EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 149 No. 10
October 2020

Research in Computing Science

Research in Computing Science, Año 19, Volumen 149, No. 10, octubre de 2020, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de octubre de 2020.

Research in Computing Science, year 19, Volume 149, No. 10, Octiber 2022, is published monthly by the Center for Computing Research of IPN.

Volume 149(10)

Advances in Computing Science

Abdiel Reyes Vera
José Ángel Avelar Barragán
Omar Velázquez López
Mitzy Gabriela Sánchez Sánchez
Elizabeth López Lozada
Juan Carlos Chimal Eguía (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2020

ISSN: in process

Formerly ISSNs: 1870-4069, 1665-9899.

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

Table of Contents

Page

System for Monitoring Fingerprint Force

Humberto Casares Hernández[1], Rolando Hernández Guerrero[2], Laura Ivoone Garay Jiménez[2]

[1] Universidad Autónoma de Guerrero,
Facultad de Ingeniería,
Mexico

[2] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas,
Mexico

humberto.casares98@gmail.com,
rhernandezg1307.alumno.ipn.mx,lgaray@ipn.mx

Abstract. Information management becomes a need when a system is used in a common area in the health services for several physicians and patients, as well as into the portable mobile systems for hand rehabilitation assigned to a home-care therapy. In both cases, tracking is required in seldom clinical dates for physical therapies. Hand injuries due to daily activities or work issues are among the main attended problems; therefore, a system capable of evaluating the physical therapy was created to reduce the physician's time to update the patient follow-up files. This work consists of two main parts, the management subsystem, and the measurement subsystem. It has the option of creating accounts for evaluators, and they can add the patients. The patient information is kept in the database, and the finger's force recordings created during the tests. The digital file of the patients can be consulted later by the assigned evaluator. It has a graphical interface connected to a relational database for storing the information of both evaluators and patients. On the other hand, the measuring system has an interface to visualize the information used to monitor the patient's finger force activity. It has a questionnaire that measures the severity of the patient symptoms in the current session. The patients' created digital file and the data can be consulted later by the evaluator to continue the follow-up session even if a new physician is assigned.

Keywords: Information management, hand rehabilitation, graphical programming, digital file.

1 Introduction

The hands are an essential part of the person's body with which they perform activities such as holding objects and therefore perform various activities in daily life. They are the most used part of the body when performing fine activities in daily activities and therefore tend to be easily injured [1].

The Secretary of Health of Mexico City, Dr. Armando Ahued Ortega, mentions that 26 percent of workplace accidents are in the wrists and hands. In 2016, the Network of the hospital of the Mexico City Government reported that there was a record of 1871 hospital admissions due to wrist and hand trauma [2]. Keeping track of so many public hospital treatments and dates is a time-consuming job of the few highly specialized people because it has not yet been fully migrated to specialized instruments [3].

For this reason, it was proposed to develop an auxiliary system for the evaluator when capturing and recording patients during the rehabilitation processes of both pressure and movement. Portable wireless system proposals that reported the bio instrumentation technology and defined the tracks of the hand's functionality within the processes of evaluation of the result of treatment and rehabilitation therapy are presented in [4, 5]. In the specific case of the fingerprint monitoring, a system based on FlexiForce force sensors with amplifier and low pass filter to send the signal to a NI myRIO® data acquisition board was proposed.

The voltage signal in the FlexiForce sensor is increased when the force is increased. These variations of voltage are normalized in a force scale. The board sends these signals wirelessly to a system similarly made in LabVIEW® as Ortegón does [6]. In this first case, the prototype was designed to be transportable, and the wireless was done with a highly specialized acquisition card. However, the main drawback is that each applicator has to separately apply the paper version of a QuickDash questionnaire into each test session and then concentrate the information manually after completing the registration protocol for both the records and the questionnaires during each rehabilitation therapy [5]. In the second case, the system is over the wrist, and further tests about the usability should be done. Therefore, in this work, the possibility of generating a tool to facilitate the questionnaires' management, evaluation of these results, administration of the signals recorded in the fingerprint evaluation system using LabVIEW® all together with a low-cost acquisition board is proposed.

2 Material and Methods

2.1 Design of the Clamp Device

The clamp devices are used for hand force measurements in clinical, industrial, research, and development; it has ergonomic properties, adjustable, and present flat surfaces where the force sensors are fixed. The mobile section of the device shown in Fig. 1 has two-cylinder sections at the extremes with holes to move the section with cylindrical aluminum bars attached to a fixed section. This section has four finger supports for a firm grip and where the force sensors will be placed. With the use of fixing screws, the appropriate distance between the two sections can be adjusted depending on the hand's size.

2.2 System Architecture

Before starting with the development of the system, various options for the management of information were analyzed, the areas of opportunity in the registration protocol, the rehabilitation doctors' feedback, and the possibility of maintaining the

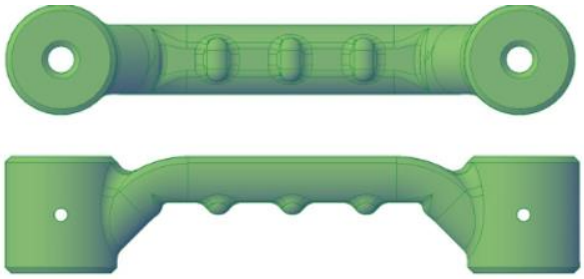


Fig. 1. Front and lateral views of the mobile system.

previous interface's graphic language were analyzed. The software used to program the system with a graphical interface was LabVIEW 2019 [7], a rapid prototype tool used in the industry to design electronics and complete systems where the handling of information acquired by hardware and software processing are integrated.

So that it can be easily used by the evaluator, preserving the possibility of data acquisition, control, analysis, and presentation using transparent methods for the end-user. NI-DAQmx 19.1 was selected for data acquisition management and MySQL for information management. In the electronic instrumentation section, the FlexiForce A201 force sensor was implemented. However, the electronic card design was modified to be in modular printed circuits for each sensor. This version could be used to implement different protocols to be evaluated according to the number of finger pressure measurements required and different measurement configurations proposed by the physicians.

2.3 Force Sensing System

The grab or "clamp" device with the sensors generates the signals from one hand either from the left or right, and the clip device only from one finger depending on the assigned configuration. Therefore, the voltage variations are generated due to the increase or decrease of force. The force measurement into the different tests is sent to the data acquisition board, stored in the patient electronic folder, and displayed in the dashboard. For this purpose, the sensor FlexiForce A201 sensor was implemented with a linear range of 0-55KgF. This sensor uses the variation of its resistance according to the inverse of the pressure exerted on the sensing area.

According to the manufacturer, a preprocessing circuit is required (see Fig. 2). It has been the base for the conditioning of the system. The circuit has been divided into two sections, an amplification circuit, and a low pass filter circuit.

2.4 Amplification Circuit

An inverse amplifier was used for the first stage, which is presented in Fig. 3. The sensors were previously characterized in a range of 0- 55 kg F (100lb) that correspond to the maximum value. It is enough to cover the hand measurements with reported maximum values around 25kgF considering age, BMI, and gender [9, 10, 11].

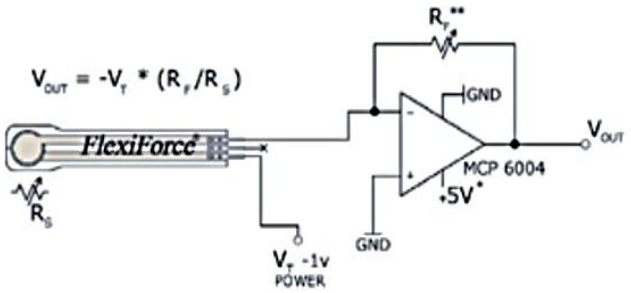


Fig. 2. Recommended circuit by the manufacturer in the datasheet [8].

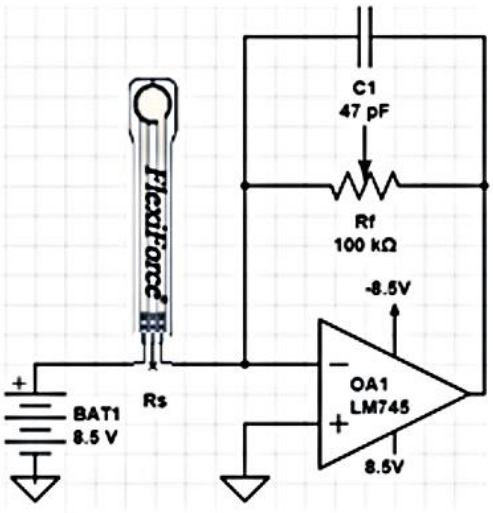


Fig. 3. Fingerprint force system amplification circuit.

The voltage provided is determined by a reference resistance Rf (sensor), so it was used a potentiometer of 100kOhms, for later adjustments and obtain the maximum value of force that can provide the sensor considering the calibration of the sensor with the maximum voltage provided. A ceramic capacitor C1 has been placed parallel to Rf to minimize voltage variations.

The LM741 operational amplifier was used in an inverter amplifier configuration. A constant dual source of ±8.5V to power the LM741 OA1 and the Rs sensor was used, and the maximum value of -7.0V was established in a circuit saturation without force applied. The gain has been calculated considering the variable resistance *Rs* of the FlexiForce sensor. *Rs* presents maximum resistance when it is not pressed, and it is diminished when a force is applied. The value of the gain is obtained from Equation 1. Therefore, the amplifier was implemented to limit the output voltage and not present differences in the sensors' readings:

$$G = -\frac{Rf}{Rs} \quad (1).$$

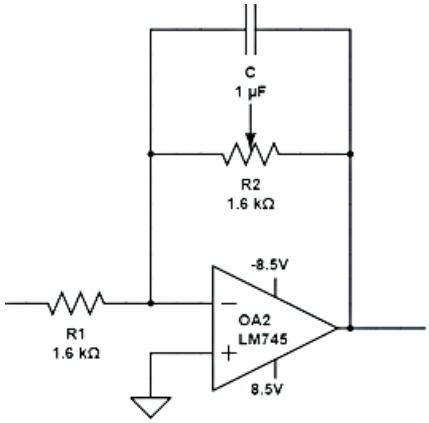


Fig. 4. Schematic of the low pass filter circuit.

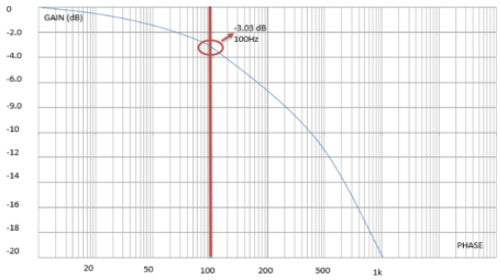


Fig. 5. Cutting the low pass filter at 100Hz from the fingerprint system.

The output voltages *Vout* has been calculated considering the reference resistance Rf. Each sensor has a different performance, so the resistance *Rf* has a potentiometer to adjust the *Vout* when the maximum force is exerted. This calibrated output voltage has been calculated from Equation 2:

$$V_{out} = -V_t * \left(\frac{R_f}{R_s}\right) \quad (2).$$

2.5 Low Pass Filter Circuit

The low pass filter was designed with a cut-off frequency of 100 Hz (see Fig. 5) to capture the full signal without interference from the disturbances present when touching the sensors or vibrations presence when the device is in use. The composition of the developed low-pass filter circuit can be seen below (see Fig. 4):

The low pass filter is a first-order inverter with a -3dB attenuation per decade from the 100Hz cut-off frequency. The component values were calculated from Equation 3:

$$f_c = \frac{1}{2\pi R_2 C} \quad (3).$$

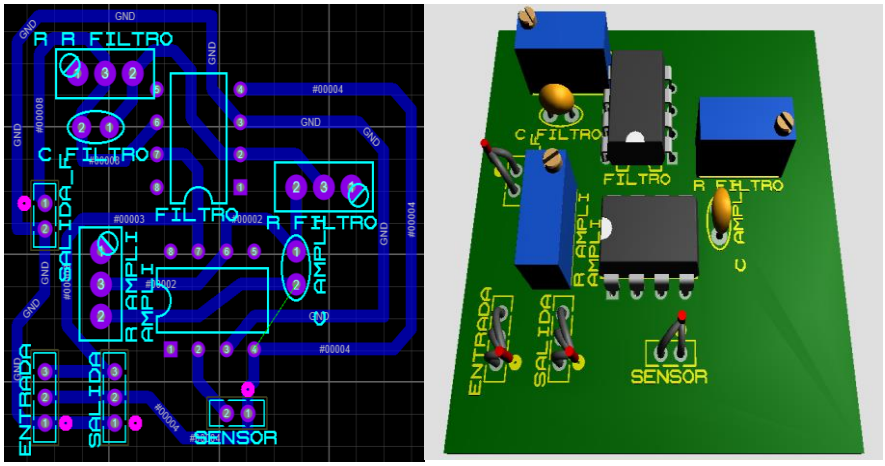


Fig. 6. Modular circuit of the fingerprint system and final circuit in 3D.

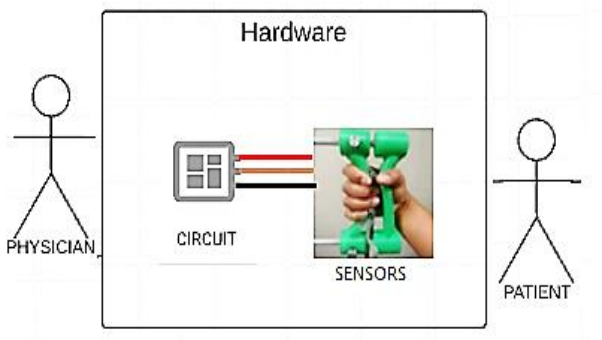


Fig. 7. System architecture for hardware.

The filter response obtained from the calculations was obtained in the implemented filter (see Fig. 4). This filter helps to eliminate electrical noise by movement and vibration.

Fig. 6 shows the circuit of the amplifier and the filter, both implemented in Proteus 8 Professional for a PCB.

Error in the measurement of 0,02% was obtained with respect to the simulation in relation to the implementation. The preliminary performance tests consisted of measuring the voltage variations associated with applying pressure to the sensor with a multi-meter STEREN MUL-600. In data acquisition, a USB-6009 board (NI, USA) [12] was used instead of the NI myRIO® because it is much more economical, light, and has the minimum required features in a therapy room with a pc or a laptop. This hardware architecture is shown below (see Fig. 7).

Finally, six modules were done, five for the grab hand test and the last for a finger's clip test. The sensors were adjusted to a range of 0-44kgF that corresponds to 80% of the maximum value available, and it is enough to cover the measurements in hand.

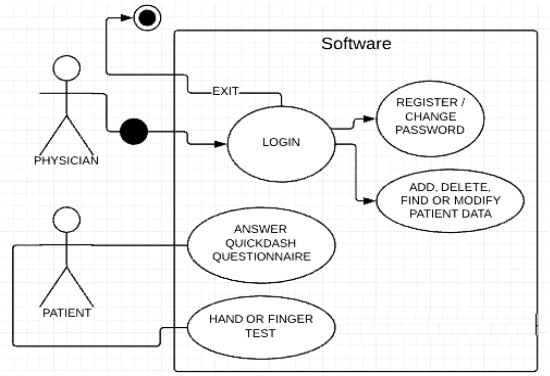


Fig. 8. System architecture for software.

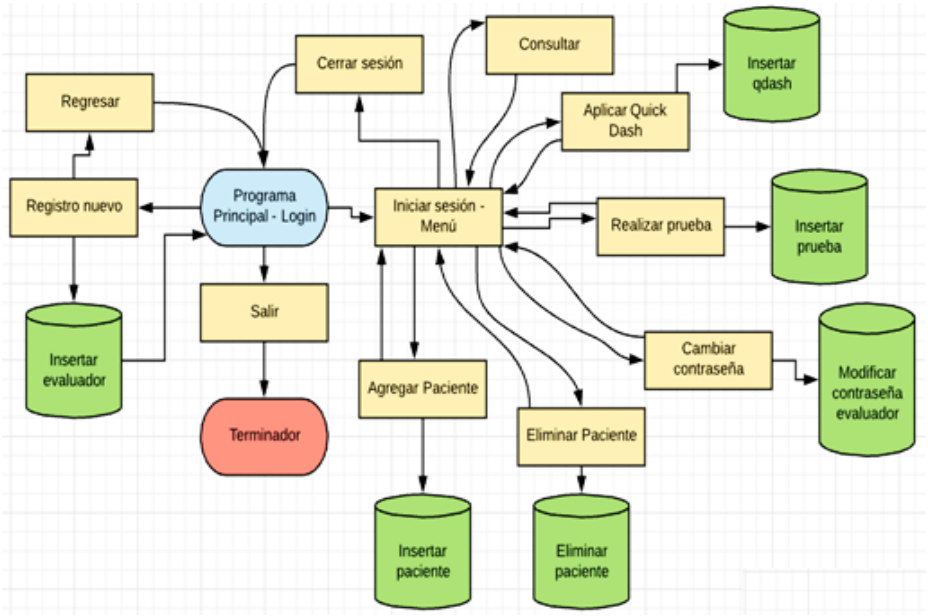


Fig. 9. The flow of processes within the management system.

2.6 Signal Acquisition

The signal pre-conditioning system design for the FlexiForce A201 sensor is retaken, and a modular version was designed based on its datasheet [8]. An upgrade version of the programming language was used considering its capabilities in handling graphical interfaces, handling questionnaires, validating responses, and modules for connecting to databases. Subsequently, the acquisition and processing programming of the obtained signals was implemented.

The software had two profiles: physicians and patients. However, the only direct end-user is the physician, with login or register of account. Once an account is created,

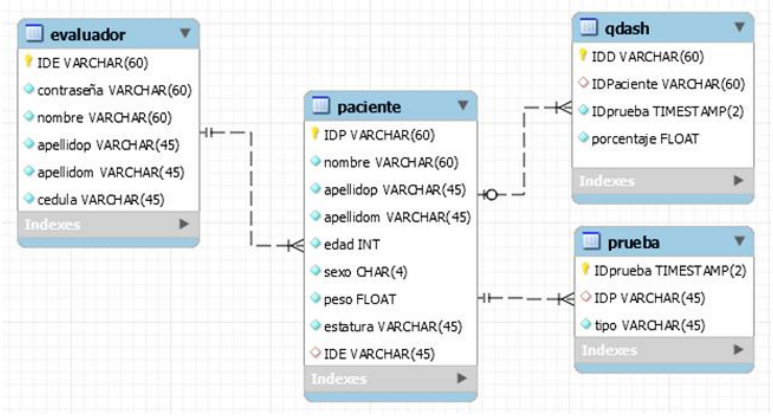


Fig. 10. Database diagram.

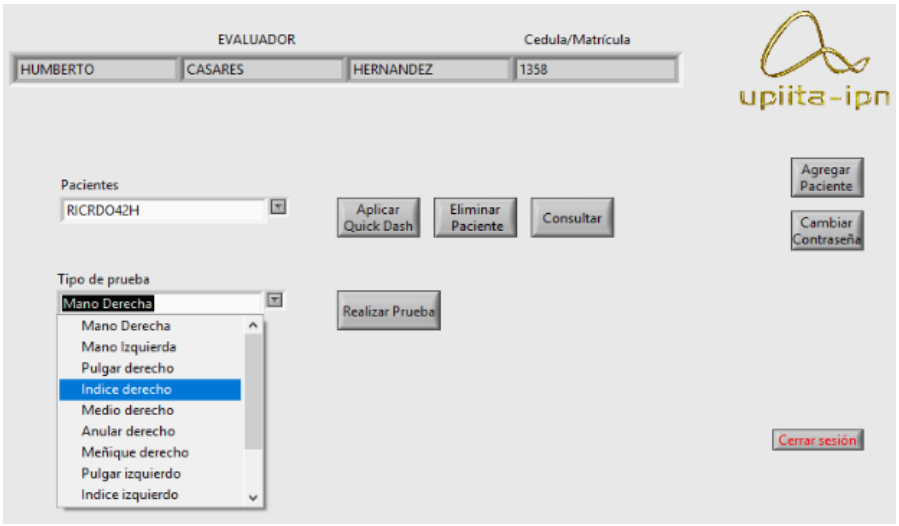


Fig. 11. Management system home screen.

in the menu, there are options to add, delete, consult, or modify patient data and change account password (see Fig. 8).

The management system for the patient data, the information from the questionnaires, and the data collected from the sensor use a database with an event-driven programming paradigm. This method consists of programs that attend events, and the functions are loaded and executed according to the selection flow. So, the program uses a menu of options and does not have a single flow of information. LabVIEW was used for the implementation of the system.

For the test module, the DAQ Assistant Object was used to measure the analog signals coming from the sensors stage. The flow diagram is shown in Fig. 9, the blue oval is the start of the program, and the red oval is the end. Each rectangle is a button and therefore has a scheduled event that could affect the database (green objects) or not.

Fig. 12. Sensor dashboard.

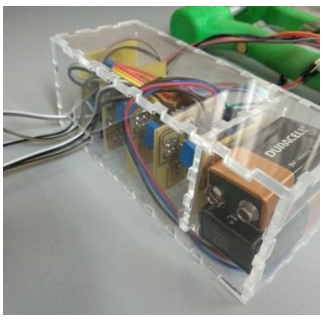


Fig. 13. System prototype version 2.

2.7 Fingerprint Database

The database implemented is shown in Fig. 10; a relational database was designed in MySQL to manipulate the information efficiently since it is a high-performance manager. It consists of 4 tables that are: evaluator, patient, test, and Qdash. These tables are used to create evaluator profiles or change their password, to add, delete, or modify patient information for their tests or Quick Dash questionnaire.

2.8 Graphical User Interface

This system considers the evaluator profile, and it can perform various functions into the patients' folder (see Fig. 11). It allows to add, delete, and consult patient information. It can also perform several tests and questionnaires for the same patient and associate them with the patient's folder.

It starts with an evaluator log in, in which he can enter or create a new account. Then, evaluator menu options are to add or remove patients, apply the digitized Quick Dash

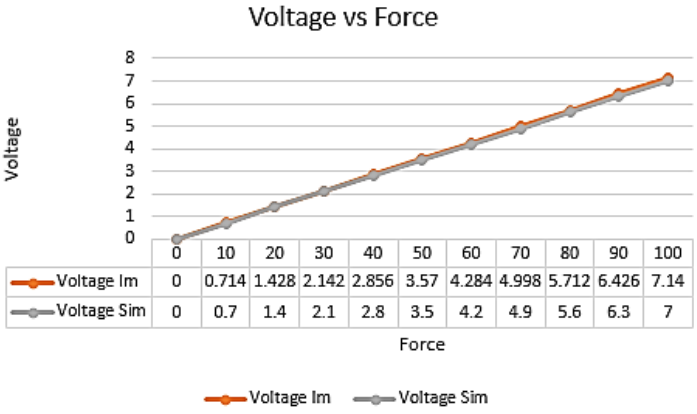


Fig. 14. Graph of voltage sensor performance against the applied force.

questionnaire, perform strength tests with real-time monitoring, and consult previously recorded results.

The tests can be of a single finger or all the fingers of a hand (see Fig. 12). The physician decides according to the diagnosis or treatment that the patient is following and his previous performance. He can reproduce previous recordings or generated new ones.

3 Results and Discussion

The sensor preprocessing stage was redesigned in a modular version, based on the one proposed by Reyes (2018). This version implements and characterizes a first-order passive low pass filter with impedance dependent amplification associated with the force sensor range with a cut-off frequency of 100 Hz [13].

The PCB boards are modular, but a case was designed to contain them in conjunction with a grip grab mechanism generated by [4].

In a clip grasp test, one PCB board was used for a single sensor. The thumb must press down; in the others, it is supported by the thumb, such as the index, middle, ring, or little finger.

In a claw grasp test, the five modular circuits are encapsulated in a box with their batteries, input, and output slots, as shown in Fig.13.

With the developed system, results have been obtained in simulation and a laboratory test of the system. The force results were accorded with the voltage obtained from the circuit, calibrating it to a voltage of 7V output for a 100kgF in the simulation and recording versions. The final maximum output voltage in the implementation was 7.14V.

The relationship between voltage and applied force was generated by the output voltage values associated with the applied force values. Therefore, the voltage values are directly proportional to the force applied to the sensor presenting linearity on the full scale of the study with an error < ±2.5% in the last part of the range of the sensor [14].

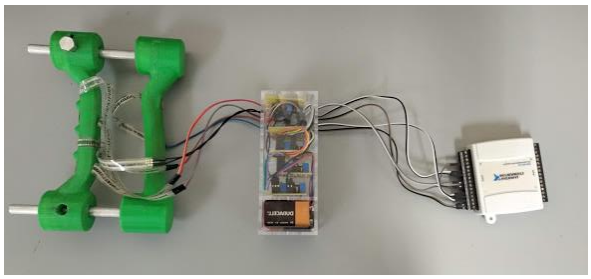


Fig. 15. Final fingerprint measurement system.

Fig. 16. A graphical user interface to consult a patient's results.

This relation of values is shown (see Fig 14), in which the results of the simulation can be observed with respect to the values of the implementation. Voltage Im belongs to the implementation, while Voltage Sim belongs to the simulation.

The event programming paradigm requires to define which event will be performed by each button and which actions are performed depending on each active window. Data flow languages like LabVIEW allow automatic parallelization of processes in programming. Unlike sequential languages like C and C++, graphical programs inherently contain information about which parts of the code they are supposed to run in parallel. The final version of the software has been tested and used in the laboratory with a no reported problem.

This system is intended for rehabilitators who can perform the force measurement and control those measurements (see Fig. 15). The hand test sensors are integrated into a device that must be pressed like a clamp; the force applied to each sensor is acquired by the software and saved in a TDMS file.

One of the project's primary purposes is to optimize the processes for the evaluators in the follow-up of the rehabilitation of the hand. Figure 16 shows the window for consulting the data of a patient. The results of the Quick Dash questionnaires are displayed as a list with date, time, and percentage in the blue box. The red box shows a graph in which files generated from the finger or hand tests can be loaded; dates/times

and the force applied to the sensor in that time interval is included. The information shown here is thanks to the database; the timestamps are saved from having better feedback on the patient for the questionnaires and hand tests. This information helps the evaluator or even another evaluator to follow up on the patient for rehabilitation.

This system has similar functionality that [4, 5]proposes, but this version can be in a deployable version. The modular approach and minimum components make this system very versatile, light, and robust. This system could be used independently of the characteristics of the user and which hand is testing.

4 Conclusions

This system can be used as a monitor in the rehabilitation process of the hand, for carrying out the force tests such as grab and clip test and reduce the physician time need to manage the electronic file of each patient in a standard rehabilitation room, giving him the possibility to associate the questionnaires, and force recordings into the session report of each of his patients, with the possibility to quickly recover the previous results from the database, during the following sessions of the therapy. The use of a low-cost data acquisition board is justified because the system only needs five analog inputs to process the signals from the sensors to a computer in this version of the system.

An advantage of using a database is the scalability of adding new functions to the system. Nowadays, the database is local, and a future job could be to run it from a web server to be manipulated from various devices and even from other rooms. In that case, the administrator's role should be included and an option to eliminate the evaluator's profile while avoiding inconsistency because there could be patients without foreign keys. In this system version, deleting the profile is not considered, but a function to assign a new evaluator could be added for patients that require a new evaluator.

The use of this system in a room of rehabilitation with several physicians along a labor day, with the possibility of recovering the information for the previous session just with his session login, could increase the following up into a moving health service to e-health. This system explores a variant of the previous version reported in [4] with improvements in managing the information and integrating the processing in a single platform. Moreover, this system could be used for force tests into other protocols because of the modularity, not necessarily hand injuries. The experimental phase was not thorough in terms of the actual version. But the concept of the system was fulfilled tested in a pilot test in [4]; several technical improvements were made, and this version could be used in the clinic because it already has roles and security access.

References

1. Latarjet, M., Ruiz-Liard, A.: Anatomía humana. Editorial Médica Panamericana (2009)
2. Secretaria de Salud: 26 por ciento de accidentes laborales, son en muñecas y manos: Armando Ahued. https://salud.cdmx.gob.mx/comunicacion/nota/26-por-ciento-de-accidentes-laborales-son-en-munecas-y-manos-armando-ahued (2017)
3. La razon: Mancera solicita digitalizar expedientes clínicos a nivel federal. https://razon.com.mx/ciudad/mancera-solicita-digitalizar-expedientes-clinicos-a-nivel-federal/ (2016)

4. Correa, F.R.M: Sistema de seguimiento dinámico de parámetros clínicos asociados a la rehabilitación de mano. Institulo Politecnico Nacional (2018)
5. Ortegón, I., Alba-Baena, N., Martínez, E., Ñeco-Caberta, R., Quezada-Carreón, A.: Prototipo inalámbrico para la medición de la fuerza de la mano. Ortegón, Cultura Científica y Tecnológica (CULCyT) (2016)
6. Hoang-Kim, A., Pegreffi, F., Moroni, A., Ladd, A.: Measuring wrist and hand function: Common scales and checklists. Injury, 42(3), pp. 253–258 (2011)
7. Ni.com: LabVIEW 2020-NI. https://ni.com/es-mx/shop/labview/labview-details. html (2020)
8. Tekscan: FlexiForce A201 Datasheet Tekscan. https://tekscan.com/resources/ datasheets-guides/flexiforce-a201-datasheet (2018)
9. Escalona, P.D'A., Naranjo-Olguin, J., Salamó, L.: Parámetros de normalidad en fuerzas de prensión de mano en sujetos de ambos sexos de 7 a 17 años de edad. Revista Chilena de Pediatría, 80(5), pp. 435–443 (2009)
10. Rojas, J.A., Valentín-Sánchez, G., Uc-Vázquez, L.C., Datta-Banik, S.: Dinamometria de manos en estudiantes de Merida, México. Revista Chilena de Nutrición, 39(3), pp. 45–51 (2012)
11. Unal, M., Kose, O., Guler, F., Arik, H.: Hand grip strength: Age and gender stratified normative data in Anatolian population. Hand Microsurg, 6(3) (2018)
12. National Instruments: Bus-Powered Multifunction DAQ USB Device (2015)
13. Arias, S., Rogeli, P., Garay, L.I., Tovar-Corona, B.: A System for Simultaneous Finger Joints Goniometric Measurements Based on Inertial Sensors. In: IEEE Latin America Transactions 15(10), pp. 1821–1826 (2017)
14. Tekscan: FlexiForce Load/Force Sensors and Systems Tekscan (2020)

A New Methodology for Solving the Vehicle Routing Problem with Time Windows Using Weighted Heuristics

Fernando Isunza Fonseca, Diego Villarreal Gutiérrez,
Santiago Enrique Conant Pablos

Instituto Tecnológico y de Estudios Superiores de Monterrey,
Departamento de Computación, Región Norte,
Mexico

{a01400624, a00820181}@itesm.mx, sconant@tec.mx

Abstract. VRPTW is an optimization problem that has been a focal point for many years by the scientific community, looking to create hyper-heuristics that are increasingly better to apply in real life. In this project, it is proposed a new hyper-heuristic based on probabilities and heuristics weights obtained through practice, after obtaining the ideal parameters in simulated annealing for each type of problem to attack. Starting by dividing and creating a categorization of VRPTW problems based on the number of clients and their distribution in space, and then dividing the problems into sections and giving weight to each heuristic in each section to improve the performance of these hyper-heuristics against other hyper-heuristics known as random, adaptive or probabilistic. It was observed that there is a behavior where this method improves compared to other hyper-heuristics when there are more clients. If it continues to develop, the optimal method for Big Data problems can be obtained.

Keywords: Hyper-heuristics, vehicle routing problem, time windows, heuristics, methodology.

1 Introduction

The Vehicle Routing Problem (VRP) is one of the most studied optimization problems. The generalization of the well-known Traveling Salesman Problem (TSP) and in this problem domain, an optimal route determination process is carried out to send products needed by a group of customers.

The most VRP researches are Vehicle Routing Problem with Time Windows (VRPTW) and Capacitated Vehicle Routing Problems (CVRP) [2]. The VRPTW involves determining an optimal set of routes on identical vehicle fleets with limited capacity to deliver customer demand items within a certain period, without violating the customer's time-window constraints and the vehicle-capacity constraints. This variant is chosen since the importance of VRPTW in many distribution systems has spurred intensive research efforts for both heuristic and exact optimization approaches.

In a real application, many shipping service companies use several similar vehicles, the shipping is divided into several shifts and there is a time limit for each shift. Modeling VRPTW as a distribution system consists of the main depot and some vehicles with the same capacity, to serve many scattered customers. Each customer has a certain time limit, their request is less than the capacity of the vehicle, and each customer is visited once by a single-vehicle. Simulated annealing as an extension of the Markov Chain Monte Carlo algorithm was first presented in 1953 [5]. Simulated annealing (SA) is a probabilistic technique for approximating the global optimum of a given function. Specifically, it is a meta-heuristic to approximate global optimization in a large search space for an optimization problem [3]. It is often used when the search space is discrete (e.g., the VRP) optimization problems [4].

For many real-world problems, an exhaustive search for solutions is not a practical proposition. The search space may be far too big, or there may not even be a convenient way to enumerate the search space [2]. Hyper-heuristics may be seen as a generalization of meta-heuristics and an easier way for categorizing a large body of literature of heuristics and meta-heuristics that was rather difficult to classify previously [3]. A hyper-heuristic is a search methodology or learning mechanism to select or generate heuristics to solve a specific combinatorial problem [1].

The overall flow of algorithms is described in the following: first, an initial solution is constructed using a constructive heuristic. Then, through a learning process, the hyper-heuristic proposes a series of perturbation, constructive, and/or improvement heuristics to improve the solution. One of the main motivations for studying hyper-heuristic approaches is that they should be cheaper to implement and easier to use than problem-specific special-purpose methods and the goal is to produce good quality solutions in this more general framework. Of course, the overall aim of the a hyper-heuristic goal goes beyond meta-heuristic technology.

- Probabilistic hyper-heuristic,
- Random hyper-heuristic,
- Adaptive hyper-heuristic,
- Interroute relocate,
- Intraroute 2opt,
- Intraroute oropt,
- Intraroute relocate,
- Interroute exchange,
- Intraroute exchange,
- Interroute 2opt,
- Cross exchange,
- Geni Exchange.

2 Motivation

The search for optimizing the number of routes and the best distance has led to the discovery of new algorithms and methodologies. The motivation for finding a new methodology to solve this type of problem is the whole improvements that this will

mean in the delivery logistic area. Having the best route has ecological and financial impacts [1]. Using the best route minimizes distance and minimizes CO_2 emissions and, more importantly, minimizes the operational costs [10].

The use of hyper-heuristics to solve optimization problems is very common today. However, it is a path that has not been completed, and finding the best hyper-heuristic for a particular problem is the ideal now. For the VRPTW problem, many hyper-heuristics, meta-heuristics, and heuristics have been studied and very good results have been obtained.

The motivation is to generate a new hyper-heuristic based on probabilities established based on the input variables of the problem and the behavior of various heuristics in the resolution of the optimal final distance in a specific number of iterations. The characteristics that will be taken into account will be the number of clients in the VRPTW and their distribution in the workspace, and the behavior of the individual heuristics Seeking to test the hypothesis that for each classification based on these parameters, a personalized hyper-heuristic would work better than the hyper-heuristics already known as probabilistic, adaptive or random. This could open the doors to new ways of finding better, faster, and more efficient results not only in the VRPTW but also in other optimization problems based on the initial conditions of the problem to be optimized.

3 Problem Description

VRP is a combinatorial optimization problem that responds to the question of the best routes to deliver packages to the company's clients. It is a generalization of the well-known Traveling Salesman Problem. VRP was introduced by George Dantzig and John Ramser in 1959 [7]. It's most general description is that "N" trucks leave a warehouse where the products bought by customers scattered throughout the city are stored. The objective is to minimize the cost of distribution, selecting the trucks that must go with each client on the shortest route [9].

It is complicated when the number of clients, warehouses, and trucks varies; when truck conditions vary, such as capacities; and when there are special orders or clients or time restrictions. In this research, the Solomon databases of 25, 50, 100, and 800 customers will be used. These datasets are well known in other VRPTW researches and many results are discovered using different hyper-heuristics. The datasets have the particularity of being separated according to the spatial distribution of the clients in three categories: grouped agglomerations of clients called *Cluster distribution*, randomly distribution of the clients in the space called *Random distribution*, and something random but conditioned, that is between the other two types called *Random clustered distribution*. There are 175 different datasets with the following variables each one:

1) Number of vehicles,
2) Customer ID,
3) Coordinate x of the customer,
4) Coordinate y of the customer,
5) Demand,

6) Ready time,
7) Due date,
8) Service time.

The principal component of the code used is the Simulated Annealing algorithm with a series of different heuristics. Another methods used are multi-class classifiers using decision trees to divide and categorize the distribution of the customers in the space (*x*,*y*) in three classes: cluster, that are grouped customers; random cluster, that are less grouped customers but not in random distribution; and random distribution. Polynomial regression is also used with *Scikit Learn* python's library to create the sections when working on the problems to construct this new hyper-heuristic.

4 Methods and Procedures

The goal is to create a hyper-heuristic that features weights for each heuristic, something similar to the probabilistic hyper-heuristic, but in this new method, there is the introduction of different sections along the run for different instances. In consequence, each heuristic will have different weights in every section. To achieve this, the first step of our method is to divide the instances based on the number of clients and their distribution in space. Then the sections were chosen by analyzing the behavior of the best distance metric along with the runs, and the calculation of the weights was by analyzing the performance of each heuristic used in each section. With the weights, it was designed as a probability rule to choose randomly a heuristic in every iteration. Each step is presented below.

4.1 Creation of Data-Frame of Independent Variables from Benchmark Problems

The database used includes the following manipulable variables: Number, demand, time windows, service time, and (x,y) coordinates for all the customers and number and capacity of the vehicles. The next calculated fields were created:Sum of all *x* coordinates:

- Sum of all *y* coordinates,
- Mean of all *x* coordinates,
- Mean of all *y* coordinates,
- Distribution of customers (Cluster, Random Cluster, or Random Distribution).

Solomon's datasets are divided into types of distribution. This research verifies and creates a new hyper-heuristic method, based on the distribution of the clients in every problem. In statistical hypothesis testing, a two-sample test is an analysis performed on the data of two random samples, each independently got from a different given population. The test aims to determine whether the difference between these two populations is statistically significant or not. The first hypothesis proposed is:

H0: The data is equal, regardless of distribution and number of clients.

Table 1. Comparison between two groups of values for the final length, when the first letter of every type is the type of distribution (cluster, random cluster, and random distribution) and the number is of the customers. α=0.05.

No. Test	Type 1	Type 2	p-value
1	C25	RC25	<0.001
2	C25	R25	<0.001
3	RC25	R25	<0.001
4	C50	RC50	<0.001
5	C50	R50	<0.001
6	RC50	RC50	<0.001
7	C100	RC100	<0.001
8	C100	R100	<0.001
9	RC100	R100	<0.001
10	C800	R800	<0.001

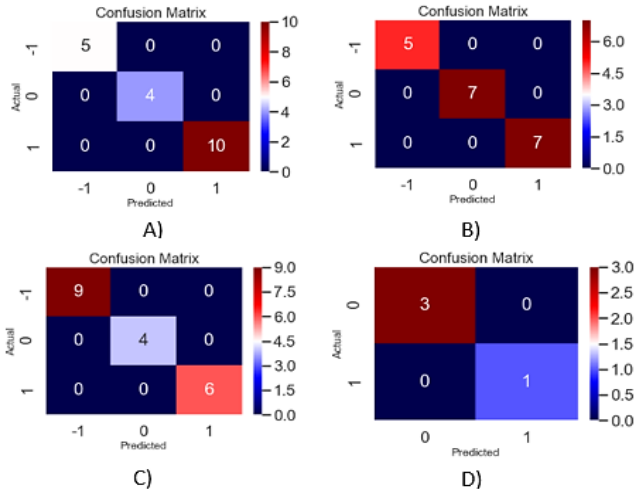


Fig. 1. Results of the test set for all types of customers. The matrix shows that the classification method succeeds in dividing for the type of distribution.

H1: A significant difference in the results exists when there are different distributions and numbers of customers.

In conclusion, the null hypothesis in every test was rejected. This means there is a statistically significant difference between the results of the problems according to their distribution and number of clients. Knowing this, it is valid to divide the problems by their distribution and number of clients. Therefore, it is possible to create the model to get a better hyper-heuristic based on the significant difference between instances. With these results, four different decision tree models were created for every class of a number of clients (25, 50, 100, and 800) to classify which type of distribution is the problem based on the spatial distribution of the clients. The results are shown figure 1.

Table 2. Sets used in the analysis. However, not all were used for every type of problem.

Set	Max Temp	Min Temp	Eq iter	Cool
C0	25	0.005	70	0.1
C1	25	0.005	90	0.3
C2	30	0.005	90	0.2
C3	40	0.005	70	0.1
C4	40	0.005	90	0.2
C5	40	0.005	110	0.3
C6	25	0.005	90	0.3
C7	30	0.005	70	0.1
C8	30	0.005	90	0.2
C9	40	0.005	70	0.1
C10	40	0.005	110	0.3
C11	50	0.005	90	0.2
C12	40	0.005	90	0.1
C13	50	0.005	90	0.1
C14	70	0.005	90	0.3
C15	70	0.005	90	0.4
C16	90	0.005	90	0.4
C17	90	0.005	90	0.5
C18	90	0.005	90	0.1
C19	100	0.005	90	0.1
C20	120	0.005	90	0.4
C21	130	0.005	90	0.3
C22	150	0.005	90	0.4

The selected models are *Decision Trees* created with *Scikit Learn Library* on Python with the default properties. With perfect precision, these models will be used to classify future datasets (different from Solomon's) to apply the resulting hyper-heuristic.

The method of Zulvia et. al. [10] was used, to select the best set of initial parameters to use in simulated annealing.

Where the Max/Min Temp is the maximum and minimum value of temperature, Eq iter is a parameter that limits the number of iterations with the same temperature value and Cool is the cooling rate of the simulated annealing.

An ANOVA one-way test was done to prove the hypothesis, the next hypothesis, and conclusion, the initial simulated annealing's parameters are statistically significant in most of the cases:

H0: The mean final length is equal regardless of the parameters used in simulated annealing.

Table 3. The best sets for each type of problem and their results.

