect_formatted": {"type": "date"},
      "origIpv4v6Addr": {"type": "ip"},
```

2.3 ETL Processing

The ETL (extract, transform and load) is the process of collecting data, adapting its fields and loading data to the base. This is done by configuring the logstash:

```

input {file {path => "C:/Users/unlam/CDR2020.txt"
      start_position => "beginning"} }
filter {csv {columns => ["cdrRecordType",...,""]}}
date { match => ["dateTimeOrigination", "UNIX"]}
target => ["dateTimeOrigination_formatted"]
date {match => ["dateTimeConnect", "UNIX"]}
target => ["dateTimeConnect_formatted"]
output {stdout {}}
elasticsearch {index => "cdr2020DBv2"} }
```

Three instances are identified. The input consists of indicating the file path to extract the data. The next step is the filter where all fields are listed in comma separated format. In the particular case of dates, the data received in UNIX format can be converted to a data type. Finally, the output indicates the indexname where the data will be loaded.

3 Results

The parameter “CalledPartyNumber” was used to classify voice traffic according to destination categories: Local, National, Emergency, International & Cellular. It allows to measure providers operating cost.

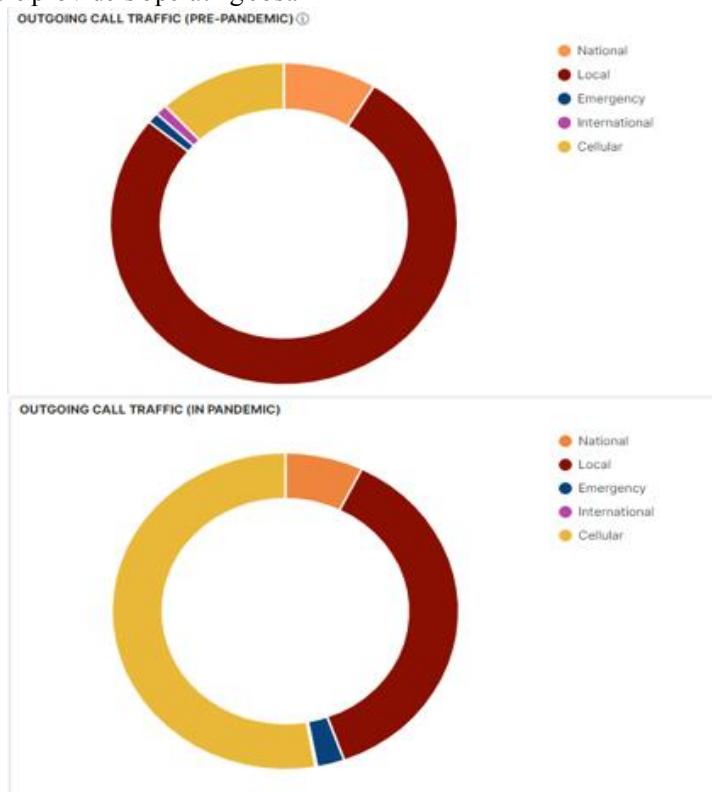


Fig. 1. Outgoing telephone traffic distribution

This filter was applied to different periods to show the changes caused by the restrictions caused by Covid 19 pandemic. The figure 1 to compare the data between the pre-pandemic period (Nov/Dec 2019) to the pandemic period (Nov/Dec 2020).

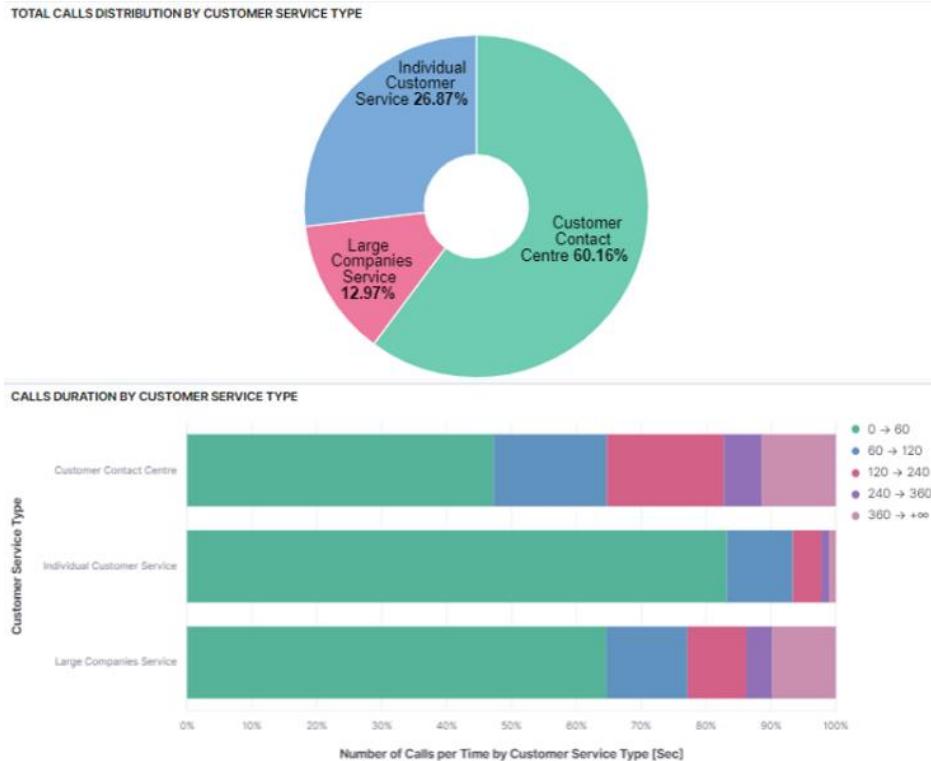


Fig. 2. Call details classified by customer service type

In the figure 2, both graphics were used to build the dashboard that measures the quality of customer service. The top graphic has the distribution of incoming calls classified by customer service type. The service is made up of three groups: customer contact center (CCC), service to individuals and large customers. The bottom graphic details the call duration time for the three customer service type.

4 Conclusions

The measurements reflected in the outgoing traffic dashboard reveal the operating costs of the telephone service. Comparing the same months of 2019 (preCovid-19) with 2020 (during Covid-19), an increase of 320% is observed for the Cellular destination category and 50% of reduction in the Local destination category. It can explain by the changes imposed by the restrictions of the pandemic that reduced the attendance of

personnel to branches in exchange for using cell phones affected. It reflects the increase in cost generated.

In the customer service dashboard, it is observed that the calls handled by the CCC represent 60% of the total calls to branches. The remaining traffic is two thirds for the individual sector and one third for large companies. At CCC, 45% of calls last less than a minute. Staff often adjust their behavior to requested productivity levels.

The calls answered by branch officials with a duration greater than five minutes, it is observed that large companies are 10%, in individuals it is only 1%. This is aligned with business objectives

The tasks of linking with the productive environment, although they require technical skills, highlight the need for soft skills for an adequate interaction and contextual interpretation.

Competing interests

The authors have declared that no competing interests exist.

Authors' contribution

"MC conceived the idea and conducted the experiments; MC, AA and AB analyzed the results and revised the manuscript. All authors read and approved the final manuscript."