Type	Best set	Mean length	Std Length	Min length
C25	C5	393.03	72.68	223.31
RC25	C5	729.36	145.48	315.27
R25	C5	693.76	127.27	340.18
C50	C10	997.45	258.09	410.5
RC50	C10	1489.97	347.51	762.89
R50	C10	1229.94	260.53	686.52
C100	C17	2731.44	775.15	948.18
RC100	C18	2975.81	551.34	1679.86
R100	C18	2346.67	431.73	1328.03
C800	C22	119161.85	3135.64	127208.45
R800	C22	109454.49	2864.01	102110.65

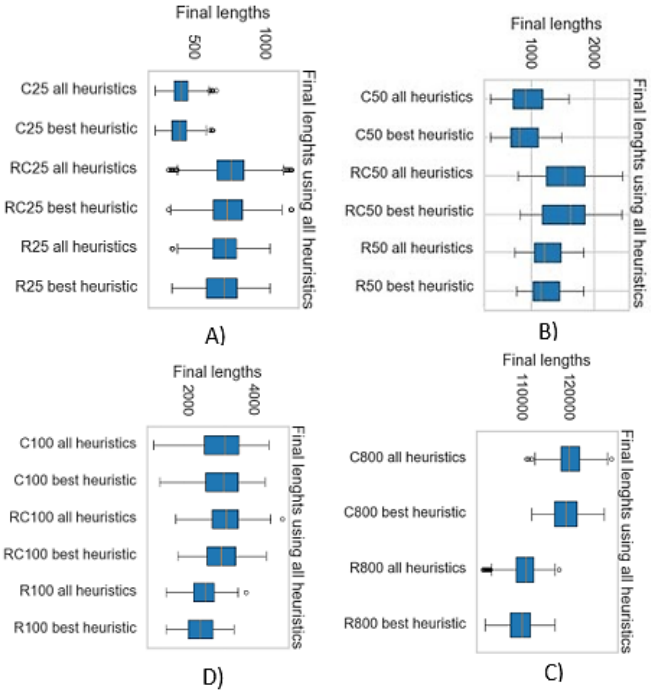


Fig. 2. Graphic comparison between one heuristic used with random parameters and with the best parameters (simulated annealing method of Zulvia [10]) for all the datasets.

H1: A significant difference in the results exists when there are different parameters in simulated annealing.

Table 4. One-way ANOVA to verify that the method improves the final length.

Problem	p-value
C25	<0.001
RC25	<0.001
R25	<0.001
C50	0.06
RC50	0.67
R50	0.15
C100	0.39
RC100	<0.001
R100	<0.001
C800	<0.001
R800	<0.001

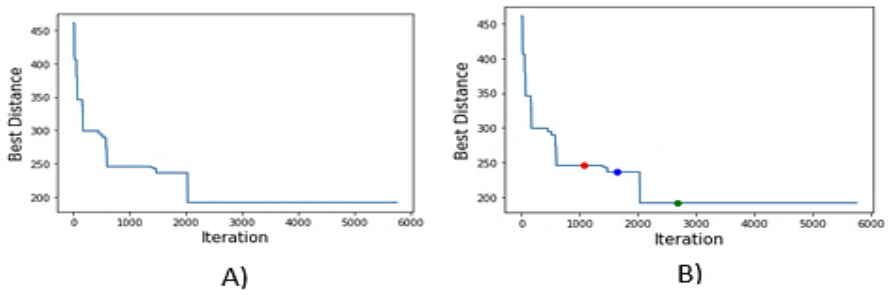


Fig. 3. The best distance as a function of iterations.

4.2 Dividing by Sections

The calculation of the length of each section was by analyzing the evolution of the best distance in every instance. Each instance was solved 30 times with the Random Saah hyper-heuristic, saving in all of them the change of the best distance. The mean evolution of the data of the best distance was calculated for each instance and then used to get the length of the sections. To achieve what has previously mentioned the following pre-programmed tools of *Sklearn Python Library* were used: *Linear Regression* and *PolynomialFeatures*. To make a polynomial regression of degree 4 [6]. This degree was selected because it was the one that obtained the best results.

Figure 9 A) shows the behavior of the best distance as a function of the iterations. To divide this function, a polynomial regression was used in different segments. The first one was the first 20 iterations and then adding intervals of 20 iterations. For each segment, the R2 was saved to know what segment fits better the regression. In other words, to know which group of iterations have similar behavioral changes at the best

Table 7. Weights for one section on instances S25.

Problem	Heuristic	Weight	Weight normalized	Cumulative Sum	Segments
C25	Intraroute_oropt	0.021467578	0.02854978	0.02854978	0 - 0.0285
C25	Intraroute_2-opt	0.066586257	0.088553211	0.117102991	0.02854 - 0.1171
C25	Intraroute_relocate	0.186704936	0.248299308	0.365402299	0.1171 -0 .3654
C25	Interroute_2-opt	0.065070949	0.086538	0.451940299	0.3654 -0 .4519
C25	Interroute_relocate2	0.191707086	0.254951678	0.706891977	0.45194 - .07068
C25	Interroute_exchange	0.179467234	0.238673872	0.945565849	0.706 -0 .9455
C25	Cross-exchange	0.03497635	0.046515126	0.992080974	0.9455 - 0.992
C25	GENI-exchange	0.005954592	0.007919026	1	0.992 - 1

distance. The first group went from 0-20, the second one from 0-40, and so on. This happened until half of the total iterations were reached (to ensure three sections minimum). In all the segments, the iteration where the section with the best R2 ends was saved.

With it, the search is started using this as the initial point, and the rest of the search range continues through the following iterations. For the previous action to stop, one of the following two conditions had to be fulfilled: the code reached the maximum capacity of sections (6) wanted or the search range remaining represented less than 10% of the total iterations.

Figure 9 B) shows an example of how the divided sections look. The first section goes from iteration 1 to the redpoint, the second from the red to the blue point, the third between points blue and green, and finally, the fourth from the green to the end of the iterations. These steps were repeated for each instance and then, their mean sections were calculated. The final results are presented below.

4.3 Weights

To calculate the weights, each problem was solved 30 times using only one heuristic, while analyzing their performance individually. In each section, the number of changes that were feasible, improvements, or best was saved for each heuristic. Then, using these numbers it was calculated the percentage of the total iterations that they represented in each section. To formulate the next equation, to assign the weights: P=X1 pf + X2 pi +X3 PB [8].

Where pf, pi, and PB represented the percentage of the feasible, the improvement, and the best changes, and Xi represented an "importance" for each percentage.

This importance was introduced to assure more heaviness to the improvement changes than the other two. One reason for this is that all the improved changes are feasible but not all the feasible ones improve the distance. The final coefficients Xi chosen were: 0.6 to the improvement changes, 0.15 to the feasible changes, and 0.25 to the best. Making the summary X1+X2+X3=1. As it can be seen in the formulas, if all the changes of a heuristic were feasible, improved, and best it corresponding weight would be 1 in that section. Making all weights to be between 0 and 1 [10]. Table 6 shows the mean weights for the instances with 25 customers.

Table 5. Sections for each type of problem.

Problem	Sections	# Sections
C25	[228,647,853]	4
R25	[101, 262]	3
RC25	[242, 656, 863]	4
C50	[297, 682, 859]	4
R50	[302, 851]	3
RC50	[348, 701, 844]	4
C100	[533, 1319]	3
R100	[400, 949]	3
RC100	[518, 1320]	3

Table 6. Weights for a section on instances S25.

Problems	Heuristic	Weights
C25	Intraroute_oropt	[0.0214, 0.0133, 0.0109, 0.0111]
C25	Intraroute_2-opt	[0.0665, 0.0399, 0.0294, 0.0286]
C25	Intraroute_relocate	[0.186, 0.098, 0.0751, 0.0747]
C25	Interroute_2-opt	[0.065, 0.043, 0.0322, 0.0319]
C25	Interroute_relocate2	[0.1917, 0.0848, 0.065, 0.0638]
C25	Interroute_exchange	[0.17946, 0.10069, 0.0903, 0.089]
C25	Cross-exchange	[0.0349, 0.02611, 0.01996 0.019766]
C25	GENI-exchange	[0.0059, 0.0014769 0.001057, 0.00102]
R25	Intraroute_oropt	[0.0439, 0.0268, 0.0234]
R25	Intraroute_2-opt	[0.0951, 0.0583, 0.0502]
R25	Intraroute_relocate	[0.212, 0.1147, 0.0962]
R25	Interroute_2-opt	[0.0941, 0.0726, 0.0641]
R25	Interroute_relocate2	[0.2039, 0.12, 0.1004]
R25	Interroute_exchange	[0.185, 0.115, 0.1047]
R25	Cross-exchange	[0.05778, 0.04375, 0.0406]
R25	GENI-exchange	[0.0209, 0.01133, 0.00985]
RC25	Intraroute_oropt	[0.035, 0.02188, 0.01754, 0.018]
RC25	Intraroute_2-opt	[0.0964, 0.05779, 0.0433, 0.0422]
RC25	Intraroute_relocate	[0.2348, 0.12873, 0.09639, 0.09412]
RC25	Interroute_2-opt	[0.0826, 0.0635, 0.055, 0.055]
RC25	Interroute_relocate2	[0.17152, 0.082, 0.06675, 0.0644]
RC25	Interroute_exchange	[0.1588, 0.098762, 0.09, 0.09]
RC25	Cross-exchange	[0.041, 0.0319, 0.02819, 0.0269]
RC25	GENI-exchange	[0.0139, 0.00776, 0.0064, 0.00608]

Table 8. Probabilities for each heuristic on *Probabilistic hyper-heuristic.*

Heuristic	Probability
Intraroute_oropt	0.04
Intraroute_2-opt	0.11
Intraroute_relocate	0.3
Interroute_2-opt	0.06
Interroute_relocate2	0.23
Interroute_exchange	0.2
Cross-exchange	0.03
GENI-exchange	0.03

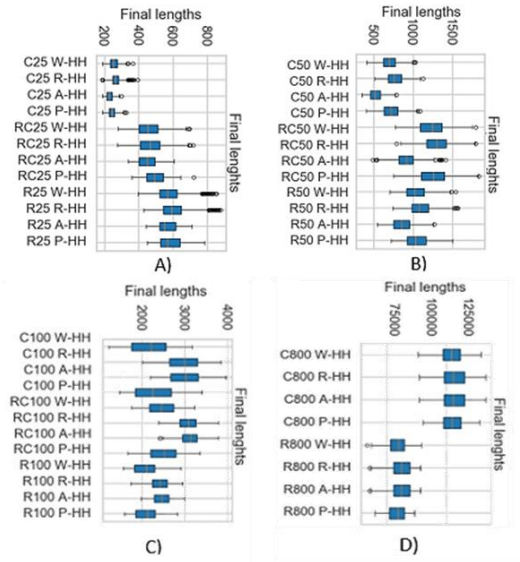


Fig. 4. Graphic comparison between the four hyper-heuristics, weighted, random choice of heuristics, adaptive, and probabilistic HH.

4.4 Probability Rule

For the selection of one heuristic in each iteration using the weights, the next step was the design of a probability rule. To achieve that the first step was to normalize the weights of every heuristics in each section, making the summary equal to 1. With the values normalized, it was used an accumulative sum to assign segments from 0 to 1, for each heuristics [13].

The first heuristic "Intra route_oropt" had the segment from 0 to the value of its weight normalized, the second one "intraroute_2-opt" had the segment between the values of the past normalized weight to this value, plus its correspondent normalized weight.

Table 9. One way ANOVA to verify that the method of the new hyper-heuristic is different from the other results in the mean of the final length and improves the final results.

Problem	p-value
C25	<0.001
RC25	<0.001
R25	<0.001
C50	<0.001
RC50	<0.001
R50	<0.001
C100	<0.001
RC100	<0.001
R100	<0.001
C800	0.0379
R800	0.0379

Table 10. Comparison of means of final lengths between methods for 25 customers. In bold the best performance of an hyper-heuristic for every type of problem.

Problem	Method	Mean Length
C25	Weighted hyper-heuristic	257.94
C25	Random hyper-heuristic	265.23
C25	Adaptive hyper-heuristic	**237.86**
C25	Probabilistic hyper-heuristic	261.78
RC25	Weighted hyper-heuristic	455.92
RC25	Random hyper-heuristic	469.60
RC25	Adaptive hyper-heuristic	**400.31**
RC25	Probabilistic hyper-heuristic	463.85
R25	Weighted hyper-heuristic	578.66
R25	Random hyper-heuristic	513.89
R25	Adaptive hyper-heuristic	**472.58**
R25	Probabilistic hyper-heuristic	582.99

Therefore, the third one was between the values of the cumulative sum and the cumulative sum, plus its correspondent normalized weight and so on. It will continue until the sum is equal to one and all the heuristics have a segment assigned. As we can infer from the procedure, if the normalized weight of a heuristics is near or equal to 0, then the probability for it to be chosen is also close to 0.

With the assigned segment from 0 and 1 to each heuristic, a random number between 0 and 1 was generated in each iteration. The random number was created using the pre-programmed python library random. Then, using this number a search was made for the segment where it falls into, thus a selection of that specific heuristic in that iteration.

The weights were actualized when the number of iterations entered the next section. An example of how does the segments look can be seen in the table 7.

Table 11. Comparison of means of final lengths between methods for 50 customers. In bold the best performance of an hyper-heuristic for every type of problem.

Problem	Method	Mean Length
C50	Weighted hyper-heuristic	703.71
C50	Random hyper-heuristic	774.52
C50	Adaptive hyper-heuristic	**528.51**
C50	Probabilistic hyper-heuristic	728.79
RC50	Weighted hyper-heuristic	1239.94
RC50	Random hyper-heuristic	1298.94
RC50	Adaptive hyper-heuristic	**917.74**
RC50	Probabilistic hyper-heuristic	1257.34
R50	Weighted hyper-heuristic	1046.7
R50	Random hyper-heuristic	1096.32
R50	Adaptive hyper-heuristic	**860.68**
R50	Probabilistic hyper-heuristic	1055.24

Table 12. Comparison of means of final lengths between methods for 100 customers. In bold the best performance of an hyper-heuristic for every type of problem.

Problem	Method	Mean Length
C100	Weighted hyper-heuristic	**2173.30**
C100	Random hyper-heuristic	2989.73
C100	Adaptive hyper-heuristic	3005.65
C100	Probabilistic hyper-heuristic	2282.08
RC100	Weighted hyper-heuristic	**2462.53**
RC100	Random hyper-heuristic	3071.33
RC100	Adaptive hyper-heuristic	3124.85
RC100	Probabilistic hyper-heuristic	2506.17
R100	Weighted hyper-heuristic	**2099.30**
R100	Random hyper-heuristic	2425.99
R100	Adaptive hyper-heuristic	2470.17
R100	Probabilistic hyper-heuristic	2115.83

5 Results and Discussion

Four hyper-heuristics are being compared: Probabilistic hyper-heuristic, Random hyper-heuristic, and Adaptive hyper-heuristic, and the other is the one proposed in this paper, the Weighted hyper-heuristic.

All of them work with the meta-heuristic algorithm: Simulated Annealing. The *Random hyper-heuristic* selects a random heuristic in each iteration, all the heuristics have the same probability of being chosen along the run. The *Probabilistic hyper-heuristic* has probabilities for each heuristics previously assigned by the programmer

Table 13. Comparison of means of final lengths between methods for 800 customers. In bold the best performance of an hyper-heuristic for every type of problem.

Problem	Method	Mean Length
C800	Weighted hyper-heuristic	**121946.45**
C800	Random hyper-heuristic	126614.63
C800	Adaptive hyper-heuristic	122837.43
C800	Probabilistic hyper-heuristic	122160.83
R800	Weighted hyper-heuristic	103416.00
R800	Random hyper-heuristic	104625.47
R800	Adaptive hyper-heuristic	104625.47
R800	Probabilistic hyper-heuristic	**103388.03**

uniform along the run. The ones assigned were the following. The last one is an *Adaptive hyper-heuristic*. It gives the same probability of being chosen to all the heuristics. As the iterations pass these probabilities are actualized.

Giving more probability of being chosen to the ones that had better performance in a particular run. The results for 100 runs per instance, with the initial simulated annealing parameters given in table 3, with the same limit of iterations (1080) per problem, are presented below in Figure 4 and tables 10, 11, 12, and 13.

Here is another hypothesis where we assume that the mean of the final distance in each hyper-heuristic differs from the others:

> **H0:** The mean of the final length is equal regardless of the hyper-heuristic used.
>
> **H1:** A significant difference in the results exists when different hyper-heuristics are used.

5.1 Results for all Type of Instances

In tables 10-13 it is shown that the *Weighted hyper-heuristic* proposed in this research, scored second place when comparing the mean final distance for cases C25 and RC25 and third place in random distribution type behind *Random Hyper-heuristic* and *Adaptive hyper-heuristic*. It is only surpassed by the *Adaptive hyper-heuristic* in all the 50 clients' classes and outperformed the rest.

However, when it surpassed 100 clients, the *Weighted Hyper-heuristic* is superior in efficiency compared to the other hyper-heuristics in this report, except for the case of R800 where it ranks second. Weighted hyper-heuristics performed well and was consistent, as it always came in first or second place in every type of problem except for one occasion. Its performance increases with a higher number of clients.

6 Conclusion

Analyzing the obtained results, it is concluded that using the *Adaptive hyper-heuristic* is better in an instance, with a small number of customers. As the cities increase, the

random and adaptive methods are less efficient. In this case, having a rule of probability or a weighted method is better. The advantage of the weighted method is that, despite the number of customers, its efficiency is constant. The *weighted hyper-heuristic* improves when the number of customers grows, as it is observed in the results, and when using statistical tests, the majority of cases are seen as better than the adaptive, random, or probabilistic hyper-heuristics.

It is important to analyze what makes each method strong and in what conditions for future approaches in order to take advantage of the potential of the weighted hyper-heuristic and improve the results obtained. The hypothesis of the hyper-heuristic has to be proved through the improvement of growth in its number of customers, the continuation of experimentation, and being able to obtain a non-supervised way to divide the problems into sections. Further implementation of the hyper-heuristic will result in better and more accurate through the law of large numbers (LLN).

References

1. Ahmed, L., Mumford, C., Kheiri, A.: Solving urban transit route design problem using selection hyper-heuristics. European Journal of Operational Research, 274(2), pp. 545−559 (2018)
2. FoodData: Agricultural Research Service. U.S.D.A. National Nutrient Data base for Standard Reference. http://ndb.nal.usda.gov/ndb/search/list (2019)
3. Caric, T., Gold, H.: Vehicle routing problem (2008)
4. Doerr, B., Lissovoi, A., Oliveto, P.S., Warwicker, J.A.: On the runtime analysis of selection hyper-heuristics with adaptive learning periods. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1015−1022 (2008)
5. Garza-Santisteban, F., Sánchez-Pámanes, R., Puente-Rodríguez, L.A., Amaya, I., Ortiz-Bayliss, J.C., Conant-Pablos, S.E., Terashima-Marín, H.: A simulated annealing hyper-heuristic for job shop scheduling problems. In: IEEE Congress on Evolutionary Com-putation (CEC), pp. 57−64 (2019)
6. Gutiérrez-Rodríguez, A.E., Conant-Pablos, S.E., Ortiz-Bayliss, J.C., Terashima-Marín, H.: Selecting meta-heuristics for solving vehicle routing problems with time windows via meta-learning. Expert Systems with Applications, 118, pp. 470−481 (2019)
7. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science, 220(4598), pp. 671−680 (1983)
8. Masand, B.M., Smith, S.J.: U.S. Patent No. 5, 251,131. DC: U.S. Patent and Trademark Office (2019)
9. Max, T.A., Burkhart, H.E.: Segmented polynomial regression applied to taper equations. Forest Science, 22(3), pp. 283−289 (1976)
10. Pillay, N. (2016). A review of hyper-heuristics for educational timetabling. Annals of Operations Research, 239(1), pp. 3−38 (2014)
11. Rembold, C.M., Watson, D.: Posttest probability calculation by weights: A simple form of Bayes' theorem. Annals of internal medicine, 108(1), pp. 115−120 (1988)
12. Tan, C.C.R., Beasley, J.E.: A heuristic algorithm for the period vehicle routing problem. Omega, 12(5), pp. 497−504 (1984)

13. van Laarhoven, P.J., Aarts, E.H.: Simulated annealing: Theory and applications. pp. 7−15 (1987)
14. Zulvia, F.E., Kuo, R.J., Nugroho, D.Y.: A many-objective gradient evolution algorithm for solving a green vehicle routing problem with time windows and time dependency for perishable products. Journal of Cleaner Production, 242 (2019)

App to Reinforce Mathematical Neurocognitive Skills During and After the COVID-19 Pandemic

Griselda Cortés Barrera, Ruth Anel Gutiérrez González,
Francisco Jacob Ávila Camacho, Mercedes Flores Flores

Tecnológico de Estudios Superiores de Ecatepec
Mexico

{gcortes, 201711669, fjacobavila,
mflores}@tese.edu.mx

Abstract. Students have returned home, and classes have switched to online instruction, because of The Epidemiological Contingency. which encourages the search for educational alternatives and strengthening teaching-learning methods. The country has made various efforts to make education relevant (mainly in mathematics). In addition, various government institutions in Mexico and the SEP have been concerned with reinforcing the teaching-learning process, a situation that motivates researchers from various areas (Psychology, Neuroscience, Pediatrics, among others) to seek strategies to strengthen children's development of calculus. mathematical (if the brain structures are mature). An important characteristic of the brain is neuronal plasticity. Therefore, the present work converges in the development of a mobile application, based on the Educational Model and the mandatory curricular proposal of the Secretary of Public Education (SEP) that contributes to the functional regeneration of nerve cells in children up to 6 years of age, achieved sensory stimulation to reinforce mathematical epistemic thinking; thus keeping neurons active, preventing the retention capacity from decreasing with age; in addition to guaranteeing Neuroplasticity, due to the implementation of techniques to stimulate and reinforce cognitive abilities (attention, perception and memory). The results obtained in the 1st and 2nd phase of tests increased by 8%, 6% and 5% for 1st, 2nd, and 3rd students; respectively.

Keywords: Artificial intelligence, cognitive neuroscience**,** mathematical cognitve skills.

1 Introduction

In Mexico, traditional basic education did not imply the use of information technologies 100%. In this century, and with the arrival of the COVID-19 pandemic, the world had to evolve [1], face new challenges, and seek new alternatives to guarantee continuity in education [2]. The way of getting an education changed dramatically. This means that the current pedagogical approach focuses on transmitting knowledge, but no longer with the approach where memorization was privileged, but by different methods where

it is contemplated: a) that information increases every day, b) discernment is generated through an unprecedented speed and that c) technology is changing dramatically.

Thus, the inclusive educational revolution begins, with the evaluation and updating of the National Educational Model [3] now known as “The New Mexican School”, that guides pedagogical practice and can serve as a reference for parents and students about the training to be achieved. The idea is that the students are instructed with quality and in a comprehensive way, generating the bases to function in their environment.

Researchers, education specialists in various sciences [4, 5] (Early Childhood Education Collective y TIC, 2014), Mexico and other institutions (SEP, UNESCO, Nokia) to mention a few; support the premise "Children First", and in compliance with the curricular plan proposed in [3], contribute to the didactic strengthening and promote the use of technological advances; helping to perfect teaching, teaching practice, as well as improve quality around the world [6]. Getting students to participate actively, autonomously in their cognition and motivation under the key competence of "learning to learn".

Recent research [7, 8, 9] has dabbled in this approach. Promoting Neuroscientific Pedagogical and Cognitive Training. They also affirm that the human can understand and remember through the continuous reproduction of patterns obtained through external stimuli, generated by memories or experience. In addition, the work of [10] indicates that the changes made in the cerebral cortex (prefrontal and hippocampus where mentalization is based) will be able to influence long-term cognitive development.

Unfortunately, our country has not explored these lines of research. Therefore, the objective is to establish a collaborative work by relating and entering mobile technology, educational cognitive Neuroscience and the national basic-level curriculum proposed by the current Pedagogical Guide of the SEP; and develop a tool on Android to generate a repository and with it train an ANN to measure the level of mathematical knowledge based on student´s cognitive skills (attention, perception and memory) of aged 3 to 6.

This article is structured as follows: Section 2. State of the art, a general and relevant panorama of the proposal is given, at the same time the recent problems are highlighted and how they have been addressed by the authors; Section 3. Proposed method, where its phases and the main components of the method to build a mathematical neurocognitive app to measure the cognitive level through an Artificial Intelligence (ANN) technique are disclosed; Section 4. The results that demonstrate the functionality of the proposal are presented. Finally, in the last section, Conclusions of the document, with directions for future work.

2 Literature Review

The nature of educational challenges has been transforming. Until a couple of decades ago, the national educational effort was focused on literacy. Although the task of educational inclusion has yet to be completed for all population groups, it is unquestionable that today the greatest challenge is to improve the quality of education, due to this, updates to educational models have emerged focusing on three learnings

(language and communication, mathematical logical thinking, and understanding of the natural and social world) [11, 12, 13].

In 2015, the effectiveness of web semantics based on cognitive skills was investigated and it was identified that the early development of the infant begins to develop their cognitive abilities through inquiring, exploring, investigating, and discovering the world around them. They also address the issue of the emerging role of educational neuroscience in educational reform for adequate educational neuroscientist training [14].

More recent research has addressed problems with cognitive deficits in adults [15]. In 2020, they solve this problem through the development of Android apps that can be used by primary school students and based on cognitive theory to help students learn [16] and improve their cognitive skills, especially in children autistic [17].

In 2016, an evaluation model was built that helps to understand, analyze, and objectively evaluate the learning levels and student´s abilities with respect to the conceptual content and the domain of specific knowledge [18]. In the same year in the research of [19], he was interested in the difficulty of correctly solving the addition and subtraction of common fractions, by means of the algorithm or by unconventional processes. The results reveal a strong positive correlation between the cognitive variables and the operations, which means that the development of these positively catalyzes the significant learning of the mathematical operations referred to and vice versa.

Recently researchers have developed work related to education and using techniques of Artificial Intelligence (AI). In 2014, the difficulties faced by students when learning mathematics were resolved, focusing on evaluating the distance education system, designed to develop mathematical problem-solving skills, through AI and expert systems, demonstrating functionality to identify whether the level of the questions is appropriate for the student and if not, adapt to it [20]. AI has contributed research to teaching and learning processes, through the use of computerized educational software and materials, to determine the student's cognitive level and help identify weaknesses and work on them to achieve a higher level of learning [21].

Based on these investigations, the realization of this research proposal is oriented using the ANN's as a tool to measure the cognitive abilities of basic-level students.

3 Proposed Method

Currently, educational technology has been used as teaching material, this justifies the need for a proposal that integrates the strategic techniques of cognitivism in the curriculum mapping, relying on mobile technology, to reinforce the mathematics Teaching-Learning Process to basic level (Fig. 1).

A. Modeling and Simulation´s Construction

The Fig. 1 show, like neuroscience, will allow the development of didactic strategies in mathematics since it assumes the task of penetrating the structure, the human brain´s functioning, and the underlying biological´s mechanisms of cognition; in such a way that it affects the individual's behavior through sapient emotions produced by the biological neural network.

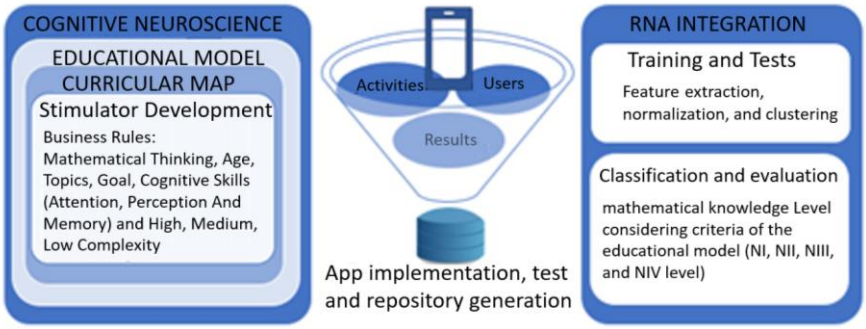


Fig. 1. Proposed method to measure the mathematical cognitive level.

Table 1. App modeling considering the axes of mathematical thinking in the Numbers topic.

General tools: Batteries and evaluation tests, graphics, images, audios, videos.
1. Represent digits with respect to the number of objects, relate to their written representation, identify between the letters, and plot in relation to the indicated direction. 2. Show collections to indicate the number of elements (greater or lesser), use movable objects to order (increasing or decreasing), remove or place as necessary; relate considering a numerical sequence and indicate its predecessor and successor. 3. Identify the equivalence relationship between the coins ($ 0.5 to $ 10), identify operators, and perform operations (addition, subtraction), reinforcing the theme of collections.

Mogollón supports this approach based on [22].

Integration of cognitive processes such as [9]:

- *Memory*: function in which the information stored in the brain is brought back to consciousness, that is, it stores and evokes assimilated content, thus allowing the execution of epistemic tasks (reasoning, understanding and problem solving).
- *Attention*: effort that individuals exert to establish themselves in a certain part of the experience such as: the ability to concentrate or stay focused on an activity, this is achieved using external means for sensory stimulation, achieving dopaminergic activation in a natural way through the motivation.
- *Perception*: mental process that transforms physical stimuli into information that passes to consciousness, forming visual environments that cause positive alterations, achieving brain plasticity.
- Integrate the educational model and knowledge receptor techniques in the stimulus generator in different topics like:

Table 2. App modeling considering the axes of mathematical thinking.

1.	Order various activities in periods (holidays, days, weeks, months, and seasons of the year).
2.	Compare lengths. Identify the order, mass quantity of objects and capacity directly or with non-conventional units.

Table 3. App modeling considering the axes of mathematical thinking.

1.	Reinforce the concepts of right, left, center, in the middle, up, down, small, medium, large, near and far.
2.	Reproduce figures and objects based on the surplus of an existing one (heart, moon, star, cross out, oval, circle, square, rectangle, and triangle). Before his acquaintance, draw and build the same things, adding colors.
3.	Reference and identify objects and classify them by some scale of measurement.

a) Table 1 show modeling of numbers axis with numbers topic (1-20), the objective is communicate, write, compare, and classify collections and identify coins.

b) Table 2 show modeling of Shape, space and measure axis with Magnitudes and measurements topic, the objective is to order the events of the day. Compare capacity direct or with unconventional drives.

c) Table 3 show modeling of Shape, space and measure axis with geometric figures topic, the objective is to establish spatial relationships and reference points. Develop geometric perception through the construction and figure´s reproduction. Compare lengths.

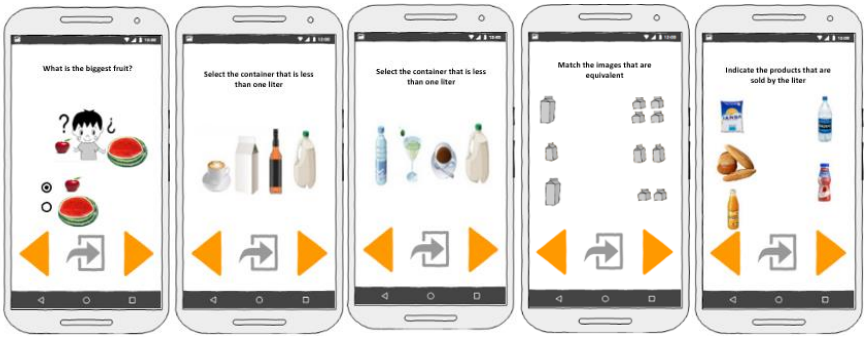
The Model proposed by the SEP contains the pedagogical approach, the reorganization of the general school system and the public policies, where the directions to follow are set out. Likewise, the curriculum foresees the benchmarks achieved in the domain of the skills expected in students, in such a way that they are perfected in their collegiate career and, through study plans and programs, contribute to the development of students [3]:

a) Guide the curriculum considering the infant's profile (digital skills and mathematical thinking)

b) Carry out the curriculum mapping in three components: key learning, curricular autonomy, personal and social development. Employ the principles and processes of the hard sciences in its main axes of the Educational Model: a) the notions of number and b) Form, space, and measure.

c) The smartphone is considered a stimulator, it will be used to motivate the learner, achieving neuronal plasticity, preparing the mind and body of infants to think critically, reflectively, autonomously; managing to retain as much information as possible.

d) Measure the knowledge acquired: here the actors intervene (the SEP, public and private schools, psychologists, neuroscientists, pedagogues, professors, parents, students, managers and businessmen; they will coexist harmoniously in the teaching process that contemplates the structure and the functioning of the



a) b) c) d) e)

Fig. 2. Measurements by a) weight, b) higher capacities, c) lower capacities. Select objects by d) equivalences e) measures and f) capacity.

mind), they contributed their knowledge and together with the different standardized assessment instruments [23, 24], it was possible to glimpse the need to estimate the percentage of success, thus:

$$x\%=ac- er / ac- om, \quad (1)$$

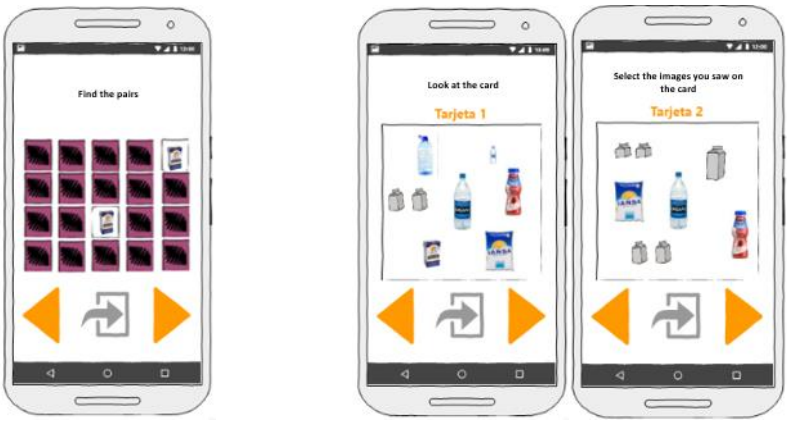
where *x%* represent the percentage obtained from the evaluated activity, *ac* the correct answers, and corresponds to the number of correct answers in each activity, *er* the number of incorrect answers and *om* the number of omitted answers. To complete the evaluation of (1), it is necessary to consider the response time in some activities: 1) of the thematic axis and 2) the level of complexity. Finally, it will be estimated with a success percentage: greater than 80 (high), greater than 60 and less than 80 and (medium) and less than 60 (low).

1) Design and construction of the stimulator and database

Each thematic axis [3] will be focused on the 3 skills, and is based on different activities that need to be evaluated and motivated to the maximum. Then, it is necessary: a) to identify the axis, b) to use the recommended instruments, achieving with them to motivate the infant in the subject to be discussed; relevant point to achieve brain plasticity, c) determine the maximum number of successful continuous reagents, to increase the degree of complexity and thus know the different prototypes that will be presented in each stage, d) to evaluate, in some cases time is decisive , e) equation (1) measures the teaching-learning process based on the ability to be reinforced manually. In this phase of the project and for an application example, the modeling of the content (magnitudes and measurements) corresponding to the axis is presented: shape, space, and measure (Fig. 2).

The objective is to assign an activity that allows making comparisons of capacity or quantity of mass that an object has, using non-conventional units of measure (Fig. 2); In order to reinforce cognitive skills, the following was considered:

- *Perception*: this is measured using evaluation batteries to ask the questions: which fruit is the heaviest? Which container has the least? and Which container that fits the least? (Fig. 2a, b and c; respectively).



(a) (b)

Fig. 3. (a) Find pairs, (b) Sample and mixed images.

- *Attention*: a series of activities are carried out, such as: relating equivalent images (Fig. 2d), selecting the objects considering their size or capacity (Fig. 2e).
- *Memory*: rely on games such as: the Memorama (Fig. 3a) or deck, the latter simulates cards where different images will be shown in a period of n seconds (assigned according to the complexity of the activity), after the time, the presents another card with new images and some of the previous ones; the infant will choose only those that were shown at the beginning (Fig. 3b) and according to the successes obtained the reinforcement will be evaluated.

The modeling of the Database (DB), see Fig. 4, contemplates the following entities: the axis contains various values: number sense, geometric figures, among others; Each of these has several associated topics: communicating and writing the first 10 numbers, comparing, matching and classifying collections based on the elements, solving subtraction and addition problems; each of these is associated with an activity, ability (attention, perception and memory) and complexity (low, medium and high).

Each registered user will have a result and details of the activity carried out, obtaining the variables: hits, errors, time, level of complexity obtained during execution. With the above, the DB is generated for the integration of the project.

Based on the analysis and design described, the application "RemaAP" available in the "Play Store" is developed and when using it, the support of parents or pedagogue is necessary for the download, installation and registration of the infant; for which, the name, sex and age are requested (Fig. 5a, b and c; respectively).

At the end of this process (Fig. 5d), it will be possible to identify and keep track of all users. To start using the app, a menu is displayed (Fig. 5e), where you will select the action to be carried out according to the ability to be reinforced (perception, attention, memory) and this is presented according to the age of the infant.

In addition, it was considered a motivating video to review the numbers in case there are still difficulties on the subject.

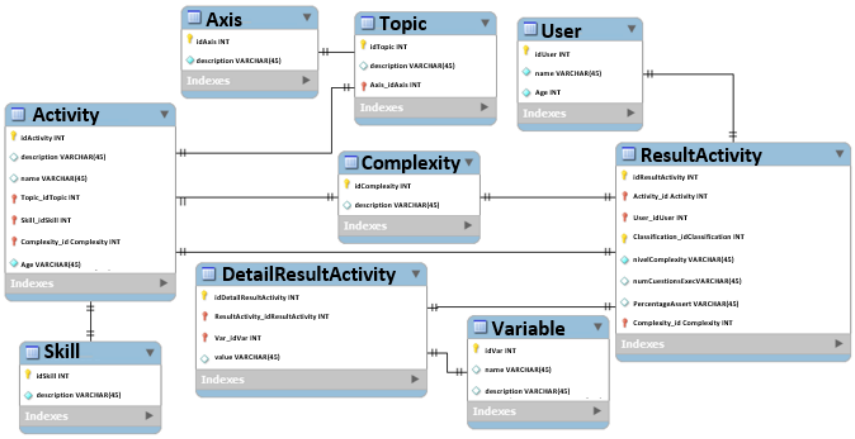


Fig. 4. Application E-R model.

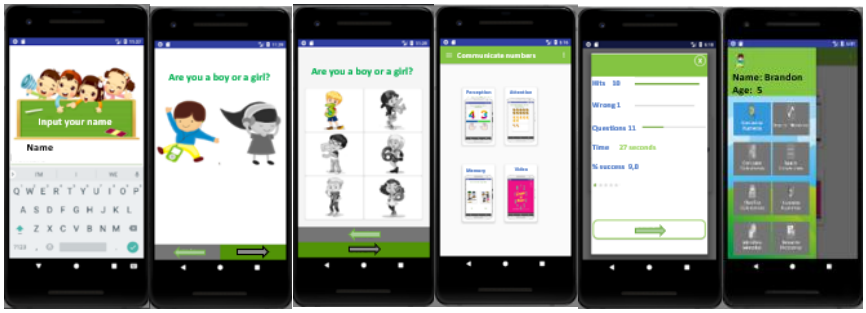


Fig. 5. Application interface (a) User registration, (b) Sex, (c) Age, (d) Complete registration, (e) Activities menu based on the 3 cognitive skills, (f) Final report, (g) Pending modules.

Once the interaction in each activity is finished, a summary is generated (Fig. 5f) of the errors, correct answers, percentage of success, among others. These are registered in the database to be analyzed.

B. Testing and Repository Generation

The tests were carried out on students of the Frida Kahlo Kindergarten with a current enrollment of 63 students in the 2018-2019 school year: 17 are 1st, 24 are 2nd and 22 are 3rd. Initially, 3 tests were carried out: two manuals (beginning and after 3 weeks) delivering 3 activities on paper and another with the app for the 3 grades; In each activity the three cognitive abilities (attention, perception and memory) are evaluated.

The tests with the application were carried out by requesting the tutors so that the students would carry out the tests using phones with the Android operating system in which the application was downloaded, for children from 3 to 6 to manipulate it, they generated 290 records in the BD.

In order for the ANN to obtain accurate results, it is necessary to perform preprocessing (debugging and normalization of the DB) to generate a repository in .csv format, for this it was necessary to carry out an analysis of the variables and records of

the DB. In addition, the most useful characteristic features were recovered: age, correct answers, errors, question´s number, success percentage, number and activity duration, executions number per user, and the mastery of the knowledge level.

The classifier attribute (proficiency level) contains the grades proposed in the SEP curriculum, the value of this criterion is obtained in the app through (1).

C. ANN for Classification by Knowledge Level

For ANN, 9 attributes were considered, one of them corresponds to the class that determines the level of knowledge (NI, NII, NIII, NIV) that the child has; and it is interpreted as mastery achieved: poor (<= 5), basic (> 5 and <= 7), satisfactory (> 7 and <= 9) and outstanding (> 9); respectively).

Once the repository is standardized, training is performed with different classifying methods (Multilayer Perceptron (MLP), Nearest Neighbor (LinearNNSearch), Radial Base Functions (RBF), Support Vector Machines SMO and Neural Network with Dendritic Processing (NNMD)) to verify that the selected attributes are adequate, as well as identify which one presents better performance and is implemented in the app.

The criteria used for each classifying method are:

- *MLP*: the backpropagation algorithm was used, the body is evaluated with a learning rate of 0.3, and applied impulse of 0.2, 500 epochs to train, a validation threshold of 20, a total of 9 hidden layers.
- *LinearNNSearch*: A Euclidean distance search algorithm and K = 1 were used for this classifier group.
- *RBF*: hybrid learning algorithm, 2 clusters (groups of clusters), with a Gaussian activation function.
- *SMO*: complexity=1.0, tolerance=0.001, Epsilon=1.0E-12, polynomial kernel=1.
- *NNMD*: proposed by [25], the data percentages for testing were: 100%, 80%, 70%, 50%, generating 94, 35, 46 and 57 hyperboxes respectively.

4 Results

The present investigation is evaluated to demonstrate its functionality. In phase one, manual tests were carried out and of the total enrollment, only 86% of the students completed them, and it was observed that 1st year surpassed 2nd and 3rd in the first skill, 2nd year students obtained the highest score in attention and the success rate of 3rd in care was 8.3 (Fig. 6). In the second manual test (Fig. 6b), it is observed that in the children of the 3 degrees the attention does not change considerably.

According to the observations made by the teachers, the 3rd graders find it difficult to follow the instructions, the 1st graders completed the exercises without problems thanks to the teachers' support and the 2nd graders were more accessible to work.

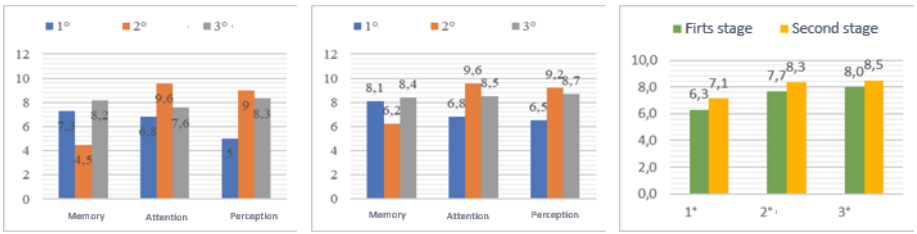


Fig. 6. Phase 1 Results: (a) First evaluation, (b) Second evaluation, (c) when using the app.

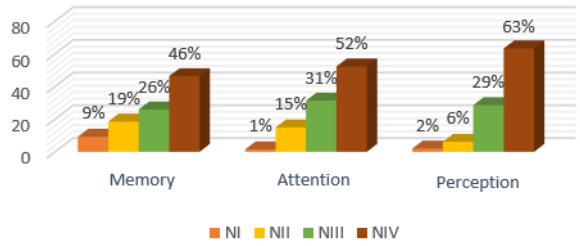


Fig. 7. Results by knowledge level and cognitive ability.

Table 2. Results of the training process with different classifier methods.

Classifier methods	Arithmetic mean	% correct classification			
		100%	**80%**	**70%**	**50%**
NNMD	99.75%	99%	100%	100%	100%
MLP	99.25%	100%	100%	99%	98%
SMO	96.25%	98%	98%	95%	94%
LinearNNSearch	96.75%	100%	98%	95%	94%
RBF	99.5%	100%	100%	100%	98%

The second phase only 68% downloaded and used the app, and compared to manual tests, its performance improved; increasing on average of the 3 skills the 0.8%, 0.7% and 1.8% in 1st, 2nd and 3rd; respectively (Fig. 6c).

So far, learning has been measured with (1), but the objective of this research is to measure the level and reinforce mathematical cognitive learning using Artificial Intelligence techniques. Therefore, the repository was evaluated with different classifying methods, using 100%, 80%, 70%, 50% of the samples for training and the remaining percentage for testing (Table 2.).

As can be seen in Table 2, the repository is well normalized, since the classification percentage of the selected methods is high, this will guarantee good precision in its implementation in the app.

Also, the level of knowledge was measured by cognitive ability, considering 290 records labeled by skill and mastery level and it was demonstrated how they were surpassing the skill obtaining until reaching an outstanding learning level (Fig. 7).

5 Conclusions and Future Perspectives

The education and comprehensive training of boys and girls will always be a transcendent factor in the world. The New Mexican School will only be a reality with the commitment and participation of all [3]. In the current context of Mexico, it is urgent to develop and apply strategies that promote the importance of complex cognitive development in students, from the basic level, so that the growth of our country is for the benefit of the whole society, and face the effects on education in Mexico by the COVID-19 pandemic.

The main contribution of this research was to achieve the interaction of Neuroscience, education, and current technology, to reinforce the basic knowledge of exact sciences, cognitive abilities (attention, perception, and memory), and infant´s skills; with this measure the mastery level of the study area through a NNMD. This work impacts public and private institutions in Mexico because the topics are focused on the current educational model proposed by the SEP.

We present a mathematical neurocognitive app that measures the level of knowledge acquired based on Artificial Intelligence techniques (DMNN). For future research, this will become a Brain-Computer Interface and by means of portable electroencephalography (EMOTIV EPOC of 14 channels), brain waves will be measured, in addition neurofeedback techniques will be implemented to stimulate the brain and with brain mapping to demonstrate that the child will reinforce learned, retaining it long term.

References

1. Khattar, A., Jain, P.R., Khurshaid-Quadri, S.M.: Effects of the disastrous pandemic covid 19 on learning styles, activities and mental health of young indian students - a machine learning approach. In: 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1190−119 (2020)
2. Alam, A.S., Lau, E., Oh, C., Chai, K.K.: An alternative laboratory assessment approach for multimedia modules in a transnational education (TNE) programme during COVID-19. In: Transnational Engineering Education using Technology (TREET), pp. 1−4 (2020)
3. Secretaría de Educación Pública: Hacia una nueva escuela mexicana. Perfiles Educativos, 41(166) (2019)
4. Toro-Carvajal, L.A., Ortíz-Álvarez, H.H., Jiménez-García, F.N., Agudelo-Calle, J.d.J.: Los sistemas cognitivos artificiales en la enseñanza de la matemática. Educación y Educadores, 15(2), pp. 167−183 (2012)
5. Borjas, M., de Castro, A., Ricardo, C., Vergara, E.: Recursos educativos digitales para la educación infantil (REDEI). Colectivo Educación Infantil y TIC 20, pp. 1−21 (2014)
6. UNESCO: http://unesdoc.unesco.org/images/0021/002196/219637s.pdf (2013)
7. Zadina, J.N.: The emerging role of educational neuroscience in education reform. Psicología Educativa, 21(2), pp. 71−77 (2015)

8. Lipina, S.J.: Introducción: Actualizaciones en neurociencia educacional. Propuesta Educativa, 46, pp. 6−13 (2016)
9. Cerda-Etchepare, G., Pérez, C.E., Romera-Felix, E.M., Casas, J.A.: Influencia de variables cognitivas y motivacionales en el rendimiento académico en matemáticas en estudiantes chilenos. Educación XXI, 20(2), pp. 365−385 (2017)
10. Carasatorrea, M., Ramírez-Amaya, V., Cintra, S.: Structural synaptic plasticity in the hippocampus induced by spatial experience and its implications in information processing. Neurología, 31(8), pp. 543−549 (2016)
11. Gob.mx: https://gob.mx/7prioridadessep/articulos/4-modelo-educativo-ypropuesta - curricular (2016)
12. Diario Oficial de la Federación: Programa sectorial de educación 2013-2018. http://dof.gob.mx/nota_detalle.php?codigo=5326568&fecha=13/12/2013 (2013)
13. Gob.mx: https://gob.mx/cms/uploads/docs/Propuesta-Curricular-baja.pdf (2016)
14. Zadina, J.: The emerging role of educational neuroscience in education reform. Psicología Educativa, 21(2), pp. 71−77 (2015)
15. Mc Elroy, A., Synnott, J., Elliot, D., Nugent, Ch.: Community-based trials of mobile solutions for the detection and management of cognitive decline. Healthcare Technology Letters, 4(3), pp. 93−96 (2017)
16. Xu, Y.F., Shi, L.S.: Research and design of APP for primary school students' safety education based on embodied cognitive theory. In: IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI), pp. 1−4 (2020)
17. Yi, C., Ruan, F., Gao, Y., Hei, X., Zhang, C.,: QiFei: Assisting to improve cognitive abilities for autism children using a mobile APP. In: Information Communication Technologies Conference (ICTC), pp. 297−301 (2020)
18. Aboalela, R., Khan, J.: Model of learning assessment to measure student learning: Inferring of concept state of cognitive skill level in concept space. In: 3er International Conference on Soft Computing & Machine Intelligence (ISCMI) (2016)
19. Karal, H., Nabíyev, V., Erümít, A.K., Arslan, S., Çebí, A.: Students' opinions on artificial intelligence based distance education system (Artimat). Procedia - Social and Behavioral Sciences, 136(2), pp. 549−553 (2014)
20. Montiel, L., Riveros, V.: Los sistemas expertos en el ámbito educativo. Omnia, 20(1) (2014)
21. González, A.C., Hernández, E.P.: Desarrollo cerebral y cognitivo. Elsevier, pp. 281−414 (2008)
22. Mogollón, E.: Aportes de las neurociencias para el desarrollo de estrategias de enseñanza y aprendizaje. Revista Electrónica Educare, 14(2), pp. 113−124 (2010)
23. Amador-Campos, J.A., Forns-Santacana, M., Kirchner, T.: Repertorios cognoscitivos de atención, percepción y memoria (2006)
24. Forner, A.: Valoración diagnóstica de la batería Piaget-Head. Infancia y Aprendizaje, 6(24), pp. 35−52 (1995)
25. Sossa, H., Guevara, E.: Efficient training for dendrite morphological neural networks. Neurocomputing, pp. 132−142 (2014)

Performance of Regression Models in the Estimation of Glucose Levels through the Analysis of FTIR Spectra of Saliva Samples

Miguel Sánchez Brito[1], Ricardo Mendoza González[1], Gustavo J. Vázquez Zapién[2], Francisco J. Luna Rosas[1], Mónica M. Mata Miranda[2], Julio C. Martínez Romo[1]

[1] Tecnológico Nacional de México, Instituto Tecnológico de Aguascalientes, Mexico

[2] Escuela Militar de Medicina, Centro Militar de Ciencias de la Salud, Secretaría de la Defensa Nacional, Mexico

{miguel_sanchezbrito, gus1202}@hotmail.com, {ricardo.mendoza.gonz, mmcmaribel}@gmail.com, {fcoluna2000, jucemaro}@yahoo.com

Abstract. According to the World Health Organization (WHO) [1], the diabetes is one of the four main non-communicable diseases that has the greatest impact on the death rate worldwide, with around 1.6 million deaths attributed to it. The American Diabetes Association (ADA) has stated that type 2 diabetes is the most common type of this condition [2]. Diabetes is a degenerative disease that has no cure, however, it is possible to adopt a set of actions that allow minimizing its effects on daily life, such as physical activities, proper nutrition, adequate medication and constant monitoring of glucose levels in order to prevent hyper/hypo glycemia. Currently, there are various methodologies that allow glucose monitoring through blood analysis, however, in this research, we present a non-invasive methodology to estimate glucose levels from the analysis of the molecular changes visible in saliva that produce the different glucose concentrations [3] by means of Fourier Transform Infrared (FTIR) spectroscopy and Artificial Neural Networks (ANN). After correctly characterizing the samples of 540 people, we infer that the proposed methodology would have a good performance to carry out this analysis.

Keywords: Saliva, Fourier transform infrared spectroscopy, artificial neural networks, type 2 diabetes, glucose monitoring.

1 Introduction

The Fourier Transform Infrared (FTIR) spectroscopy involves the study of the interaction of radiation with molecular vibrations [4]. The aforementioned interaction

Table 1. Population information.

Gender		Age (average)
Male	312	61±11
Female	228	60±12

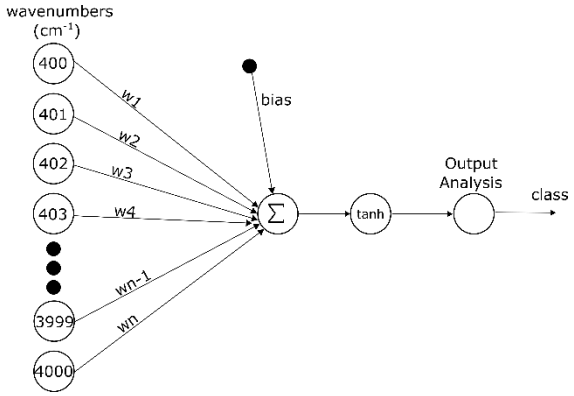


Fig. 1. ANN configuration.

refers to the vibration produced by the molecular bonds that make up a sample when impacted by an electromagnetic wave with a specific frequency, the vibration of the link is stored in a vector known as the FTIR spectrum; as might be expected, for the case of FTIR spectroscopy, these frequencies belong to the region of the infrared spectrum, from 0.11992 to 419.70944 THz approximately.

Based on the frequencies of the electromagnetic wave, the infrared spectrum is divided into three regions: near, medium and far; the middle region (11.9 to 119.9 THz approximately) being the most suitable for analyzing organic samples due to the links that make it up [5].

Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys, and nerves. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin [1]; while some people can control their blood sugar levels with healthy eating and exercise, others may need medication or insulin to help manage it. One of the main techniques for monitoring glucose levels according to the ADA is the A1C test, which focuses on the analysis of glycosylated hemoglobin.

Hemoglobin, is a protein that links up with glucose, is found inside red blood cells, its job is to carry oxygen from the lungs to all the cells of the body. Glucose enters your red blood cells and links up (or glycates) with molecules of hemoglobin. The more glucose in your blood, the more hemoglobin gets glycated. By measuring the percentage of A1C in the blood, you get an overview of your average blood glucose control for the past few months [6].

The largest component of saliva is water (99%); however, it is also possible to find proteins, inorganic ions, and enzyme cofactors metabolites, RNA, and DNA [7].

Although it is not possible to find hemoglobin in saliva, it is possible to find other proteins that can bind with glucose molecules [8].

Through the analysis with regression models of Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Multivariable Linear Regression Models (MLRM) of FTIR spectra of saliva samples from people diagnosed with type 2 diabetes, a non-invasive methodology is proposed to estimate the glucose values of 540 patients, the results indicate that ANN models are the most suitable for estimating since they presented the lowest root-mean-square-error for the validation subset.

2 Materials and Methods

With the consent of the patients and with the approval of the research protocol 001/2019 by the ethics committee of the Unidad de Especialidades Médicas (UEM) of the Secretaría de la Defensa Nacional (SEDENA), approximately 1ml of saliva from patients previously diagnosed with type 2 diabetes was collected in the laboratory area after taking a blood sample for their routine examination's glucose control. Our database was made up as indicated in Table 1.

The glucose values recorded by the UEM obtained range between 47 and 503 mg / dL. By pipetting, 3 µl of saliva was deposited in a Jasco FTIR-6600 spectrometer. After drying the sample using an incandescent lamp, the spectrum was captured using a resolution of 4 cm^{-1} and 120 scans as is suggested by [9] for liquid samples. Once all the spectra were obtained, they were normalized by Standard Normal Variate (SNV) (1):

$$z = \frac{(x - \mu)}{\sigma}. \tag{1}$$

The ANN's initial configuration is a single hidden layer with a neuron using the hyperbolic tangent as the activation function.

SVM (2) was initially configured using a grade 2 polynomial kernel *d* in (3), the initial cost of constraints violation was 1, and the epsilon in the insensitive-loss function 0.1:

$$\min 0.5 \left\| \vec{w} \right\|^2 + C \sum_{i=1}^{n} \xi_i \tag{2}$$

$$\text{s.t. } y_i(\vec{w} * \vec{x} - b) \geq 1 - \xi_i, \forall \vec{x}_i, \xi_i \geq 0 ,$$

$$K(\vec{x_i}, \vec{x_j}) = (\vec{x_i} * \vec{x_j} + 1)^d . \tag{3}$$

3 Results

After drying the spectra, it is possible to appreciate the three main macromolecular groups mentioned by [10]: Lipids 3000-2800 cm^{-1}, Proteins in the range of 1700-1600 and 1560-1500 cm^{-1} (Amide I and Amide II respectively), and nucleic acids in the region of 1250-1000 cm^{-1}. In Fig. 2, the morphological changes of the FTIR spectrum of the sample derived from the drying process can be appreciated, in the same way, the

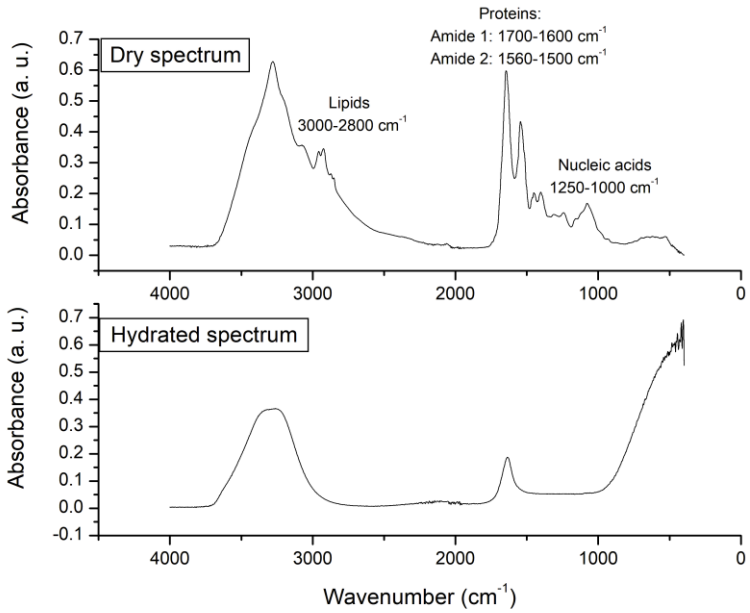


Fig. 2. Morphological changes in the spectrum due to the drying process. In the dry spectrum, the three main macromolecular groups are indicated: lipids, proteins, and nucleic acids.

main macromolecular groups are indicated. Once the spectra of the samples that make up our database were captured, it was normalized according to SNV.

In Fig. 3, the frequencies of the glucose values recorded by the patients contemplated in the present study are presented. We use the Leave One Out Cross Validation (LOOCV) methodology to estimate the glucose value from a certain spectrum, this means that we use n-1 samples to train the ANN, SVM, and MLRM models and we use this model to estimate the glucose value of the spectrum that it was omitted in the training process [11].

It is possible to think about the idea that by capturing an FTIR spectrum of glucose it could indicate the optimal region to analyze without the need for complex methodologies such as ANN or SVM. In Fig 4 a), we present the full FTIR spectra of injectable water solutions with different percentages of glucose. In their research [12] indicates the region 1180-950 cm^{-1} as the optimal region to detect vibrations associated with glucose in saliva, this region is presented in Fig. 4 b).

Considering what is indicated by [12], we selected the wavenumber 1075 cm^{-1} as an indicator of the glucose level. From Fig. 4 b), we can see an inverse behavior of the absorbance to the glucose concentration, so, it would be natural to expect to see similar behavior in saliva spectra.

However, in Fig. 5, we present the distribution of the absorbances of the FTIR spectra at wavenumber 1075 cm^{-1}, after calculating the Pearson correlation coefficient (r: 0.056), we determined that the relationship between absorbance and glucose level was null [13].

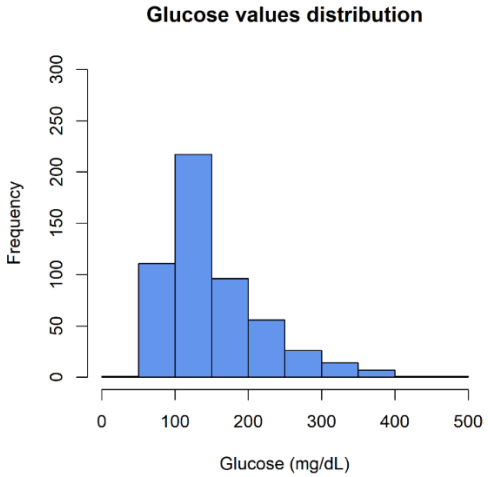


Fig. 3. Distribution of the glucose values recorded.

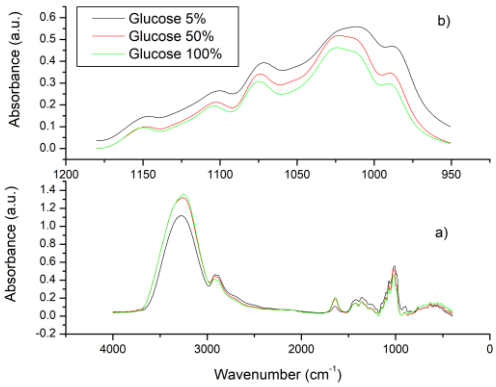


Fig. 4. Variations in the FTIR spectrum of glucose and injectable water solution.

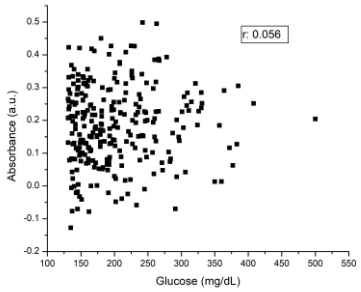


Fig. 5. Correlation between glucose level and absorbance of the FTIR spectra.

Table 2. Calculation of the RMSE-V for the methodologies analyzed.

Methodology	r	RMSE-V
ANN	-0.99	67.68
SVM	0.170	91.28
MLRM	-0.043	21646.21

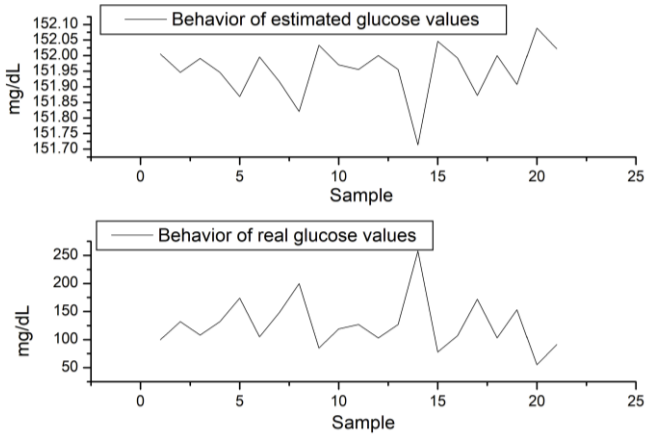


Fig. 6. The behavior of the glucose values of the first 20 samples in the database.

The null relationship presented in Fig. 5 is due to the complexity in the saliva constitution, while in Fig. 4, the spectrum involves only water and glucose, saliva includes several other components [14], this is the main obstacle in the use of FTIR spectra: the complexity of the samples [9]: "*Infrared spectroscopy works best on pure substances since all bands can be assigned to a single molecular structure. If a sample's composition is complex, its spectrum will be complex and it will be hard to know which infrared bands are due to which molecules.*", therefore, the use of techniques such as ANN and SVM are of great interest for the analysis of FTIR spectra.

After implementing regression models for ANN, SVM, and MLRM following the LOOCV methodology, the root-mean-square-error and *r* were calculated for the validation process (RMSE-V), obtaining the following results:

From Table 2, we can see that the best performance is obtained by ANN, however, the RMSE-V obtained is considerable. Analyzing of the amounts of glucose estimated by the ANN models built, we note that they all oscillate between 151 and 153 mg/dL when real values range from 47 to 503 mg/dL.

This allows us to infer that a large number of FTIR spectra of people who presented a glucose value of between 100 and 150 mg / dL as seen in Fig. 3 affects the performance of the ANN models, so it is advisable to have more homogeneous groups in terms of quantity to have a better estimate. The estimates of the model are considerably different, in several cases, from the real value obtained, however, evaluating the behavior of the estimates made, we note that it is very similar to the real one but inverted, this can be seen in Fig. 6.

Table 3. RMSE-V and r obtained through k-fold.

Methodology	r	RMSE-V
ANN	-0.928	78.8

After implementing *r* to evaluate the similarity of the signals, we obtained a value of -0.99, which indicates a strong inverse relationship [13]. Although LOOCV is a commonly used technique for the evaluation of some machine learning model, it is common that the k-fold cross validation (k-fold) methodology is also used to have a better perspective of the performance of the model [15]. The k-fold methodology consists of forming *k* groups from all the samples that make up the database distributed equally and using *k-1* groups for the model training process, which will be used to evaluate the group that did not participate in this process. Considering the results of Table 2, we present in Table 3 the results obtained after evaluating the database with ANN and the k-fold methodology with k = 10.

It is possible to appreciate from the results obtained in Table 3, that the r obtained, despite being reduced, is still a strong inverse correlation.

4 Conclusions

In the present work, the results obtained from the proposal of a model for the estimation of glucose in the blood through the analysis of FTIR spectra of saliva samples using machine learning techniques were presented.

After analyzing more than 500 samples, it is possible to infer that the best technique to estimate the glucose value with the proposed methodology is ANN since it allowed obtaining the lowest RMSE-V.

The RMSE-V obtained could be associated with the number of spectra that correspond to glucose values between 100 and 150 mg / dL, so it is of great interest to continue with the collection of spectra to balance the sample. Additionally, as presented in Fig. 6, probably the output emitted by the ANN model could be entered for one more subsequent process that allows the RMSE-V to be further reduced since, as can be seen in the mentioned figure, the behavior of the ANN estimates is similar but vertically inverted to the behavior of the actual glucose values. Such post-processes could be, in the first instance, the multiplication by a rotational matrix to invert the signal as well as some normalization process to scale the obtained estimates.

References

1. Who: https://who.int/news-room/fact-sheets/detail/noncommunicable-diseases (2011)
2. Diabetes.org: Treatment & Care|ADA. https://diabetes.org/diabetes/treatment-care (2020)
3. Stanford, K.I., Goodyear, L.J.: Exercise and type 2 diabetes: Molecular mechanisms regulating glucose uptake in skeletal muscle. Advances in Physiology Education, 38(4), pp. 308–314 (2014)
4. Larkin, P.: Infrared and Raman Spectroscopy: Principles and spectral interpretation. Elsevier (2011)
5. Sharma, B.K.: Spectroscopy. Krishna Prakashan Media, Goel Publishing House (1981)

6. Diabetes.org: A1C and eAG ADA. https://diabetes.org/diabetes/a1c-test-meaning/a1c-and-eag (2020)
7. Panta, P.: Oral Cancer Detection: Novel strategies and clinical impact. Springer (2019)
8. Shetty, P.K., Pattabiraman, T.N.: Salivary glycoproteins as indicators of oral diseases. Indian Journal of Clinical Biochemistry, 19(1), pp. 97–101 (2004)
9. Smith, B.C.: Fundamentals of Fourier transform infrared spectroscopy. CRC Press (2011)
10. Bel'skaya, L.V., Sarf, E.A., Makarova, N.A.: Use of Fourier transform IR spectroscopy for the study of saliva composition. Journal of Applied Spectroscopy, 85(3), pp. 445–451 (2018)
11. Jarvis, S., Crossley, S.A.: Approaching language transfer through text classification: Explorations in the detectionbased approach. Multilingual Matters (2012)
12. Scott, D.A., Renaud, D.E., Krishnasamy, S., Meriç, P., Buduneli, N., Çetinkalp, Ş., Liu, K.Z.: Diabetes-related molecular signatures in infrared spectra of human saliva. Diabetology and Metabolic Syndrome, 2(1) (2010)
13. PhD, M.S., Dontje, K.J.: Statistics for advanced practice nurses and health professionals. Springer Publishing Company (2014)
14. Wong, D.T.: Salivary diagnostics. John Wiley & Sons (2009)
15. Jarvis, S., Crossley, S.A.: Approaching language transfer through text classification: Explorations in the detection based approach. Multilingual Matters (2012)

Image Classification via Quantum Machine Learning

Héctor Iván García-Hernández, Raymundo Torres-Ruiz, Guo-Hua Sun

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

gsun@cic.ipn.mx

Abstract. Quantum Computing and especially Quantum Machine Learning, in a short period of time, has gained a lot of interest through research groups around the world. This can be seen in the increasing number of proposed models for pattern classification applying quantum principles to a certain degree. Despise the increasing volume of models, there is a void in testing these models on real datasets and not only on synthetic ones. The objective of this work is to classify patterns with binary attributes using a quantum classifier. Specially, we show results of a complete quantum classifier applied to image datasets. The experiments show favorable output while dealing with balanced classification problems as well as with imbalanced classes where the minority class is the most relevant. This is promising in medical areas, where usually the important class is also the minority class.

Keywords: Quantum machine learning, image classification, quantum computing, computational intelligence, imbalanced classification.

1 Introduction

Image classification is of utmost importance in several areas of science and technology such as medical diagnosis and prognosis, face detection, or multiple object detection for autonomous cars. By classical models, this task can be solved using Convolutional Neural Networks [1] but it is notorious the enormous number of parameters needed to train, as seen, for example, in [2]. Nevertheless, exploiting the prowess of Quantum Mechanics such as interference, superposition, and entanglement, which promises great power of computation and in compass with the recent implementation of several quantum computers, it is worth to propose and evaluate quantum models for machine learning. Although these models are, in essence, simple and with performances lower than the state of the art, they serve as stepping stones for the construction of increasingly complex models with much better performance.

The **advantage** of the quantum models is the inherent parallelism in their execution, the speed at which they are executed, and even more important

is the exponential reduction of the number of qubits necessary to encode the information compared to the classical models, for example, only 6 qubits are needed to encode a 64-dimensional pattern and with just 30 qubits we could encode a 32768 x 32768 binary image, which is more than a billion-dimensional flatten vector. This reduction is possible by exploiting superposition states and quantum entanglement.

Let us present some high-level descriptions of models proposed by various research groups, either purely quantum or hybrid combining classical and quantum processing. Yamamoto et al. [3] proposed a quantum perceptron model that allows classifying non-linearly separable data. Maria Schuld et al. [4] proposed another quantum perceptron model using unitary operators acting on two qubits and the inverse quantum Fourier transform. Maxwell Henderson et al. [5] put forward a quantum convolutional layer model for the extraction of features in images. Sebastien Piat et al. [6] proposed a preprocessing with auto-encoders, a restricted Boltzmann machine (RBM) is trained in a quantum computer. This RBM is used to initialize a classical neural network which is subsequently trained in a classical way.

Iris Cong et al. [7] utilized another model for a quantum convolutional network. Zhao et al. [8] proposed a swap-based red neural quantum test. Dang et al. [9] proposed a KNN-based quantum classifier, with a classical model for feature extraction. Francesco et al. [10] proposed a new model for a quantum neuron implemented in a real quantum processor. Using Qiskit [11] and Pytorch, arbitrarily large hybrid models can be generated [12]. Despite having various proposals and with various applications [13], each of the models omits a feasible implementation in a real quantum processor, they lack a proof in a real dataset or in their most extreme case they do not correctly use quantum mechanics [14].

With this work, we aim to show the potential application of Quantum Machine Learning to real-world problems in image classification validating and testing a quantum classification model beyond the theoretical realm and the standard proof of concept.

In this work we first explore two real image datasets, some performance measures are discussed, and the implementation of a quantum classifier is described both at a theoretical and at a high level in Section 2. The performance of the classifier in the two datasets is presented in Section 3, evaluating its performance when facing a problem of both, balanced classes and highly unbalanced classes. Finally, the conclusions and future work are presented in Section 4.

2 Datasets and Algorithms

In this section, we summarize two digits images datasets. The performance measures used to evaluate the classifier are discussed and the classifier itself is analyzed. Before starting to introduce them, it is necessary to propose the theoretical description to be used in this work.

2.1 Theoretical Model

Prior to describing the practical steps in the algorithm, a theoretical approach must be taken, so with a little more in-depth description, and following [10] almost verbatim, we start with the binary pattern $\vec{x}^T = (x_0 \dots x_{m-1})$ and the weight vector $\vec{\Omega}^T = (\Omega_0 \dots \Omega_{m-1})$ with $x_j, \Omega_j \in \{0, 1\}$ and then we map them to:

$$\vec{i} = \begin{pmatrix} i_0 \\ i_1 \\ \vdots \\ i_{m-1} \end{pmatrix}, \vec{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{m-1} \end{pmatrix}, \tag{1}$$

with $i_j, w_j \in \{-1, 1\}$ and with them we can define two quantum states:

$$|\psi_i\rangle = \frac{1}{\sqrt{m}} \sum_{j=0}^{m-1} i_j |j\rangle \text{ and } |\psi_w\rangle = \frac{1}{\sqrt{m}} \sum_{j=0}^{m-1} w_j |j\rangle \ . \tag{2}$$

The states $|j\rangle \in \{|00\dots00\rangle, |00\dots01\rangle, \dots, |11\dots11\rangle\}$ form the computational basis of a quantum processor. If N qubits are used in the register, there are $m = 2^N$ basis states labeled $|j\rangle$ and we can use factors ± 1 to encode the m -dimensional classical patterns and weights into a uniformly weighted superposition of the computational basis.

The first step is to prepare the state $|\psi_i\rangle$ by encoding the input values of $\vec{i}$. With the qubits initialized in the zero state $|00\dots00\rangle \equiv |0\rangle^{\otimes N}$, we perform a unitary transformation U_i:

$$U_i |0\rangle^{\otimes N} = |\psi_i\rangle \ . \tag{3}$$

The second step computes the inner product between $\vec{w}$ and $\vec{i}$ using the quantum register. This can be done defining a unitary transformation, U_w, such that:

$$U_w \, |\psi_w\rangle = |1\rangle^{\otimes N} = |m-1\rangle \ . \tag{4}$$

If we apply U_w after U_i, the quantum state becomes:

$$U_w \, |\psi_i\rangle = \sum_{j=0}^{m-1} c_j |j\rangle \equiv |\phi_{i,w}\rangle \ . \tag{5}$$

Using Eq. (4), the scalar product between the two quantum states is:

$$\begin{aligned} \langle \psi_w \mid \psi_i \rangle &= \left\langle \psi_w \left| U_w^\dagger U_w \right| \psi_i \right\rangle \\ &= \langle m-1 \mid \phi_{i,w} \rangle = c_{m-1} \ , \end{aligned} \tag{6}$$

and from the definitions in Eq. (2) we see that the scalar product of input and weight vectors is $\vec{w} \cdot \vec{i} = m \, \langle \psi_w \mid \psi_i \rangle$. Hence, the desired result is contained, up to a normalization factor, in the coefficient c_{m-1} of the final state $|\phi_{i,w}\rangle$.

In order to extract such an information, an ancilla qubit (a) initially set in the state $|0\rangle$ is toggled by a multi-controlled NOT gate between the N encoding qubits, this leads to:

$$|\phi_{i,w}\rangle\,|0\rangle_a \rightarrow \sum_{j=0}^{m-2} c_j|j\rangle|0\rangle_a + c_{m-1}|m-1\rangle|1\rangle_a\ . \tag{7}$$

The nonlinearity required by the threshold function at the output of the perceptron is immediately obtained by performing a quantum measurement. By measuring the state of the ancilla qubit produces the output $|1\rangle_a$ with probability $|c_{m-1}|^2$.

Even though general advantages or disadvantages cannot be outlined, we can mention some of the differences between this quantum model and a classical one: an execution of a classical model gives an activation directly comparable to a threshold whereas the quantum model gives a binary output which must be repeated several times in order to acquire a measure approximately to theory and comparable to the threshold. Another difference is that since the quantum measure is binary, it can be readily interpreted as the assigned class given the nonlinearity function is already satisfied. And lastly, in general, a classical model does not require input patterns with numerical attributes to be codified into other representations, although it can be useful, however, most quantum models, including this one, are limited to patterns with binary attributes or to be coded as such.

2.2 Datasets

Two databases are used for the experiments. They are both images of digits. The first is the "Optical Recognition of Handwritten Digits Data Set (Digits Dataset)", which contains 64 attributes in a range from 0 to 16 with 5620 total instances [15]. The test set with 1797 instances, is available directly on the scikit-learn Python package. The second database is the "Semeion Handwritten Digit Data Set (Semeion Dataset)" which contains 1593 instances each with 256 binary attributes [17]. Both datasets are balanced, containing approximately the same number of instances per class.

Table 1. Summary of both used digits images datasets.

Dataset	Classes	Imbalance Ratio	Total Instances
Digits	10	1.032	5620
Semeion	10	1.045	1593

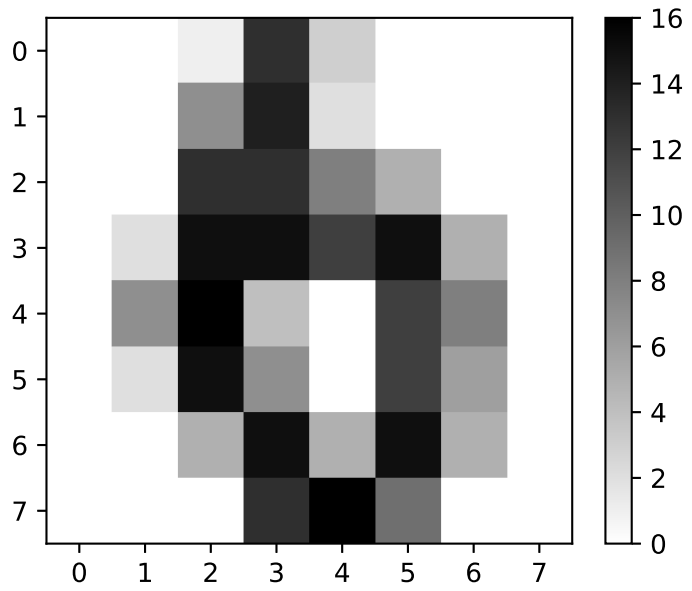


Fig. 1. An instance for class 6 in the Digits Dataset. Each pixel has a value varying from 0 up to 16. As can be seen, the low resolution can lead to misclassification, since this instance can be easily mistaken for an instance of class 0.

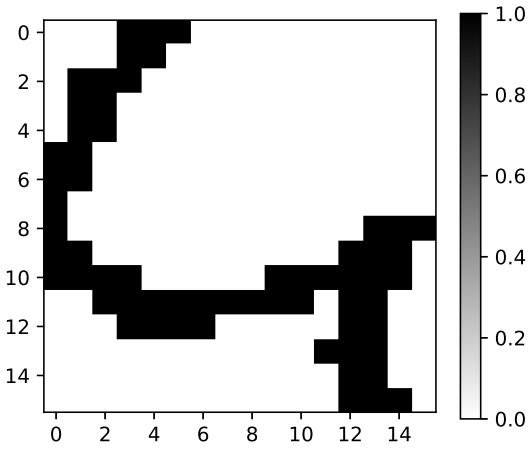


Fig. 2. An instance for class 4 in the Semeion Dataset. Each attribute, or pixel, has a binary value. These patterns have a bigger dimensionality, and thus are less likely to be misclassified, but also the complexity of the processing needed to build the quantum circuit is increased.

2.3 Performance Measures

When each class contains roughly the same number of instances in a dataset, it is known as a balanced dataset. In these cases, most of the performance measures are adequate, as long as there is no bias towards any class. However, depending on the application of the classifier and the relevance of any of the classes, some other performance measure may be chosen.

Table 2. Confusion matrix for a two-class dataset.

		Real Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The most common is accuracy, which measures the ratio of instances correctly classified to the total number of instances:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \ . \tag{8}$$

A dataset is unbalanced when one or more classes is poorly represented in the dataset. Most classical performance measures produce a majority class bias in an unbalanced class problem. In these cases the True Positive Rate (TPR):

$$TPR = \frac{TP}{TP + FN} \ , \tag{9}$$

which is also known as Recall or Sensitivity can be used to measure the ratio of the number of positive instances correctly classified to the total number of positive instances.

We also keep track of the Positive Predictive Value (PPV) (also known as Precision):

$$PPV = \frac{TP}{TP + FP} \ , \tag{10}$$

which measures the ratio of the number of positive instances correctly classified to the total number of positive classified instances. With TPR and PPV the $F1$ score can be obtained, which is the harmonic mean of these performance measures:

$$F1 = 2 * \frac{PPV * TPR}{PPV + TPR} = \frac{2 * TP}{2 * TP + FP + FN} \ . \tag{11}$$

When classes contain insufficient instances to be partitioned in a traditional validation method, it is common to use all instances for training and testing process. Accuracy under this validation method is known as Resubstitution Error.

2.4 Quantum Classification Algorithm

The model used for the classification task is an implementation of the one described in [10]. In this model, an instance with binary attributes is encoded by means of a method called *hypergraph states generation subroutine* [10, 16]. The weight vector is randomly initialized, which will be updated accordingly to a set of hyper-parameters, which will regulate its rate of change. At the end of the execution of the dynamically generated quantum circuit through the process described in [10], a measurement is performed on the ancilla qubit, which will take the value 0 or 1. By means of several repetitions of the circuit, the proportion of measurements with result 1 over the total measurements can be obtained. The more measures are made, the closer the result is to the real one.

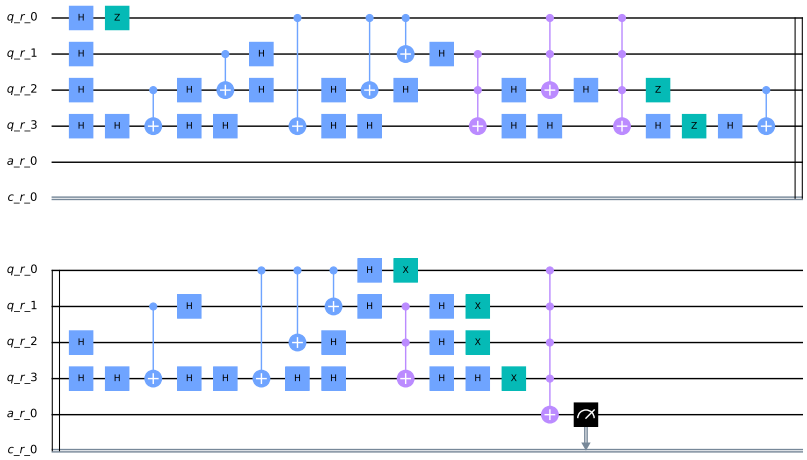


Fig. 3. Quantum circuit programmatically generated encoding both a pattern and a vector weight. The circuit is encoding an image of size 4x4 using 4 qubits for processing, one qubit as ancilla, and one classical bit storing the measurement. This relatively simple circuit already contains 34 layers of quantum gates, this can convey the magnitude of the circuits needed to process the 16x16 images. Currently bounded by the physical implementation of real quantum computers, the model is not limited by the pattern dimensions it can process.

This proportion, which we called *readout*, is compared with another hyper-parameter, which is called threshold. This threshold is used to assign the class. If the readout is less than the threshold the positive class is assigned, otherwise the negative class is assigned to the pattern in turn.

As it is a supervised classification task, during the training step, it will be evaluated if the assigned class is correct. If it is, we simply continue with the next pattern. In the case it is incorrectly classified, depending on its real class,

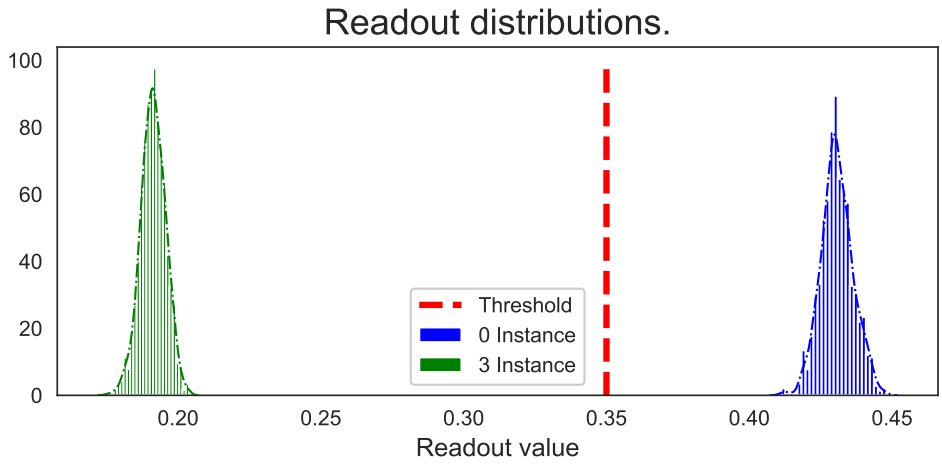


Fig. 4. Readouts distributions after a thousand iterations of the same circuit using the final weight vector for the 3/0 binary classifier. The distribution at the left is for an instance of the class 3 and is well below the threshold. The distribution at the right is for an instance of class 0, this time, above the threshold.

a hyper-parameter will be used, in a similar fashion to the traditional learning rate in neural networks, which defines the proportion of change in the weight vector. There is a learning rate for the positive class and another one for the negative class.

As usual, this procedure can be repeated for an arbitrary number of epochs, where one epoch means that the classifier has seen the entire training set. Or also, as it was done, the training can be finished earlier if a critical value has been reached in a certain metric that we seek to optimize.

The whole process, aiming for efficiency, is simulated using Qiskit [11], however, the entire process is suitable and ready to be executed on a real quantum machine.

Although we implemented and followed the model described by [10], we diverge from them in the sense that we do not test the model in an *ad hoc* dataset. Furthermore, the *ad hoc* dataset was split by means of a previously and arbitrarily selected weight vector. In other words, the classification task was already known to be solvable because it was constructed to do so, but in this work, we do not assume or fabricate such property and let the model converge to a solution.