5 References

- [1] Chen, M.: Use cases and challenges in telecom big data analytics. In: APSIPA Transactions on Signal and Information Processing, vol. 5, p. e19, Cambridge University Press (2016)
- [2] Verma, J., Agrawal, S., Patel, B.: Big Data Analytics: challenges and applications for text, audio, video and social media data. In: International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1 (2016)
- [3] Elagib, S., Hashim, A., Olanrewaju, R.: CDR Analysis using Big Data Technology. International Conference on Computing, Networking, Electronics and Embedded Systems Engineering (2015). <https://ieeexplore.ieee.org/document/7381414>, last accessed 2021/01/21.
- [4] Morato, J., Sanchez Cuadrado, S., Fernández, B.: Trends in the technological profile of information professionals. 25(2), 169-178 (2016). <https://doi.org/10.3145/epi.2016.mar.03>, last accessed 2021/04/02.
- [5] Dobre, C., Xhafa, F.: Parallel programming paradigms and frameworks in Big Data Era. In: International Journal of Parallel Programming, 42(5), 710–738 (2014). <https://doi.org/10.1007/s10766-013-0272-7>, last accessed 2021/03/28.
- [6] Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U.: Challenges and Opportunities with Big Data. USA, Cyber Center Technical Reports (2011). <http://docs.lib.psu.edu/cgi/viewcontent.cgi?article=1000&context=cctech>, last accessed 2021/03/17.

Proposal of a Data Warehouse for Scholarly Institutions built on Institutional Repositories

Pablo C. de Albuquerque^{1,2} [0000-0001-5277-1665], Gonzalo L. Villarreal^{1,2} [0000-0002-3602-8211], Marisa R. De Giusti^{1,2} [0000-0003-2422-6322]

¹ PREBI-SEDICI, Universidad Nacional de La Plata. La Plata, Argentina.

² CESGI, Comisión de Investigaciones Científicas. La Plata, Argentina.

pablo@sedici.unlp.edu.ar, gonzalo@prebi.unlp.edu.ar, marisa.degiusti@sedici.unlp.edu.ar

Abstract. A Data Warehouse (DW) is a tool that integrates and unifies information from multiple data sources and is used to assist decision making. In academic institutions, a Data Warehouse oriented to scientific and academic intellectual production could provide valuable information to understand, optimize and promote the processes involved in intellectual production. This work proposes to use the data sources that conform the Institutional Repositories to start developing a DW.

Keywords: Data Warehouse, Institutional Repositories, Business Intelligence

1 General Data Warehouse Concepts

A Data Warehouse (DW) is a tool that integrates and unifies information from different data sources of an organization, and serves and is useful for decision making. Data sources are usually heterogeneous, both from the point of view of the technological support (e.g., relational databases, NoSQL databases, spreadsheets, text files, etc.) and also from the point of view of the purpose of each source (for example transactional management systems, monitoring services, server access logs, etc.). The integration of these data sources into the DW is done by retrieving or extracting data from those sources, which are then transformed and finally integrated into a centralized database; this process is known as ETL: Extract-Transform-Load.

One of the DW design premises is to keep a simplified data model, requiring simple queries to retrieve useful information. This simplicity ease the integration of the DW with different Business Intelligence (BI) and/or reporting systems, such as Power BI, Google DataStudio or even MS Excel, and also promotes the exploitation of the data by users who have elementary concepts but are not necessarily database experts.

The volume of data in a DW usually grows rapidly and in many cases at an accelerated rate, reaching the order of GB, TB or even EB in short time. Despite its size, the DW must be able to execute queries and return results in optimal response times. To achieve these requirements, many actions linked to the optimization of the underlying tools (server, database engine, network, etc.) must be combined with the design of the

DW database itself, usually based on a denormalized star model, based on facts and dimensions associated with the facts. [1]

2 Data Warehouse in Scholarly Institutions

In scholarly institutions, a Data Warehouse focused on scientific and academic outputs could provide valuable information to understand, optimize and promote the processes involved in their production: what type of resources are produced, what are the areas of research, who conducts the research, where they are produced from (research centers, departments, editorial teams), when the different resources are generated, what mechanisms are used to produce or publish the resources, and how they are used both internally (research projects, working groups, theses, etc.) and externally (citations, visualizations, downloads, mentions, etc.). [2]

Like any organization, most scholarly institutions have a wide diversity of systems that manage, host and publish different resources produced by the institution. Like expected, each system organizes and manages its data based on its needs and the availability of technological resources at the time of the development. That is why the diversity of data sources that make up the ecosystem of technologies around an institution can make certain tasks more difficult such as assisting in decision making, since it is necessary to integrate these sources in a single place. Some of the typical systems used in scholarly institutions include institutional repositories, current research information systems (CRIS), journal portals, conference portals, digital book portals, among others. It is important to be clear that not only the information directly associated with the function of each system is important (for example, books and their authors in a Books Portal), but also much information generated by the system itself: users' access, server logs or even security reports linked to each system.

Following is a description of some of these systems, with emphasis on what data they manage, how reliable they are, and what processes should be implemented to integrate these data into the Data Warehouse:

- Institutional Repository (IR):
 - Advantages: data are already standardized through the use of multiple controlled vocabularies, reviewed by staff dedicated to ensuring compliance with repository policies, and adoption of guidelines that allow integration into repository networks. The organization into collections and communities provides valuable information. The use of persistent identifiers makes it possible to identify a resource univocally on the web, facilitating interoperability with other systems. As mentioned above, a repository can be part of repository networks that provide services and increase the visibility of the scholarly production. Repositories can also participate in different agreements with other institutions, which gives access to standardized resources that, after being reviewed, can be incorporated into a particular collection. The adoption by the institution's users is also important since many are already using these services. [3]

- Disadvantages: authors do not always deposit their production in the IR which could generate a partial view. Besides, many repositories include "less interesting" resources such as learning object or internal lecture notes.
- CRIS System
 - Advantages: the amount of data these systems usually store tends to be very complete, since these systems are generally used for institutional evaluation tasks and therefore the different actors of the institution must ensure that the results of their work are there for the evaluators.
 - Disadvantages: the information is uploaded by the author who is generally not skilled in describing the resources that are submitted, and there are usually no instances of review of this data. Many data will be repeated among authors, in part because no identifiers are used to create relationships between resources and people (authors, editors, etc.). In general, the data are not standardized.
- Books and Journals Portals
 - Advantages: Similar to the IR, these portals have reliable data, uploaded by the authors and in this case corrected by the different editorial teams. Their organizational structure is usually simple: numbers, volumes, articles, in the case of journals; academic units, thematic areas in the case of book portals. These systems use persistent identifiers, which improve interoperability. In many cases they provide data on how these resources were generated.
 - Disadvantages: not all editorial teams will necessarily have the same policies and quality in their metadata, nor will they follow the same workflows. It should also be noted that systems based on standardized metadata schemas are not always used.

While there may be many other systems in these institutions, it seems clear that the IR is a great candidate to begin the development of a DW: the volume of information that an IR can handle, the reliability of the stored data, the available interoperability tools, and the existing services and infrastructure provide an interesting starting point.