3 Experimental Results and Discussion

In this section we present the experimental results of the classification model described above. The model was tested in both Optical Recognition of Hand-

written Digits Data Set and Semeion Handwritten Digit Data Set. In case of Digits, Resubstitution Error and Hold-out were used as validation method. For Hold-out, we used the partition offered by the authors [15]. For Resubstitution Error, only the test set was used, acting as both training and test sets. This dataset requires processing, as each attribute has a value between 0 and 16, and the model only works with binary values a threshold was applied to binarize the patterns as follows:

$$\text{binarized pixel} = \begin{cases} 0, & \text{if original pixel value} < 10 \\ 1, & \text{otherwise} \end{cases} \tag{12}$$

Fig. 5. At the left, there is an original instance from the Digits Dataset. In the middle, there is the result of the binarization threshold. And at the right, there is the final weight vector for the 0/3 classifier. The vector tends to take the form of sort of a mask for one of the classes.

This threshold is itself a new hyper-parameter that can be optimized. In the case of Semeion, Resubstitution Error was used due to the relatively small number of available patterns. This dataset does not need processing as the attributes are already binary. In each dataset, binary classifiers of two styles are trained and tested: class vs class, also known as One vs One (OvO), which in this case represents a balanced classes problem, and class vs the rest also known as One vs All (OvA) which represents an unbalanced class problem. The results are shown in several tables.

In Table 3, the diagonal represents the trivial classification of one single class. It is interesting to note although we can use the upper or lower half to classify the reflexive class, i.e. use the trained positive/negative classifier to try to classify the negative/positive problem, we did not get good results in this scenario. This makes sense when the vector weight for each binary classifier is inspected since it tends to take the form of a sort of mask resembling the instances of the negative class.

In Table 4, we keep track of some performance measures, including the Area Under the Curve (AUC). It is evident from the Recall measure that the quantum model can distinguish the positive class from all the other instances with good accuracy. The ratio of minority class against the rest is approximately 1:100. This result is promising for medical applications where datasets are generally

heavily imbalanced. We show, in Table 5 the increase in accuracy performance when a different validation method is used. It is also notable the latent power of generalization because in this case the evaluation was recorded upon a test set containing instances not seen during the training set.

In Table 6, we show the usual metrics, where is notorious the benefit gained by the Hold-out method for this classification model in this dataset. We noted an improvement in seven out of the ten classes. This validation method gives us more confidence in the generalization power that this model might have.

In Table 7 and in Table 8, we give the already known performance metrics. Acceptable performance can be seen in almost every classification task, but it is notorious the decrease compared to the Digits dataset. This drop can be explained by the fact that each instance lives in a bigger dimensional space, and therefore the quantum circuit needed to process each pattern is also bigger and more complex quantum-gates-wise. Nevertheless, we must keep in mind although both datasets might seem similar they are in fact different and we should not expect the same performance in both of them.

Table 3. Resubstitution Error by balanced class classification.

Digits Resubstitution Error										
	Negative Class									
	0	1	2	3	4	5	6	7	8	9
Positive Class										
0	1.0	0.725	0.794	0.714	0.671	0.769	0.353	0.834	0.678	0.589
1	0.377	1.0	0.660	0.849	0.749	0.780	0.898	0.833	0.609	0.759
2	0.833	0.710	1.0	0.730	0.804	0.894	0.631	0.764	0.564	0.820
3	0.933	0.753	0.833	1.0	0.780	0.887	0.807	0.676	0.624	0.652
4	0.974	0.567	0.837	0.758	1.0	0.785	0.792	0.780	0.749	0.764
5	0.886	0.785	0.830	0.715	0.779	1.0	0.953	0.695	0.800	0.640
6	0.635	0.807	0.843	0.807	0.682	0.785	1.0	0.883	0.540	0.828
7	0.983	0.653	0.803	0.776	0.875	0.797	0.905	1.0	0.728	0.832
8	0.752	0.480	0.706	0.792	0.783	0.651	0.830	0.597	1.0	0.666
9	0.564	0.812	0.815	0.696	0.883	0.527	0.678	0.813	0.788	1.0

4 Conclusions and Future Work

In this work, the performance of a fully quantum machine learning model in real datasets was tested. The evaluation is generally favorable, providing positive feedback that QML is promising.

Table 4. Resubstitution Error by heavily imbalanced class classification.

Digits Resubstitution Error					
Positive Class	Recall	Accuracy	Precision	F1	AUC
0	0.983	0.303	0.122	0.218	0.605
1	0.857	0.399	0.129	0.224	0.602
2	0.966	0.377	0.133	0.234	0.639
3	0.814	0.328	0.112	0.198	0.543
4	0.955	0.420	0.143	0.249	0.657
5	0.873	0.359	0.123	0.216	0.587
6	0.845	0.368	0.121	0.212	0.580
7	0.837	0.417	0.128	0.222	0.604
8	0.724	0.388	0.107	0.186	0.538
9	0.861	0.419	0.132	0.229	0.615

Table 5. Accuracy measure by balanced class classification.

Digits Hold-out										
	Negative Class									
	0	1	2	3	4	5	6	7	8	9
Positive Class										
0	1.0	0.844	0.833	0.900	0.896	0.833	0.924	0.831	0.732	0.731
1	0.969	1.0	0.852	0.863	0.818	0.826	0.914	0.853	0.775	0.779
2	0.895	0.824	1.0	0.841	0.810	0.860	0.877	0.828	0.740	0.843
3	0.883	0.819	0.730	1.0	0.920	0.854	0.964	0.823	0.851	0.719
4	0.933	0.815	0.899	0.840	1.0	0.876	0.790	0.819	0.819	0.811
5	0.933	0.892	0.793	0.802	0.856	1.0	0.964	0.872	0.811	0.765
6	0.827	0.809	0.843	0.925	0.825	0.953	1.0	0.902	0.839	0.739
7	0.974	0.878	0.876	0.859	0.794	0.869	0.977	1.0	0.813	0.871
8	0.889	0.682	0.860	0.826	0.814	0.823	0.842	0.804	1.0	0.824
9	0.849	0.850	0.907	0.746	0.822	0.850	0.955	0.835	0.836	1.0

With this work, we provided a real validation to the quantum model beyond the theoretical correctness and the usual proof of concept. This showed the potential real-world application of the QML in the near future.

Table 6. Accuracy by heavily imbalanced class classification.

Digits Hold-out					
Positive Class	Recall	Accuracy	Precision	F1	AUC
0	0.955	0.206	0.107	0.192	0.539
1	0.972	0.218	0.112	0.201	0.552
2	0.858	0.212	0.098	0.176	0.500
3	0.803	0.188	0.093	0.167	0.460
4	0.994	0.176	0.108	0.195	0.539
5	0.950	0.249	0.114	0.204	0.560
6	1.0	0.154	0.106	0.192	0.529
7	0.843	0.346	0.116	0.204	0.567
8	0.965	0.204	0.105	0.190	0.544
9	0.966	0.193	0.107	0.193	0.536

Table 7. Resubstitution Error by balanced class classification.

Semeion Resubtitution Error										
	Negative Class									
	0	1	2	3	4	5	6	7	8	9
Positive Class										
0	1.0	0.851	0.806	0.890	0.788	0.840	0.593	0.843	0.835	0.680
1	0.947	1.0	0.728	0.869	0.352	0.763	0.832	0.700	0.608	0.650
2	0.937	0.707	1.0	0.767	0.793	0.707	0.706	0.716	0.601	0.637
3	0.912	0.797	0.672	1.0	0.759	0.757	0.853	0.712	0.525	0.624
4	0.981	0.801	0.793	0.837	1.0	0.793	0.751	0.749	0.515	0.664
5	0.890	0.794	0.672	0.657	0.693	1.0	0.721	0.703	0.528	0.570
6	0.757	0.842	0.800	0.881	0.636	0.809	1.0	0.805	0.506	0.692
7	0.946	0.734	0.624	0.779	0.702	0.722	0.724	1.0	0.530	0.667
8	0.854	0.728	0.786	0.866	0.655	0.671	0.781	0.683	1.0	0.610
9	0.905	0.809	0.763	0.769	0.689	0.611	0.780	0.737	0.607	1.0

However, it is in its early stages and in order for it to be competitive against traditional Machine Learning, there is still a gap which this work seeks to bridge.

Some points that would contribute to moving the QML to more mature stages are an increment in the number of available qubits to perform computation,

Table 8. Resubstitution Error by heavily imbalanced class classification.

Semeion Resubtitution Error					
Positive Class	Recall	Accuracy	Precision	F1	AUC
0	0.770	0.193	0.090	0.161	0.449
1	0.962	0.230	0.113	0.202	0.555
2	0.849	0.222	0.1	0.178	0.500
3	0.937	0.212	0.106	0.191	0.534
4	0.807	0.183	0.092	0.166	0.460
5	0.937	0.202	0.105	0.189	0.528
6	0.925	0.197	0.105	0.189	0.520
7	0.873	0.230	0.102	0.183	0.516
8	0.864	0.210	0.097	0.175	0.502
9	0.873	0.211	0.100	0.180	0.506

reduce noise during the application of quantum gates, and mainly a feasible, scalable, and robust codification to map classical inputs to their quantum representation.

As future work, we would propose to extend the biclass to multiclass classification by means of the naive extension One vs One and One vs All as a baseline. A modification in the quantum circuit generation would be proposed to allow the coding of images at three channels depth, that is, in color. It would also be interesting to implement one of the proposals for quantum convolutional layers for features extraction.

References

1. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86, pp. 2278–2323 (1998)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv: 1512.03385.
3. Yamamoto, A.Y., Sundqvist, K.M., Li, P., Harris, H.R.: Simulation of a multidimensional input quantum perceptron. Quantum Inf Process, 17, pp. 128 (2018)
4. Schuld, M., Sinayskiy, I., Petruccione, F.: Simulating a perceptron on a quantum computer. Physics Letters A., 379, pp. 660–663 (2015)
5. Henderson, M., Shakya, S., Pradhan, S., Cook, T.: Quanvolutional neural networks: Powering image recognition with quantum circuits, arXiv: 1904.04767 (2019)
6. Piat, S., Usher, N., Severini, S., Herbster, M., Mansi, T., Mountney, P.: Image classification with quantum pre-training and auto-encoders. International Journal of Quantum Information. 16, 1840009 (2018)
7. Cong, I., Choi, S., Lukin, M.D.: Quantum convolutional neural networks. Nature Physics, 15, pp. 1273—1278 (2019)

8. Zhao, J., Zhang, Y. H., Shao, C. P., Wu, Y.C., Guo, G.C., Guo, G.P.: Building quantum neural networks based on a swap test. Phys. Rev. A., 100, 012334 (2019)
9. Dang, Y., Jiang, N., Hu, H., Ji, Z., Zhang, W.: Image classification based on quantum K-Nearest-Neighbor algorithm. Quantum Information Processing, 17, 239 (2018)
10. Francesco, T., Chiara, M., Dario, G., Daniele, B.: An artificial neuron implemented on an actual quantum processor. NPJ Quantum Information, 5(1) (2019)
11. Aleksandrowicz, G., Alexander, T., Barkoutsos, P., Bello, L., Ben-Haim, Y., Bucher, D., Cabrera-Hernández, F.J., Carballo-Franquis, J.: Qiskit: An Open-source Framework for Quantum Computing (2019)
12. Qiskit.org: Hybrid quantum-classical Neural Networks with PyTorch and Qiskit, `https://qiskit.org/textbook/ch-machine-learning/machine-learning-qiskit-pytorch.html` (2019)
13. Jeswal, S.K., Chakraverty, S.: Recent developments and applications in quantum neural network: A review. Archives of Computational Methods in Engineering, pp. 1–15 (2018)
14. Zhou, J., Gan, Q., Krzyżak, A., Suen, C.Y.: Recognition of handwritten numerals by quantum neural network with fuzzy features. IJDAR, 2, pp. 30—36 (1999)
15. Alpaydin, E., Kaynak, C.: Cascading Classifiers. Kybernetika, 34, pp. 369–374 (1997)
16. Rossi, M., Huber, M., Bruß, D., Macchiavello, C.: Quantum hypergraph states. New Journal of Physics, 15, 113022 (2013)
17. UCI Machine Learning Repository: Semeion Handwritten Digit Data Set. `https://archive.ics.uci.edu/ml/datasets/semeion+handwritten+digit` (2019)

Mitigating Gender Bias in Knowledge-Based Graphs Using Data Augmentation: WordNet Case Study

Claudia Rosas Raya, Ana Marcela Herrera Navarro

Universidad Autónoma de Querétaro,
Mexico

crosas17@alumnos.uaq.mx, claudiarosas16@gmail.com

Abstract. WordNet ontology is examined in order to show how it reflects historical gender biases in the semantic relationships between terms in a knowledge-based graph. We define a general benchmark to diagnose the gender bias in the WordNet ontology. Subsequently, we evaluate a set of words that have the equivalent in masculine and feminine and found their semantic-related terms according to WordNet. We then propose a technique to mitigate bias by data augmentation in order to create a neutral vector that combines both features without any distinction between genders. The results were compared with the Wu-Palmer semantic metric to validate the results and corroborate that gender bias was mitigated.

Keywords: Ontology, semantic similarity, data augmentation, WordNet, gender bias.

1 Introduction

Gender bias is exhibited repeatedly in the field of Natural Language Processing (NLP), including in training data, pre-trained models and algorithms themselves. Gender bias can be defined as the unfair difference in the way women and men are treated. The propagation of gender bias in Artificial Intelligence algorithms poses a danger of perpetuating stereotypes in real-world applications. Several problems have been already reported in different fields like machine learning (ML) [1].

Specifically, machine translation [2], where problems where found when translating sentences including words that historically had belonged to men or women like “The doctor” is more likely to be interpreted as masculine when translating and “The nurse” is more likely to be feminine.

Also, working with word embeddings [3] had shown issues like in [4] where results such as “man is to computer programmer as woman is to homemaker” is one of the problems detected when working with analogies. Or in [5] where it was found that training data contains significantly more male than female entities.

Based on Crawford categorization [6], bias can include harms of allocation, harms of representation and politics of classification. In terms of NLP applications, allocation bias is reproduced when models often perform better on data associated with majority. Detection of gender bias in Artificial Intelligence applications is a nascent field, and one of the first works is [7] defining gender bias as the correlation between the magnitude of the projection onto the gender subspace of a word embedding representing a gender-neutral word and that word's bias rating.

In [8] a new method called GN-GloVe is proposed. The authors train the word embeddings by isolating gender information in specific dimensions and maintaining gender-neutral information in another dimension.

The Implicit Association Test (IAT) is applied in psychology to measure subconscious gender bias in humans. In [9] the IAT's core concept is adopted, measuring gender bias through the difference in the strength of association of concepts, to measure bias in word embeddings using the Word Embedding Association Test (WEAT).

We show with new experiments that this methodology can be used to mitigate gender bias when using an ontology and show promising results to keep working with knowledge graphs.

The paper is organized as follows: Section 2 presents a brief survey of different works related to manipulation of an ontology. Section 3 explains in detail our data augmentation approach and the validation of our method. Section 4 includes the results and their interpretation. Finally, section 5 concludes the paper.

2 Related Work

2.1 Knowledge-based Graphs

Ontologies (a type of knowledge-based graphs) are a way to systematize knowledge providing semantic context in a method that a machine can handle it. An ontology is a "specification of a conceptualization" [10]. It models the classification of entities and the relationships between said entities. They are used to attempt to understand what exists in unstructured data in order to help systems to overcome semantic heterogeneity and facilitate them to interchange knowledge. They permit to transform unstructured text to structure data for a computer to understand it as it would be processing knowledge not only symbols. Semantic models contribute to build systems with more human-like behavior.

Wordnet [11, 12] is a semantic network widely investigated for NLP because of its accessibility. It has la large number of representations of semantic relationships, making it more appropriate for natural language understanding applications.
The main relation among words in WordNet is synonymy. Synonyms refer to words that denote the same concept and are interchangeable in many contexts. Additionally, a synonym contains a brief definition.

The most frequently encoded relation among terms is the super-subordinate relation (also called hyperonymy, hyponymy or ISA relation).

It links more general concepts to increasingly specific ones. All noun hierarchies ultimately go up the root node entity. Hyponymy relation is transitive [11, 12].

2.2 Semantic Similarity Measure

In general, semantic metrics can be classified into two groups: metrics which use only a thesaurus (e.g., WordNet) and those which use a thesaurus and probabilistic information from distributions in corpora [13].

The Wu-Palmer metric (WUP) weights the edges based on distance in the hierarchy. Namely, jumping from inanimate to animate is a larger distance than jumping from Feline to Canine. Using the same logic, the word senses of love and hate, while antonyms, are very related since they essentially belong to the same semantic type. Thus, it would be expected that the metrics give a higher similarity to them, than to the tuple love-romance; romance is very similar to love, but its type is not as close as say hate or dislike [13]. E.g., *hate-love* is 0.857 and *love-romance* is 0.615.

Other metrics based on thesaurus are: Path similarity, Leacock-Chodorow Similarity that work similarly to Wu-Palmer. But for experimentation purposes Wu-Palmer was chosen.

Wu-Palmer similarity [14] proposes a measure that takes into account the position of concepts in a taxonomy relative to the position of the Least Common Subsumer. Based on the edge counting method. Assuming that the similarity between two concepts is in function of the path length and depth. The range is between 0 and 1. It is continue and it normalizes the data; the score can never be zero, it is heavily dependent on the quality of the graph. It was first created to achieve machine translation between Chinese and English. The similarity measure of Wu and Palmer is defined by the following expression:

$$Sim_{wup} = 2 * \frac{N}{N1 + N2},$$

where N represents the distance to the closest common ancestor, N1 and N2 stand for node 1 and node 2 that are being compared.

2.3 Data Augmentation

Data augmentation has been shown to be flexible. It can mitigate gender bias in different applications.

Frequently a data set has a disproportionate number of references to one gender. To reduce this, [15] proposed to create an augmented data set identical to the original data set but biased towards the opposite gender and to train on the union of the original and data-swapped sets. The augmented data set was generated using gender-swapping information. The target of data augmentation is to reduce the biased predictions by training the model on a gender-balanced data set.

In [15] when creating a gender-balanced data set, data augmentation works as follows: for every sentence in the original data set, a sentence with the gender swapped is created.

Table 1. Extract of terms after obtaining the semantic terms related to the tuple and its definition.

Term	Hypernym	Hyponym	Definition
woman	adult, person, physical entity, organism, living thing	cinderella, madame, divorcee, dominatrix, geisha, girl, belle, bimbo, maid, sex kitten, tomboy, gold digger, gravida, prostitute, trophy wife	adult, female, person, wife, mistress, girlfriend.
man	adult, person, physical entity, organism, living thing	Adonis, bachelor, boy, bull, ejaculator, gentleman, guy, iron man, patriarch, peter pan, Tarzan, womanizer, Casanova, don Juan	adult, male, person, manservant, attendant, employer

Following, name-anonymization is applied to every original sentence and its gender-swapped equivalent. Name anonymization consists of replacing all named entities with anonymized entities, such as "E1". This deletes gender associations with named entities in sentences. Next, the model is trained on the union of the original data set with name-anonymization and the augmented data set. The identification of gender-specific words and their equivalent opposite gender word requires lists of terms associated to genders.

In [8, 16] independently designed GBETs based on Winograd Schemas. The corpus consists of sentences which contain a gender-neutral occupation (e.g., doctor), a secondary participant (e.g., patient), and a gendered pronoun that refers either the occupation or the participant. For each sentence, [16] considered three types of pronouns (female, male, or neutral), and [8] considered male and female pronouns. They designed metrics to analyze gender bias by examining how the performance difference between gender with respect to each occupation.

In hate speech detection [17], data augmentation reduced the False Positive Equality Difference (FPED) and False Negative Equality Difference (FNED) between male and female predictions of a Convolutional Neural Network (CNN) by a broad margin.

Data augmentation without name-anonymization has also been used to debias knowledge graphs built from Bollywood movie scripts [18] by swapping the nodes for the lead actor and actress, but metrics evaluating the success of gender-swapping were not provided.

3 Methodology

Schiebinger [19] suggested that scientific research fails to take gender issues into account, claiming that the phenomenon of male defaults on technology enables an asymmetry and WordNet is not an exception to this.

This proposal was implemented on MacBook Pro (retina, Mid 2012), with 8 Gb 1600 MHz DDR3 RAM memory and four-core Intel i7 2.3 GHz processor. Programmed in Python 3.8.1

Table 2. General results of the dataset analysis.

Female		Male	WuP	Male Words	Female words	Augmented	Common words
1.	Woman	Man	0.667	312	149	448	13
2.	Queen	King	0.571	48	68	90	26
3.	Mother	Father	0.923	60	46	95	11
4.	Girl	Boy	0.631	42	62	84	20
5.	Aunt	Uncle	0.600	16	15	22	9
6.	Actress	Actor	0.952	45	12	47	10
7.	Princess	Prince	0.900	21	17	28	10
8.	Waitress	Waiter	0.957	17	12	20	9
9.	Hen	Rooster	0.090	12	25	37	0
10.	Mare	Stallion	0.909	16	17	23	10
11.	Spinster	Bachelor	0.571	15	12	20	7
12.	Bride	Bridegroom	0.909	13	18	18	13
13.	Sister	Brother	0.545	29	28	42	15
14.	Countess	Count	0.133	65	33	68	30
15.	Duchess	Duke	0.72	11	13	17	7
16.	Goddess	God	0.824	108	8	109	7
17.	Heroine	Hero	0.125	38	22	42	18
18.	Madam	Sir	0.600	14	14	21	7
19.	Witch	Wizard	0.667	17	31	40	8
20.	Mummy	Daddy	0.857	9	14	15	8
21.	Girl guide	Boy scout	0.944	74	30	86	18
22.	Conductress	Conductor	0.545	82	10	85	7
23.	Chairwoman	Chairman	1.000	14	11	14	11
24.	Lady	Gentleman	0.600	18	22	33	7
25.	Headmistress	Headmaster	0.917	11	11	13	9
26.	Hostess	Host	0.947	69	27	76	20
27.	Wife	Husband	0.600	20	30	39	11
28.	Handlady	Landlord	0.957	9	10	11	8
29.	Lady	Lord	0.125	47	22	61	8
30.	Nun	Monk	0.600	13	19	24	8

Based on the principle of Data Saturation [20], a selected set of 30 tuples of words, were obtained and analyzed in WordNet ontology version 3.0 The tuples were compound by one male and one female word to extract the different semantically related words contained in the ontology hierarchy: hyponyms, hypernyms and the text

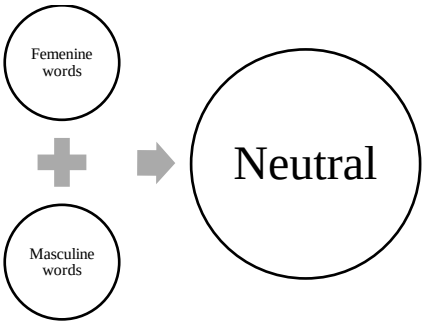


Fig. 1. The masculine terms plus the feminine terms create the gender-neutral result.

in the definition provided to every synset(synonym) of each word in the tuple. An extract of this phase is shown in Table 1.

In table 1, the example uses the tuple man-woman, getting their semantic related terms and their definitions. The text from the definition of the synonym's set were considered. Preprocessing the text included the removal of stop words (words with no semantic information like articles, prepositions, etc.), numbers and punctuation to obtain the words for further analysis.

Wu-Palmer similarity score was calculated and only the terms greater or equals that 0.5 were considered for further analysis. The automatization of noun gender detection is out of the scope of this paper. Therefore, the tuples used for experimentation were manually organized in male and female categories.

After the aforementioned analysis of the results are shown in Table 2 where seven columns can be observed. The first two contain the tuple of words, then the Wu-Palmer similarity between them is calculated. After obtaining all the semantic terms related with each word, the fourth and fifth column encloses the count of the total terms related with the tuple words.

The implementation of the Data Augmentation technique consists in adding to a neutral vector both contents of the vectors in the masculine and feminine lists of terms and deleting the duplicates to include the vocabulary that is missing and probably containing the gender bias information since the words should be describing practically the same topic (sixth column).

Once the list of words was obtained for each word, the female and the male counterpart, both lists were joined to form a neutral set of words that included both words (Fig 1). Based on the fact that both words were describing almost the exact same noun and that the definitions should be very alike, if they do not accomplish this, it may be a situation of gender bias in the description of each word regarding its counterpart.

3.1 Validation

To validate the augmented vector, a comparison between the words belonging to each original vector were compared to those contained in the neutral vector that were not covered in the original ones. Namely, if a word appeared in the neutral vector was not considered in the feminine vector, said word is compared with the feminine word in the

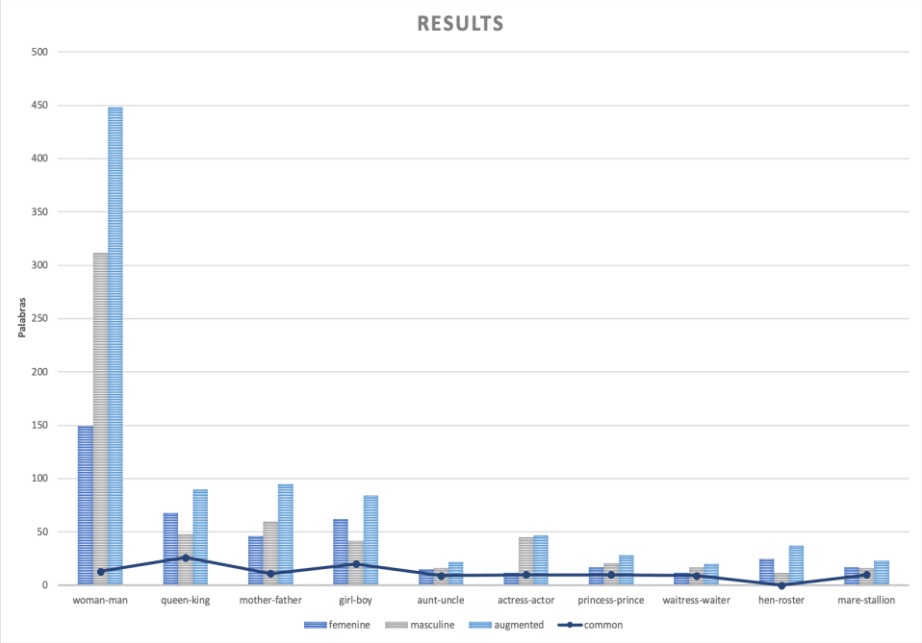


Fig. 2. Related words according to each tuple.

tuple that generated the vocabulary, e.g., the word employer in table 1 was not in the words related to woman and the neutral vector contains it then a comparison between woman and employer is applied to obtain the Wu-Palmer score.

4 Results

After extracting related words from the ontology, each tuple obtained its own list of words. In figure 2, an extract of 10 examples is plotted to show the data set size (in word quantity) individually and after the data augmentation. The number of related words between terms is also shown. The results after the process of data augmentation are shown in table 2. Where the tuples are shown individually and also, the words that previously had in common.

In table 3 it can be observed that 17 out of 30 tuples shown that the average Wu-Palmer similarity of the evaluated terms the validation process was above 0.5. That indicates that the terms contained in each word equivalent can be related to the other.

It can be interpreted as if there is a node that represents the semantic relationship between these words that were initially not considered by the ontology and thus the gender bias mitigation exists, e.g., Wu-Palmer score may be above 0.5 because the node person englobes the words woman and man and even though woman (or man) did not contain everything its equivalent counterpart had, the node *person* essentially considers the whole.

Table 3. Validation results.

	Tuple		Wu-Palmer average
1.	Woman	Man	0.50995519
2.	Queen	King	0.45074813
3.	Mother	Father	0.40816809
4.	Girl	Boy	0.60335059
5.	Aunt	Uncle	0.65689612
6.	Actress	Actor	0.73613838
7.	Princess	Prince	0.71450283
8.	Waitress	Waiter	0.51829822
9.	Hen	Roster	0.13443386
10.	Mare	Stallion	0.59177404
11.	Spinster	Bachelor	0.59299314
12.	Bride	bridegroom	0
13.	Sister	Brother	0.58704386
14.	Countess	Count	0.22238513
15.	Duchess	Duke	0.72356838
16.	Goddess	God	0.77088293
17.	Heroine	Hero	0.34991724
18.	Madam	Sir	0.64572908
19.	Witch	Wizard	0.39590476
20.	Mummy	Daddy	0.51336914
21.	Girl guide	Boy scout	0.22495529
22.	Conductress	Conductor	0.42019613
23.	chairwoman	Chairman	0
24.	lady	Gentleman	0.59448325
25.	headmistress	headmaster	0.65679842
26.	hostess	Host	0.44464883
27.	Wife	Husband	0.36695316
28.	landlady	Landlord	0.76397516
29.	lady	Lord	0.55205839
30.	nun	monk	0.22885582

4.1 Validation Results

Validation results are presented in Table 3.

5 Conclusions

Gender bias detection and mitigation are not easy tasks but the consequences of not start doing it are going to be important due to the more present artificial intelligence applications. Many of said application are based on trained data and the data is labeled somehow. In this work, the main target was one of the tools that are in charge of tagging data. In this particular scenario, we worked with the WordNet ontology, but the theoretical fundamentals can be applied to any ontology or knowledge-based graph with similar structure.

The main intention was to create a vector of neutral gender to try to stablish a start point into a space of non-binary gender approaches.

Text related bias depends not only on individual words, but also on the context in which they appear but as demonstrated, individual words are associated with historical biases posed on a binary gender reality. A possible interpretation of the biases showed by WordNet or any other model with similar problems would be that the model is no more, or less, biased than the real world. Assessing an accurate degree of prejudice of a model, requires the establishment of an ideal set of rules for language and that is a still-going discussion.

In the perspective of computing, data augmentation is easy to implement but it can be expensive if there is high variability in the data or if the data set is large. Furthermore, data augmentation tends to double the size of the training set or origin data, which can increase analysis or training time by a factor specific to the task at hand. Finally, blindly gender swapping in data augmentation can create nonsensical sentences or relationships, in a world with binary gender logic, man to godmother could obtained a high score if the graph organization admits it and therefore influence in the results of application like text clustering.

5.1 Future Work

To improve this work, random tuples could be selected from raw text. In order to prove the relevance in application-oriented works, text clustering and text indexing application may be benefited from approaches like this.

Data Augmentation techniques inevitably generate bigger sets of data and working with high-dimension vectors or datasets inevitably poses some restrictions in how to deal with it. But whenever possible, it is important that new techniques be developed because the results are increasingly having more impact in real lives.

We believe that our work can help to keep the debate going about the different machine bias. Non-binary genders as well as racial biases have largely been ignored in NLP and this may allow to perpetuate social struggles.

Gender bias in NLP is a compound issue, requiring interdisciplinary communication, teaming with social scientist and overall Computer Science must start to ask questions as NLP systems have been increasingly integrated with our daily life because it carries real consequences for people that are using the NLP developments.

References

1. Hellström, T., Dignum, V., Bensch, S.: Bias in machine learning what is it good (and bad) for? (2020)
2. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: A case study with google translate. Neural Computing and Applications, pp. 1−19 (2019)
3. Kurpicz-Briki, M.: Cultural differences in bias? Origin and gender bias in pre-trained german and french word embeddings. In: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS) (2020)
4. Nissim, M., van Noord, R., van der Goot, R.: Fair is better than sensational: Man is to doctor as woman is to doctor (2020)
5. Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., Chang, K.W.: Gender bias in contextualized word embeddings (2019)
6. Crawford, K.: The trouble with bias. Keynote at Neural Information Processing Systems (NIPS‘17) (2017)
7. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Neural Information Processing Systems (NIPS‘16) (2016)
8. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning gender-neutral word embeddings. In: Empirical Methods of Natural Language Processing (EMNLP‘18) (2018)
9. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like bi-ases. Science, 356(6334), pp. 183–186 (2017)
10. Miller, G.: WordNet: A lexical database for english. Communications of the ACM, 38(11), pp. 39−41 (1995)
11. Fellbaum, C.: WordNet: An electronic lexical database. Cambridge, MA: MIT Press (1998)
12. Miller, G.: WordNet: A lexical database for english. In: Communications of the ACM, 38(11), pp. 39−41 (1995)
13. Jurafsky, D., Martin, J.H.: Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. Upper Saddle River, Prentice Hall, pp. 664−682 (2008)
14. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: North American Chapter of the Association for Computational Linguistics (NAACL‘18) (2018)
15. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Gender bias in coreference resolution: Evaluation and debiasing methods. In: North American Chapter of the Association for Computational Linguistics (NAACL‘18) (2018)
16. May, Ch., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. In: North American Chapter of the Association for Computational Linguistics (NAACL‘19) (2019)
17. Park, J.H., Shin, J., Fung, P.: Reducing gender bias in abusive language detection. In: Empirical Methods of Natural Language Processing (EMNLP‘18) (2018)

18. Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., Saxena, M.: Analyze, detect and re- move gender stereotyping from bollywood movies. In: Conference on Fairness, Accountability and Transparency (FAT‘18), pp. 92–105 (2018)
19. Schiebinger, L.: Scientific research must take gender into account. Nature, 507(7490), pp. 9−9 (2014)
20. Faulkner, S.L., Trotter, S.P.: Data saturation. The International Encyclopedia of Communication Research Methods, pp. 1−2 (2017)

Assessment of Wiener Method Applied to Filtering of Bio-Ultrasonic Signals

Carlos Alberto López Hernández, Ivonne Bazán Trujillo, Alfredo Ramírez García

Centro de Ciencias de la Ingeniería,
Departamento de Ingeniería Biomédica,
Universidad Autónoma de Aguascalientes,
Mexico

{ivonne.bazan, alfredo.ramirez, al223348}@edu.uaa.mx

Abstract. Ultrasound has a varied field of applications, especially in the evaluation of biological tissues providing important information of the analyzed structures. Whether in one dimensional signals or images, in Ultrasound is common to find noise in them. The Linear FIR Wiener method proposed in this work is used to denoise the bio-ultrasonic signals simulated considering a wide range of white noise. Evaluation parameters such as SNR and FFT were used to assess the denoising method effectiveness applied to different levels of noise. Results showed consistent decreased noise, and a SNR enhancement.

Keywords: Wiener filtering, signal processing, ultrasound, SNR, FFT.

1 Introduction

Ultrasound is widely used as a noninvasive technique for obtaining structural information of materials, especially in biological tissues. The main modalities used are A-scan, a single dimension signal showing echoes in a timeline; and B-scan, a series of A-scan signals forming an image [1].

As in any other acquisition system, in ultrasound is common to find along with the desired signal, unwanted components that make it hard to analyze or process the signal, those components are commonly known as noise [2]. The noise can vary from signal to signal, even when obtained under similar conditions and its origin may be diverse like from electromagnetic interference, a power line interference acoustic noise, etc. For ultrasound signals and images, the most common kind of noise found is structural noise, generated by the interference of several echoes originated in small structures inside the analyzed material. For A-scan evaluations, that kind of noise is known as back scattering and as speckles in the case of B-Scan evaluations [1, 2].

The methods of filtering ultrasound signals are varied, from non-adaptive FIR an IIR filters, wavelet transform [1, 3], denoising method based on matching pursuit [4] and several forms of Wiener filtering. [5, 6] In this article the filtering method applied is a linear FIR Wiener method, looking to eliminate most of the white noise present in ultrasound signals, and to evaluate the effectiveness considering different levels of

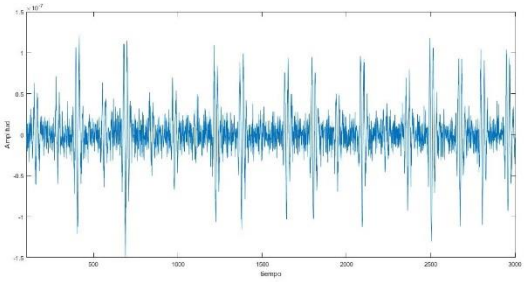


Fig. 1. Simulated signal with an SNR of 5.93dB.

noise. One of the advantages of working with Wiener Filter is that it is not required to know the frequency components of the desired signal and of the noise, such as in traditional techniques e.g. Butterworth or Chebyshev filters. In Wiener filtering the signal and the noise are considered stochastics processes, therefore the solution is obtained using statistical analysis, estimating the best filter response [7].

2 Methodology

2.1. Simulation of Data

The simulation of the ultrasound signal from biological tissue, used in this work, is made by the superposition of single echoes. These echoes are generated by the interaction of an ultrasonic emitted pulse and the tissue characterized by the presence of dispersive behavior. These structures can be associated with cells, blood vessels and any tissue with dispersive properties.

Using the model described in [8], biological tissues can be described as a series of scatterers separated by as distance "d". When the pulse mentioned above travels through this path of scatterers the result is several single echoes from each scatterer, which summed, depending on the distance each echo travels, can be described accordingly to:

$$x(t) = \sum_{m=1}^{n} A_k S_k \left(t - \left(\frac{2p_k}{v} \right) \right), \quad (1)$$

where A_k is the echo amplitude caused by the k^{th} reflector, S_k (.) is the shape of that echo caused by the k^{th} reflector, t is time, p_k is the position vector of the k^{th} reflector and v is the speed of sound.

In this case, the ultrasonic pulse mentioned before is modeled accordingly to:

$$u(t) = -te^{-4\sigma^2 t^2} \sin(2\pi f_c t), \quad (2)$$

where σ is the bandwidth, and f_c is the central frequency of the transducer.

The white noise was calculated by an algorithm based on the signal power. These noises were calculated to obtain an SNR between 1 dB and 120 dB and then added to the simulated signals. 120 signals with different levels of noise were obtained to be processed using the wiener filter. In Fig. 1, an example of simulated signal with an SNR=5.93dB is shown.

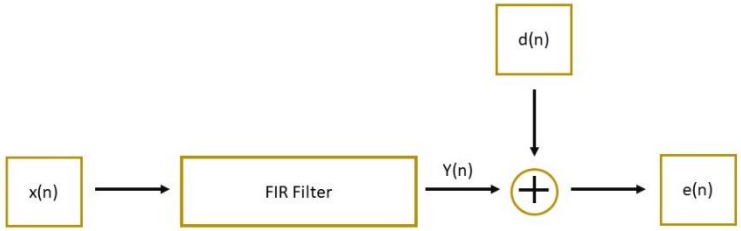


Fig. 2. Diagram representing the working of the Wiener filter.

2.2. Wiener Filter

Wiener filtering uses properties of a signal that is considered to have the characteristics desired in the data to be processed, and the properties of the noise added to the signal, considering both as linear and stochastic processes [7, 9].

Considering a finite impulse response (FIR) filter, where *x(n)* is the signal corrupted with noise, and G_k are the coefficients of the filter, the output *y(n)* of the filter can be defined as:

$$y(n) = \sum_{k=0}^{N} G_k \, x(n-k). \tag{3}$$

The objective of the Wiener algorithm is to find the most adequate value to the coefficients to minimize the error between the desired signal *d(n)* and the filtered signal *y(n)* [9]. In Fig. 1 a diagram explaining that logic can be observed.

Calculating the optimal coefficients is achieved through the Wiener-Hopf equation, obtained according to [7]. This equation involves the autocorrelation, *r*, of the input signal, and the cross correlation, *p*, of the desired and the input signals [7]:

$$\sum_{j=0}^{a} G_k \; r(i-k) = p(-k). \tag{4}$$

Expressed in vectorial terms we have:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(N-1) \\ r(1) & r(0) & \dots & r(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(N-1) & r(N-2) & \dots & r(0) \end{bmatrix} \begin{bmatrix} G_0 \\ G_1 \\ \vdots \\ G_{N-1} \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_{N-1} \end{bmatrix}. \tag{5}$$

2.3. Data Analysis

Two parameters are used to analyze the results. The first is the Fast Fourier Transform (FFT), which is applied to the signals prior and post filtering, and then compared.

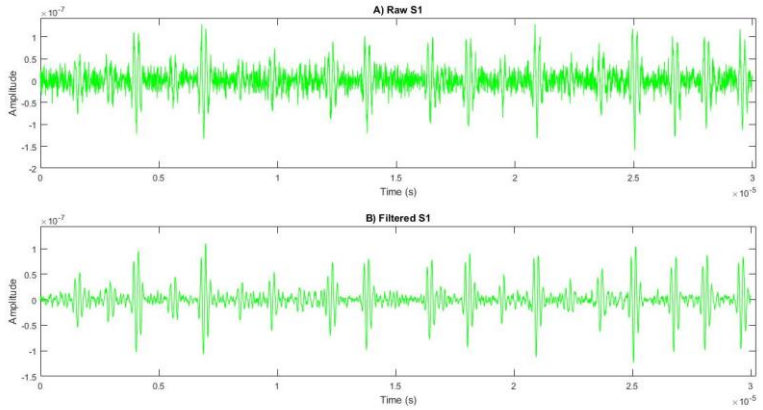


Fig. 3. (a) Original bio-ultrasonic signal, (b)Bio-ultrasonic filtered signal.

The second parameter is the signal to noise ratio (SNR), which in a similar way to the FFT is obtained prior filtering and post filtering and later the data is compared. The SNR was calculated according to:

$$SNR = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right), \tag{6}$$

where P_{signal} and P_{noise} are the energy of the signal and the noise, respectively and were obtained following the equation:

$$P = \frac{\sum_{i=0}^{N-1}(X_i)^2}{N}, \tag{7}$$

where P is the energy, N is the number of samples in the signal and X_i is the i-th sample of the signal.

3 Results and Discussion

From the 120 simulated bio-ultrasonic signals, those corresponding to a SNR between 1 and 119 dB were filtered using the Wiener method, the signal with the SNR of 120dB was used as the desired signal, here the result data of 1 of those signals is shown. Fig. 3 shows the data of the signal before and after filtering. Fig. 4 shows the FFT of the signals in Fig. 3.

Averaging the FFT Amplitudes and comparing them between the cases before and after the filtering, result in a reduction of the average energy of 53% for S-1, a 40% for S-5, 24% for S-10 and 2% for S-25. Meanwhile, for S-49, S-73, and S-97 the reduction is close to 0%, an indicator that the level of noise in those signals is minimal which allows to suppose that the applied Wiener Filter does not affect the frequency band of interest and performs an efficient elimination of components mainly in the undesired frequencies band. This aspect is relevant in the conditioning stage of the bio-ultrasonic signals.

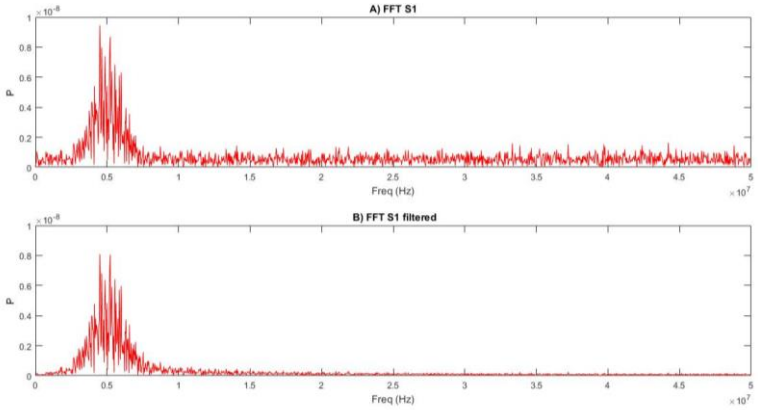


Fig. 4. a) FFT of the original bio-ultrasonic signal, b) FFT of the bio-ultrasonic filtered signal.

Table 1. Energy reduction in the FFT for the signals in different frequency intervals.

Signal	Overall	0-10MHz	10MHz-50MHz
S-1	53%	14.9%	80%
S-5	40%	5.1%	74%
S-10	24%	1.2%	65%
S-25	2%	0%	30%
S-49	0%	0%	0%
S-73	0%	0%	0%
S-97	0%	0%	0%

Table 2. SNR parameter results before and after filtering of the signal.

Signal number	Initial SNR	Final SNR
S-1	3.8403	9.8757
S-5	8.1045	13.3407
S-10	13.1333	17.1846
S-25	28.0637	29.6697
S-49	52.0623	52.0480
S-73	75.9167	75.7463
S-97	100.0126	99,8557

Now making the same analysis but separating the frequency bands, the results show that for S-1dB from 0 to 10Mhz the reduction of energy in FFT is 14.9% and for the rest of the spectrum is 80%.

Similar results are found for S-5, S-10, and S-25. For S-49, S-73, and S-97 the changes are minimal for the same reasons explained before. Results in detail are shown in Table 1.

Regarding the second parameter, the signal to noise ratio, it was obtained before and after filtering the signals. The SNR data can be found in Table 2. If we compare SNR results these show that for the first three signals, the increase of the SNR is considerable. For the S-1 the increase is 157%, for the S-5 is 64.6% and for S-10 is 30.84%.

Meanwhile for S-25 the increase is merely 5%, and in the rest of signals the SNR practically stays the same, showing that the applied filter is more efficient for small to medium values of SNR than for large values.

Comparing the results obtained with those of methods such as denoising based on wavelet frames[3] and on matching pursuit[4], a similar behavior is found regarding the SNR improvements: To a higher SNR, the improvement is worse than those with lower SNR.

Regarding the numerical improvement of the SNR, results show similar improvements to denoising with wavelet methods, for example: for a signal with an original SNR of 5dB, the Wiener filter shows a 5dB improvement while in [3], depending on the order of the wavelet SNR improvements vary from 4 to 6dB; for a signal with an original SNR of 10dB the Wiener Filter shows an improvement of 4dB while the wavelets show improvements from 2 to 5 dB for that same original SNR.

Making the same comparison showed above for the matching pursuit method[4] it is clear that the method in [4] provides better SNR improvements since for a 4dB original SNR level it shows an 8dB improvement while the Wiener filter shows an improvement of 6dB. Other example is that for a 10dB original SNR, [4] shows a 7dB improvement while the Wiener filter shows a 6dB improvement.

Although there are better results found with other more complex methods, it is important to highlight that the Wiener Filter still provides good results while being a less complex method. And when compared with other more traditional methods of filtering Wiener Filter still provides the advantage that it is not required to know the frequency components of the desired signal and of the noise, such as in Butterworth or Chebyshev filters.

4 Conclusions

There are some important conclusions to be drawn from these results, the first one is the Wiener Filtering technique used in this article observed good results when applied to eliminate the white noise in simulated ultrasound signals. The results showed that the filter eliminates a considerable amount of the frequency components corresponding to the noise, leaving those components where the ultrasound signal information is.

The second is that in those signals where from the beginning there were not much noise, the signal obtained after filtering remains practically the same, which was also confirmed from the analysis of the FFT.

The third is that though the increase in SNR percentage is considerable for the signals with a higher level of noise, it still is nowhere near the SNR of the signals with lower level of noise. Thus, for those signals with the higher level of noise it is possible that not all the noise was filtered.

References

1. Pardo Gómez, E.: Transformadas Wavelet no diezmadas para reducción de ruido y detección de señales: Aplicaciones en evaluación no destructiva por ultrasonidos. Universidad politécnica de Valencia (2011)

2. Granados-Castro, L.: Método de procesamiento de señal para la evaluación ultrasónica de dimensiones anatómicas del globo ocular. Universidad Autónoma de Aguascalientes (2020)
3. Zhang, Y., Wang, Y., Wang, W., Liu, B.: Doppler ultrasound signal denoising based on wavelet frames. In: IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency control, 40(3), pp. 709−716 (2001)
4. Zheng, Y.: Denoising of quadrature ultrasound Doppler signal from bi-directional flow based on matching pursuit. Ultrasonics, 49(1), pp. 19−25 (2009)
5. Izquierdo, M.A.G, Hernández, M.G., Graullera, O., Ullate, L.G.: Time-frequency Wiener filtering for structural noise reduction. Ultrasonics, 40, pp. 259−261 (2002)
6. Neal, S., Speckman, P., Enright, M.: Flaw signature estimation in ultrasonic nondestructive evaluation using the Wiener filter with limited prior information. In: IEEE Transactions On Ultrasonics, Ferrolectrics And Frequency Control, 40(4), pp. 344−353 (1993)
7. Cervantes A.: Filtros Wiener. Polibits, 9(1), pp. 19 (1998)
8. Bazan, I., Ramos, A., Calas, H., Ramirez, A., Pintle, R., Gomez, T.E., Negreira, C., Gallegos, F.J., Rosales, A.J.: Possible patient early diagnosis by ultrasonic noninvasive estimation of thermal gradients into tissues based on spectral changes modeling. Hindawi Publishing Corporation (2012)
9. Manju, B.R., Sneha, M.R.: ECG Denoising using Wiener filter and Kalman filter. Procedia Computer Science, 171, pp. 273−281 (2020)

Phases of a Cryptographic Protocol for Body Area Networks in a Medical Application

Kevin A. Delgado Vargas[1], Gina Gallegos-Garcia[1,2], Fernando Hernández Pérez[3], Gualberto Aguilar Torres[4]

[1] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Laboratorio de Ciberseguridad,
Mexico

[2] Instituto Politécnico Nacional,
Escuela Superior de Cómputo,
Mexico

[3] Tecnológico de Monterrey Campus Santa Fe,
Departamento de Computación, División de Ingeniería y Ciencias,
Mexico

kdelgadov1200@alumno.ipn.mx, ggallegosg@ipn.mx,
hernandezf@live.com.mx, gualberto.aguilar@tec.mx

Abstract. Currently there are different applications that allow to maintain secure communications within medical applications. In this kind of applications, heterogeneous devices interact with each other within a body area network to transmit private patient information. This paper presents the design of five phases of a cryptographic protocol that focus on confidentiality and authentication. The design of such phases can be an option to use in medical applications that consider the use of heterogeneous devices. Proposed protocol uses a key encapsulation mechanism (KEM) scheme and a key derivation function (KDF) to obtain a symmetric key. Both of them together with the digital signature and encryption schemes are used to achieve security services aforementioned.

Keywords: Body area network, cryptographic protocol, medical approach.

1 Introduction

Doctors, hospitals and health systems maintain communication with patients using different devices in which different data are processed. These data can be general of the patients such as: Name, disease, treatments or the information that is handled in these kind of systems. Depending on the disease that the patient presents, data that can be obtained from different sensors, such as: Motion or glucose sensors.

Some of the diseases that must be monitored require access to reliable information, because people's lives are at stake. Therefore, any attack on the system would endanger people who use it. Cryptographic protocols help to secure information transmitted in these systems [1]. Moreover, they minimize the amount of trust required between each participating entity [2]. For these reasons, a cryptographic protocol that can resist different attacks is required. To achieve this, it is necessary to carry on through a structured interaction between different cryptographic primitives, which are strategically located, so that all the characteristics of the participating entities can be used.

This paper presents the design of four phases of a cryptographic protocol for receiving and sending information through heterogeneous devices and some of them with limited resources like used in medical applications.

2 Related Work

Within the state of the art, there are different works in which protocols and cryptographic schemes are used to add security to medical applications. From the secure communications point of view, there are different solutions that start from the use of a single cryptographic algorithm to encrypt information to more complete solutions such as: Cryptographic protocols. All these solutions can be classified into those that focus solely on improving schemes, others that focus on the medical system and those that focus on improving or defining a protocol.

2.1 Solutions Focused on Cryptographic Schemes and Systems

In 2000 in [3], an authentication scheme was shown. It was based on Hwang and Li scheme [4] with the main difference that it used smart cards and an one way function. Authors performed two tests on Hwang's system and on their own to make a comparison between both of them. They made a security and efficiency analysis. Results showed that the proposal provides greater security compared to the original.

In 2010 in [5], different techniques were proposed and used to monitor patients effectively. They focus on provide confidentiality to patient information. It was achieved by using asymmetric encryption schemes and a key agreement, such as: RSA [6] and a key agreement based on electrocardiogram signals called EKG (electrocardiogram) -based key agreement (EKA) [7]. It was applied to all body area sensors network. It was proposed to the medicine scenario. It had two main entities that communicate with each other: the body sensors and the doctor.

In 2014 in [8], an authentication system was proposed. It was based on the Rabin algorithm [9]. In this system, public key cryptographic algorithms were used to provide authentication in a Wireless Body Area Network (WBAN). Used schemes in this system were lightweight because the authors tested their system in devices with constrained resources. Proposed system had as application scenario a sensors network and actuators in the body of patients with different diseases. The system had 4 main entities: sensor, coordinator node, doctor and

the actuator. They interacted with each other to execute the phases of the system.

The work in [10] was carried out in 2015. It proposed an improvement to the scheme proposed by Lu *et al.'s* [11]. It was an authentication scheme for the Telecare Medical Information System (TMIS). However, the authors showed that the scheme was not resistant to different attacks, such as: Replay attack, impersonation attack, privileged insider attack, man-in-middle attack and off line password guessing attack. That's the reason why they proposed some improvements to Lu's work. As result, the scheme could resist all those attacks made before. Since the application scenario is the TMIS, this proposal only presented two entities, the patient and the TMIS server.

2.2 Solutions that Focus on Cryptographic Protocols

In 2015 in [12], a lightweight key management protocol was presented. This protocol established a secure communication between a node with limited resources and a remote server. This protocol provide integrity, authentication and confidentiality. It was distributed in 5 main phases: Initial exchange, securing connections between parties, proving third parties's representativeness of constrained node (CN) or the sensor to unconstrained node (UN) or the remote server, secret generation and delivery and termination phase. It worked in applications of *E-health* and worked with 3 main entities: the constrained nodes, the unconstrained nodes and the third parties. The information traveled from the nodes of the patient's body to a server and finally reached a third entity.

In 2019 in [13], an authentication protocol was presented, which used the protocol proposed by Das *et al.* [14]. It was designed to be used in TMIS scenarios. The protocol had 4 participating entities: Doctor, a physical server, patient and the medical record server. This protocol provided at least 11 interactions between entities, where different messages were sent. Their proposed protocol had 9 phases: Setup, medical-server registration, physician-server registration, patient registration, login, authentication and session key negotiation, new physician-server addition, password renewal, and biometric renewal. In the same way, the authors showed a security analysis of their protocol. 6 cases of seizures were shown and discussed.

In 2020 in [15], the authors proposed a protocol for implantable medical devices (IMDs). The protocol facilitated the security services of confidentiality, integrity, non-repudiation, access control and user authentication. The protocol had 3 main entities that interact each other: User smart card (C), reader (R) and implant (I). The protocol had 4 main phases: $R \leftrightarrow C$ mutual authentication, user authentication, session-key (K'_{RI}) establishment and the main phase.

3 Body Area Network Devices

The different devices that are currently used in various fields or applications are compatible with x86 and ARM processors. On the one hand the x86 processor

is used principally on desktops and portables computers. It is a processor that is based on a type of architecture such as the high-end (able to run a full fledged operating system). This type of processors are capable of performing a large number of processes but having a high energy cost. Something that in principle is not a problem in computers that are permanently connected to the electricity grid.

On the other hand, ARM processors are based on the low-end architecture. They perform a small amount of instructions, reason why they are faster than high-end based devices [16]. Their small size and low power consumption make them perfect for use in small devices. This means that they are ideal for tablets or mobile phones. In addition to that, there are also some devices that are only for reading information, for example, those that produce a digital signal from an analog or mechanical signal. Examples of them are body or biometric sensors.

In fact, one of the application scenarios where is possible to observe this type of devices is in the medical scenarios, with the well-known sensor networks, since they allow monitoring the status of patients through the Wireless Body Area Networks (WBAN) [17, 18]. These networks involve different sensors that are interconnected with each other and placed in the human body. Fig. 1 shows the graphic representation of a WBAN.

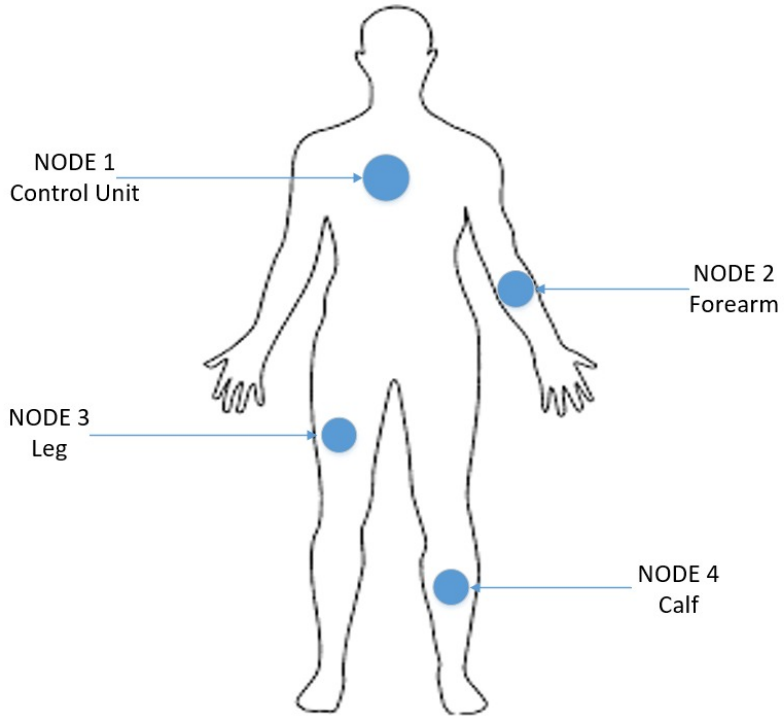


Fig. 1. Distribution of reading sensors within a Wireless Body Area Network.

4 Design of the Phases in our Protocol

4.1 Participating Entities

The design of the 5 phases in our protocol considers 4 main entities: Server, coordinator node, sensors and actuators. They maintain communication in the following way:

1. *Server* denoted by A will be in charge of receiving all the data. Such data will be reviewed by the doctor.
2. *Coordinator node* denoted by B is the intermediary between what the sensors read and what is sent to the server.
3. *Sensors* denoted by C are all those devices that collect specific data depending on their function.
4. *Actuators* denoted by D are those that receive the instructions coming from the server, which they will work with.

4.2 Cryptographic Description of the Phases of our Protocol

Entities listed in Section 4.1, are all of them presented in a WBAN with medical application. In this type of application, all data transferred on the network must be protected since the patient's life is at stake. According to the aforementioned, the phases in our protocol focus on authentication and confidentiality. Fig. 2 shows the sequence diagram of the phases of the protocol. In such figure is possible to observe the messages sent between each of the entities as well as the functions that each one of them performs.

Each of the phases shown in the diagram are specified in the following subsections. They describe the cryptographic primitives that are going to be needed in each one of them, as well as some of the inputs and outputs that should be gotten from each phase.

Phase 0 - Setup. In this phase, the registration, the security level agreement and the system parameters are carried out. All sensors must be registered with the coordinator node and the latter must do the same with the server. In this way, the server will be the one that stores the public key of the coordinator node and this in turn will have the public keys of all the sensors.

Phase 1 - Key Generation. In this phase, the different keys that are used in the following phases of the protocol must be generated. For this reason, different variables are needed to key generation, that is why it generally has 3 values as input: p and q are prime numbers and g is a generator. These values are previously agreed by the entities. The first key to obtain is the symmetric key for the encryption scheme. This key must be agreed through a Key encapsulation mechanism (KEM) scheme. Such scheme must be executed between the server (entity A) and the coordinating node (entity B). The process

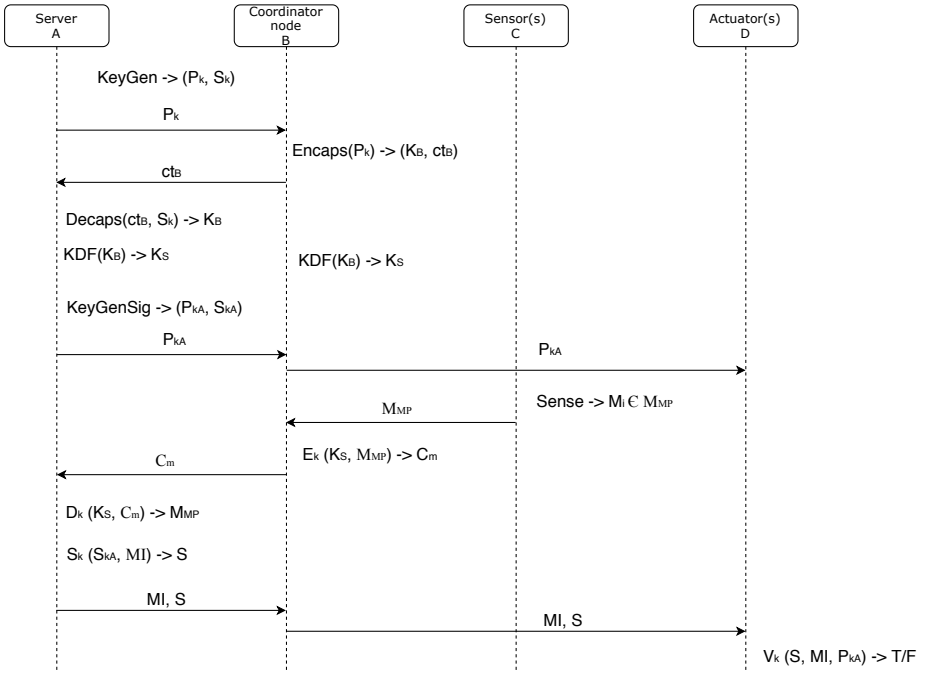


Fig. 2. General diagram of the interaction between the entities and the phases of the protocol.

begins with the $KeyGen$ function of the KEM scheme, this function will return a key pair (P_k, S_k), where P_k is sent to B. With the received key B uses the $Encaps(P_k)$ function that will return two values: K_B and ct_B. From these values, the value ct_B must be sent back to A. It will be used as input in the function $Decaps(S_k, ct_B)$, which will output K_B. This value should be the same value that B get. Since both entities have the same value, both must use a Key Derivation Function ($KDF(K_B, Label, Context, L)$) where $Label$ is a string that identifies the purpose of the KDF, $Context$ is a binary string that contains information related to K_B, and L is the length of the key that will be obtained at the output. All of them in order to get the same symmetric key K_s.

Subsequently, the Server (entity A) must generate a key pair (P_{kA}, S_{kA}) through the $KeyGen$ function of the digital signature scheme. In this way it will be possible to maintain the authentication of the entity. Once the key pair is gotten, the P_{kA} is sent to the coordinator node (entity B) and will be in charge of distributing it to the different actuators (entity D). Fig. 3 represents the input to the phase and obtained output.

Phase 2 - Sensing, encryption and sending of information. Once all the sensors (entity C) have been registered and the keys have been generated, phase 2 begins. In this phase, C collects all the patient's data. The input of the phase

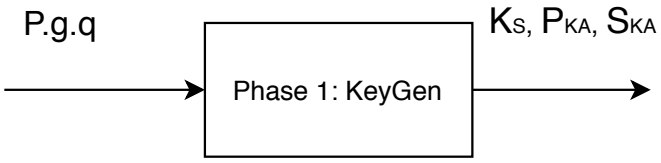


Fig. 3. Variables needed to obtain as output the keys that will be used in the following phases.

is the data that can be shown on the left side of the Fig. 4. For example, it could be data from a cardiac signal or patient movement data. The data is collected per unit of time, shown as m_i, in such way that the complete message that is produced as the output of the phase is denoted by M_{MP}, which is the set of data, such that $M_{MP} = m_0, m_1, m_2, m_3, ..., m_n$. When C has the complete message, it is sent to entity B.

The second input of the phase is the symmetric key K_s. This key is used to encrypt the data M_{MP} with the function E_k, which delivers the encrypted message C_m, where $E_k(K_s, M_{MP}) \rightarrow C_m$. It can be seen in Fig. 4. After that, C_m is sent from entity B to entity A. The sending of this data marks the end of phase 2 and begins phase 3 of the protocol.

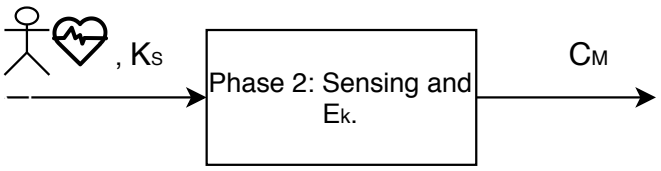


Fig. 4. Phase 2 takes as input the measurements delivered by the sensors and the symmetric key from the previous phase.

Phase 3 - Information decryption, signing and sending instructions. The input and output of phase 3 can be seen in Fig. 5, where the inputs are: The symmetric key K_s, the private key S_{kA}, the encrypted message C_m and a message with medical instructions MI, where MI depends on what the doctor considers based on the decrypted data. Firstly, K_s and C_m are used as input to the decryption function, where $D_k(K_s, C_m) \rightarrow M_{mp}$. In this way entity A obtains the original message that was generated by entity C.

Secondly, once M_{mp} is known, the doctor must generate MI. This data is signed with the function S_k of the digital signature scheme. The function can be viewed as $S_k(MI, S_{kA}) \rightarrow s$. The output of the function gives a value s, which is the digital signature. It is sent together with the value MI to entity B. Finally, after B receives this information, it is responsible for sending it to entity D which

corresponds. When entity D receives signature and instructions, phase 3 ends and phase 4 begins.

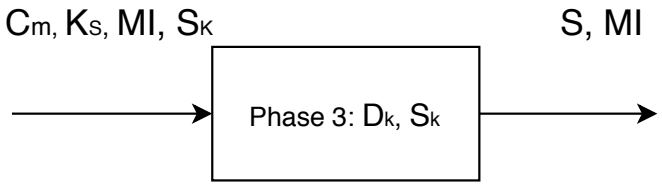


Fig. 5. Phase 3 input is made up of the encrypted information, the symmetric key, a public key of the target actuator and the instructions that it must execute.

Phase 4 - Verification of the signature and instructions. The last phase of the protocol begins by taking as input parameters the signature s, the public key of entity A (P_{kA}) and the medical instructions MI, as shown in Fig. 6, the output of this phase is determined by the verification function V_k of the digital signature scheme. Therefore $V_k(s, MI, P_{kA}) \rightarrow T/F$ is applied, the output is a boolean value, which determines if the signature is authentic. In case the value is true, the instructions received must be applied by the actuators on the patient's body. In case the signature is not valid, then entity D will not apply any instructions. The output shown in Fig. 6 indicates an action by entity D. In this case represented by a syringe.

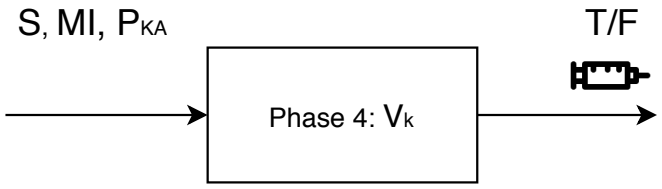


Fig. 6. The input of the phase 4 is formed by the set of instructions and the digital signature associated with these instructions, as well as the server's public key.

5 Conclusions

In this paper the design of five phases that make up a cryptographic protocol for Body Area Networks in a medical application were described. For the design, not only the interaction that the entities will have is taken into account, but also the use of cryptographic schemes. Taking into account that the application scenario

is the wireless body area network (WBAN) in a medical field, the requirements of such scenario were also analyzed. Some of them are: The necessary bandwidth, the sizes of the packets that are going to be sent in the protocol, as well as the format that they must have. Moreover, due to in medical field the heterogeneous devices are used, we do not have to forget they have enough capacity to perform what corresponds to them. It could give us the opportunity to estimate energy consumption required by the phases in our protocol.

It is well known that times in medicine are essential, therefore, the protocol must have the shortest execution times without neglecting safety. For all the above reasons, the use of different cryptographic algorithms to test the protocol could be take as future work. It could include a comparative of performance between classical and postquantum algorithms. This last one in order to analyze if they are capable of working on heterogeneous devices for Body Area Networks in medical applications.

Acknowledgments. The authors thank the Instituto Politécnico Nacional for the support granted for the realization of this work through the projects SIP 20201754.

References

1. Menezes, A., van Oorsschot, P., Vanstone, S.: Handbook of applied cryptography. CRC press (1996)
2. Ferguson, N., Schneier, B.: Practical cryptography. 141, New York: Wiley (2003)
3. Hung-Min, S.: An efficient remote use authentication scheme using smart cards. IEEE Transactions on Consumer Electronics, 46(4), pp. 958–961 (2000)
4. Hwang, M.S., Li, L.H.: A new remote user authentication scheme using smart cards. IEEE Transactions on Consumer Electronics, 46(1), pp. 28–30 (2000)
5. Ren, Y., Werner, R., Pazzi, N., Boukerche, A.: Monitoring patients via a secure and mobile healthcare system. IEEE Wireless Communications, 17(1), pp. 59–65 (2010)
6. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM, 21(2), pp.120–126 (1978)
7. Venkatasubramanian, K.K., Banerjee, A., Gupta, S.K.S.: EKG-based key agreement in body sensor networks. In: Proceedings INFOCOM Wksps (2008)
8. Thaier, H., Athanasios, V., Ghada, A., Bassam, M., Muhammad, I., Muhammad Zeeshan, S., Khalid, Q.: Public-key authentication for cloud-based WBANs (2014)
9. Rabin, M.: Digitalized signatures and public key functions as intractable as factorization. Massachusetts Institute of Technology, Reading, Massachusetts (1979)
10. Ashraf Chaudhry, S., Mahmood, K., Naqvi, H., Khurram Khan, M.: An improved and secure biometric authentication scheme for telecare medicine information system based on elliptic curve cryptography. Springer Science Business Media New York (2015)
11. Lu, Y., Li, L., Peng, H. et al.: An enhanced biometric-based authentication scheme for telecare medicine information systems using elliptic curve cryptosystem. Journal Med. Syst., 39, pp. 32 (2015)

12. Abdmeziem, M., Tandjaoui, D.: An end-to-end secure key management protocol for e-health applications. Computers & Electrical Engineering, 44 (2015)
13. Amin, R., Islam, S.H., Gope, P., Choo, K.R., Tapasn, N.: Anonymity preserving and lightweight multimedical server authentication proto-col for telecare medical information system. IEEE Journal of Biomedical and Health Informatics, 234, pp. 1749–1759 (2019)
14. Das, A.K., Odelu, V., Goswami, A.: A secure and robust user authenticated key agreement scheme for hierarchical multi-medical server environment in TMIS. Journal Med. Syst., 39(9) (2015)
15. Siddiqi, M.A., Doerr, C., Strydis, C.: Imdfence: Architecting a secure protocol for implantable medical devices. arXiv preprint arXiv:2002.09546 (2020)
16. Rouse, M.: What is ARM processor? - Definition from Whatis.Com. https://whatis.techtarget.com/definition/ARM-processor (2015)
17. Latre, B., Braem, B., Moerman, I., Blondia, C., Demeester, P.: A survey on wireless body area networks. Wireless Networks, 17(1), pp. 1–18 (2011)
18. Ullah, S., Higgins, H., Braem, B., Latre, B., Blondia, C., Moerman, I., Saleem, S., Rahman, Z., Kwak, K.S.: A comprehensive survey of wireless body area networks. Journal of Medical Systems, 36(3), pp. 1065–1094 (2012)

Intelligent Time Use Suggestions for Wellbeing Enhancement

Mario E. Marin, Julio C. Ponce

Universidad Autónoma de Aguascalientes,
Centro de Ciencias Básicas,
Mexico

marin@ia.org.mx, jcponce@correo.uaa.mx

Abstract. An intelligent application is developed to automatically suggest to users alternative time assignations to activities, based on data analysis of INEGI's ENUT (national time use survey), with the purpose of enhancing wellbeing. The predictive value of the datasets was ascertained with a comparative analysis of feedforward deep neural networks, support vector machines, logistic regressions and random forest assembles. Logistic regression used the less computational time, and the random forest assembles had the best accuracy. Respondents and users were profiled using K-means clustering, and a non-linear optimization model was developed to find the best datapoints to take suggestions from.

Keywords: Intelligent systems, data analytics, wellbeing, time use, clustering, optimization.

1 Introduction

Time can be considered a scarce resource "whose use largely determines the progress, achievement and wellbeing of individuals, families, communities and societies" [1], therefore managing it to put it to good use should be a priority for individuals and organizations. Despite most time management applications being focused on improving work performance, there is ample evidence that time management practices have a positive correlation with individual wellbeing [2]. However, the research in this field is scarce and distributed among many disciplines [3] and a review of the current applications resulted in very few automated or Artificial Intelligence (AI) approaches, with this field dominated by automated scheduling tools, and some tools for health assessments and wellness. No AI or automated tool was concerned with first suggesting what kind of activities, beyond the preexisting ones, could enhance wellbeing for the user except for tools specifically centered on wellness or health that can hardly take the place of current time management tools in organizations or workplaces where people spend most of their waking hours.

The question this work intended to answer was whether, given an adequate time use dataset, intelligent tools could be developed that automatically consider alternative activities in multiple areas of life from data of similarly positioned individuals in terms of roles and responsibilities, with a wellbeing enhancement criterion.

Using the datasets from a national time use survey that included subjective wellbeing data, the data was analyzed, and using a select subset of attributes and a clustering algorithm, 30 personal time use profiles were developed of which a user could be considered a member.

Then, a program was developed to suggest users to what activities they could give more or less time in their weekly routine to make a low-cost change in their membership from their original group to a similar one with a greater wellbeing average level. This was accomplished without resorting to data from most costly and cumbersome approaches like sensors and surveillance software.

The value of using time use survey datasets was proven by answering the question posited above in the affirmative; and by the success with predictive algorithms in estimating, from time use and demographic data, subjective attributes such as satisfaction and objective ones like gender or income.

These valuable time use datasets can not only be obtained from surveys, but also from other low cost, low nuisance methods like preexisting software calendars and Internet of Things (IoT) devices that can record the activities done in different time intervals.

2 Wellbeing Theory

A broad definition of wellbeing would be "an individual's ability to live the kind of life he or she values, on a sustainable basis" [4]. Wellbeing is a multifaceted concept, usually modeled by way of constructs that include both objective and subjective measures of wellbeing. It is difficult to obtain comprehensive datasets of objective wellbeing measures such as those indicating health, income, skills, job situation, and even characteristics of the environment in which a person lives.

Even then, a further problem remains: weighting each measure according to the subjective importance a user may give them. Subjective measures of wellbeing can be more convenient.

Subjective wellbeing is a "broad category of phenomena that includes people's emotional responses, domain satisfaction and global judgments of life satisfaction" [5] and it has been found to have positive correlations to various measures of objective wellbeing [6–8].

This makes subjective wellbeing a powerful proxy for overall wellbeing, and its being a self-reported measure [9] that has been shown to respond according to relevant circumstances [10], despite some of its components having correlation with personality traits [11], facilitates the generation of useful datasets.

The choice of time use datasets is because of the definition of wellbeing as an ability that people may have in relationship to the environment in which they act. Therefore, an agency based definition of wellbeing in relation to time use is more adequate [12].

From this perspective, a clearer definition of wellbeing is that it is the state in which people can act while gaining greater ability to engage with their environments. Given the complexities of studying environmental and interpersonal factors of wellbeing, it is hoped that behavior of people obtained through time use datasets, and subjective wellbeing evaluations can offer insights into wellbeing enhancement in general.

3 Wellbeing Analytics

The analysis approach in this work is that of data analytics, which is a set of "theories, technologies, tools and processes that enable an in-depth understanding and discovery of actionable insight" [13]. In the context of wellbeing, the datasets available and necessary for an automated approach and for many kind of analyses are such that the implications of the model of the 5 Vs of big data must be considered: volume as in very big datasets; variety as in different kinds of data of which time use surveys are just a few slices; velocity, with many data even generated in real time from large groups of people; value [14], which in the case of wellbeing analytics must be linked to an agency based definition of wellbeing; and veracity, as in underlying accuracy of the data specifically impacting the ability to derive actionable insights and value out of it [15].

While this work is centered in time use datasets obtained through surveys, the same kind of data can be obtained from IoT devices, portable sensors [16], digital traces [17], and from data generating activities. Thought the interest of this work is mostly centered in data generated with a low cost, low nuisance approach, hence one-time surveys, the insights of this project can be carried over to analysis of other varieties of data.

This work can be considered as one of data analytics, and the analyses in the next sections as different levels of data analytics [14]. From the descriptive, with the correlation analyses, to the predictive with a comparative analysis of different methods of prediction; and to the prescriptive in which from the very beginning the user is aided by all the information and knowledge accumulated in the system developed.

4 Datasets and Methods

The datasets used for this work are from the microdata repository of the National Time Use Survey in Mexico performed by its National Institute of Statistics and Geography (INEGI) in 2014 (ENUT-2014) [18]. Respondents were asked about their weekly time use, meaning, their time allocations to different activities during a week; as well as some demographic data. This time use survey was the first in Mexico to also ask about subjective wellbeing. Eight ENUT-2014 datasets with data from Spanish speaking people were preprocessed, transformed, and unified into a single dataset.

The ENUT-2014 datasets had 58,187 entries considering 42,117 people who took the complete time use survey, 1,914 people not living in the same house but who donated their time in it, and data about other people who did not answer the time use survey as they were less than 12 years old or had a disability that prevented it. The 42,117 complete respondents were from among 15,500 homes or family units living in 15,058 houses, with some of these housing more than one home or family unit. Data from different sections of the survey went into different datasets as we see in Table 1.

The subjective wellbeing attribute "happiness" is the one being enhanced in this work, because of its simpler description and high correlation to some of the other options, but the model can work with any subjective wellbeing attribute in the dataset. Reindexing, ordering, transformation, and fusion procedures were made in the datasets to create a single unified dataset with 42,117 entries and 322 attributes which nonetheless had all the relevant data from the other datasets included in each entry. A diagram of this process can be seen in Fig. 1.

Table 1. Datasets from the ENUT-2014.

Dataset	Survey section	Data from	Entries	Features
TVivienda	I	Houses	15,058	24
THogar	II, III, VIII	Home or family units	15,500	51
TSDem	III	Sociodemographic data	56,273	7
TModulo1	IV	Time use	42,117	182
TModulo2	V, VII	Time use	42,117	192
TModulo3	VI, VII	Time use and subjective wellbeing	42,117	178
TNoResidente	VIII	Nonresidents	1,914	33
Metadatos	I-VIII	Entries	58,187	4

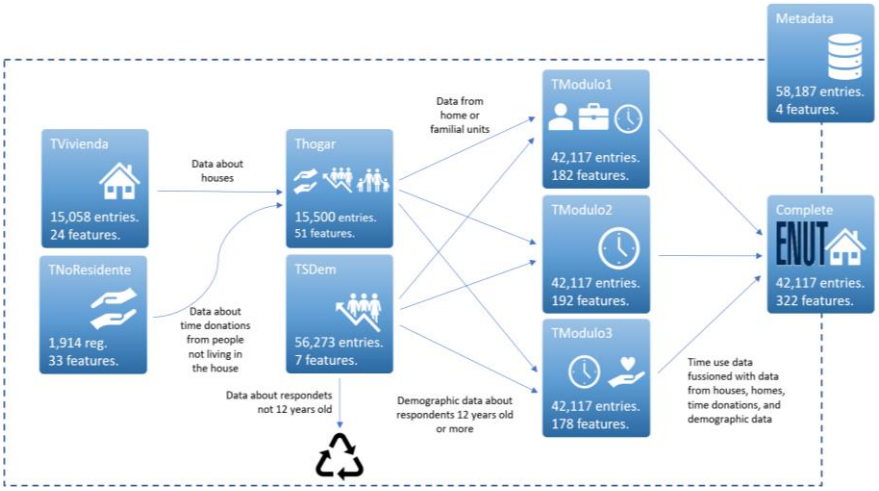


Fig. 1. Data preprocessing process.

In order to ascertain the value contained in the unified dataset, four different predictive algorithms were used to predict subjective wellbeing and other relevant attributes using the unified ENUT-2014 dataset and compared: feedforward deep neural networks, support vector machines (SVM), random forest assembles, and logistic regression.

The visualization technique for high dimensional data FreeViz [19] was used to evaluate the structure of the unified dataset and the K-means clustering algorithm [20] was used to create different clusters that could represent groups of people with similar time use assignations.

A non-linear optimization model was presented with two different objective functions for different approaches: change in minutes, and cost of these changes, whether they are increments or decrements.

Table 2. Predictive analysis comparison.

Attribute	Method	Accuracy	Weighted F1	C. Time (s)
Satisfaction with studying time use* (Values: 0,1)	Random forest	**0.9602**	**0.9601**	014.56
	SVM	0.8962	0.8961	016.15
	Neural Net	0.9158	0.9158	101.67
	Log. Regression	0.9130	0.9130	**000.55**
Satisfaction with work time use* (Values: 0,1)	Random forest	**0.8146**	**0.8082**	057.02
	SVM	0.7823	0.7800	254.73
	Neural Net	0.7698	0.7693	636.34
	Log. Regression	0.7471	0.7467	**001.37**
Satisfaction with life in general* (Values: 1-5)	Random forest	0.2448	0.2337	004.43
	SVM	**0.2750**	**0.2549**	001.57
	Neural Net	0.2467	0.2346	014.06
	Log. Regression	0.2957	0.2906	**000.15**
Income (Values: 1-5)	Random forest	**0.5716**	**0.5725**	004.71
	SVM	0.4159	0.3852	002.59
	Neural Net	0.3964	0.4022	015.4
	Log. Regression	0.4145	0.4061	**000.21**
Gender (Values: 0,1)	Random forest	**0.8872**	**0.8872**	070.57
	SVM	0.8302	0.8295	373.28
	Neural Net	0.8569	0.8569	850.42
	Log. Regression	0.8595	0.8594	**001.81**
Indigenous self-identification (Values: 0,1)	Random forest	**0.6604**	**0.6602**	079.32
	SVM	0.5710	0.5645	514.21
	Neural Net	0.5659	0.5656	736.62
	Log. Regression	0.5950	0.5949	**001.77**

*All subjective wellbeing attributes were dropped from the training dataset.

5 Predictive Value of the Dataset

As previously stated, in practice applications using time use datasets would fall into the big data analytics field; therefore, it is relevant to ascertain the value of the dataset [14] before using it in an application. The approach decided was to apply four different predictive algorithms to the dataset and evaluate their performance in predicting relevant attributes.

All subjective wellbeing attributes were dropped when predicting a subjective wellbeing attribute as some are highly correlated. This analysis was performed with an

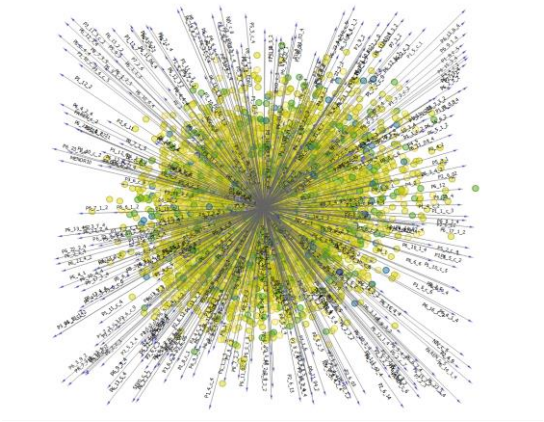


Fig. 2. FreeViz visualization of the completely unified dataset of the ENUT-2014.

Anaconda Python distribution in a laptop with an Intel Core i5-9300H processor, and 32GB in RAM.

Some relevant results are in Table 2, where the performance metrics confirm the value of the unified ENUT-2014 dataset. In almost every case, random forest assembles have the better accuracy, while logistic regression uses the less computational time. Because the dataset is unbalanced, undersampling was used taking an equal, or close to equal, number of entries for every class.

As can be seen, classifiers for satisfaction in specific domains related to time use in work and academic settings have high performance, whereas satisfaction in general just manages to be above a random classifier.

Other general subjective wellbeing attributes suffer from similar low performance too, and other domain specific time use satisfaction measures without as much related times use attributes as work and study have hover between 0.6 and 0.7 accuracy which is still notable for attributes with 5 classes.

It is also notable the high accuracy predicting gender and income, and a moderate one predicting indigenous self-identification which points to time use and demographic patterns linked to these attributes.

6 Time use Profiles Creation

Using the unified dataset, a FreeViz [19] visualization was generated, and it was clear from it that the dataset had no easily separable structure as can be seen in Fig. 2. Given this lack of structure, the K-means algorithm was used to partition the dataset in 30 clusters with the intention of having them represent groups with centroids well distanced from other groups' centroids and relatively little variance within groups which this clustering technique algorithm [20] is useful for.

Also, given the success of predictive algorithms with sociodemographic attributes, it was judged prudent to do a correlation analysis of the dataset with the class assigned by the K-means algorithm. If the groups were mainly defined by their

Table 3. First eight Pearson correlation of all attributes with the K-Means class.

Corr.	Attribute
0.818	Income by means of pension
0.294	User/respondent is retired
0.277	There are elderly people in need of special care at home.
0.253	Income by means of rent of some property
0.169	House has a landline.
0.157	Number of rooms in the house.
0.149	Age
0.142	Kitchen has a sink.

Table 4. First eight Pearson correlation of all attributes with happiness.

Corr.	Attribute
-0.15	Age
0.146	Toilet has running water.
0.144	Hours of checking e-mail, social networks, or chat apps.
0.141	Kitchen has a sink.
0.141	House has Internet connection.
0.141	There is a laptop computer in the house.
0.13	There is washing machine in the house.
0.128	Respondent has a car or pick-up truck.

sociodemographic attributes or others not easily changeable, this would mean that most suggestion to pass from one group to another would be problematic, impractical, or implausible.

6.1 Correlation Analysis

Various correlation analysis, using Pearson correlation as endorsed by the literature of wellbeing and quality of life studies [21], were done for the dataset and the class of the clusters generated. The first using all the attributes of the dataset showed that attributes with the highest correlation to the class of the group were indeed sociodemographic ones, such as the ones seen in Table 3. More worryingly, a correlation analysis with the happiness attribute shed light in that most of the attributes with high correlation with happiness were about having appliances and basic services available at home, age and just one time use attribute, as seen in Table 4.

If groups were allowed to be constructed in this way, the suggestions the application would give the user would be to change attributes like those shown in Table 3 and Table 4 and others like them, which would be of limited value in its context.

Table 5. First eight Pearson correlation of time use attributes and K-Means class.