3 Users and roles

As mentioned above, around a repository there are actors with different responsibilities and needs that periodically require access to information to assist them in their decision making. The IR provides data that allows them to prepare reports, analyses and dashboards that reflect the reality of a part of the repository at a given time. Some examples of data requirements to an IR are repository managers may need to know the impact factor that was generated by the import of a new collection into the repository; technical staff may want to know how many requests are served by the web server in the last month and of that total, which is the flow of malicious bots identified by a third-party service and then, based on this information, make decisions that allow filtering and maintenance of the infrastructure; the administrative staff working in the

repository may want to know the status of the resources imported in the last year, in order to know if they should perform revision tasks on them; the authorities of the institution may need to see the growth in the number of items by typology in the last semester, by institution, department or academic unit; the authors of the resources stored in the repository may want to know from where the resources they have participated in have been accessed; visitors who search and download resources from the repository may want to see where the research lines of the different academic units are progressing. Although these are just a few examples, it can be seen that some IRs are already fulfilling the functions of a DW. However, as they only have their own internal data, they do not provide an overall picture of the entire institution.

4 Putting it all together

In this work, we have reviewed some of the functions and requirements typically served by an IR, with emphasis on the information requests and reports that may be solicited periodically. As we have mentioned, each system or data source structures its information to respond to its own needs, so some data may not be directly available and may require a special process to be inferred or calculated.

While many of the above tasks can be automated and scheduled, it is important to keep in mind that they always involve data from the repository itself, but it is often necessary to combine data from other data sources to get a complete picture.

To solve this, it may be necessary to use other sources, so it is no longer sufficient to define tasks that process information from one source to infer other data, but rather to define processes that unify and standardize various sources in one place.

A Data Warehouse would solve these problems, gathering in one place the necessary information to have a broader view of the academic situation of an institution, simplifying the tasks of data integration and normalization, with the aim of answering queries to users, such as institutional authorities, technical and administrative staff and the general public, in order to assist them in their decision making. The development of a DW implies a great effort, both on the part of the team responsible for its design, implementation and maintenance, as well as on the part of those responsible for the different areas of the institution whose data must be periodically integrated. For this reason, the success of such a project requires the commitment of the entire institution, from the highest authorities to the technical staff responsible for managing each database. However, the potential for obtaining useful, quality and instantaneous information from this kind of tool suggests that perhaps academic institutions should seriously consider investing resources in its implementation.

References

1. Kimball, Ralph, y Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. John Wiley & Sons, 2013.
2. Rójas-Muñoz, C., & Saquicela, V. (2017). Sistema de ayuda a la decisión basado en un data warehouse. *Maskana*, 8, 175–187. Recuperado a partir de <https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/view/1461>

3. De Giusti, Marisa Raquel. «Los nuevos roles del repositorio institucional». Visión Conjunta 10, n.º 18 (agosto de 2018). <http://sedici.unlp.edu.ar/handle/10915/72087>.

Semantic Web

Semantic Web for interoperable food safety legislation data: A case study

Carlos Enrique Pintor¹, Carlos Francisco Ragout¹, Diego Torres^{1,2}, and Alejandro Fernandez^{1,3}

¹ LIFIA, Facultad de Informática, UNLP, Calle 50 y 120, La Plata, Argentina.

² Depto CyT, Universidad Nacional de Quilmes. Roque Saenz Peña 352, Bernal, Argentina.

³ Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, La Plata, Argentina.

Abstract. Food safety legislation plays a central role in regulating the levels of chemicals used in agriculture practices in order to prevent potential risks to consumers' health within a certain region or country. Public Health organizations publish these regulations as recommendations on allowed quantities of chemicals residues for different types of crops. These documents pose a major challenge for automatic processing as their format is not normalized nor the terminology used is uniform in any way. Semantic Web technology tools offer a solution as these documents may be published as linked data which would allow computers to process them automatically, so that further analysis and interoperability would be possible. In this paper we introduce MRL-O, an ontology for describing data on allowed levels of residues present in commodities of agricultural origin. MRL-O serves as a standardized framework for sharing interoperable data and to provide tracking metadata about its sources and transformation processes. We also describe a step-by-step procedure to obtain MRL-O linked data from real non-normalized documents. Also, we applied this procedure on data published by Argentina and Brazil with promising results. Consequently, we argue that the proposed ontology is sufficient to model the domain of MRL regulation and serves as the basis for tools that support interoperability in this domain.

Keywords: Maximum Residue Limits, Agriculture, Health, Regulation, Semantic Web, Linked Open Data

1 Introduction

Agrochemical substances and its derivatives are used throughout agricultural processes to prevent and control the presence of pests. As a result the products obtained from these practices may contain certain levels of chemical residues potentially harmful to human health. A mechanism to monitor and control the maximum concentration of pesticide residues (MRL) in food commodities is therefore required [7]. Governments and Health Organizations determine and

publish recommended values of MRL periodically, as these values have a significant impact in human health [6] and in international food trade [5][8].

Given the lack of official or standardized guidelines regarding how this data should be produced and published, a wide range of methods and supporting media for publishing documents on MRL are used, involving different formats (e.g. pdf, xml, csv, etc.), content types (tables, graphics, lists, etc.) and language. There is no formal curation process on the data itself to prevent inaccurate terms, syntax errors, omissions, synonyms and proprietary data structures.

The diversity in publication formats makes it difficult to process and analyze the datasets by using computers due to incompatibility issues among documents from different sources, or even between versions of the same document. We believe the Semantic Web [1] offers an alternative to address this interoperability challenge.

In this paper we apply Semantic Web technologies and tools to design and create MRL-O (Section 3), a specific ontology to represent MRL-related data. We propose a semantic pipeline (Section 4) to transform non-normalized data into MRL-O semantic datasets ready to be consumed and processed by computers without any regards about formats of origin.

2 Background and related work

The concept of Semantic Web was introduced by Berners-Lee [1] to encompass a set of technologies that provides a better knowledge representation with the use of ontologies, software agents, and logic rules. It is an extension to the World Wide Web where the information is described in a machine-readable format. Data in the Semantic Web is modeled using RDF (Resource Description Framework) [9, Chapter 2]. RDF models are built around web resources and triples.

This work builds upon several existing developments, ontologies and vocabularies.

AGROVOC [3] is a multilingual open dataset about agriculture concepts and relationships which is used to identify resources covering all areas of interest of United Nations FAO (Food and Agriculture Organization).

ChEBI [2] is a database and ontology specialized in small chemical compounds of biological interest developed by the European Bioinformatics Institute. ChEBI has been widely adopted by numerous bioinformatics projects and as ontological reference in several semantic-web projects.

The Units Ontology [4] is also part of the ontology network of MRL-O. It is used to express quantities and proportions of agrochemical components under standard terms.

3 MRL-O

MRL-O (Maximum Residue Limit Ontology) is an ontology that models the domain of MRL regulation. Following the best practices of the Semantic Web,

MRL-O borrows elements from other existing ontologies. In particular MRL-O relies on AGROVOC, ChEBI, and the Units Ontology to express the information contained in a single record of an MRL-O dataset. Similarly, it relies on PROV-O, Dublin Core, and Wikidata to describe the process of applying the transformation pipeline (presented in the following section) to datasets published by a given organization.