Corr.	Attribute
0.13	Time taking care of kids while doing something else (weekdays)
0.12	Time taking care of kids while doing something else (weekends)
0.079	Time dedicated to eating meals (weekdays)
0.078	Time dedicated to work (weekends)
0.077	Time dedicated to commutes (weekdays)
0.075	Time dedicated to work (weekdays)
0.068	Time dedicated to cleaning the interior of the house
0.067	Time taking care of babies or toddlers (weekdays)

6.2 Profile Generation

To avoid generating unusable groups for current purposes, they were generated only considering time use attributes. The correlations of the attributes of these groups with the class assigned by the clustering algorithm were analyzed and the attributes with high correlation to the class are time use assignations as can be seen in Table 5.

These attributes tend to distinguish different people with different roles; for example, parents or children's tutors do have big time use assignations of childcare, whereas people without children do with far less regularity. The same happens with students having time assignments for classes and homework, and so on. This is the kind of groups that were sought: people with similar roles and responsibilities.

6.3 Characterization of Profiles and Users

As can be seen in Table 6, the 30 groups generated have been characterized by happiness level, age group, work situation and the distance they have to travel to their work or school, and whether they have children under their responsibility or not. The average of each attribute was used as per practices used by the Organization for Economic Cooperation and Development (OECD) [22] and then converted to a label.

It is to be noted that some users may be assigned, or suggested to change, to a group with certain roles and responsibilities that do not exactly match their own. However, as the groups were calculated using only time use attributes, this only means that some users may have similar time use assignments to those of people with different roles or may have "something to learn" from people with different roles or responsibilities.

7 Optimization Model

So far, the model has users grouped with other similarly positioned people in terms of roles and responsibilities as indicated by their time use assignations. The issue remains of how to suggest to users a low-cost change in their weekly activities that is backed by the data to have a probable wellbeing enhancement.

Table 6. Characterization of groups, 5 selected attribute.

Group	Happiness	Age	Children	Work	Commutes
1	Low	Medium	No	Part time	Far
2	High	Y. Adults	Yes	Full time	Very far
3	Medium	Y. Adults	No	Full time	Very far
4	High	Teenagers	No	No	Very close
5	High	Y. Adults	Yes	No	Close
6	Medium	Y. Adults	No	Extra hours	Very far
7	Low	Medium	No	No	Very close
8	High	Y. Adults	No	Full time	Ver far
9	High	Y. Adults	Yes	No	Close
10	High	Teenagers	No	No	Very close
11	Low	Medium	No	No	Ver close
12	Low	Medium	No	No	Close
13	Low	Medium	No	Few hours	Close
14	High	Y. Adults	No	Full time	Very far
15	Medium	Medium	No	No	Close
16	High	Y. Adults	Yes	No	Very close
17	Low	Medium	No	No	Very close
18	High	Teenagers	No	No	Very close
19	High	Teenagers	No	Part time	Far
20	Medium	Medium	No	Part time	Very far
21	High	Y. Adults	No	Full time	Very far
22	Medium	Medium	No	Part time	Far
23	High	Medium	No	Full time	Very far
24	High	Y. Adults	Yes	Full time	Far
25	Very low	Medium	Yes	Few hours	Close
26	Medium	Y. Adults	No	Extra hours	Very far
27	Medium	Medium	No	No	Very close
28	Very high	Teenagers	No	No	Very close
29	High	Y. Adults	No	Extra hours	Very far
30	Medium	Y. Adults	No	Full time	Adjacent

There are two approaches. To minimize the changes suggested in minutes assigned to certain activities, and to minimize the costs of such increases or decreases. The cost of increasing a particular activity may be different from that of decreasing it, and big costs to decrease can be associated to activities deemed essential by the user.

These costs would need to be asked to the user. Therefore, an optimization model is presented that has two options for objective function: one for change and one for costs of change.

For both approaches, the easiest path would be to suggest the time use assignation of a close datapoint with better wellbeing level in the same group as the user. This would have the advantage of taking the suggested changes from a person similarly positioned in terms of roles and responsibilities. However, there is no reason to think that just recommending another person's time use assignments, even if that persons has better wellbeing levels, will do something to improve wellbeing levels for the user.

A better path is to have information about a group of thousands of people with similar time use patterns linked to a greater average wellbeing level than that of the user, then we would have reason to recommend a datapoint within this group. Provided that the user does not have a maximum level of wellbeing already, this is exactly what the model shown in this work already provides. The process would be to find the closest datapoints, using Manhattan distance, with greater wellbeing level that belong to groups with greater average wellbeing than the current level of the user. This set of datapoints then can be presented ordered as recommendations to the user to decide between them.

The optimization model is proposed thus:

Indexes:

$$i \in \{1, \dots, n\},$$

n: Number of groups.

$$j \in \{1, \dots, m\}.$$

m: Number of activities considered.

Sets

Dataset composed of n groups:

$$D = \{G_1, \dots, G_n\}.$$

Parameters:

Current time use assignment to activity *j*:

$$u_j \in \mathbb{N}_o.$$

Current wellbeing level of user/decision maker:

$$f \in \mathbb{N}_o.$$

Cost/Value of increasing time assignment to activity *j:*

$$a_j \in \mathbb{R}.$$

Cost/Value of decreasing time assignment to activity *j:*

$$b_j \in \mathbb{R}.$$

Time use assignment for the centroid of group *i* in activity *j*:

$$c_{ij} \in \mathbb{R}_0^+.$$

Average wellbeing of group *i*:

$$h_i \in \mathbb{R}_0^+.$$

Original group to which the user/decisor was assigned to:

$$g \in \{1,2, \dots n\}.$$

Variables:
Time use assignment recommended for activity j taken from data from group i:

$$x_{ij} \in \mathbb{N}_o.$$

Recommended group to switch to:

$$y_i \in \{0,1\}.$$

Restrictions:
A group must be selected:

$$\sum_{i=1}^{n} y_i = 1 .$$

Time use assignments for all activities must not exceed minutes in a week:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \leq 10080.$$

Recommended time assignments for all activities must be a datapoint belonging to the group being considered:

$$\{x_{i1}, x_{i2}, \dots, x_{im}\} \subseteq \{G_i | y_i = 1\}.$$

Objective function for minimizing change:

$$C(y_i, x_{ij}) = \sum_{i \in \{1,\dots,n | f < h_i\}} \sum_{j=1}^{m} y_i |x_{ij} - u_j|.$$

Objective function for minimizing cost:

$$C(y_i, x_{ij}) = \sum_{i \in \{1,\dots,n | f < h_i\}} \sum_{j=1}^{m} y_i \left\{ \frac{a_j}{2} [sign(x_{ij} - u_j) + 1](x_{ij} - u_j) + \frac{b_j}{2} [sign(u_j - x_{ij}) + 1](u_j - x_{ij}) \right\}.$$

8 Results

The correlation and predictive analysis done demonstrate the value contained in the ENUT-2014 dataset, not only as it provides insights about differences in time use by different demographic groups, but also about their relationship to subjective wellbeing. Therefore, it is possible to extract knowledge about how time use patterns affect

Table 7. Suggestions summary for 5 randomly selected respondents.

# ID	Change in minutes	Group change	Wellbeing difference	Greatest increase	Greatest decrease
40902	9308	8 → 5	+0.214304	Taking classes	Caring for adults
	9323	8 → 5	+0.214304	Taking classes	Caring for kids*
	9494	8 → 5	+0.214304	Taking classes	Caring for adults
7815	4725	12 → 5	+0.214304	Taking classes	Growing food
	4935	12 → 5	+0.214304	Taking classes	Growing food
	4990	12 → 5	+0.214304	Taking classes	Growing food
5091	5679	15 → 9	+0.165179	Caring for kids*	Collecting firewood
	5729	15 → 9	+0.165179	Caring for kids*	Cooking
	5929	15 → 9	+0.165179	Caring for kids*	Collecting firewood
99	4275	0 → 5	+0.214304	Sleep, weekdays	Homework
	4445	0 → 5	+ 0.214304	Sleep, weekdays	Listening to audio
	4795	0 → 4	+0.301020	Sleep, weekend	Listening to audio
27055	6250	17 → 25	+0.168627	Caring for kids*	Tend to livestock
	6315	17 → 25	+0.168627	Caring for babies*	Tend to livestock
	6340	17 → 5	+ 0.214304	Taking classes	Tend to livestock

*Activity done while doing something else.

subjective wellbeing. The clustering, and subsequent characterization of the groups of people similarly positioned in terms of roles and responsibilities that dictate their time use patterns, allow the use of this knowledge to develop applications to enhance wellbeing by recommending behavior changes with no prior data of the user before the input of its weekly time use data, as well as some sociodemographic data.

The optimization model was solved for randomly chosen datapoints within the ENUT-2014 dataset by exhaustive search to guarantee the results for this test, and to obtain a set of ordered recommendations of theoretically actionable suggestions to change time use assignation to activities, as can be seen in Table 7 with three options presented to five respondents from the ENUT-2014.

Also, by asking the user information to generate a vector of costs for increasing or decreasing time assignations, the model can take into consideration activities that the user thinks are more easily changeable or those that are essential, thus including information about the agency limitations of the user. As the ENUT-2014 does not contain data of this kind, no attempt was done to use the model to minimize cost. But as we see in Table 7, some users could label tasks such as “caring for adults” or “caring for kids” as essential and therefore assign high costs for reducing them. This would alter the changes suggested by considering the agency limitations of the users such as those that preclude a particular change in time use that would enhance subjective

wellbeing because, overall, the capacity to change environment and conditions is limited, and thus overall wellbeing is limited as well.

9 Conclusions, Discussion, and Future Work

There are two main interests that directed this work. The first was to stablish the possibility and value of working with time use datasets in the context of developing intelligent applications to enhance wellbeing, despite established links of some subjective wellbeing components to personality [11]. If wellbeing enhancement applications can work with behavior change instead of personality traits, the task of creating these intelligent applications would be easier, as behaviors in the context of organizations, but also individually, can create data automatically that can more easily be recovered to form the datasets to power the applications. Also, data obtained from behaviors is less subject to distorted interpretations. The conclusion is that this work is indeed possible and valuable, and a data analytics approach is useful, particularly in the context of big datasets that can lead to prescriptive analysis automated applications that aid users backed by knowledge obtained from these datasets.

The other interest is about finding sources of data that are low cost, low nuisance to users. As we have seen, useful time use datasets can be created with one-time surveys like the ENUT-2014, but they can also be created by time management applications already in wide use in organizations as well as by the increasing number of everyday objects that are part of the IoT ecosystem. This means that in many organizations and households there is already a wealth of data waiting to be used for applications of wellbeing enhancement and time management.

Future work can consist of obtaining new time use datasets with a more streamlined set of attributes, datasets that include costs of increasing or decreasing each time commitment to an activity in order to apply and study the results of the optimization model which takes into account the cost of the changes. It remains to be seen if new indicators of wellbeing can be inferred from patterns in the time use and subjective wellbeing datasets that link to the definition of wellbeing based on agency and if these result in better suggestions. Also, the agency of the users, their capabilities to change their time use and wellbeing levels if they decide so, should be given a bigger role in the model and have the impact of these modifications ascertained. Linking an application like the one presented to a time management one is also an option, as well as analyzing other varieties of data, such as emotional analyses in text and images.

Finally, care should be taken to treat the datasets of time use as already describing organizational and societal biases that might need to be addressed before an application can be in use. Some of these biases can be inferred from the prediction algorithms used having high accuracy levels for linking time use patterns to sociodemographic attributes such as gender or indigenous self-identification, as well as to income.

References

1. Ironmonger, D.: Time use. The new palgrave dictionary of economics. Palgrave Macmillan UK, pp. 13676–13681 (2018)
2. Aeon, B., Aguinis, H.: It's about time: New Perspectives and insights on time management.

Academy of Management Perspectives, 31(4), pp. 309–330 (2017)
3. Claessens, B.J.C., van Eerde, W., Rutte, C.G., Roe, R.A.: A review of the time management literature. Personnel Review, 36(2), pp. 255–276 (2007)
4. Karacaoglu, G., Krawczyk, J.B., King, A.: Introduction and Overview. In: Intergenerational Wellbeing and Public Policy: An Integrated Environmental, Social, and Economic Framework, Springer Singapore, pp. 3–26 (2019)
5. Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L.: Subjective well-being: Three decades of progress. Psychological Bulletin, 125, pp. 276–302 (1999)
6. OECD: Why does subjective well-being matter for well-being? in How's Life?: Measuring well-being. OECD Publishing, pp. 266–267 (2011)
7. Diener, E., Ryan, K.: Subjective well-being: A general overview. 39(4), pp. 391–406 (2008)
8. Diener, E., Biswas-Diener, R.: Happy people function better. Oxford: Blackwell Publishing, pp. 27–88 (2008)
9. Diener, E.: Subjective well-being. Psychological Bulletin, 95(3), pp. 542–575 (1984)
10. Lucas, R.E.: Long-term disability is associated with lasting changes in subjective well-being: Evidence from two nationally representative longitudinal studies. Journal of Personality and Social Psychology, 92(4), pp. 717–730 (2007)
11. Emmons, R.A., Diener, E.: Personality correlates of subjective well-being. Personality and Social Psychology Bulletin, 11(1), pp. 89–97 (1985)
12. Granqvist, N., Gustafsson, R.: Temporal institutional work. The Academy of Management Journal, 59(3), pp. 1009–1035 (2016)
13. Cao, L.: The data science era: In data science thinking the next scientific, technological and economic revolution. Springer International Publishing, pp. 3–28 (2018)
14. Chen, M., Mao, S., Zhang,Y., Leung, V.C.M.: Introduction in big data: Related technologies, challenges and future prospects. Springer International Publishing, pp. 1–10, (2014)
15. Pendyala, V.: The big data phenomenon in veracity of big data: Machine learning and other approaches to verifying truthfulness, Berkeley, pp. 1–15 (2018)
16. Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schüssler-Fiorenza Rose, S.M., Perelman, D., Colbert, E., Runge, R., Rego, S., Sonecha, R., Datta, S., McLaughlin, T., Snyder, M.: Digital health: Tracking physiomes and activity using wearable biosensors reveals useful health-related information. PLoS Biology, 15(1) (2017)
17. Luhmann, M.: Using big data to study subjective well-being. Current Opinion in Behavioral Sciences, 18, pp. 28–33 (2017)
18. INEGI: Encuesta nacional sobre uso del tiempo (ENUT) 2014. Programas INEGI, 2015 (2019)
19. Demšar, J., Leban, G., Zupan, B.: FreeViz-An intelligent multivariate visualization approach to explorative analysis of biomedical data. Journal of Biomedical Informatics, 40(6), pp. 661–671 (2007)
20. Jin, X., Han, J.: K-Means Clustering. Encyclopedia of Machine Learning. In: Sammut C., Webb, G.I. (Eds.) Springer, pp. 563–564 (2010)
21. Ark, N.: Zero-order relationships. Encyclopedia of Quality of Life and Well-Being Research. In: Michalos, A.C. (Ed.), Springer Netherlands, pp. 7311–7313 (2014)
22. OECD: Overview, in how's life? Measuring well-being. OECD Publishing, pp. 13–36 (2011)

Sistema computacional aumentativo y alternativo de comunicación con interfaz pictográfica dinámica

Oscar Alejandro Delgadillo Martinez[1], Ismael Díaz Rangel [2], Alejandra Morales Ramírez [1], Cuauhtémoc Hidalgo Cortés[1], Alejandro Andrés Serapio Carmona[2]

[1] Universidad Autónoma del Estado de México, México

[2] Universidad Nacional Autónoma de México, Facultad de Estudios Superiores Aragón México

otarito@gmail.com, {alejandra_25_22,ismael1099, andreserapio}@hotmail.com, chidalgoc@uaemex.mx

Resumen. Existen padecimientos ocasionados por accidentes o enfermedades neurodegenerativas que ocasionan una discapacidad del habla y motora, afectando significativamente la comunicación del paciente, ocasionando un decremento de su calidad de vida. Los mejores métodos para solventar este problema son muy costosos y los métodos tradicionales son económicos, pero requieren de una asistencia para generar sus oraciones, además de ser métodos de comunicación lentos. Debido a esto, el siguiente trabajo presenta el desarrollo de un sistema aumentativo y alternativo de comunicación utilizando una microcomputadora, que muestra una interfaz pictográfica, y como entrada utiliza un interruptor mecánico. El sistema se conforma con diversas plantillas de pictogramas que, al ser seleccionadas, permiten al paciente formar oraciones, y estas son emitidas a través de un sintetizador de voz.

Palabras clave: Sistemas aumentativos y alternativos de comunicación, esclerosis lateral amiotrófica, pictograma.

Augmentative and Alternative Communication Computer System with Dynamic Pictographic Interface

Abstract. There are conditions occasioned by accidents or neurodegenerative diseases that cause speech and motor disability, significantly affecting the patient's communication, causing a decrease in their quality of life. The best methods to solve this problem are very expensive and traditional methods are inexpensive, but require assistance to generate sentences, besides they are slow communication methods. Due to this, the following work presents the development of an augmentative and alternative communication system using a

microcomputer, that showing a pictographic interface, and as input uses a mechanical switch. The system has several pictogram templates and when they are selected, allow to the patient form sentences, and these are emitted through a speech synthesizer.

Keywords: Augmentative and alternative communication systems. amyotrophic lateral sclerosis. pictogram.

1. Introducción

Existen padecimientos ocasionados por accidentes o enfermedades que ocasionan una discapacidad del habla y motora, afectando significativamente la comunicación del paciente, algunos ejemplos de estos padecimientos son: la Esclerosis Lateral Amiotrófica (ELA), se estima afecta a nivel mundial 500,000 personas; de las cuales, en México existen 6,000 casos diagnosticados [3] [4]. Otra condición es el traumatismo craneoencefálico (TCE), se estiman anualmente 5.48 millones casos a nivel mundial [5].

Ambos padecimientos (ELA y TCE) pueden causar al paciente una escasa movilidad y trastornos del habla, provocando una comunicación deficiente en diversas actividades del día y un decremento en su calidad de vida [1].

Una solución es el uso de Sistemas Alternativos Aumentativos de Comunicación (SAAC), los cuales ayudan a satisfacer sus necesidades de comunicación, brindando un apoyo a los pacientes y familiares durante los procesos de recuperación [1].

Los SAAC comerciales de alta tecnología pueden implicar un elevado costo, por lo que no es un artículo accesible para la mayoría de los pacientes; es por esto, que esta propuesta consiste en el desarrollo de un sistema computacional de comunicación de bajo costo que implementa técnicas de navegación mediante pictogramas. El funcionamiento consiste en pictogramas agrupados en plantillas temáticas; un cursor se desplaza automáticamente por cada pictograma de la plantilla de manera cíclica, cuando este pasa por el icono deseado, el paciente debe seleccionarlo, lo que lleva a una siguiente plantilla para ir conformando una oración compleja, o puede seleccionar la opción de síntesis de voz para transmitir el mensaje.

2. Estado del arte

Existen trabajos que han propuesto diferentes SAAC, los cuales implementan tecnologías para facilitar la comunicación del paciente. En [6] se describe un sistema electrónico que consta de un teclado virtual y un cursor que se desplaza automáticamente, éste se detiene al detectar un movimiento en la comisura lateral del ojo, a partir de este movimiento se van conformando las oraciones mediante la selección de letras, al terminar de crear la oración se puede realizar una síntesis de voz, adicionalmente el proyecto cuenta con un prototipo domótico, el cual podría ser implementado en un futuro. El sistema está conformado por una microcomputadora, un sensor óptico TCRT 5000, una pantalla y bocinas.

Una desventaja de este sistema es que en la interfaz se presentan más de 90 elementos a seleccionar, lo que podría provocar un decremento en la velocidad para la

creación de oraciones y también un agotamiento visual; además de requerir una pantalla grande para visualizar sin problema los elementos.

Por otro lado, la investigación [2] ha desarrollado una aplicación móvil que simula el comportamiento de un cuadro E-Tran, el cual es un método muy utilizado para la comunicación de personas con ELA; esta aplicación utiliza la cámara del teléfono para detectar hasta seis diferentes comandos generados por la posición del ojo; además, la aplicación cuenta con un predictor de texto que agiliza la creación de oraciones y un sintetizador de voz.

Una de sus desventajas es la necesidad de un asistente, la detección del rostro falla si el paciente utiliza mascarilla de oxígeno, y esta aplicación solo está disponible para dispositivos con sistema operativo iOS; sin embargo, este sistema es rápido para la creación de oraciones.

El proyecto desarrollado está conformado por un sistema de navegación, basado en categorías pictográficas, agilizando la creación de oraciones específicas, este prototipo tiene la función de cubrir temáticas como: necesidades fisiológicas, malestares comunes (fiebre, alergia, comezón, dolor, etc.), higiene personal, confort dentro de una habitación y malestares o dolores en partes específicas del cuerpo, todas estas temáticas están pensadas para comunicarse con los médicos o familiares.

3. Sistemas aumentativos y alternativos de comunicación

Los sistemas aumentativos y alternativos de comunicación se pueden definir como estrategias creadas para personas con diferentes problemas del habla y motoras que permiten expresar sus necesidades y deseos de forma temporal o permanente, permitiendo mejorar su calidad de vida, autonomía y participación en la sociedad [7].

Estos sistemas se conforman de elementos gráficos como: fotografías, pictogramas, letras o palabras; sin embargo, cada sistema utiliza diversas técnicas para la interpretación o conformación de ideas, la elección de un sistema depende de sus ventajas y desventajas como puede ser la agilidad, composición o forma de empleo y costo. El método que se utilizó para este prototipo fue un "Sistema Pictográfico de Comunicación (SPC)"; este se basa en imágenes, las cuales tienen un significado específico y conforman oraciones utilizando la estructura "sujeto, verbo y complemento" (figura 1).

Su modo de empleo es fácil y se pueden comunicar ideas complejas utilizando pocos pictogramas; su principal desventaja radica en la cantidad de pictogramas que se deben utilizar para tener un sistema robusto en términos de cobertura de lenguaje.

4. Metodología

El sistema propuesto está conformado por diferentes elementos que se observan en el diagrama de bloques de la Figura 2.

Fig. 1. Ejemplo de una oración utilizando un SPC.

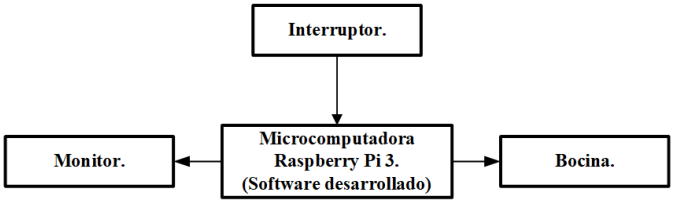


Fig. 2. Diagrama de bloques del sistema.

Microcomputadora: Este es el elemento principal, requiere de un sistema operativo, bibliotecas y algunos periféricos para permitir la interacción hombre- máquina.

Monitor: Se requiere un monitor con una resolución mínima de 720 puntos por pulgada y un conector HDMI para comunicarlo con la microcomputadora.

Interruptor: Es el elemento que permite la comunicación hombre-máquina, y debe consistir en un generador de flancos; para este trabajo se hizo la implementación con un *push button* que se conecta a la microcomputadora mediante los puertos de propósito general (GPIO). La ventaja de este método de interacción entre sistema y paciente es que se puede incorporar cualquier dispositivo que sea capaz de enviar flancos; por ejemplo, se puede acondicionar la señal de un sensor mioeléctrico para la interacción con el sistema.

Bocina: Este periférico tiene la función de emitir la oración que es procesada por el sintetizador de voz; si el monitor cuenta con bocinas integradas se puede omitir.

Interfaz pictográfica: El usuario podrá visualizar diversos pictogramas, opciones de navegación y un cuadro de texto para visualizar la oración creada.

4.1. Diseño de sistema

Para el diseño de navegación se realizó una búsqueda de diversos catálogos y materiales creados por la comunidad ARASAAC (Aragonese Center of Augmentative and Alternative Communication) [8]. Se recabaron un total de 170 pictogramas, 8 temáticas, 36 plantillas y 370 oraciones. Este conjunto de elementos se organizó en un mapa de navegación donde se indican las relaciones entre ellos. El mapa ayuda a comprender la conexión entre todos los elementos; así mismo, podría funcionar como guía para los usuarios para conocer las diferentes oraciones que se pueden crear, y la

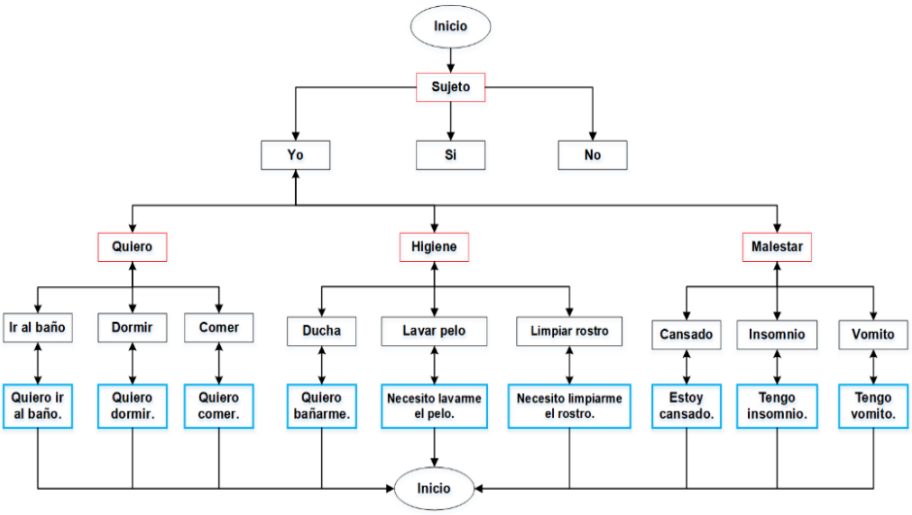


Fig. 3. Sección del diagrama de navegación.

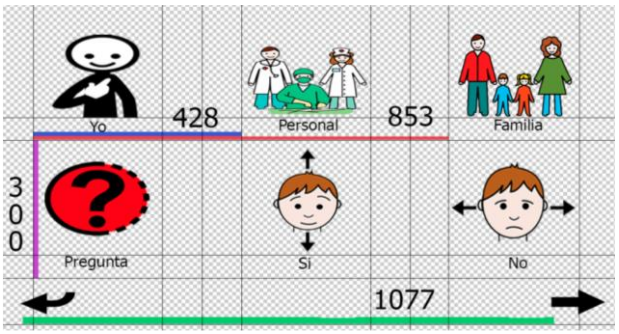


Fig. 4. Estructura base para la creación de plantillas.

secuencia de pictogramas a seguir para conformar las oraciones. En la figura 3 se muestra una sección del mismo.

El mapa está compuesto por tres elementos: temáticas (color rojo), pictogramas (color negro) y oraciones (color azul); estos elementos están ligados con la finalidad de crear oraciones. El mapa también permite conocer la posición previa, que es útil sí se desea regresar a una temática o pictograma diferente; también facilita la codificación, la creación de nuevas plantillas y su actualización.

La función de las plantillas es agrupar los pictogramas que pertenecen a una misma temática. Se diseñó una estructura base (figura 4) para generar un conjunto de plantillas homogéneas, éstas se conforman por un máximo de seis pictogramas acompañadas de una palabra para indicar su significado y flechas de navegación, que tienen la función de regresar a la plantilla anterior o mostrar una nueva que contenga más pictogramas de la misma temática.

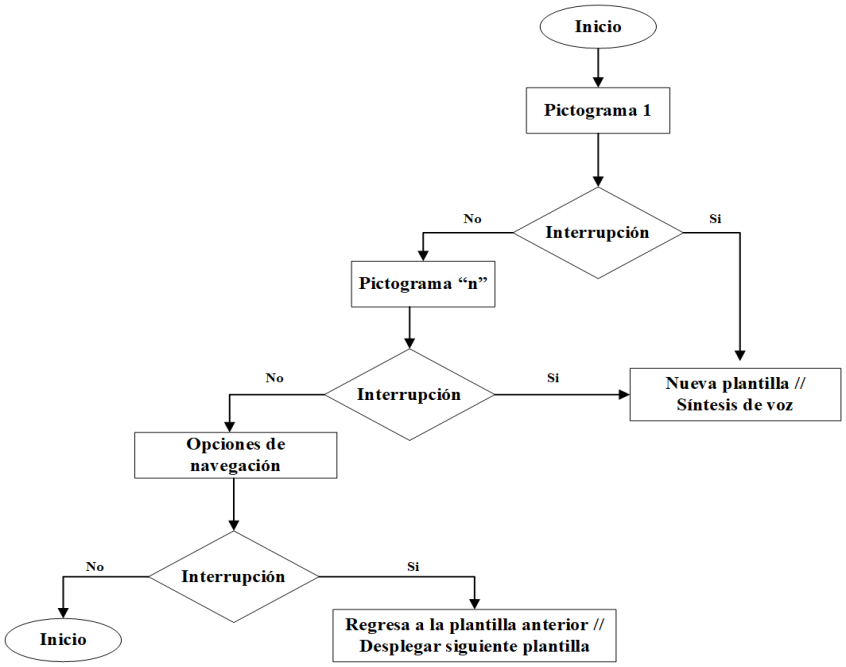


Fig. 5. Diagrama de flujo del cursor.

Por último, se muestra el diagrama de flujo (figura 5) que describe el comportamiento del cursor dentro de la interfaz gráfica.

El diagrama de flujo inicia posicionando el cursor en el primer pictograma de la plantilla, si el usuario realiza una selección mediante el push button (interrupción) el sistema hará una síntesis de voz o el despliegue de una nueva plantilla, estas acciones se realizarán siguiendo el diagrama de navegación; sin embargo, si el sistema no recibe una interrupción el cursor señalizará un nuevo pictograma. El proceso se repite hasta que el cursor señaliza el último pictograma, y si este no es seleccionado, el ciclo se repite, posicionando el cursor en el primer pictograma de la plantilla; este proceso se repetirá en cada plantilla del sistema.

4.2. Construcción

El sistema se ejecuta sobre una microcomputadora, la cual necesita de un sistema operativo, para este proyecto se eligió instalar Raspbian [9], se cargó en una memoria microSD clase 10 con una capacidad de 16 GB y se descargaron las bibliotecas y paquetes necesarios(sintetizador de voz de uso libre eSpeak) [10].

El hardware de entrada (push button) tiene dos terminales y se conecta a la Raspberry Pi utilizando un pin del GPIO y el pin de 3.3V, ya que el sistema está diseñado para responder a flancos de subida. Basándose en interruptores comerciales se diseñó una base para montar un push button de 100 mm de diámetro, 1.5 m de cable de dos polos y un conector jack 3.5 mm de 2 polos (figura 6).

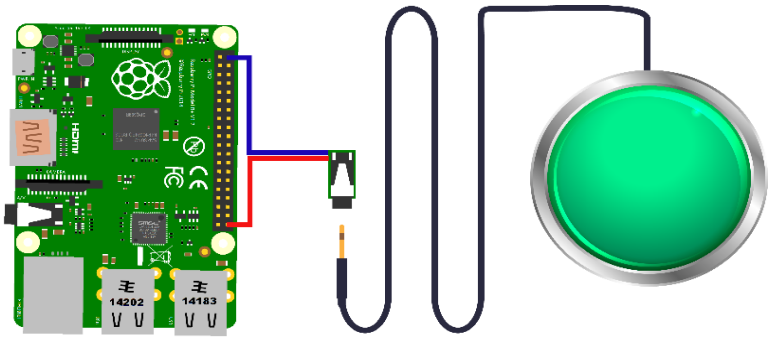


Fig. 6. Esquema del prototipo.

Tabla 1. Velocidad de escritura "Sistema pictográfico".

Participante	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Promedio [s]	42.6	35.3	34.8	29.8	39.8	110.3	120.3	62.2	55.3
Tiempo promedio en general por oración [s]:									65.6

5. Pruebas y resultados

Para comprobar el funcionamiento del sistema se realizó una prueba de usabilidad del mapa de navegación, la cual tuvo como propósito conocer si la distribución de los pictogramas es adecuada y entendible para los usuarios; otra prueba corresponde a la cobertura de oraciones, con el objetivo de conocer las limitantes del lenguaje del sistema con sus 370 oraciones incorporadas; por último, se presentan comparaciones de velocidad de escritura contra un sistema de comunicación tipo cuadro de E-Tran, con la finalidad de conocer la diferencia de velocidad de generación de oraciones respecto al sistema computacional propuesto.

Usabilidad del mapa de navegación

Para esta prueba se les solicitó a 9 participantes -sin experiencia previa en usar el SACC- conformaran 12 oraciones, las cuales fueron seleccionadas para cubrir las ramificaciones principales del sistema pictográfico. La tabla 1 muestra el tiempo promedio en segundos que los participantes demoraron en conformar las oraciones y el promedio general de los participantes en conformar una oración.

Como resultado de esta prueba, se observó que todos los usuarios lograron completar las oraciones, y considerando que cada una de ellos fueron elegidas para explorar las principales ramificaciones del sistema, así como la inexperiencia previa para el manejo por parte de los usuarios, se puede argumentar un buen diseño del mapa de navegación, logrando un sistema de manejo intuitivo; esto aunado a que el tiempo promedio de generación de una oración, es alrededor de un minuto en su primer acercamiento por parte de los usuarios, nos lleva a considerar que el sistema pictográfico tiene una buena usabilidad en términos de navegación para la generación de oraciones.

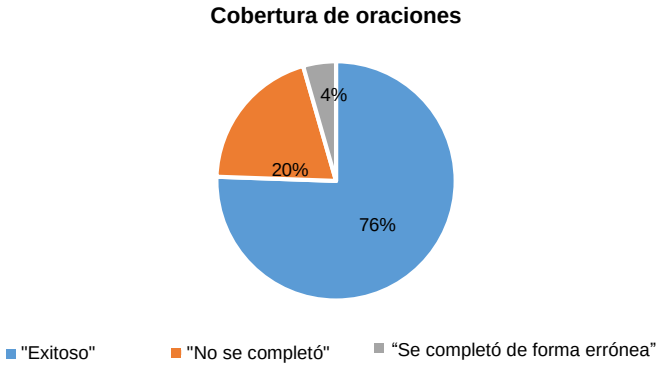


Fig. 7. Resultados de la prueba “Cobertura de oraciones”.

Cobertura de oraciones

La finalidad de esta prueba era determinar que tanto el sistema permitía a los usuarios generar oraciones de su elección. Para ello, se le pidió a cada uno de los 9 participantes sugerir 5 oraciones y se establecieron tres posibles resultados:

- Se completó con éxito; esto es cuando la oración es literalmente igual a la de la interfaz o que reflejar una idea equivalente.
- Se completó de manera errónea; esto es cuando se generó una oración, pero con una idea diferente a la indicada a la oración planteada.
- No se completó; el usuario no logró generar una oración.

La figura 7está dividida en tres secciones, cada una corresponde al porcentaje de los posibles resultados considerando las 45 oraciones propuestas en total.

Cualquier sistema pictográfico presenta limitaciones ya que no son capaces de producir un número infinito de oraciones a diferencia de las alfabéticas. Este sistema presentó una cobertura del 76%, respecto a las oraciones sugeridas por los usuarios, mismas que no tuvieron restricciones. No se logró generar de manera exitosa el 24% de las mismas.

Estos resultados indican que el sistema pictográfico tiene una buena cobertura de oraciones, aunque solo está conformada por 370 opciones en total; sin embargo, es tema de estudio analizar más a fondo la incorporación de nuevas plantillas y pictogramas para mejorar la cobertura, pero sin afectar la usabilidad. También es importante no perder de vista que este tipo de interfaz siempre tendrá limitaciones de cobertura.

Comparación de sistemas de comunicación

La finalidad de esta prueba fue comparar la velocidad de escritura de los participantes sin experiencia en conformar una oración. Para ello se pidió a los participantes de la prueba anterior, realizaron la interacción con el cuadro E-Tran conformando las mismas 12 oraciones.

Tabla 2. Velocidad promedio de escritura por oración.

Oración	Sistema pictográfico [s]	Cuadro E-Tran [s]
1	79.0	132.1
2	33.6	116.0
3	62.7	195.1
4	78.7	322.0
5	97.9	165.3
6	42.0	99.5
7	116.6	138.6
8	45.6	249.8
9	43.2	101.1
10	73.9	225.9
11	52.6	65.9
12	61.8	173.5
	Promedio general [s]	
	65.6	165.4

La tabla 2 muestra el promedio que tomó a los participantes en generar cada oración sobre cada una de las interfaces; además, se muestra el promedio general de tiempo de generación de oraciones.

Con los resultados de esta prueba, se determinó que los participantes sin experiencia lograron conformar las oraciones de una manera más rápida utilizando la interfaz pictográfica; lo cual refleja una de las características importante de un SAAC pictográfico; sin embargo, como se analizó en la prueba anterior, un sistema pictográfico siempre tendrá una limitación del lenguaje, ya que no es capaz de representar un número infinito de oraciones.

Por otro lado, se logró observar que un cuadro E-Tran es un sistema lento comparado con una interfaz pictográfica ya que es un 152% más lento en promedio. Por lo tanto, se puede afirmar que el SAAC computacional pictográfico presentado, permite conformar oraciones de una manera más rápida que un tradicional cuadro E-Tran.

6. Conclusiones

Se desarrolló un sistema de comunicación aumentativa en una microcomputadora Raspberry Pi 3, el cual utiliza un interruptor mecánico como hardware de entrada, y un sintetizador de voz para verbalizar la oración. Se creó un sistema que tiene una buena usabilidad con respecto al mapa de navegación, lo que vuelve al sistema intuitivo para su manejo; además se considera un sistema robusto de 370 oraciones, el cual según las pruebas realizadas tienen una cobertura de 76% de las oraciones.

Al comparar el sistema creado con un cuadro E-Tran, se observó que un sistema computarizado es una alternativa para los usuarios de un cuadro E-Tran convencional.

Adicionalmente, la estructura de programación permite expandir fácilmente la cantidad de pictogramas y por lo tanto se podrían sumar más oraciones.

También se tiene el potencial para agregar una plantilla alfabética, que permitiría al usuario transmitir cualquier idea y no limitar la comunicación a las oraciones predefinidas del sistema pictográfico.

Otro aspecto importante, es la implementación de un método de interrupciones por flanco sobre un pin de la Raspberry, lo que facilita la incorporación de diversos sistemas de interacción hombre-máquina, para adecuarlo a condiciones particulares de los pacientes.

Referencias

1. Gómez-Taibo, M.L., Pérez, E.M.: La intervención de la comunicación aumentativa y alternativa en el traumatismo craneoencefálico. Revista de Investigación en Logopedia, 8(1), pp. 43–62 (2018)
2. Zhang, X., Kulkarni, H., Ringel-Morris, M.: Smartphone-based gaze gesture communication for people with motor disabilities. In: CHI '17 CHI Conference on Human Factors in Computing Systems, pp. 2878–2889 (2017)
3. Gob.mx: https://gob.mx/conadis/articulos/la-esclerosis-lateral-amiotrofica-ela?Idiom = es (2019)
4. Fundación Francisco Luzón: La ELA: Una realidad ignorada. Consejo General de Colegios Oficiales de Farmacéuticos, pp. 48 (2017)
5. Dewan, M.C., Rattani, A., Gupta, S., Baticulon, R.E., Hung, Y.C., Puchak, M., Agrawal, A., Adeleye, A.O., Shrime, M., Rubiano-Escobar, A.M., Rosenfeld, J.V., Park, K.B.: Estimating the global incidence of traumatic brain injury. Journal of Neurosurgery, 130(43), pp. 1–18 (2018)
6. Collaguazo, J.H.C.: Sistema electrónico para facilitar la comunicación y actividades cotidianas de los enfermos con esclerosis lateral amiotrófica (ELA) (2018)
7. Toscano, S.V.C.: Lenguaje y parálisis cerebral: El uso de los SAAC como medio de comunicación. Montevideo (2016)
8. Arasaac.org: Centro aragonés para la comunicación aumentativa y alternativa. Gobierno de Aragón. http://arasaac.org/index.php (2020)
9. Raspberry Pi Foundation: https://raspberrypi.org/downloads/raspbian/ (2019)
10. Github: Acorn/RISC_OS computers. Github.com, https://github.com/espeak-ng/ espeak-ng/ (2019)

Diseño y construcción de un sistema monoaxial semi-automático para la fabricación de recubrimientos en sustratos estáticos

Luis Edgar Alanís Carranza, Moisés Vicente Márquez Olivera,
Ricardo Cuenca Álvarez, Octavio Sánchez Olivera,
Héctor Abraham Flores Avalos, Viridiana G. Hernández Herrera

Instituto Politécnico Nacional,
Centro de Investigación e Innovación Tecnológica,
México

samyo19@hotmail.com,
{mvmarquez, rcuenca, vhernandezhe}@ipn.mx
{osanchez0112, steenguitar777}@gmail.com

Resumen. Se presenta el desarrollo de un sistema mono axial de manipulación de una pistola de rociado térmico por flama para la manufactura de recubrimientos sobre sustratos estáticos. Esta técnica presenta alrededor de 45 variables, las cuales condicionan las características del recubrimiento que se construye. En la actualidad, el control de estas variables dependen directamente de la experiencia del operador debido que esta técnica se realiza de manera manual, produciendo heterogeneidades en el producto final. Dada esta premisa, se presenta la necesidad de construir un sistema de un eje cartesiano que controle el desplazamiento de la pistola de rociado térmico, controlando variables que mejoren la calidad de los recubrimientos permitiendo asegurar la homogeneidad de las características microestructurales de los sustratos empleados. Para llevar a cabo el diseño y construcción del mecanismo de sujeción y desplazamiento de la pistola de proyección, se determinaron las variables de estudio del proceso. Posteriormente, se implementó el controlador INTECMX14 al sistema monoaxial para la regulación de la velocidad de giro del motor sin escobillas y consecuentemente, el desplazamiento del mecanismo. Finalmente, el diseño de una interfaz GUI permite tanto la interacción del usuario con el dispositivo como el monitoreo del recorrido de la pistola. Se construyeron recubrimientos por aplicación manual y con el mecanismo propuesto. Los resultados muestran que la aplicación manual conduce al depósito de partículas sobre trayectorias erráticas, mientras que el sistema monoaxial deposita partículas sobre bandas de forma regular que siguen trayectorias rectilíneas, controlando el calentamiento del sustrato. Se concluye que el dispositivo mecatrónico permite realizar el rociado térmico en flama bajo las condiciones impuestas por el operador.

Palabras clave: Rociado térmico, recubrimiento, automatización.