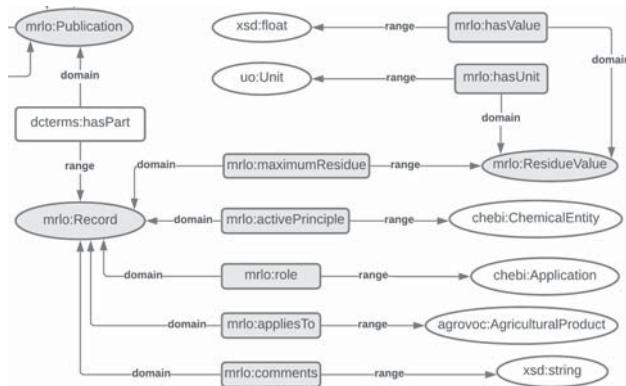


Fig. 1. Part of the MRL-O ontology that represents MRL records

Figure 1 uses the graph notation of RDF to provide an overview of the terms used to express the information contained in a record. The elements with a gray background are those introduced by MRL-O, whereas elements with a white background are adopted from other vocabularies.

4 Transformation pipeline

At the heart of our proposal lies the transformation pipeline. Figure 2 provides an overview showing its main activities namely, Clean, Align, and Transform. Following, we discuss each of these activities with more detail.

Clean

The “Clean” activity takes data files as inputs and produces a single file of clean data rows representing a unique statement involving one crop, one active principle, and one application. Then, normalization takes place to trim blanks, collapse consecutive blanks, and standardize capitalization. The output of this first activity is a clean table, with one row per MRL record, and four columns:

- Active principle: Name of a chemical substance (e.g., 2,4-D)
- Role: Role or usage of the chemical substance (e.g., herbicide)
- Product: Crop or agricultural product or commodity (e.g., tomato)
- MRL: Maximum residue level in mg/Kg (e.g., 0.05)

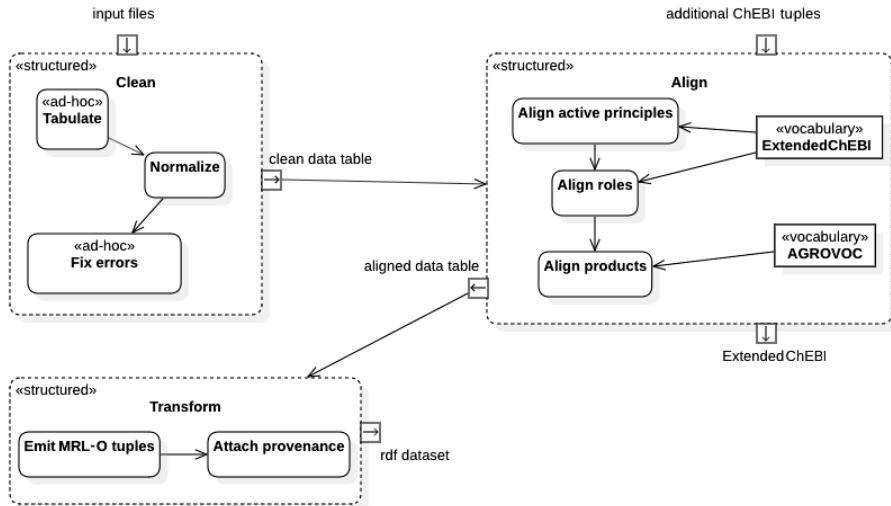


Fig. 2. Semantic transformation pipeline

- Comments: Any comments regarding this record (possibly empty)

Align

The goal of the “Align” activity is to replace all references to chemical substances and to agricultural products or crops in the data with the corresponding resources in the Semantic Web using the previously mentioned reference ontologies.

Transform

The final activity transforms the clean table into an RDF dataset. After all rows in the table have been processed, the process adds provenance information triples for the publication resource. The final result is an MRL-O based dataset.

5 Case study

As a proof of concept we applied our vocabulary and the pipeline on two real world datasets from Argentina and Brazil. These documents were published in 2020, and the pipeline was implemented using OpenRefine.

Regarding the Argentinean case study, we found that the reference document on MRL is published by SENASA as an excel worksheet. On the other hand, the Brazilian government through its national sanitary agency (ANVISA) publishes information regarding phytosanitary legislation on their official website from which a csv file can be downloaded.

5.1 Inter-operable queries

The most interesting part of generating semantic datasets for us was creating usage scenarios where useful information could be extracted. For example, we created a set of SPARQL queries to answer some common questions comparing food legislation in Argentina and Brazil. It is worth mentioning that this exercise, although plausible, would have been extremely complex to achieve without the support of the Semantic Web tools we applied.

6 Conclusions

The two case studies in this article show that MRL-O is rich enough to cover the basic requirements to express meaning within the MRL domain. The results obtained from the pipeline execution proved to be effective as well. The set of sample SPARQL queries shows how simple it is to extract meaningful information from MRL-O data, and that more complex combinations between records are also possible. The idea of having a food safety legislation based upon the Semantic Web is feasible and valuable.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. *Scientific american* **284**(5), 34–43 (2001). Publisher: JSTOR
2. Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research* **36**(suppl.1), D344–D350 (2007)
3. FAO: AGROVOC. FAO (2021). DOI 10.4060/cb2838en
4. Gkoutos, G.V., Schofield, P.N., Hoehndorf, R.: The Units Ontology: a tool for integrating units of measurement in science. *Database* **2012**(0), bas033–bas033 (2012). DOI 10.1093/database/bas033
5. Li, Y., Xiong, B., Beghin, J.C.: The Political Economy of Food Standard Determination: International Evidence from Maximum Residue Limits. In: Nontariff Measures and International Trade, *World Scientific Studies in International Economics*, vol. Volume 56, pp. 239–267. World Scientific (2016). DOI 10.1142/9789813144415_0014
6. Li, Z.: Evaluation of regulatory variation and theoretical health risk for pesticide maximum residue limits in food. *Journal of Environmental Management* **219**, 153–167 (2018). DOI 10.1016/j.jenvman.2018.04.067
7. WHO: Principles and Methods for the Risk Assessment of Chemicals in Food. World Health Organization (2009)
8. Xiong, B., Beghin, J.C.: Stringent maximum residue limits, protectionism, and competitiveness: The cases of the us and canada. In: Nontariff Measures and International Trade, pp. 193–207. World Scientific (2017)
9. Yu, L.: A developer’s guide to the semantic web. Springer, Berlin (2011). OCLC: 700066210

Towards Ubiquitous and Actionable Augmented Reality Browsers by using Semantic Web Technologies

Martín Becerra ^{1[0000-0002-8084-5091]}, Jorge Ierache ^{1[0000-0002-1772-9186]} and María José Abasolo ^{2,3[0000-0003-4441-3264]}

¹ National University of La Matanza, Engineering Department, Applied Augmented Reality Team,
1754, San Justo, Buenos Aires, Argentina.

² National University of la Plata, School of Computer Sciences, III-LIDI, 1900, La Plata, Buenos
Aires. Argentina.

³ Commission for Scientific Research of Buenos Aires., 1900, La Plata, Buenos Aires, Argentina.

{mabecerra, jierache}@unlam.edu.ar
mjabasolo@lidi.info.unlp.edu.ar

Abstract. In this paper we describe the preliminary results in the context of PhD thesis work which expands the PROINCE C231 2020-2021 project Voice Commands and Face Recognition for Augmented Reality applications. Our work aims to develop a framework to assist users to do their daily tasks through the creation and exploitation of reusable procedures for AR apps using the ontology of our research. In the second instance we aim to contribute to actionable and interoperable data sources using semantic web technologies where apps consume data regardless of the application that generates them. This paper will explain the basis of the experimental framework and the preliminary design of a web service called Semantic middleware that performs all the semantic operations necessary to make procedures interoperable with other third-party applications using Semantic Web technologies

Keywords: Augmented Reality, Semantic Web, Linked Data, Linked Data Cloud.

1 Introduction

Augmented Reality (AR) adds virtual elements to the real environment, enriching the perception of reality with virtual information [1]. In recent years, AR has expanded to different application fields such as education, healthcare, industry, tourism, marketing and entertainment. Currently there are several popular augmented reality browsers (AR Browsers) on the market such as LayAR[2], wikitude[3] to provide augmented reality experiences. These are limited because they allow users to passively consume a delimited set of functions. Different alternatives were researched like ARAMA3D [4], T. Matuszka et. al. [5] y SmartReality [6] which offer a ubiquitous experience through integration of semantic web technologies to add information from the Linked Data Cloud [7] to enrich the description of point of interest near the user position. Although these applications allow the creation of contents, they are consumed statically. In other words, they can only be applied to view descriptions without being able to perform any

action on them. However, it is useful for users to have a dynamic interaction with the contents available through the definition of procedures composed of a set of actions to be performed in an environment enriched by augmented reality. In this order the presented framework is positioned in the interception of the areas of Augmented Reality browsers, users, and ontologies (Semantic web) (Figure 1), providing a dynamic interaction (DI), through interoperable procedures about augmented objects using AR applications and semantic web technologies. These capabilities will impact in several areas in Industry 4.0 contexts, such as the creation of a sequence of tasks to be performed by an intelligent operator in his workstation in a Smart factory, in the augmentation of tasks to be done with an IoT equipment of the plant or in augmented home devices integrations as well.

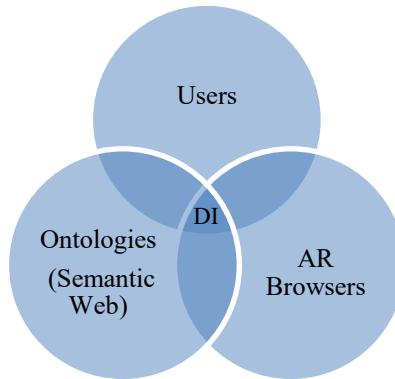


Fig. 1. Integration of knowledge areas

This paper will explain the basis of the experimental framework and the preliminary design of a web service called Semantic middleware that performs all the semantic operations necessary to make procedures interoperable with other third-party applications using Semantic Web technologies.

2 Framework

This work aims to develop a framework to assist users to do their daily tasks through the creation and exploitation of reusable procedures for AR apps, using the ontology of our research. Secondly, it is aimed to contribute to actionable and interoperable data sources using semantic web technologies where apps consume data regardless of the application that generates them [8]. The general architecture is divided in three parts: A Procedure editor, a semantic middleware, and an Augmented Reality Browser (Figure 2). The procedure editor will allow content creator users to create and edit procedures composed of a set of steps/actions to be performed in the physical environment, using augmented reality technologies. Each step can involve object manipulations, so the editor will allow to search and relate data about that objects from Liked data cloud if it is available. For this task data will be mainly fetched from DBpedia [9]. Semantic

Procedure Editor will allow procedures to be associated to virtual catalogs thanks to the Semantic Augmented Catalog System service, through a semantic layer which will apply our ontology "Semantic Catalog System Extension Ontology". In Addition, the semantic editor of procedures will allow our augmented virtual catalog system [10] to be an interoperable data source with the framework proposed as a doctoral thesis work. The use of semantic technologies will allow other augmented reality applications to consume procedures and discover augmented virtual catalogs for integration and exploitation on their platform for their own purposes.

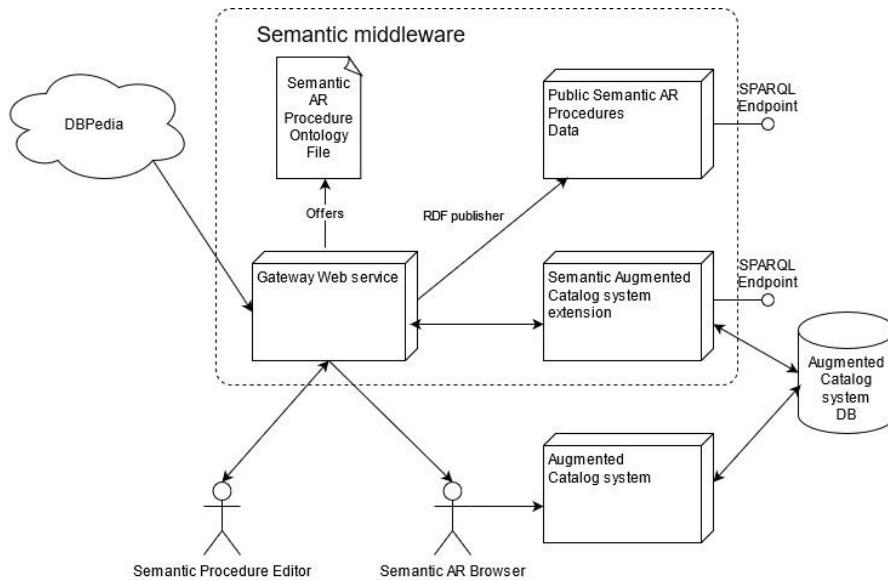
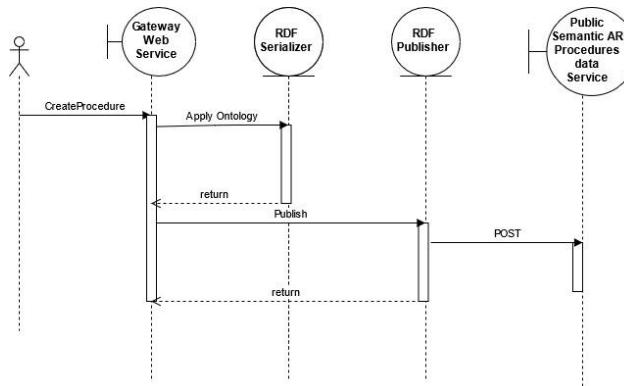


Fig. 2. Conceptual architecture of the framework

The semantic middleware is composed of a main web service called Gateway web service that has as main responsibilities to maintain the Semantic Augmented Reality Procedures ontology and to be the entry point of requests for the creation of procedures and search of procedures to be added by the augmented reality browsers. A RDF triple-store acts as a Public Semantic AR Procedures data service that is responsible for storing the procedures created and providing a SPARQL endpoint so that other applications can consume the data generated by the system. Finally, the Semantic Augmented Catalog System extension service that as mentioned above, works as the semantic layer of our augmented catalog system.