Design and Construction of a Semiautomatic Monoaxial System for Coatings Manufacturing in Static Substrates

Abstract. In this work, the development of a monoaxial system is presented, which enables manipulation the flame spray pistol to manufacture coatings on static substrates. The flame spraying technique presents about 45 variables that condition the coating‘s characteristics built. Nowadays, controlling these variables depends directly on operator experience because this technique is performed manually, which produces heterogeneities on the final product. Under this idea, there is a need to build a Cartesian axis system that controls the displacement of the flame spray gun, controlling variables that improve coatings' quality, this allows ensure the homogeneity of the microstructural characteristics of the substrates. To make the mechanism design and construction for holding and moving the projection pistol, the variables of the process study were determined. Subsequently, an INTECMX14 driver was implemented in a monoaxial system to regulate the rotation speed of the brushless direct current motor and consequently, the movement of the mechanism. Finally, the GUI interface allowed the interaction between the user-machine and monitoring the pistol's path. The coatings were built by manually application and the device proposed. The results show that manual application produces deposition of particles on erratic trajectories, while the monoaxial system deposits particles on bands with a regular shape that follow rectilinear trajectories, controlling the substrate heating. In conclusion, the flame spraying process is performed, under the conditions imposed by the operator trough a mechatronic device.

Keywords: Thermal spraying, coating, automation.

1. Introducción

En la actualidad existen sistemas automatizados que eliminan o reducen la intervención humana durante los procesos de manufactura. Tal es el caso del rociado térmico, el cual consiste en el depósito de un recubrimiento sobre la superficie de un material. Durante este proceso, son formadas pequeñas gotas fundidas o semi-fundidas desde un polvo para generar un recubrimiento, generalmente, el nombre que reciben estas gotas son "splats" [1]. Existen diferentes tipos de rociado térmico (por flama, plasma y eléctrico) siendo una de sus ventajas la gran variedad de materiales que se pueden emplear para la fabricación de recubrimientos [2]. Este articulo está enfocado en el rociado térmico por flama ya que esta técnica es de las más utilizadas en la industria por su bajo costo y fácil uso. No obstante, el rociado térmico por flama es realizado de manera manual, por lo que, la calidad del producto depende de la experiencia del operador.

Durante los últimos años, la mecatrónica ha desarrollado distintos robots cartesianos para la manipulación controlada de herramientas usadas en la fabricación de productos, los más comunes son los sistemas de Control Numérico por Computadora o por sus siglas en inglés CNC. Acuña et al. [3] diseñaron y construyeron un sistema automático de 4 ejes (X, Y, Z y B) para el transporte de materia prima y productos maquinados, haciendo uso de una interfaz desarrollada en Labview, un microcontrolador Atmega328

y procesamiento digital de imágenes para el reconociendo de material con forma circular o rectangular. Rangel y Castro [4] desarrollaron el control de un sistema de coordenadas X, Y con el fin de identificar los puntos de una superficie y la transporta, por medio del uso de motores a pasos y de una placa Arduino. Con el programa que tiene cargado el microcontrolador, el usuario coloca los valores de X,Y y el mecanismo comienza a moverse.

Jhan y Ricardo [5] diseñaron e implementaron una fresadora de 3 ejes coordenados para la fabricación de placas PCB utilizando Arduino y un sistema HDI que monitorea los datos. Su sistema redujo tiempo de maquinado y costos.

Aguirre et al. [6] construyeron un equipo de tres grados de libertad con la finalidad de darle el uso de manipulador en aplicaciones de pick y place o como CNC para la fabricación de piezas. Utilizaron Arduino UNO y una interfaz GUI desarrollada en Matlab para la interacción con el usuario y la máquina. Su sistema puede utilizarse para la realización de prácticas en laboratorios de mecatrónica.

Jayachandraiah et al. [7] desarrollaron un CNC de bajo costo que permite el fabricado de piezas pequeñas utilizando la tarjeta Arduino. El sistema puede moverse en los tres ejes cartesianos (X,Y,Z) de manera simultánea.

El enfoque de este documento es el desarrollo de un sistema de un solo eje cartesiano (X) para la manipulación de la pistola de rociado térmico por flama y fabricar recubrimientos en la superficie del material. El diseño, adquisición, manufactura, ensamble, etapa de control y potencia, además del desarrollo de la interfaz GUI son presentados en la sección II, sección III examina los resultados obtenidos, las conclusiones son mostradas en la sección IV. referencias o trabajos citados en sección V.

2. Materiales y métodos

Se consideran 4 etapas para la construcción y puesta en marcha del sistema propuesto: (*i*) identificación de variables, (*ii*) diseño mecánico, (*iii*) manufactura y ensamble, y (*iv*) etapa de potencia y control.

2.1. Identificación de variables

El proceso de rociado térmico está influenciado por alrededor de 45 variables, siendo imposible de variarlas simultáneamente. Algunas de estas variables, que definen la calidad del recubrimiento, son: las propiedades de la flama, características del polvo, inyección del material de recubrimiento y los parámetros de trabajo [8], [9].Después de realizar el rociado térmico de partículas sobre sustratos estáticos, en forma manual, la información obtenida fue la siguiente: la distancia de proyección es cambiante, dado que influye en la adhesión de las partículas en el sustrato, el desplazamiento de la pistola describe variaciones tanto en la trayectoria como en la velocidad, por lo que se presentan diferencias en el tiempo de exposición a la flama. Por estas razones, la geometría y el espesor del recubrimiento depositado son heterogéneos. Entonces, la *distancia* de tiro con el sustrato, la *velocidad* y el *movimiento rectilíneo* de la pistola son escogidos como las variables a controlar y que rinden la condición homogénea del recubrimiento.

Tabla 1. Características requeridas para diseño del dispositivo mecatrónico.

Parámetro	**Unidad**
Longitud del dispositivo	70 cm
Ancho máximo del sustrato	30 cm
Grados de libertad	1
Carga	Pistola de rociado térmico de 4,3 kg
Estructura mecánica	Rígida y liviana

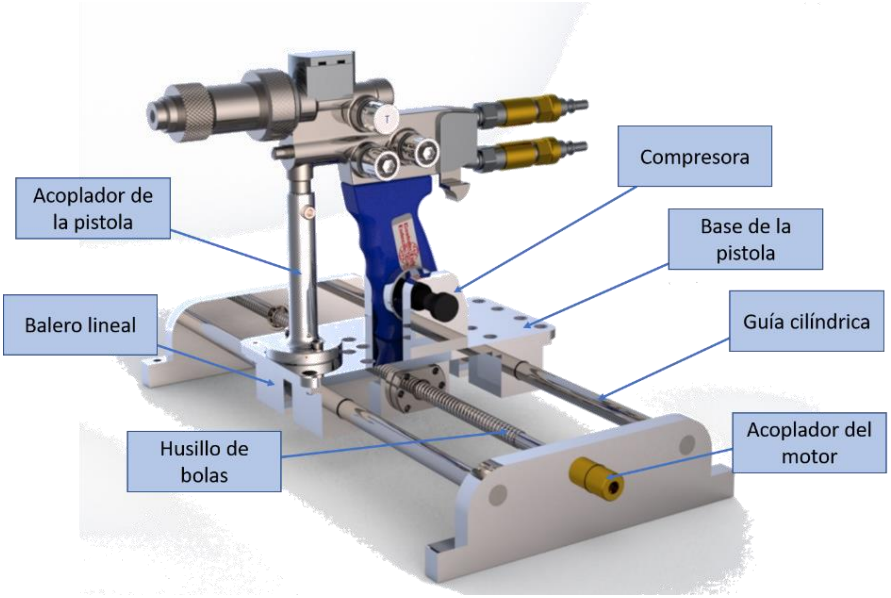


Fig. 1. Diseño del dispositivo.

2.2. Diseño mecánico

Los parámetros que arroja el análisis anterior permiten la fabricación de un recubrimiento homogéneo y son considerados para el diseño del mecanismo. Las características requeridas del dispositivo se muestran en Tabla 1.

El diseño del manipulador fue desarrollado en SolidWorks, mostrado en la Fig. 1. está compuesto por dos guías cilíndricas, cuatro deslizadores, un acoplador para la pistola y motor, un husillo de bolas con paso de 4 mm y un sujetador manual.

La tabla 1 indica que la estructura debe ser ligera y transportable, así que, el aluminio fue seleccionado como material para la construcción del eje.

2.3. Manufactura y ensamble

El material fue adquirido para el ensamble de las partes del dispositivo. No obstante, algunos componentes con medidas específicas no son comerciales, por lo que fueron

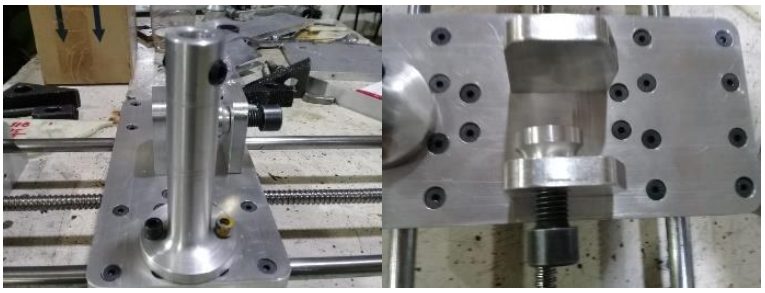


Fig. 2. Piezas maquinadas en CNC.

Fig. 3. Subensambles del dispositivo.

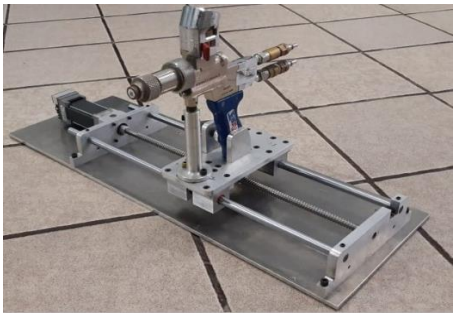


Fig. 4. Sistema mono axial.

manufacturadas en CNC. La (Fig. 2) muestra algunas de estas piezas fabricadas. En la (Fig. 3) algunos subensambles son mostrados y requeridos para generar el desplazamiento de la pistola por medio del motor.

Los deslizadores están sujetos a una base que permite colocar todos los elementos necesarios para sujetar la pistola ver (Fig. 2). La base sostiene un acoplador de forma cilíndrica y su función es sujetar a la pistola de rociado. El acoplador tiene dos opresores de cada lado para asegurar el apriete. También, el dispositivo cuenta con un sujetador de mango que permite la disminución de vibraciones y asegura el disparo sin tener cambios de ángulo de la boquilla. Finalmente, los subensambles de la (Fig. 3) permiten el desplazamiento de la base por medio del husillo de bolas. La Fig. 4. muestra el ensamble de todo el material adquirido y maquinado.

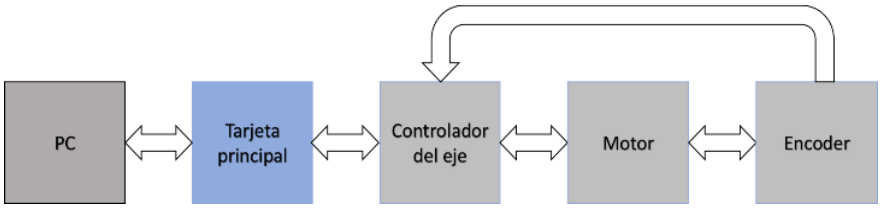


Fig. 5. Diagrama general de control de un motor.

2.4. Etapa de control y potencia

La etapa de control del motor debe cumplir con los siguientes requerimientos:

- Variar la velocidad angular,
- Controlar la posición del eje rotario,
- Variar la aceleración/desaceleración.

Por lo general, el control de la velocidad y posición de un motor DC requiere de una tarjeta principal y un controlador, disponibles en el mercado y de origen extranjero. En la Fig.5 se muestra el diagrama general de control del motor.

No obstante, el controlador de origen mexicano INTECMX14 [10] [11] [12] fue implementado para la manipulación del actuador. Este controlador ofrece ventajas tales como:

- La manipulación de 2 o más motores al mismo tiempo,
- Dentro de la tarjeta incluye el módulo de potencia y control,
- Es menos costoso que otros controladores.

Considerando el uso del controlador INTECMX14, el diagrama general se modifica y resulta como se muestra en la Fig. 6, en la que se muestra que solo se requiere de una tarjeta para la manipulación del motor, ya que en el controlador se encuentran los módulos de control y potencia.

Control: el usuario establece los parámetros de velocidad, aceleración y posición desde la interfaz. El módulo supervisor recibe los valores para generar un perfil de trayectoria, mostrado en la Fig. 7.

Todo sistema con desplazamiento de un punto A hacia un punto B debe de contar con un perfil de trayectoria para evitar desplazamientos con la menor cantidad de vibraciones posibles y daños en el mecanismo. El controlador INTECMX14 tiene agregado un algoritmo PID. Este utiliza operaciones básicas de suma, resta y multiplicación, facilitando su programación en el FPGA, por lo que se considera más fácil de implementar que otros algoritmos de control. El algoritmo PID recibe la señal calculada por el perfil de trayectoria y el valor real del sensor óptico de posición (Encoder) calculando el error y en consecuencia, regular la energía que se requiere para poder llegar a la posición deseada. Finalmente, la carta de conmutación electrónica genera el movimiento del motor por medio del cambio de bobinas A, B y C. El diagrama general del módulo de control es mostrado en la Fig. 8.

Fig. 9. Interfaz GUI del dispositivo.

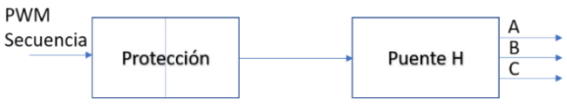


Fig. 10. Diagrama a bloques de la etapa de potencia.

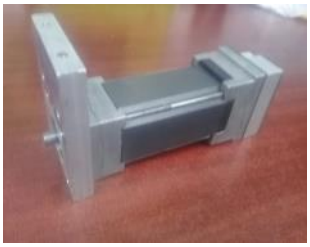


Fig. 11. Motor BLDC DMA0204024B101, marca Microchip.

La interfaz GUI (Fig. 9) permite activar y desactivar el dispositivo, mandar parámetros (tamaño del sustrato a rociar, la aceleración y la velocidad deseada), cuenta con botones de arranque y paro de emergencia, una visualización del recorrido que va a tener la pistola en el sustrato, y finalmente incluye etiquetas que indican el tipo de sustrato y polvo a utilizar.

Potencia: se implementó una protección entre el módulo de control y potencia, aislando a las dos etapas, previniendo el caso de que algún problema generado en una etapa no afecta a la otra. Luego, el puente H permite el cambio de sentido del flujo de la corriente con el fin de regular y producir el movimiento del motor. El diagrama general es mostrado en la Fig. 10.

El motor utilizado es el modelo DMA0204024B101, marca Microchip (Fig. 11) de la familia de motores BLDC (motor de corriente directa sin escobillas) y, en comparación con otros motores de corriente directa, tiene las siguientes ventajas:

- Torque máximo desde la primer RPM,

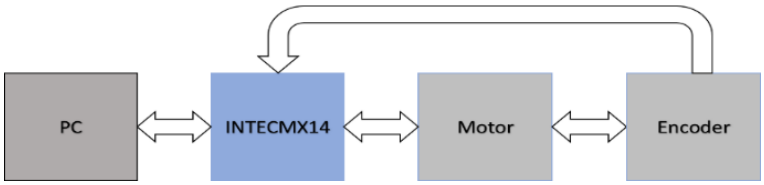


Fig. 6. Diagrama general de control de un motor utilizando INTECMX14.

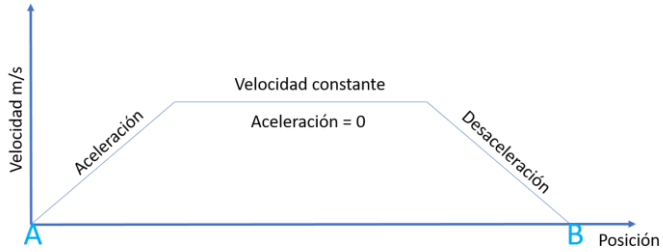


Fig. 7. Perfil de trayectoria.

Fig. 8. Diagrama a bloques de la etapa de control.

- Mayor durabilidad,
- Mínimo ruido eléctrico,
- Mínimo mantenimiento.

3. Resultados

Los recubrimientos fueron manufacturados manualmente y con el dispositivo. El sustrato utilizado para estas pruebas fue acero AISI 1018 con longitud de 30 cm. El polvo utilizado para los experimentos fue cobre. En la Fig. 12, son mostrados algunos recubrimientos fabricados manualmente por operadores, donde se aprecia que el ancho de la banda del recubrimiento no es constante. Esto conduce a identificar zonas con mayor cantidad de material depositado, además, las regiones con recubrimiento depositado pueden presentar oxidar con mayor rapidez. Finalmente, la prueba con el dispositivo muestra los siguientes resultados Fig. 13.

La banda del recubrimiento de la Fig.13. es constante en todo momento. No obstante, las esquinas muestran diferentes tonalidades de color, a consecuencia de que la velocidad no fue constante debido al perfil de trayectoria. Se puede considerar que el desempeño del dispositivo es homogéneo y repetible, ya que el recubrimiento de la Fig. 13 A) bajo los parámetros de la tabla 2.

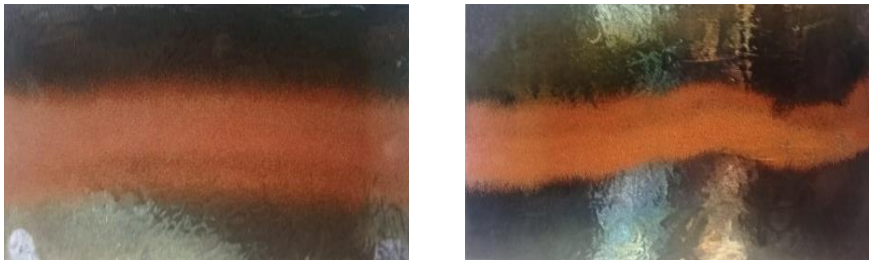
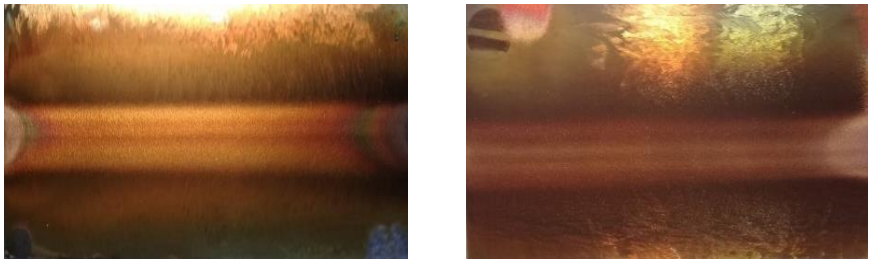


Fig. 12. Recubrimientos de aplicación manual.



A) B)

Fig. 13. Recubrimientos fabricados por el dispositivo A) prueba 1 y B) prueba 2.

Tabla 2. Parámetros para recubrimiento de la prueba 1.

Distancia Boquilla/sustrato	Relación De gases	Velocidad de tiro	Temperatura del sustrato	Número de pasadas	Largo de la banda
12.5 cm	O-25 C_2H_2-50	0.01 m/s	85-90 °C	1	≈ 3.8 cm

Tabla 3. Parámetros para recubrimiento de la prueba 2.

Distancia Boquilla/sustrato	Relación De gases	Velocidad de tiro	Temperatura del sustrato	Número de pasadas	Largo de la banda
12.5 cm	O-25 C_2H_2-50	0.02 m/s	85-90 °C	1	≈ 3.8 cm

Por lo tanto, una banda de recubrimiento aproximada a 3.8 cm es obtenida al utilizar los mismos parámetros.

Los parámetros de la tabla 3 permiten la manufactura de un recubrimiento con características de la figura 13 B) y una franja del mismo tamaño. Sin embargo, el cambio de tonalidad a más obscura puede perjudicar en el siguiente pase de recubrimiento. El sustrato estuvo expuesto a mayor tiempo ante la flama en esta prueba.

4. Conclusiones

En este trabajo se realizó el diseño y construcción de un dispositivo que permite el control del desplazamiento de la pistola de rociado térmico por flama en un eje coordenado. Los mecanismos de sujeción y desplazamiento fueron diseñados en SolidWorks y posteriormente manufacturados en máquinas CNC. La aceleración y velocidad constante de recorrido se genera por medio del motor BLDC y este es manipulado por el controlador INTECMX14. Además, el dispositivo tiene una interfaz GUI capaz de recibir parámetros de desplazamiento y tiene una visualización previa del recorrido de la pistola.

Los resultados muestran que el dispositivo propuesto tiene un desplazamiento constante a diferentes velocidades y manufactura recubrimientos más homogéneos. No obstante, las características del recubrimiento dependen de la velocidad de tiro, por lo que el estudio del sustrato y del polvo a rociar debe ser previo para obtener las velocidades requeridas para cada material y evitar modificaciones en el recubrimiento como se mostró en la figura 13. B).

Referencias

1. Champagne, V.K.: The cold spray materials deposition process. Elsevier (2007)
2. Pawlowski, L.: The science and engineering of thermal spray coatings. John Wiley & Sons (2008)
3. Capilla-Falcon, C.A., Pulloquinga-Zapata, J.L.: Diseño y construcción de un prototipo de sistema automatizado de almacenamiento/recuperación (AS/RS), para sistemas flexibles de manufactura en el Laboratorio CNC de la ESPE Extensión Latacunga, utilizando procesamiento digital de imágenes. BS tesis Universidad de las Fuerzas Armadas ESPE Extensión Latacunga (2014)
4. Rangel-Domínguez, H.E., Castro-Sánchez, R.: Control de coordenadas en xy de un mecanismo de motores a pasos y un microcontrolador. Jóvenes en la Ciencia, 3, pp. 2273–2277 (2019)
5. Ore, J., Palomares, R.J.: Diseño e implementación de una fresadora CNC de 3 gdl para fabricación de tarjetas electrónicas usando arduino y ubuntu linux (2016)
6. Aguirre, A., Gomez-Suarez, F.J., Ramirez-Bernal, G., Aguirre-Carillo, F.A.: Manipulador mecatrónico de tres grados de libertad: Sistema pick and place y CNC. I Encuentro de Jovenes Investigadores de Sinaloa (2013)
7. Jayachandraiah, B., Krishna, O.V., Khan, P.A., Reddy, R.A.: Fabrication of low cost 3-Axis CNC router. International Journal of Engineering Science Invention, 3, pp. 1–10 (2014)
8. Rodríguez-Santos, E.G.: Recubrimientos compósitos: Alternativa para reducir el impacto ambiental ocasionado por corrosión húmeda. Tesis, Instituto Politécnico Nacional (2018)
9. Posada, B.A., Gamboa, D., Marulanda-Arevalo, J.L.: Protección contra la corrosión por sales fundidas de un acero al carbono por rociado térmico. Scientia et Technica, pp. 251–256 (2009)
10. Sánchez-García, O.: Controlador Senoidal para motor PMBLDC en tracción de vehículos eléctricos. Tesis Instituto Politécnico Nacional (2016)
11. García-Sotuyo, E., Hernández-Herrera, V.G., Sánchez-García, O., Márquez-Olivera, M.V.: Diseño e implementación de un hexápodo robótico empleando el driver “intecmx14” para el control paralelo de motores. Pistas Educativas, 39(125) (2017)

12. Martínez-Fernández, G., Cruz-Contreras, A., Hernández-Herrera, V.G., Márquez-Olivera, M.V.: Sistema de control distribuido embebido en FPGA para motores de CD en aplicaciones multieje. Pistas Educativas, 39(125) (2017)

Diseño de un nuevo controlador no lineal con aplicación al modelo de un biorreactor para producción de microalgas

Omar S. Castillo Baltazar[1], Pablo A. López Pérez[2],
J. Marcelino Gutiérrez Villalobos[1], Ricardo Aguilar López[3],
Vicente Peña Caballero[1]

[1] Universidad de Guanajuato,
Departamento de Ingeniería Agroindustrial,
México

[2] Universidad Autónoma del Estado de Hidalgo,
Escuela Superior de Apan,
México

[3] Instituto Politécnico Nacional,
Centro de Investigación y de Estudios Avanzados,
Departamento de Biotecnología y Bioingeniería,
México

{omar.castillo, vicente.caballero,
jmgutierrez@ugto.mx}@ugto.mx, esave1991@yahoo.com.mx,
raguilar@cinvestav.mx

Resumen. En este trabajo, se presenta el diseño de un nuevo controlador no lineal (CNL) para regular la concentración residual de la fuente de nitrógeno (nitrato de sodio, $NaNO_3$) implementado sobre el modelo de un biorreactor a fin de incrementar la producción de biomasa de *Isochrysis galbana*. El CNL consiste en una estructura no lineal y se caracteriza por su facilidad en la sintonización de sus parámetros. El CNL es evaluado a diferentes concentraciones de referencia, set points, de $NaNO_3$. El controlador se utilizó para regular la concentración residual de sustrato $NaNO_3$ en el volumen del biorreactor a partir de la medición del error entre la dinámica de la concentración de $NaNO_3$ medida y su valor de referencia. La medición de $NaNO_3$ fue seleccionada por ser más económica y fácil de implementar en línea, en contraste con a la medición de biomasa o lípidos. El desempeño del CNL se evaluó frente a un controlador clásico PI a través del criterio del error ITEC (la integral del tiempo multiplicada por el error al cuadrado). Los resultados de la simulación sugieren que el desempeño del CNL propuesto frente a diferentes condiciones de referencia de la variable regulada es adecuado para controlar el proceso de producción de microalgas.

Palabras clave: Diseño de control no lineal, cultivos de *isochrysis galbana*, producción de microalgas.

Design of a New Non-Linear Controller with Application to the Model of a Bioreactor for Microalgae Production

Abstract. In this work, the design of a new non-linear controller (CNL) is presented to regulate the residual concentration of the nitrogen source (sodium nitrate, NaNO_3) implemented on the model of a bioreactor in order to increase the biomass production of Isochrysis galbana. The CNL consists of a non-linear structure and is characterized by its ease in tuning its parameters. The CNL is evaluated at different reference concentrations, set points, of NaNO_3. The controller was used to regulate the residual concentration of NaNO_3 substrate in the bioreactor volume from the measurement of the error between the dynamics of the measured NaNO_3 concentration and its reference value. The NaNO_3 measurement was selected as being cheaper and easier to implement online, in contrast to the measurement of biomass or lipids. CNL performance was evaluated against a classical PI controller through the ITEC error criterion (the time integral multiplied by the squared error). The simulation results suggest that the performance of the proposed CNL under different reference conditions of the regulated variable is adequate to control the microalgae production process.

Keywords: Nonlinear control design, *isochrysis galbana* cultures, microalgae production**.**

1. Introducción

Recientemente, el campo de investigación de las microalgas marinas se ha incrementado por su amplio espectro de usos y aplicaciones. En la producción de microalgas en reactores (a cielo abierto o cerrados) con diferentes configuraciones, los nutrientes nitrogenados en cultivos de microalgas son para generar biomasa, mantener la actividad metabólica y para la síntesis de los productos de interés económico. Por lo que la fuente de nitrógeno es uno de los factores clave que rige el crecimiento y la acumulación de lípidos en las microalgas.

La importancia de la producción de microalgas y sus productos derivados del metabolismo se relaciona con producción de biodiesel con impacto ambiental reducido de tercera generación [1, 2], captura del CO2 y la mitigación del cambio climático [3, 4]. Por otro lado, también para la eliminación de metales pesados, especialmente plomo y cromo, presentes en aguas residuales procedentes de la industria textil [5-7]. En el área de alimentos, provee proteínas y otros nutrientes para consumo animal y suplementos alimenticios para humanos [8-10].

Además, los lípidos de microalgas pueden ser compuestos como suplemento en alimentos para otros organismos [11-13]. En relación a *Isochrysis sp*., Liu y Lin [14] señalan que esta es una de las microalgas que más atención ha tenido por su capacidad para producir ácido graso poliinsaturado decosahexaenoico que provee beneficios saludables. Este ácido y sus derivados ayudan a prevenir y tratar algunas patologías en humanos. Por ejemplo, el ácido graso C22-poliinsaturado y sus derivados ayudan a prevenir y tratar patologías asociadas a la enfermedad cardíaca coronaria y la

aterosclerosis, además de inflamatorios y algunos tipos de cáncer [15]. Y se cree que desempeñan un papel en la nutrición infantil [16].

En lo referente al modelado del bioproceso de cultivos de microalgas, algunos trabajos se han enfocado a desarrollar el modelo matemático del proceso para describir el crecimiento de la microalga [17-20]. Para el caso específico de *I. galbana* [21] demostraron la alta producción de lípidos en esta microalga marina, en un proceso discontinuo en condiciones limitantes en la fuente de nitrógeno y propusieron un modelo fenomenológico discontinuo de tres estados.

En general, el notable crecimiento del interés por los cultivos de microalgas es lo que ha llevado a la ciencia a la necesidad entender y controlar los procesos de biotransformación de sustratos a productos. En particular para el caso de control de estos bioprocesos (para la producción de biomasa microalgal y en consecuencia la optimización de sus productos metabólicos; de manera tal que el proceso se desarrolle en condiciones controladas) con el fin de mejorar los rendimientos de biomasa y sus productos. Para tal fin, diferentes biorreactores se han diseñado y/o configurado para mejorar la transferencia de masa, distribución de nutrientes, incremento de la disponibilidad de luz, entre otras variables [22-24]. Sin embargo, la producción de microalgas, por su propia naturaleza no lineal en sus procesos metabólicos, enfrenta muchos retos importantes de control para garantizar alta densidad de biomasa [25, 18]. Por lo que en los procesos de producción de microalgas es indispensable estimar (observar) y controlar variables (estados), por ejemplo, temperatura, consumo de sustratos, generación de productos, suministro de intensidad luminosa, agitación, entre otras [15]. Esto debido a que el control de estas variables está relacionado directamente con la productividad y economía del proceso.

En los bioprocesos, el controlador más utilizado se refiere al controlador clásico Proporcional-Integral-Derivativo (PID), el cual en muchos de los casos opera simplemente como un controlador PI (Proporcional-Integral) debido a las dificultades encontradas normalmente en la utilización del modo derivativo en aplicaciones donde las señales pueden contener ruido de medición [26]. Recientemente, el diseño de nuevos controladores es una alternativa al control clásico para el control de procesos y bioprocesos en las áreas de química y de bioquímica. Por ejemplo, nuevos controladores no lineales basados en estructuras de PI y modos deslizantes de alto orden se han propuesto como una alternativa para controlar procesos químicos [27]. Para el caso de procesos biológicos, en [28] diseñaron un controlador no lineal adaptativo para regular el consumo de sulfato para un proceso sulfato reductor. En [29] se diseñó un controlador no lineal para regular la producción de hidrógeno. En cuanto a procesos de microalgas, en [30], se presenta el diseño de un controlador no lineal para regular la concentración residual de la fuente de nitrógeno en el reactor para incrementar la de producción de lípidos.

Por lo tanto, en este trabajo, el objetivo es presentar el diseño de un nuevo controlador no lineal (CNL) para regular la concentración residual de la fuente de nitrógeno (nitrato de sodio, $NaNO_3$) implementado sobre el modelo de un biorreactor a fin de incrementar la producción de biomasa de *Isochrysis galbana*.

Este trabajo está organizado como sigue. En la sección 2 se presenta de manera breve la estructura del modelo fenomenológico del reactor discontinuo y continuo para un bioproceso de producción de microalgas para *Isochrysis galbana*. Se continua con la presentación del controlador no lineal propuesto y el desarrollo de la prueba de

convergencia de éste. En la sección 3 se presenta los resultados numéricos acompañados de una breve discusión. Finalmente, en la sección 4 se presentan las conclusiones.

2. Metodología

2.1. Modelo fenomenológico de I. galbana

El modelo no lineal del cultivo de microalgas en discontinuo para la producción de lípidos en *I. galbana* se considera como un modelo de referencia y con fines de implementar numéricamente controlador propuesto. El modelo fenomenológico se define por el conjunto de ecuaciones diferenciales ordinarias, para una mejor precisión en la deducción del modelo se remite al lector a [21].

Formación de la biomasa:

$$\frac{dX}{dt} = \mu_m X\left(1 - \frac{X}{X_m}\right), \tag{1}$$

donde: $\frac{dX}{dt}$ es la tasa de crecimiento de microalgas (mg L^{-1} día^{-1}), μ_m es la tasa máxima de crecimiento específico de la microalga (día^{-1}), X es la concentración de biomasa de la microalga (mg L^{-1}) y X_m es la concentración máxima de biomasa de la microalga (mg L^{-1}).

Formación del producto:

$$\frac{dP}{dt} = \alpha\frac{dX}{dt} + \beta X\,, \tag{1a}$$

donde: $\frac{dP}{dt}$ es la tasa de formación del producto (mg L^{-1} día^{-1}), P es la concentración del producto (mg L^{-1}), α es el coeficiente de correlación de crecimiento (mg mg^{-1}), y β es el coeficiente de no-correlación de crecimiento (mg mg^{-1}).

Consumo del nitrato de sodio:

$$-\frac{dS}{dt} = \frac{1}{Y_{\frac{x}{s}}}\frac{dX}{dt} + mX\,, \tag{1b}$$

donde: $\frac{dS}{dt}$ es la tasa de consumo de nitrato de sodio (mg L^{-1} día^{-1}), S es la concentración de nitrato de sodio (mg L^{-1}), $Y_{x/s}$ es el coeficiente máximo de crecimiento microalgal (mg mg^{-1}), y m es el coeficiente máximo de mantenimiento (mg mg^{-1}) [31].

En la tabla 1 se muestran los valores de los parámetros del proceso de producción de *I. galbana* que se analiza en este trabajo [21].

Tabla 1. Parámetros del modelo [21].

$(\mathbf{NaNO_3})_0$	$\mathbf{50\ mg\ L^{-1}}$	$\mathbf{75\ mg\ L^{-1}}$
	Biomasa	
μ_{max} (día^{-1})	$0.586 \pm 1.75 \times 10^{-2}$	$0.448 \pm 1.89 \times 10^{-2}$
X_0 (mg L^{-1})	62.15 ± 3.31	75.77 ± 5.10
X_m (mg L^{-1})	538.28 ± 4.07	670.58 ± 10.57
	Lípidos	
P_0 (mg L^{-1})	5.82 ± 2.15	8.35 ± 3.98
α (mg mg^{-1})	$3.43 \times 10^{-2} \pm 8.85 \times 10^{-3}$	$3.91 \times 10^{-2} \pm 1.91 \times 10^{-3}$
β (mg mg^{-1})	$1.35 \times 10^{-2} \pm 1.0 \times 10^{-3}$	$1.39 \times 10^{-2} \pm 2.28 \times 10^{-3}$
	$NaNO_3$	
S_0 (mg L^{-1})	44.33 ± 1.75	63.82 ± 3.13
Y_{XS} (mg mg^{-1})	13.64 ± 1.35	7.02 ± 0.41
m ($mg\ mg^{-1}$)	$3.2 \times 10^{-4} \pm 8.1 \times 10^{-4}$	$-3.14 \times 10^{-3} \pm 1.04 \times 10^{-3}$

2.2. Diseño del controlador

La dinámica del cultivo de *I. galbana* en (1) se extiende a continuo con fines de implementar numéricamente el controlador propuesto. La dinámica en (1) se representa en su forma compacta, es decir, en forma matricial en (2):

$$\begin{gathered} \dot{x} = f(x) + g(x)u;\ x(0) = x_0; \\ y = h(x) = Cx, \end{gathered} \tag{2}$$

donde:

$$\boldsymbol{f(x)} = \begin{bmatrix} \mu_m X\left(1 - \frac{X}{X_m}\right) \\ \alpha \frac{dX}{dt} + \beta X \\ -\frac{1}{Y_{\frac{X}{S}}} \frac{dX}{dt} + mX \end{bmatrix}; \boldsymbol{g(x)} = \begin{bmatrix} -X \\ -P \\ (S_{in} - S) \end{bmatrix}; u = D; \boldsymbol{y} = \begin{bmatrix} X \\ P \\ S \end{bmatrix}. \tag{2a}$$

Siendo $u = D = F/V$ la tasa de dilución o la variable de control.

Proposición 1: La siguiente entrada de control es un controlador para el sistema en la Ec. (2):

$$u\big(e(t)\big) = \varphi_1 \frac{e(t)}{e(t) + \varphi_2 \left(\frac{e(t)}{\varphi_3} - 1\right)^2}, \tag{3}$$

donde: $e(t)$ es la señal de error, $\varphi_i, i = 1, 2, y\ 3$ son parámetros del controlador que se ajustan a prueba y error por el usuario.

Para la magnitud de la entrada de control, $\|u\big(x(t)\big)\|$, se puede obtener mediante el uso de los límites de la función de control, es decir, $\lim u(t)$ cuando $e(t) \to 0$:

$$\lim_{e(t)\to 0} \varphi_1 \frac{e(t)}{e(t) + \varphi_2 \left(\frac{e(t)}{\varphi_3} - 1\right)^2} = \frac{\varphi_1}{\varphi_2} = \varphi_4\,, \tag{4}$$

$$\|u\big(x(t)\big)\| \leq sup\varphi_1 \frac{e(t)}{e(t) + \varphi_2 \left(\frac{e(t)}{\varphi_3} - 1\right)^2} \leq \left\| \varphi_1 \frac{e(t)}{e(t) + \varphi_2 \left(\frac{e(t)}{\varphi_3} - 1\right)^2} \right\| \leq \varphi_4\,, \tag{5}$$

donde la regulación del error es definida como sigue considerando que $x(t) = S(t); x_{sp} = S_{set\ point}$:

$$e(t) = x(t) - x_{sp}\,. \tag{5a}$$

Prueba de estabilidad del controlador

Para demostrar la estabilidad del controlador propuesto se considera la dinámica del error ($\dot{e}(t)$):

$$\dot{e}(t) = f(x) + g\big(x(t)\big)u\big(x(t)\big), \tag{6}$$

y se propone la siguiente función de Lyapunov:

$$V = e^T Q e = \|e(t)\|_Q^2, Q^T > 0\,, \tag{7}$$

Ahora si se considera la derivada de la función de Lyapunov a lo largo del tiempo para (6):

$$\dot{V} = \dot{e}^T Q \dot{e} = \dot{e}^T Q e + e^T Q \dot{e}\,, \tag{8}$$

$$= \dot{e}^T Q[f(x) + g(x(t))] + e^T Q[f(x) + g(x(t))]\,, \tag{9}$$

$$= 2\dot{e}^T Q f\big(x(t)\big) + 2\dot{e}^T g(x) u\big(x(t)\big), \tag{9a}$$

$$\leq 2\left[\left\|\overbrace{\dot{e}^T Q f\big(x(t)\big)}^{a}\right\| + \left\|\overbrace{\dot{e}^T g(x) u\big(x(t)\big)}^{b}\right\|\right]. \tag{10}$$

Para a) considerando que la matriz Q se puede expresar como $Q = MM^T$, entonces:

$$\|\dot{e}^T Q f\big(x(t)\big)\| = \dot{e}^T M M^T f\big(x(t)\big) = \tilde{e}^T \tilde{f}\,, \tag{11}$$

donde:

$$\tilde{e}^T = \dot{e}^T M \text{ y } \tilde{f} = M^T f\big(x(t)\big)\,, \tag{12}$$

entonces:

$$\|\tilde{e}^T\| = \sqrt{(\tilde{e}^T\tilde{e})}\,, \tag{13}$$

$$= \sqrt{(e^T M M^T e)}\,, \tag{14}$$

$$\|\tilde{e}^T\| = \|e\|_Q\,, \tag{15}$$

por analogía para $\tilde{f}$se define que:

$$\square = \square\square\,, \tag{16}$$

por lo tanto, se puede demostrar que:

$$\|\dot{e}^T Q f\| = \left\|\tilde{e}^T\tilde{f}\right\| \leq \|e\|_Q \left\|\tilde{f}\right\|_p\,, \tag{17}$$

para b tomado en cuenta a:

$$\dot{e}^T g(x)u\big(x(t)\big) \leq \|e\|_Q \|g(x)\|_Q \|u\|\,, \tag{18}$$

entonces de (a) y (b),

$$\dot{V} \leq 2\left[\|e\|_Q \left\|\tilde{f}\right\|_Q + \|e\|_Q \|g(x)\|_Q \|u\|\right], \tag{19}$$

ahora considerando la proposición 1:

$$\dot{V} \leq 2\left[\|e\|_Q \left\|\tilde{f}\right\|_Q + \|e\|_Q \|g(x)\|_Q \varphi_4\right], \tag{20}$$

$$\dot{V} \leq 2\left[\left\|\tilde{f}\right\|_Q + \|g(x)\|_Q \varphi_4\right] \|e\|_Q\,, \tag{21}$$

$$\dot{V} \leq 2[L + G\varphi_4]\|e\|_Q\,, \tag{22}$$

finalmente, si seleccionamos $\varphi_4 < 0$, y $G\varphi_4 > L$, $\varphi_4 > -\frac{L}{G}$:

$$\dot{V} \leq 2[L + G\varphi_4]\|e\|_Q \leq 0\,. \tag{23}$$

Lo anterior define que la función de Lyapunov propuesta con su derivada, es definida negativa, con lo que se concluye que el controlador propuesto es estable.

3. Resultados y discusión numérica

Para probar el controlador propuesto, se tomaron los datos reportados por [21]. Según esto, la producción de *I. galbana* se realizó en un cultivo en lote en un tiempo de 13 días para diferentes condiciones iniciales de la concentración de $NaNO_3$ ($25\ mg\ L^{-1}, 50\ mg\ L^{-1}$, y $75\ mg\ L^{-1}$) en el medio de cultivo, observando mayor producción de lípidos a concentraciones bajas de la fuente de nitrógeno $NaNO_3$. Para afectos de análisis de la dinámica del proceso con la implementación numérica del controlador propuesto en (3), la dinámica del proceso de producción de microalgas en

(1) se extendió a continuo en (2), considerando una tasa de dilución de $0.15\ \text{días}^{-1}$, un valor aceptable para un proceso de producción de microalgas [32]. Dos concentraciones para $NaNO_3$ en la corriente del influente de a) $44.33/2\ \text{mg L}^{-1}$ y b) $63.82/2\ \text{mg L}^{-1}$ fueron seleccionadas para fines de estudios numéricos, con condiciones importantes para la producción de lípidos en cultivos de *I. galbana*.

Se simuló el proceso de $0\ \text{días} < t < 2\ \text{días}$ a lazo abierto y de $2\ \text{días} < t < 13\ \text{días}$ a lazo cerrado. La acción del controlador disminuyó inmediatamente la concentración de $NaNO_3$ de $41.09\ \text{mg L}^{-1}$ a $36\ \text{mg L}^{-1}$a los dos días, después al día seis disminuyó a $33\ \text{mg L}^{-1}$, y por último a los 12 días disminuyó a $25\ \text{mg L}^{-1}$. En la simulación se utilizó un método de integración de Runge-Kutta de cuarto orden, con un paso de 0.01 días. Las condiciones iniciales $\boldsymbol{x_0}$, fueron para a) $[62.15\ \text{mg L}^{-1} \quad 5.82\ \text{mg L}^{-1} \quad 44.33\ \text{mg L}^{-1}]^T$ y para *b*) $[75.77\ \text{mg L}^{-1} \quad 8.35\ \text{mg L}^{-1} \quad 63.82\ \text{mg L}^{-1}]^T$. Los parámetros del controlador propuesto para ambos casos de estudio fueron $\varphi_1 = 100$, $\varphi_2 = 0.1\ \text{mg L}^{-1}$, y $\varphi_3 = 1\ \text{L mg}^{-1}$. El controlador PI fue sintonizado mediante el método de [33]; vía perturbación de 5% en la entrada de control, es decir, la tasa de dilución ($D_0 = 0.15\ \text{días}^{-1}$) a través de una función escalón; la ganancia en estado estable fue calculada como $K = 1400\ \text{mg L}^{-1}\ \text{día}^{-1}$, el tiempo característico $\tau = 170\ \text{días}^{-1}$. El desempeño del controlador propuesto se evaluó en términos del índice de funcionamiento ITEC [34].