At the time of procedure creation, the gateway web service works as a mediator between the editor and the RDF triplestore (Public Semantic AR Procedures data service) to store the created procedures. In Figure 3 we can observe in a sequence diagram, when the Gateway web service receives the created procedure, applies the ontology from our research and redirects this structured data to the triplestore for its corresponding storage for later search and consumption by the augmented reality browsers.

**Fig. 3.** Procedure creation sequence diagram

3 Conclusion

The present paper presents general architecture of the framework and the preliminary design of the semantic web service that will allow users to incorporate linked data in their daily actions through the use of augmented reality browsers in their real physical environment, allowing to generate specific procedures for the physical instruments to be augmented; this last capability concentrates the main contribution of this paper, allowing a new way to exploit the augmented content, where the user interacts dynamically through procedures that are interoperable with other augmented reality browsers using Semantic Web technologies.

References

- Yee C., Abásolo M. J., Más Sansó R. y Vénere M. (2011). “Realidad virtual y realidad aumentada. Interfaces avanzadas.” ISBN 978-950-34-0765-3.
- LayAR. <https://www.layar.com/>. Last accessed 2021/4/3.
- Wikitude. <https://www.wikitude.com/>. Last accessed 2021/4/3.
- Aydin B., Gensel J., Genoud P. Extending Augmented Reality Mobile Application with Structured Knowledge from the LOD Cloud. <https://tinyurl.com/3x87c6ss>. 2021/4/3.
- Matuszka T. et. al. The Design and Implementation of Semantic Web-Based Architecture for Augmented Reality Browser. <https://tinyurl.com/nf5jxswx>. Last accessed 2021/4/3.
- Nixon L., Grubert J. Reitmayer G. SmartReality: Integrating the Web into Augmented Reality. <https://tinyurl.com/46hzw8b9>. Last accessed 2021/4/3.
- Linked Data Cloud. <https://lod-cloud.net/>. Last accessed 2021/4/3.
- Vert S., Vasiu R. Integrating Linked Data in Mobile Augmented Reality Applications. <https://tinyurl.com/4uynbb95>. Last accessed 2021/4/3.
- DBpedia. <https://www.dbpedia.org/>. Last accessed 2021/4/3.
- Ierache J., Mangiarua N., Bevacqua S., Verdicchio N., Becerra M., Sanz D., Sena M., Ortiz F., Duarte N., Igarza S. (2015). “Development of a Catalogs System for Augmented Reality Applications”. Science Index 97, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9(1), 1 - 7. ISSN 1307:6892.

Smart Cities and Emerging Topics

Thermodynamic Dissipative Systems and Information Theory to Study the Social Component of a Smart City

Gabriele De Luca^{1[0000-0001-9728-9581]}, Thomas J. Lampoltshammer^{1[0000-0002-1122-6908]}, and Felipe Vogas^{12[0000-0001-5087-7845]}

¹ Department for e-Governance and Administration, Danube University Krems, Krems, Austria

² COPPEAD Graduate School of Business, UFRJ, Rio de Janeiro, Brazil

gabriele.deluca@donau-uni.ac.at

Abstract. In this article, we discuss the application of information theory and the theory of thermodynamic dissipative systems to smart cities. Specifically, we study how to model the interaction between a society and a smart city, under an information-theoretic approach. Because the smart city comprises both a social and a technological component, it then becomes possible to use information theory to study them both. In this paper, we discuss a model that applies the constraints from thermodynamic dissipative systems theory in order to study smart cities, and their associated social system, in their information processing capacity and in their evolution over time. Within the context of our model, we are allowed to study under what conditions a smart city would expand or contract, or to state that the smart city shrinks if its output greatly exceeds its input.

Keywords: Smart cities, socio-technical systems, information theory, dissipative systems.

1 Smart cities as an open thermodynamic systems

Social systems are non-equilibrium, open thermodynamic systems that exchange energy, matter, and information, with the rest of the environment [1]. They inherit this property from the set of biological organisms that comprise them; i.e. the humans, whose interaction with one another constitutes the social system [2]. Because it is possible to frame social systems as thermodynamic systems, it is therefore also possible, in principle, to apply information theory for their analysis. This implies, for example, the possibility to apply entropic methods from statistical thermodynamics and information theory, in order to study the trajectory of the dynamic evolution of a social system in its environment [3]. In using this approach, however, the problem that we face is that the description of the state of a social system implies the arbitrary definition of a finite and small set of variables, as it is common in the application of agent-based modeling to social systems [4]. This, in turn, draws the most complex system in the universe, the social system, in a rather cartoonish form that grossly trivialize most of its aspects and characteristics, and therefore reduces the complexity that is studied. An-

other possible research direction is to apply not entropic methods, but dissipative systems theory [5], which has already been applied to study cities and their growth [6]. This is promising but does not directly allow the researchers to model the internal mechanisms of the social system, which is best treated as a problem of entropy and its variation. This inability by traditional thermodynamic approaches to represent social systems sufficiently well is, however, typically not a characteristic of purely technological systems such as computers or information networks [7].

Social sciences provide a useful frame to study hybrid systems comprising a social and a technical component, under the theory of socio-technical systems [8]. This type of approach is useful, for example, to delimit and analyze the system that we call a smart city. A smart city is, in this context, a cyber-physical system for the intelligent management of space and agents in an urban environment [9]. Because it is a technical system, a smart city can be studied under thermodynamic and information-theoretic approaches, insofar as we study its computational or information-network components. Because it is a social system, it is also possible, in principle, to study a smart city under a thermodynamic or information-theoretic approach. In this paper, we propose to use the consideration that a smart city is a hybrid socio-technical system that operates in non-equilibrium as an open thermodynamic system, to extend the information-theoretic approach for the study of its technological component to the study of the social system in which the latter is embedded.

2 Modeling of the system

We argue that, if the social system can be considered as a dissipative system, then a smart city that comprises that social system also can. This means that a smart city can be modeled as a system with a state $X(t)$ at time t , that receives a thermodynamic or information input $u(t)$ from the environment and responds with an output $y(t)$. If the supply rate for that smart city is $w(u(t), y(t))$, then the following inequality holds:

$$\dot{S}(X(t)) \leq w(u(t), y(t)) \quad (1)$$

In here, $S(X(t))$ is a storage function that indicates the energy held by the system at time t , with $S(0) = 0$ and $\forall t: S(X(t)) \geq 0$, which is the condition required for the system to exist. We also assume S to be continuously differentiable. Further, we also assume that, even though this is not so obvious for real world systems, the input, and the output $u(t), y(t)$ to the system consist of known measurables. This means that there is a finite-dimensionality vector comprising all the variables that describe the input from the environment into the smart city, and another one comprising the output.

In this model, we can apply information theory to describe $X(t)$, the state of the system, and its variation over time. If X comprised only of computers and networks, we could then describe it as we do for those systems, and the theoretical problem we indicated earlier would not exist. The state X however comprises also the state of the n individuals that populate the social system, and we call these states x_i with $1 \leq i \leq n$. For simplicity, we imagine that the state of the technological component x_{sc} of the

smart city comprises a single computing system, that represents the multitude of computers and IoT devices that, in a real-world smart city, characterize the latter's technological component [10]. If we do that, then the set of states of the elements of the system is the state of the system:

$$X = \{x_{sc}, x_1, x_2, \dots, x_n\} \quad (2)$$

This, however, does not take into account the existing relationships between the humans in the social system with one another, and all humans individually with the technological component of the smart city. The rest of the modeling can then be conducted by constructing a directed graph $G = \{V, E\}$, where V and E can be described as:

$$V = \{sc, 1, 2, \dots, n\}, E = \{e_{sc}, e_1, e_2, \dots, e_n\} \quad (3)$$

All human elements of this graph connect to a finite and small subset of V . The vertex sc , corresponding to the technological component of the smart city, possesses undirected edges with all elements of V other than itself. In other words,

$$e_{sc} = \{(x_{sc}, x_1), (x_{sc}, x_2), \dots, (x_{sc}, x_n)\} \quad (4)$$

The smart city also distinguishes itself from a standard computing system because its technological component makes decisions by considering the decisions that its human components would make. This means that there exists some kind of decision function f such that:

$$x_{sc}(t+1) = f((x_1(t), e_1), (x_2(t), e_2), \dots, (x_n(t), e_n)) \quad (5)$$

Whereas all humans make decisions according to their decision functions g_i :

$$x_i(t+1) = g_i(x_i(t), e_i(t)) \quad (6)$$

This gives us a skeleton model on which we can apply entropic methods to study the variation in complexity of the system, as the supply or the decision functions change.

3 The measurement of the input and the output, and why does all of this matter

We can measure the input $u(t)$ from the environment by measuring the bits that enter the smart city system or its database at any given time. The measurement of the output $y(t)$ can also be analogously conducted. This is because, in order not to clutter any information system, information must be deleted from it. Deletion of information, according to Landauer's principle, generates heat [11]. The information deleted from a smart city, or its corresponding heat, therefore comprise the thermodynamic output from the system. This consideration allows us to treat smart cities as dissipative systems, and supports the validity of the abstract model we propose. If we are allowed to frame smart cities as dissipative systems, we are also allowed to ask additional questions that pertain the latter, in its application to the former. We can, for example, study

under what conditions would a smart city expand or contract, under the constraint indicated in (1), above. For instance, a simple consideration would be to state that the smart city shrinks if its output greatly exceeds its input, such that the supply rate $w(u(t), y(t))$ is very low. If we assume that most of the input of the smart city comprises data concerning the human population, this brings us to ask questions such as “how does the variation in the human population affect the growth of a smart city?”. We can, in fact, imagine a smart city that keeps growing its database even in absence of a human population, as a consequence of input from IoT devices. Albeit undesirable, this is in principle possible. If, instead, we have a way to relate the information originating from the social component of the smart city to the one originating from IoT, this gives us a language for studying trade-offs between the increase or reduction in the two, and the growth of the smart city as a whole.

4 Acknowledgements

This research was supported by the Erasmus+ Programme of the European Union, project reference number 598273-EPP-1- 2018-1-ATEPPKA2-CBHE-JP.

5 References

1. Schweitzer, F.: Modeling complexity in economic and social systems. World scientific (2002)
2. Michail, P.: The mechanisms operation of thermodynamic system of a human organism. European Journal of Biophysics 2, 29-37 (2014)
3. Preiser, R., Biggs, R., De Vos, A., Folke, C.: Social-ecological systems as complex adaptive systems. Ecol. Soc. 23, (2018)
4. Schulze, J., Müller, B., Groeneveld, J., Grimm, V.: Agent-based modelling of social-ecological systems: achievements, challenges, and a way forward. Journal of Artificial Societies and Social Simulation 20, (2017)
5. Gallopin, G.C.: Cities, sustainability, and complex dissipative systems. A perspective. Frontiers in Sustainable Cities 2, 54 (2020)
6. Bristow, D., Kennedy, C.: Why do cities grow? Insights from nonequilibrium thermodynamics at the urban and global scales. Journal of Industrial Ecology 19, 211-221 (2015)
7. Anand, K., Bianconi, G.: Entropy measures for networks: Toward an information theory of complex topologies. Phys. Rev. E 80, 045102 (2009)
8. Winby, S., Mohrman, S.A.: Digital sociotechnical system design. The Journal of Applied Behavioral Science 54, 399-423 (2018)
9. Camarda, D.: Complexity, governance and the smart city. In: REAL CORP 2019-IS THIS THE REAL WORLD? Perfect Smart Cities vs. Real Emotional Cities. Proceedings of 24th International Conference on Urban Planning, Regional Development and Information Society, pp. 171-181. CORP–Competence Center of Urban and Regional Planning, (Year)
10. Gaur, A., Scotney, B., Parr, G., McClean, S.: Smart city architecture and its applications based on IoT. Procedia computer science 52, 1089-1094 (2015)
11. Talukdar, S., Bhaban, S., Melbourne, J., Salapaka, M.: Analysis of heat dissipation and reliability in information erasure: A gaussian mixture approach. Entropy 20, 749 (2018)

Videogames and virtual assets exchange.

Flavio A. Garrido^{1,3}, Hernán D. Merlino^{1,2,3}

¹ Programa de Maestría de Ingeniería en Sistemas de Información. Escuela de Posgrados – Universidad Tecnológica Nacional (UTN) – Facultad Regional de Buenos Aires – Argentina

² Laboratorio de Sistemas de Información Avanzados (LSIA) Facultad de Ingeniería, Universidad de Buenos Aires (FIUBA) - Argentina

³ Grupo de Estudio de Metodologías para la Ingeniería en Software y Sistemas de Información
ing.flaviogarrido@gmail.com, hmerlino@fi.uba.ar

Abstract. With the increase of the online videogame industry and the acceptance of the players to invest time and real money in the games, the developers create new business models for selling virtual assets. This work is part of a postgraduate thesis in development and examines the business of video games, virtual assets, why players spend real money and the advantage of a common market between different games would have to increase the profits of the developers and allow the time invested by the players to also give them profits.

Keywords: Videogames, virtual assets, currency exchange, common market

1 Videogames

The video game industry is one of the most has grown today [1] due to the increased possibilities to connect to the Internet and the growth of social networking through interactivity that exists between users of these networks. This is reflected in participatory online games, where several people connect to play together, either within groups such as clans or joining to meet goals collaborating with each other, these games are called "Massively Multiplayer Online Game" (MMOG) [2].

We have different business architectures in online game, some where the player pays a monthly amount to access all the content and others where it is played for free, generating the need to spend real money to progress in the game, these business model is called "Free to play" (F2P) [3]. This model has become an integral part of online services, but more quickly in games [4], where there is a need to spend real money, to advance faster; either by purchasing resources or eliminating spam through subscription accounts or premium currencies [5], even if a small group of users spend money it seems to be a successful revenue model [5].

The F2P model is used especially for casual games, those that can be easily learned and are played occasionally, as well as video games available on social networks [5]. However, it is being implemented in more complex videogames such as Cross-Fire, which is among the best sellers worldwide [5]. The success of this modality continues to call on developers to create more and more video games that implement it, produc-

ing a great offer, which reduces the user base that a game can attract. Thus reducing retention and increasing the expense required to bring in new players [5], therefore, developers must identify the most profitable users [5] and thus be able to keep them.

As mentioned by Hamari, Hanner & Koivisto [6] of the analysis on the 300 best applications of the Apple store reveals that this business model has become the main option of many virtual services; similar results were obtained from the store of Google. Based on the great demand, developers must face the problem of a balance between creating a main system with the highest possible quality and at the same time producing the need for premium content to obtain benefits [6].

1.1 Why players spend real money?

The successful of this model make us question “Why players spend money in intangible items such as armors, esthetic or collectable items”. This question can be answered with these factors list: 1) Eliminate Spam, 2) the customization of players characters, 3) shows different social status, 4) Advance faster in-game and 5) Avoid repetitions. These factors were obtained from the analysis of several studies [4] [5] [7] [8] [9] [10] and it can be seen that there is a generated need to invest money in or out of the game and thus obtain benefits.

2 Virtual Assets Exchange.

Another point to take into account is the exchange of virtual assets for other virtual assets, real money, goods and services. The exchange of virtual items first emerged in 1999, through exchange between players, in games such as Ultima Online and EverQuest, where users listed their items on eBay and others bid for them [7].

In the study by Bi and Shu [19] is indicated that the implementation of an official platform for the exchange of virtual money depends not only on the demands of consumers, but also on the will of issuers. As an example of an exchange platform we have GameUSD.com [19], this platform allows players to sell and buy virtual currencies, the price often being lower than the game provider; which disrupts the normal pricing system of the virtual currency and damages the earnings of the game provider. It can also be seen [19] the importance of implementing an official currency exchange platform, which leads to a demand increasingly strong of reverse change by consumers, which promotes the emergence of third-party platforms.

In the work of Siira et al. [1] an implementation of a common market between two games is proposed, with which items could be bought with the currency of one game in the other, and there must be a commitment from the game providers for the exchange rate. And in the case of mobile videogames, Apple and Google also come into play, who receive a commission from the transactions carried out in videogames for real money, for which a platform for purchases between videogames must be implemented in compliance with the rules that they stipulate [1].

3 Common Marketplace

Based on the analysis, the video game market is booming and will continue to grow. This creates an overpopulation of video games and developers must get creative when developing them. On the other hand, when players leave a game, the progress and resources obtained remain in the account until it is reused and if real money was invested, it is considered a loss. To avoid this, the option of a common market is proposed where the resources obtained in a game can be traded for a virtual currency and it is used for the purchase in other games. Where developers will be benefited both by the sale of assets within the system and by the possibility that a player has to recover the investment in another video game within this ecosystem, therefore it will be more likely to select a new video game that works with this platform.. Expanding the architecture proposed by Siira et al. [1] the following common market theoretical architecture is proposed, to which new modules are added for the management of virtual currencies. Where the element in this architecture are 1) Bank, will be in charge of managing user accounts and transactions records, 2)Games A to N, are the different games, 3)Mobile platforms, such as Google and Apple, are the interaction mechanism for purchases in mobile video games, 4) New virtual currency, new common currency to use within the system, 5)Virtual currency wallets, it is the payment method within the new market., 6)Third Party Web Stores, allows third parties to sell products and services and 7)Store platform, allows transactions to be carried out.

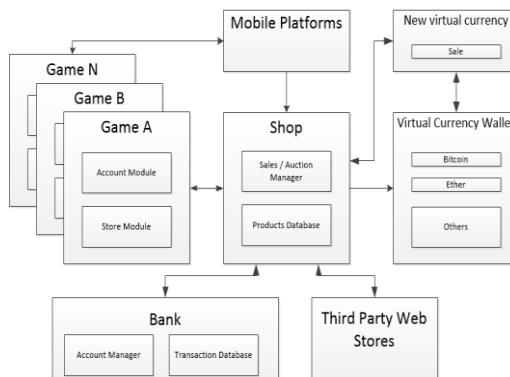


Fig. 1. Theoretical architecture of the common market

As mentioned, it is a theoretical model since various legal implications must be taken in visits within the system, such as: 1) the regional laws where users and developers are located, on the profits obtained, 2) it must have a method of control over developers to prevent them from unbalancing the price of crypto assets. 3) In the case of mobile platforms, the commissions they receive for the sales of crypto-assets generated by stores within video games must be taken into account. This is a model under study that continues to be analyzed and validated through surveys of players, developers and staff of mobile platforms. Likewise, the analyses of the laws of each of different countries to validate that local regulation are not violated.

4 Conclusion

As can be seen with the expansion of networks, and especially mobile networks, the video game business has grown, adopting new architectures and models to attract and keep gamers. At the same time the players accept them and are willing to spend real money on virtual goods, accessories.

Finally, the need is seen to be able to exchange virtual goods with each other, inside or outside the game, generating profits for developers and players. Proposing as a solution to this a common market where the elements obtained in one game can be exchanged within another or through a common means of exchange.

Therefore, the study of a market continues that allows the exchange of virtual goods between different games or platforms, using a common currency between them, in which the largest number of developers and players can converge.

5 References

1. Siira, E., Annanperä, E., Simola, O., Heinonen, S., Yli-Kantola, J., & Järvinen, J. (2017). Designing and Implementing Common Market for Cross-Game Purchases between Mobile Games. In 30th Bled eConference: Digital Transformation: From Connecting Things to Transforming Our Lives, Bled 2017 (pp. 531-544).
2. Keegan, B., Ahmed, M. A., Williams, D., Srivastava, J., & Contractor, N. (2010, August). Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. In Social Computing (SocialCom), 2010 IEEE Second International Conference on (pp. 201-208). IEEE.
3. Alha, K., Koskinen, E., Paavilainen, J., Hamari, J., & Kinnunen, J. (2014). Free-to-play games: Professionals' perspectives. Proceedings of nordic DiGRA, 2014.
4. Hamari, J., Alha, K., Järvelä, S., Kivikangas, J. M., Koivisto, J., & Paavilainen, J. (2017). Why do players buy in-game content? An empirical study on concrete purchase motivations. Computers in Human Behavior, 68, 538-546.
5. Hanner, N., & Zarnekow, R. (2015, January). Purchasing behavior in free to play games: Concepts and empirical validation. In System Sciences (HICSS), 2015 48th Hawaii International Conference on (pp. 3326-3335). IEEE.
6. Hamari, J., Hanner, N., & Koivisto, J. (2017). Service quality explains why people use freemium services but not if they go premium: An empirical study in free-to-play games. International Journal of Information Management, 37(1), 1449-1459.
7. Hamari, J., & Lehdonvirta, V. (2010). Game design as marketing: How game mechanics create demand for virtual goods.
8. Wang, Q. H., & Mayer-Schonberger, V. (2010, January). The monetary value of virtual goods: An exploratory study in MMORPGs. In System Sciences (HICSS), 2010 43rd Hawaii International Conference on (pp. 1-11). IEEE.
9. Cheung, C. M., Shen, X. L., Lee, Z. W., & Chan, T. K. (2015). Promoting sales of online games through customer engagement. Electronic Commerce Research and Applications, 14(4), 241-250.
10. Hamari, J., & Keronen, L. (2016, January). Why do people buy virtual goods? A literature review. In 2016 49th Hawaii International Conference on System Sciences (HICSS) (pp. 1358-1367). IEEE.