La fig.1C muestra la concentración de nitrato de sodio regulada por el controlador propuesto y el controlador PI con las mismas condiciones de simulación del sistema en (2). El CNL prtesenta un mejor desempeño frente al error entre la variable controlada y el valor de la señal de referencia deseada, es decir, el controlador propuesto frente al ajuste de sus parámetros alcanza un valor mínimo del ITEC frente al valor del controlador PI.

A lazo cerrado el error converge a cero inmediatamente, esto indica que el CNL es capaz de operar para regular su acción para diferentes valores de referencia para $NaNO_3$. No así para el controlado PI. La fig.1D1 muestra el criterio integral ITEC a la variable $NaNO_3$, cuyos valores son bajos frente a los valores del controlador PI a lazo cerrado. Para este caso, la producción de lípidos se incrementa en el sistema operando en ambos lazos abierto y cerrado (figB). Esta respuesta del sistema concuerda con los resultados de [21], ya que al abatir la concentración de $NaNO_3$ en el medio de cultivo se favorece la producción de biomasa y también un incremento en la producción de lípidos. Esto último se ha reportado ampliamente en varias investigaciones, debido que las células cuando están creciendo a un ambiente limitado en la fuente de nitrógeno, responden con un incremento en la producción de lípidos. Una respuesta contraria se observa cuando las microalgas crecen en ambientes con altas concentraciones de la fuente nitrógeno. Para el caso b), en la fig.2C se muestra la concentración de lípidos como respuesta de la regulación de la concentración del nitrato de sodio (fig.2A).

Para este caso, la condición inicial de la concentración del sustrato S_0, fue mayor que el caso a). De igual forma la concentración del sustrato en la corriente del influente al biorreactor fue mayor ($63.82/2\ \text{mg L}^{-1} > 44.33/2\ \text{mg L}^{-1}$) (ver tabla 1). En lazo abierto se incrementa la concentración del sustrato hasta $68.06\ \text{mg L}^{-1}$, esta concentración se puede considerar como una concentración alta de la fuente de nitrógeno en el medio de cultivo y en consecuencia las células de *I. galbana*, podrían bajar la síntesis de lípidos. Este efecto se muestra en los valores de simulación de la

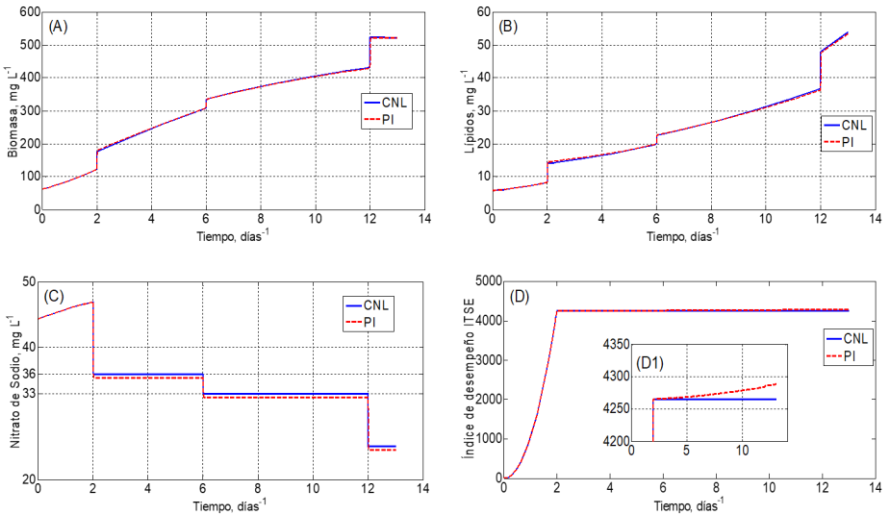


Fig. 1. Valores de concentración de: biomasa (A), lípidos (B), nitrato de sodio (C).

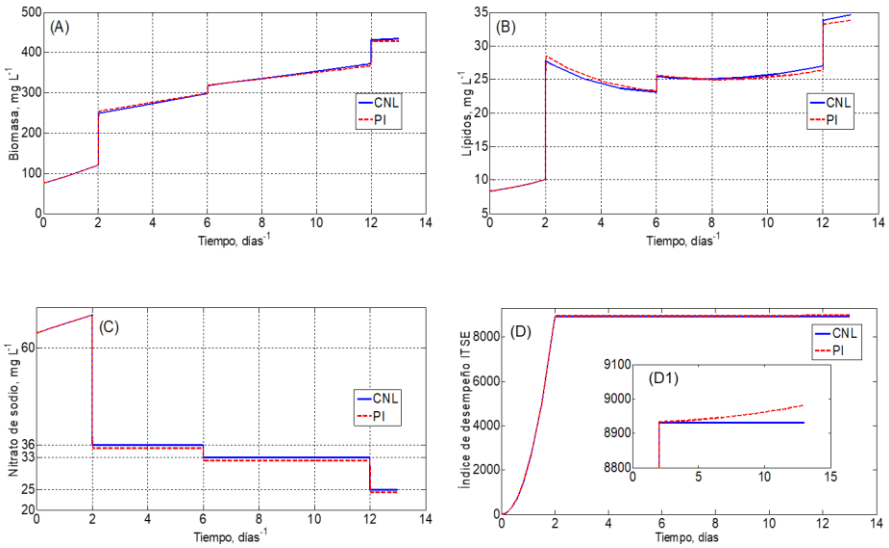


Fig. 2. Valores de concentración de: biomasa (A), lípidos (B), nitrato de sodio (C), e índice de desempeño (D).

concentración de lípidos [21], en el dominio de acción del controlador en 2 días $< t <$ 6 días. El abatimiento en la producción de lípidos para este dominio es suave.

Al regular la concentración del sustrato de 33 a 25 mg L^{-1}, la producción de lípidos se incrementa. Sin embargo, el incremento en b) es menor en comparado con el caso anterior donde al entrar en acción el controlador se alcanza una concentración menor

de sustrato en el medio de cultivo, es decir, de $42.81\ \mathrm{mg\ L^{-1}}$ en contraste con la concentración de $68.06\ \mathrm{mg\ L^{-1}}$ para el caso b). De igual manera el controlador propuesto presento un mejor desempeño que el controlador PI (figs.2D-C) y la dinámica cero, es decir las variables que no se controlan, para ambos casos son estables (figs.2A-B). Por lo tanto, el CNL numéricamente es un buen controlador para regular la concentración de $NaNO_3$cuando para diferentes valores de referencia en la concentración de $NaNO_3$.

Con los resultados anteriores se puede considerar que el modelo en (1) a su extensión a continuo, Ec. (2), predice numéricamente el comportamiento del cultivo de *I. galbana* en continuo.

4. Conclusiones

En este trabajo se presentó el diseño de un nuevo controlador no lineal y se aplicó al modelo de producción de microalgas de *I. galbana* en régimen continuo. El controlador propuesto presentó un buen desempeño numérico para regular la concentración del sustrato limitante en el proceso de producción de microalgas en contraste con el menor desempeño del controlador clásico PI.

Referencias

1. Scott, S.A., Davey, M.P., Dennis, J.S., Horst, I., Howe, C.J., Lea-Smith, D.J., Smith, A.G.: Biodiesel from algae: Challenges and prospects. Current Opinion in Biotechnology, 21(3), pp. 277–286 (2010)
2. Sánchez-Bermudez, A., Maceiras, R., Cancela, M.A., González-Pérez, A.: Culture aspects of isochrysis galbana for biodiesel production. Applied Energy, 101, pp. 192–197 (2013)
3. Lam, M.K., Lee, K.T.: Microalgae biofuels: A critical review of issues. Problems and the Way Forward. Biotechnology Advances, 30(3), pp. 673–690 (2012)
4. Demirbas, A.: Use of algae as biofuel sources. Energy Conversion and Management, 51(12), pp. 2738–2749 (2010)
5. Abdel-Raouf, N., Al-Homaidan, A.A., Ibraheem, I.B.M.: Microalgae and wastewater treatment. Saudi Journal of Biological Sciences, 19(3), pp. 257–275 (2012)
6. Mehta, S.K., Gaur, J.P.: Use of algae for removing heavy metal ions from wastewater: Progress and prospects. Critical Reviews in Biotechnology, 25(3), pp. 113–152 (2005)
7. Kumar, K.S., Dahms, H.U., Won, E.J., Lee, J.S., Shin, K.H.: Microalgae–A promising tool for heavy metal remediation. Ecotoxicology and Environmental Safety, 113, pp. 329–352 (2005)
8. Becker, E.W.: Micro-algae as a source of protein. Biotechnology Advances, 25(2), pp. 207–210 (2007)
9. Chisti, Y.: Biodiesel from microalgae. Biotechnology Advances, 25(3), pp. 294–306 (2007)
10. Mata, T.M., Martins, A.A., Caetano, N.S.: Microalgae for biodiesel production and other applications: a review. Renewable and Sustainable Energy Reviews, 14(1), pp. 217–232 (2010)
11. Coutteau, P., Castell, J.D., Ackman, R.G., Sorgeloos, P.: The use of lipid emulsions as carriers for essential fatty acids in bivalves: A test case with juvenile placopecten magellanicus. AGRIS, 15, pp. 259–264 (2013)

12. Knauer, J., Southgate, P.C.: Evaluation of microencapsulated squid oil as a substitute for live microalgae fed to pacific oyster (crassostrea gigas) spat. Journal of Shellfish Research, 16, pp. 137–141 (1997)
13. Knauer, J., Southgate, P.C.: Growth and fatty acid composition of pacific oyster (crassostrea gigas) spat fed a microalga and microcapsules containing varying amounts of eicosapentaenoic and docosahexaenoic acid. Journal of Shellfish Research, 16, pp. 447–453 (1997)
14. Liu, C.P., Lin, L.P.: Ultrastructural study and lipid formation of isochrysis sp. CCMP1324. Botanical Bulletin of Academia Sinica, 42(3), pp. 207–214 (2001)
15. Nordøy, A., Hansen, J.B.: ω-3 Fatty acids and cardiovascular risk factors. World Review of Nutrition and Dietetics, 76, pp. 51–54 (1994)
16. Conner, W.E., Neuringer, M.: Importance of dietary omega-3 fatty acids in retinal function and brain chemistry. In: Morley, J.E., Sterman, M.B., Walsh, J.H. (eds.) Nutritional modulation of neural function. Academic Press, New York, pp. 191–201 (1987)
17. Baquerisse, D., Nouals, S., Isambert, A., dos Santos, P.F., Durand, G.: Modelling of a continuous pilot photobioreactor for microalgae production. Journal of Biotechnology, 70(1–3), pp. 335–342 (1999)
18. Bernard, O.: Hurdles and challenges for modelling and control of microalgae for CO2 mitigation and biofuel production. Journal of Process Control, 21, pp. 1378–1389 (2011)
19. Straka, L., Rittmann, B.E.: Light-dependent kinetic model for microalgae experiencing photoacclimation, photodamage, and photodamage repair. Algal Research-Biomass Biofuels and Bioproducts, 31, pp. 232–238 (2018)
20. Packer, A., Li, Y., Andersen, T., Hu, Q., Kuang, Y., Sommerfeld, M.: Growth and neutral lipid synthesis in green microalgae: A mathematical model. Bioresource Technology, 102, pp. 111–117 (2011)
21. He, Y., Chen, L., Zhou, Y., Chen, H., Zhou, X., Cai, F., Huang, J., Wang, M., Chen, B., Guo, Z.: Analysis and model delineation of marine microalgae growth and lipid accumulation in flat-plate photobioreactor. Biochemical Engineering Journal, 111, pp. 108–116 (2016)
22. Mesquita, T.J.B., Sargo, C.R., Neto, J.R.F., Paredes, S.A.H., Giordano, R.D., Horta, A.C.L., Zangirolami, T.C.: Metabolic fluxes-oriented control of bioreactors: A novel approach to tune micro-aeration and substrate feeding in fermentations. Microbial Cell Factories, 18(1) (2019)
23. Shomal, R., Hisham, H., Mlhem, A., Hassan, R., Al-Zuhair, S.: Simultaneous extraction-reaction process for biodiesel production from microalgae. Energy Reports, 5, pp. 37–40 (2019)
24. Taher, H., Giwa, A., Abusabiekeh, H., Al-Zuhair, S.: Biodiesel production from Nannochloropsis gaditana using supercritical CO2 for lipid extraction and immobilized lipase transesterification: Economic and environmental impact assessments. Fuel Processing Technology, 198 (2019)
25. Tebbani, S., Lopes, F., Becerra-Celis, G.: Nonlinear control of continuous cultures of Porphyridium purpureum in aphotobioreactor. Chemical Engineering Science, 123, pp. 207–219 (2015)
26. Arrieta-Ramos, O., Alfaro-Ruiz, V.M.: Sintonización de controladores PI y PID utilizando los criterios integrales IAE e ITAE. Ingeniería, 13, pp. 31–39 (2003)
27. Aguilar-López, R., Martínez-Guerra, R., Maya-Yescas, R.: Temperature regulation via pi high-order sliding-mode controller design: Application to a class of chemical reactor. International Journal of Chemical Reactor Engineering, 7, pp. 1–16 (2009)
28. Peña-Caballero, V., López-Pérez, P.A., Neria-González, M.I., Aguilar-López, R.: A class nonlinear adaptive controller for a continuous anaerobic bioreactor. Journal of Scientific and Industrial Research, 71, pp. 480–483 (2012)

29. López Pérez, P.A., Neria-González, M.I., Aguilar-López, R.: Increasing the bio-hydrogen production in a continuous bioreactor via nonlinear feedback controller. International Journal of Hydrogen Energy, 48, pp. 17224–17230 (2015)
30. López-Pérez, P.A., Peña-Caballero, V., Ruiz-Camacho, B., Aguilar-López, R.: Increasing of lipid productivity in microalgae cultures via dynamic analysis and closed-loop operation. European Chemical Bulletin, 6, pp. 145–150 (2017)
31. Luedeking, R., Piret, E.L.: A kinetic study of the lactic acid fermentation. Batch process at controlled pH. Biotechnology and Bioengineering, 1, pp. 393–412 (1959)
32. Loubiére, K., Olivo, E., Bougaran, G., Pruvost, J., Robert, R., Legrand, J.: A new photobioreactor for continuous microalgal production in hatcheries based on external-loop airlift and swirling flow. Biotechnology and Bioengineering, 102, pp. 132–147 (2009)
33. Rivera, D.E., Morari, M., Skogestad, S.: Internal model control 4: Pid controller design. Industrial and Engineering Chemistry Process Design and Development, 25, pp. 25–265 (1986)
34. Ogunnaike, B.A., Ray, W.H.: Process dynamics, modeling, and control. Oxford University Press, pp. 1260 (1994)

Panorama de la gestión de la transformación ágil

Yadira Jazmín Pérez Castillo, Sandra Dinora Orantes Jiménez

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

yaddy9011@gmail.com,dinora@cic.ipn.mx

Resumen. El enfoque ágil ha tenido gran éxito en los últimos años, sin embargo, el proceso de transformación entre un enfoque ágil y uno tradicional aún cuenta con mucha incertidumbre y no ha logrado consolidarse totalmente; por lo que este trabajo de investigación, intenta proveer una revisión de los problemas a los que se puede enfrentar un equipo de trabajo a la hora de emprender una transformación ágil, a lo cual, se le denomina anomalías para la transformación ágil. Por otra parte, este trabajo también abordará un modelo conocido como *Agileroadmap* que se ha venido consolidando desde hace un par de años y que permite apoyar la gestión de transformación ágil, con la finalidad de guiar en el camino a un equipo de trabajo a implantar prácticas acordes al contexto del equipo, solventando algunas de las anomalías antes mencionadas. Derivado del análisis del modelo, se determinan las bases para desarrollar un constructo que soporte esta gestión de transformación.

Palabras clave: Enfoque ágil, transformación ágil, anomalías de implantación ágil, constructo.

Agile Transformation Management Overview

Abstract. The agile approach has been very successful in recent years, however, the transformation process between an agile approach and a traditional one still has a lot of uncertainty and has not been fully consolidated; Therefore, this research work attempts to provide a review of the problems that a work team may face when undertaking an agile transformation, which is called anomalies for agile transformation. On the other hand, this work will also address a model known as Agileroadmap that has been consolidating for a couple of years and that allows supporting agile transformation management, in order to guide a work team on the way to implement practices according to the context of the team, solving some of the aforementioned anomalies. Derived from the analysis of the model, the bases are determined to develop a construct that supports this transformation management.

Keywords: Agile approach, agile transformation, agile implementation anomalies, construct.

1. Introducción

Debido a la gran aceptación que ha tenido el enfoque ágil en los últimos años, muchas empresas dedicadas al desarrollo de software y en general sus equipos de trabajo (desarrolladores), han optado por realizar un cambio de paradigma a través de prácticas y/o metodologías ágiles como lo son Scrum, XP (extreme Programming), Lean, DSDM (Dynamic Systems Development Method), Crystal y Kanban, entre otras; dicho enfoque ágil, se encuentra presente en múltiples organizaciones a nivel mundial, tal como se expresa en el *14th Annual State Of Agile Report* [1].

Lo anterior derivado de que en 2001, Bob Martin realizó una reunión con 16 líderes del movimiento ágil, para escribir el denominado Manifiesto Ágil [2]. La idea era englobar los modelos como "*Metodologías de Desarrollo de Software de peso liviano*", siendo estás una alternativa a las metodologías tradicionales, las que consideraban pesadas y rígidas por su carácter normativo y fuerte dependencia con las planificaciones detalladas previas al desarrollo de software.

Como resultado de esta reunión se obtuvieron los valores principales sobre los que se basan los métodos ágiles y que quedan establecidos en cuatro postulados, a los que se ha llamado *Manifiesto Ágil*, descrito a continuación:

"Estamos poniendo al descubierto mejores métodos para desarrollar software, haciéndolo y ayudando a otros a que lo hagan. Con este trabajo hemos llegado a valorar:

A los individuos y su interacción, por encima de los procesos y las herramientas.

El software que funciona, por encima de la documentación exhaustiva.

La colaboración con el cliente, por encima de la negociación contractual.

La respuesta al cambio, por encima del seguimiento de un plan.

Aunque hay valor en los elementos de la derecha, valoramos más los de la izquierda"[2].

El Manifiesto Ágil define la filosofía del enfoque ágil, dando un marco a cómo se espera que se desarrolle el trabajo, los procesos y actividades que pretendan ser ágiles, estos postulados vienen acompañados de *12 principios del Manifiesto Ágil* (se pueden encontrar en [2]). Estos principios van guiados de la mano por características tales como: mejora continua, calidad desde el primer día, colaboración continua, incorporación al cambio, priorización de requerimientos, entre otras.

Tomando datos recopilados en [1] es posible afirmar que acelerar la entrega de software y mejorar la capacidad de gestionar el cambio, siguen siendo actualmente las principales razones indicadas para realizar una transformación ágil, tal como se aprecia en la Fig. 1 en donde se reporta el año 2020.

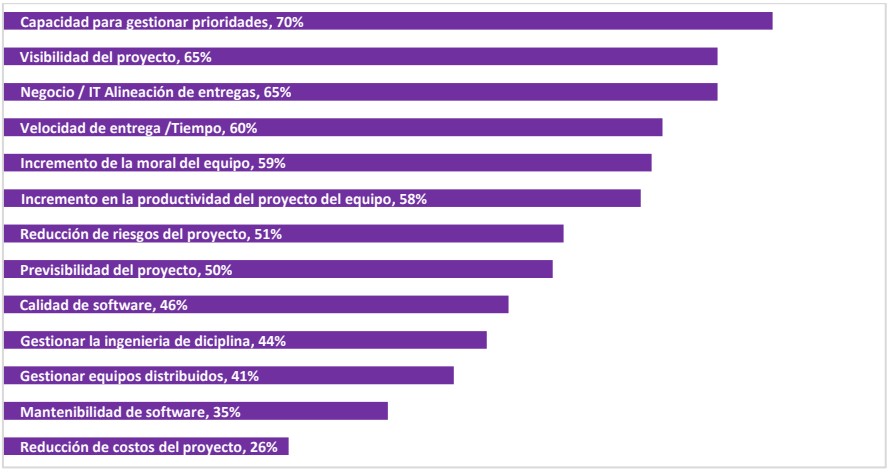


Fig. 1. Beneficios de iniciar una adopción ágil, State of Agile Survey, 2020 [1].

Por consiguiente, el uso de este nuevo enfoque ágil brinda muchos beneficios a las empresas que los adoptan, especialmente para sus equipos de trabajo.

Sin embargo, a pesar del rotundo éxito que tiene el enfoque ágil, lograr su incorporación a un equipo de trabajo puede ser complicado, por diversas peculiaridades a las que se enfrente el equipo u organización al momento de emprender el camino de transformación ágil, mismas que no permiten alcanzar una implantación de prácticas ágiles completamente y acorde. A estas situaciones en adelante se le denominan como "anomalías" para la transformación ágil.

Por consiguiente, este trabajo tiene la finalidad de dar a conocer algunas de estas anomalías que se han encontrado dentro de algunas empresas/organizaciones que requieren realizar un proceso de transformación ágil, además de esbozar una de las estrategias posibles para hacerles frente.

Este documento se encuentra organizado de la siguiente manera: en la *sección 2*, se dan a conocer lo que son y qué se consideran anomalías dentro de proceso de transformación ágil, tomando en cuenta algunas de las que se han encontrado en la actualidad. En la *sección 3*, se da a conocer una estrategia que ha tenido éxito en los últimos años y que apoya el proceso de transformación ágil, así como sus pros y contras. Esta estrategia tiene la finalidad de apoyar a mitigar las anomalías abordadas en la sección 2. Asimismo, en la *sección 4*, se abordan los resultados a manera de discusión sobre la estrategia planteada, dando también un breve resumen del trabajo realizado, destacando los aportes que trae consigo esta investigación. Finalmente, dentro de la *sección 5* se establecen las conclusiones del trabajo realizado y comentarios acerca de posibles trabajos futuros.

2. Anomalías en el proceso de transformación ágil

El proceso de incorporación de un método ágil, no es una tarea sencilla, tiene sus desafíos, ya que no basta con conocer a fondo sus prácticas al pie de la letra, por ejemplo, las correspondientes al marco de trabajo SCRUM.

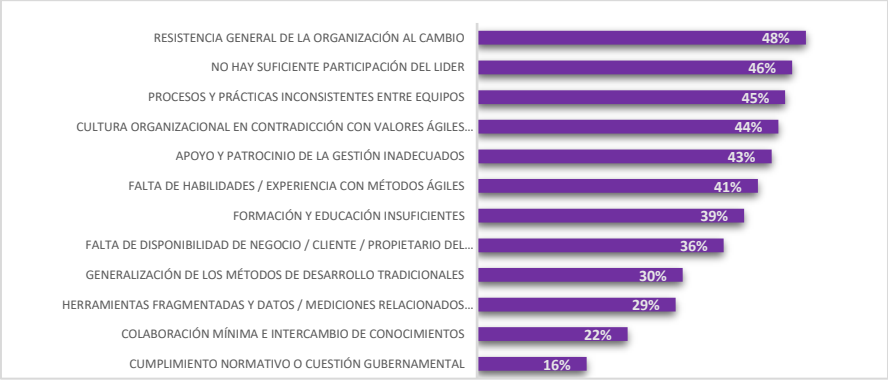


Fig. 2. Obstáculos para la adopción de la agilidad, State of Agile Survey, 2020 [1].

En general se trata de un conjunto de características que hacen que su incorporación a una empresa o equipo de trabajo se de manera clara y concisa.

En [1], se ha logrado analizar mediante una encuesta a diversas empresas, cuales son las principales barreras para lograr un transformación ágil. Estas barreras se muestran en Fig. 2 y refleja que los problemas culturales de la organización, siguen siendo las principales dificultades para adoptar y escalar ágilmente. La resistencia general al cambio, el apoyo administrativo o por parte de dirección, procesos y prácticas inconsistentes entre los equipos, se clasifican como los tres principales desafíos.

Los obstáculos para una adopción ágil vienen dados por diversas anomalías dentro del proceso de transformación ágil, estas de alguna manera, recaen en el éxito o fracaso de la organización que quiere emprender un camino ágil. Algunas de estas anomalías fueron clasificadas y analizadas en [3] como anomalías en la implantación de prácticas ágiles y se detallan en Tabla 1.

Por otro lado, un claro ejemplo que podría entrar dentro de la clasificación de las anomalías mencionadas anteriormente y que ha sonado últimamente, es el termino *Flaccid Scrum* [4], que viene siendo una especie de "ser ágil, pero…" y es representado de la siguiente manera:

> *"Quieren usar un proceso ágil y eligen Scrum,*
> *Adoptan las prácticas de Scrum y tal vez incluso los principios.*
> *Después de un tiempo, el progreso es lento porque la base del código es un desastre* [4]*."*

Lo que sucede es que no se ha prestado suficiente atención a la calidad interna del software. Si se comete este error, pronto se encontrará que la productividad se verá reducida porque es mucho más difícil agregar nuevas funciones de lo que se piensa [4]. Esta anomalía recae, en resumen, en la calidad del producto, ya que no se trata solo de tomar las prácticas tal cual, con tal de entregar un producto funcional rápidamente, sino también se deben de considerar la calidad con la que se está desarrollando, prestando atención al producto, quizás tomando en consideración el MPV (Mínimo Producto Viable), el cual tiene como objetivo satisfacer la funcionalidad mínima del producto.

Estos ejemplos de anomalías que se han abordado quedan sustentados en que el 95% de los encuestados, tanto en el reporte de agilidad [1], en donde los encuestados

Tabla 1. Anomalías en la implantación de prácticas ágiles [3].

Implantación "anecdótica".	El comentario suele ser como el siguiente "hicimos un proyecto de forma ágil, nos fue bien y hasta resulto entretenido, pero ahora seguimos trabajando como siempre".
Implantación "inconsciente"	Cuando se asegura que se está trabajando de forma ágil (incluso indicando un método ágil específico) pero se tiene muy poca argumentación respecto a qué prácticas se están aplicando y en qué nivel de intensidad o cuáles prácticas explícitamente se han postergado en su aplicación o cuáles simplemente se han descartado.
Implantación "limitada"	Cuando solo se ha implantado una o muy pocas prácticas ágiles, sin la intención de continuar con la iniciativa. Por ejemplo, muchos equipos se conforman con implantar un proceso iterativo e incremental para autodenominarse ágiles. El proceso iterativo e incremental es fundamental en Scrum y en Extreme Programming, pero ni es una exclusividad del enfoque ágil (la metodología tradicional RUP también sigue un proceso iterativo e incremental) ni tampoco es lo único que ofrece el enfoque ágil en cuanto a prácticas, el agilismo es mucho más que desarrollo iterativo e incremental. Es más, el proceso aplicado podría NO ser iterativo (como el propuesto por el método Kanban) y aun así ser totalmente válido como enfoque ágil.
Implantación "todo o nada"	Cuando un equipo trabaja algunos proyectos con enfoque ágil y otros con enfoque tradicional, de forma totalmente alternativa. Si bien existen prácticas excluyentes entre el enfoque ágil y el tradicional, no es muy sensato plantear el enfoque ágil como un todo o nada. Hay muchas prácticas ágiles que podrían aplicarse de forma complementaria en un enfoque tradicional. Además, en general es inviable aplicar todas las prácticas ágiles a la vez, o al menos hacerlo en un plazo relativamente corto pues muchas prácticas requieren un esfuerzo de preparación considerable antes de su aplicación (por ejemplo, aplicar pruebas automatizadas).
Implantación "la vida sigue igual"	Cuando en teoría ya se ha implantado el enfoque ágil, pero en opinión de los involucrados parece que no ha habido grandes cambios; por ejemplo, "nuestro jefe sigue repartiendo faena", "se sigue planificando con diagramas Gantt", "seguimos especificando todo y en detalle al principio, incluso lo que no se implementará a corto plazo", etc. Esta anomalía es en gran medida consecuencia directa de una "Implantación limitada".

expresaron que al menos algunos de sus proyectos ágiles han tenido éxito y el 48% informó que la mayoría o todos sus proyectos ágiles tuvieron éxito. Confirmando que, aunque la gestión ágil de equipos de trabajo es una clara tendencia en la industria, no se ha alcanzado un porcentaje óptimo de éxito en los proyectos en los cuales se utiliza un enfoque ágil en los últimos años.

Las anomalías dentro de la transformación ágil también podrían resumirse en que, como se expresa en [5], se está “Haciendo Ágil” cuando solo se siguen las prácticas, y se está “Siendo Ágil” cuando se actúa con una mentalidad ágil, lo que hace notar que existe una gran barrera entre hacer y ser ágil.

Lo anterior puede derivar de que no se tienen claros dos conceptos: adopción y transformación ágil. Es importante que un equipo que quiere alcanzar la agilidad tenga clara diferenciación entre estos dos conceptos, por lo que citando a [6]: “ágil se describe

mejor como una sombrilla de valores y principios y debajo de ella son muchos conjuntos diferentes de *frameworks* que ayudan a una organización a lograr la agilidad. Así, "la agilidad es la meta", mientras que Scrum, Lean, Kanban u otro tipo de *framework* es a menudo el "cómo". "Adopción" es el acto de tomar o poner en práctica alguna cosa y "Adopción ágil" es "hacer" ágil.

Por otro lado, cuando se habla de "Transformación ágil" es "ser, convertirse o cambiar el carácter o condición de agilidad". Esto es mucho más difícil de lograr. Se trata de un cambio de mentalidad en todas las personas de una organización que puede ser incómodo para la mayoría".

Otro concepto relevante, que se destaca en fuentes y trabajo como en [7] y [8], que vendría siendo el futuro de la agilidad, es el post-agilismo que simplemente se refiere a "hacer lo mejor para ti", por encima de seguir una metodología ágil al detalle. Seguir los principios base de los métodos ágiles sin seguir una metodología ágil al pie de la letra, enfatizando que el post-agilismo no es anti-agilismo, sino es su evolución.

Lo anterior contribuye a la diferenciación entre adopción y transformación, concluyendo que no se trata simplemente de adoptar unas cuantas prácticas, sino un cambio de pensamiento completo, lo que incita a tomar en cuenta las mejores prácticas de cada metodología ágil, considerando que sean las más adecuadas al equipo de trabajo.

Por lo tanto, una gestión de transformación debe cumplir ciertas características para poder apoyar en la reducción de las anomalías dentro de la implantación del enfoque ágil, por lo que estas particularidades serán analizadas dentro de la sección 3.

3. Estrategia para la gestión de la transformación ágil

Dentro de las metodologías ágiles es posible encontrar varias estrategias para conseguir implantarlas, como por ejemplo las prácticas a seguir, además de establecer reuniones, seguir ciertos lineamientos, entre algunas otras. Sin embargo, no abordar este enfoque correctamente conllevaría tener ciertas anomalías dentro de su incorporación a un ambiente de trabajo.

Por lo tanto, con base en que se ha demostrado dentro de la sección 2, se afirma que se necesita de una gestión de transformación ágil adecuada y renovada para hacer frente a las anomalías antes abordadas.

Conocer una metodología ágil, como por ejemplo Scrum, no debería implicar que solo se tome en cuenta está metodología y sus prácticas, lo ideal sería poder tomar lo mejor de todas las metodologías ágiles existentes, manteniendo un catálogo de prácticas ágiles que tome lo mejor de todos ellos, ya que no siempre pueden tomarse en cuenta todas las prácticas de uno solo, tal vez porque no aplican todas para un entorno de trabajo.

Es sumamente importante conocer el contexto de equipo de trabajo, como, por ejemplo, el tipo del proyecto que realizará, el ambiente de trabajo en donde se llevará a cabo el proyecto, las habilidades de cada persona correspondiente al equipo, el área a la que pertenece al equipo, ya que como se ha comentado, no siempre tiene que ver con software o TI (Tecnologías de la Información). Destacando que lo más importante es hacer una retrospectiva completa del equipo antes de comenzar un proceso de transformación ágil.

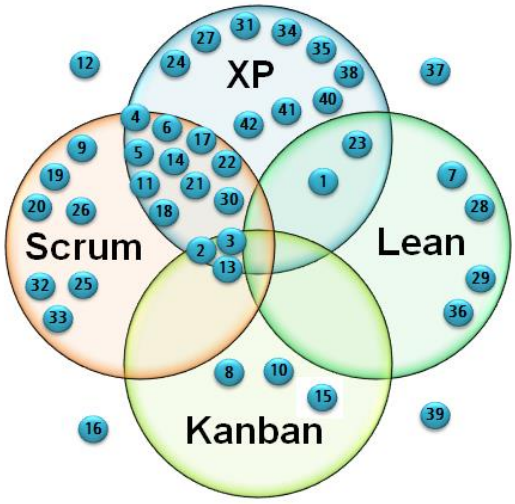


Fig. 3. Solape de prácticas entre los métodos ágiles más populares. Fuente: [11].

Para apoyar el análisis del contexto y dar inicio a una implantación de prácticas ágiles acordes, es necesario contar con un *roadmap* (mapa de ruta) que apoye en la sección de prácticas ágiles.

Esta elección de prácticas puede darse de distintas maneras, como, por ejemplo, de forma intuitiva y/o manual. No obstante, para no realizar una selección sin sentido se requiere de una estrategia de apoyo, que incluya objetivos de mejora que un determinado equipo de trabajo requiera alcanzar. Además, para estos objetivos es deseable se valore el nivel de agilidad o de progreso que se va alcanzando, como parte de la estrategia de apoyo al *roadmap*.

Esta estrategia que se ha abordado requiere de un constructo/herramienta que soporte este proceso de gestión de implantación planteado, por lo que en el estudio realizado en [9] se dan a conocer algunas que proveen un apoyo o soporte para dicho proceso. De este estudio se analizaron aquellas herramientas, métricas y modelos que ayudan a determinar un nivel de agilidad de partida como apoyo a la gestión de transformación ágil, sin embargo, no todas estas cumplen con las características deseables para una gestión de implantación adecuada.

No obstante, del estudio se destacan un modelo, *Agileroadmap*[10] y dos herramientas, *Agileroadmap+* [11] y un *formulario para determinar nivel de agilidad* [12] (FDNA), estas últimas dos herramientas se complementan, sin embargo se encuentran en entornos separados.

El modelo *Agileroadmap* se basa en un catálogo de 42 prácticas ágiles seccionadas de los métodos ágiles más comunes, como lo son SCRUM, KANBAN, XP Y LEAN, su relación se puede contrastar en la Figura 3. Su incorporación ha tenido éxito en varios casos de estudio aplicados, la información más a fondo puede ser localizada en [11].

La estrategia del modelo se basa principalmente en: la selección de prácticas ágiles acodes al equipo de trabajo y el nivel de aplicación de cada práctica candidata; así como los desafíos que pueda presentar su implementación; además considera objetivos de mejora y su importancia, con la finalidad de apoyar en la selección de prácticas.

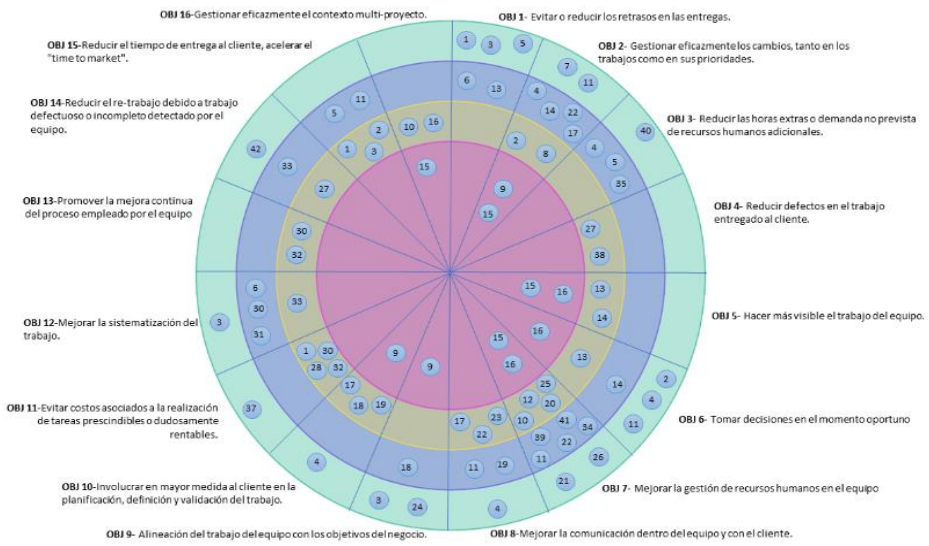


Fig. 4. Solapamiento de la relación entre objetivos y prácticas. Fuente: [11].

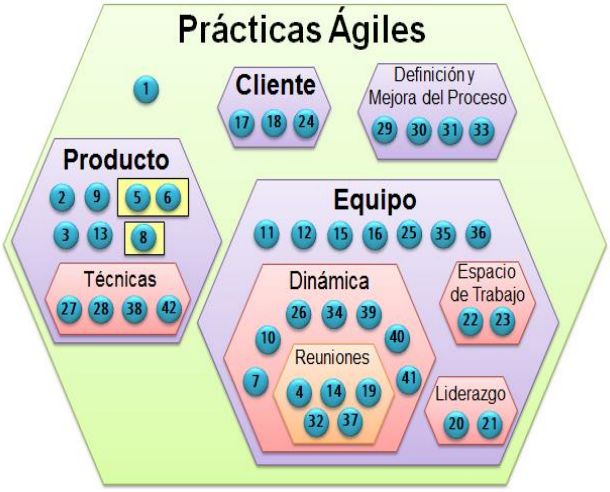


Fig. 5. Agrupación de prácticas ágiles por categorías. Fuente: [11].

Para lo anterior se considera una relación entre prácticas y objetivos, misma que se puede visualizar en la Fig. 4.

Por otro lado, las prácticas correspondientes al modelo, también se encuentran clasificadas en áreas de trabajo, como se muestra en la Fig. 5, esta relación permite ayudar también a identificar qué áreas y sus prácticas correspondientes serían las ideales para poder comenzar a implantar en un equipo de trabajo. Para conocer más en detalle sobre la relación de estas prácticas y objetivos, además de ver el catálogo de las mismas, se puede acudir al sitio proporcionado por [11].

Como se ha abordado, existen dos herramientas que soportan el modelo de *Agileroadmap*. La primera *AgilevRoadmap+* [11], toma como base el modelo presentado anteriormente y se encuentra disponible para todo usuario.

La herramienta principalmente se basa en proveer un *roadmap* de apoyo para la sección de prácticas ágiles de manera manual, se ha probado en varios casos de estudio igualmente y permite ir modificándolo a parecer del usuario.

Por otro lado, se cuenta con la herramienta FDNA[12], que es un tipo formulario de *Google Form* que permite conocer el nivel de agilidad de un equipo de trabajo, basándose en los objetivos y áreas de implementación, el reporte es enviado vía correo electrónico. También toma en cuenta el modelo de *Agileroadmap*.

Al analizar estas herramientas se puede detectar que ambas se complementan, pero se encuentran en diferentes ambientes de trabajo, encontrando algunas oportunidades de mejora/deficiencias que podrían dar apertura a una nueva herramienta mejorada en un futuro. Estas deficiencias quedan englobadas en:

1. Ambas herramientas se encuentren separadas en ambientes tecnológicos y desarrollados con tecnologías diferentes.
2. Al generar una evaluación en el formulario de Google Form, el usuario debe de ingresar los mismos datos que ingreso en *Agileroadmap+*, como por ejemplo el nivel de aplicación de cada práctica, el nivel de desafío, etc.
3. No existe un histórico de evaluaciones dentro del formulario, el usuario solo se queda con un documento PDF, lo deseable sería que en una sola herramienta se le puede ir mostrando un avance de progreso de implantación de sus prácticas.
4. El usuario no genera varios *roadmap* dentro de la herramienta *Agileroadmap+*, lo que también sería deseable, puesto que una organización puede tener varios equipos de trabajo con diferentes líneas de trabajo.

Con la información contenida hasta el momento ha podido analizarse un modelo y dos herramientas complementarias que se encuentran disponibles y que ayudan en una gestión de trasformación ágil tal como se ha venido planteando, dependiendo del contexto del equipo y de los objetivos que este quiera alcanzar.

Resumiendo, que una integración de las herramientas abordadas podría dar apertura a una mejora en el proceso de transformación ágil, solventando las deficiencias de las herramientas actuales.

4. Aportes de investigación

De acorde a lo analizado en las secciones anteriores, como parte de los resultados y tomando como referencia el modelo ofrecido por *Agileroadmap*, es importante que una organización o equipo de trabajo que quiere comenzar su camino hacia la agilidad tome en consideración realizar una gestión de transformación ágil, tomando en cuenta lo siguiente:

1. Conocimiento del contexto donde se implantarán las prácticas ágiles.
2. Una colección de las mejores prácticas ágiles correspondientes al enfoque ágil en general.

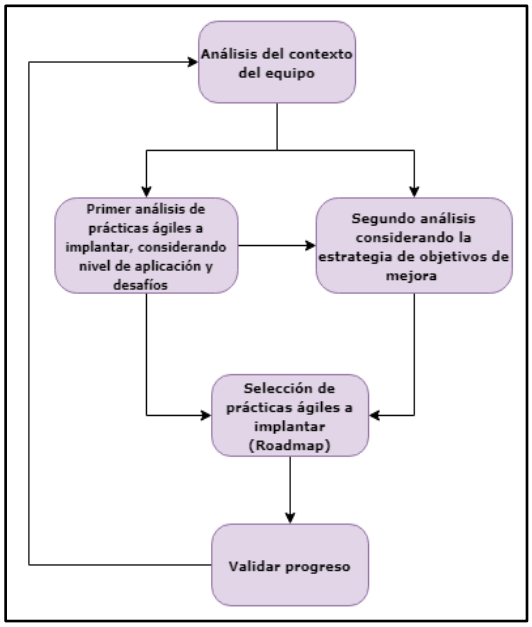


Fig. 6. Estrategia para una gestión de implantación ágil. Elaboración propia-basada en [11].

3. Un roadmap para apoyar la selección de las prácticas ágiles a implantar.

4. El roadmap requiere de una estrategia de apoyo, que incluya objetivos de mejora que un equipo de trabajo requiera alcanzar.

5. Para estos objetivos es deseable se valore el nivel de agilidad o de progreso que se va alcanzando, como parte de la estrategia de apoyo al roadmap.

Resumiendo lo anterior en el diagrama de la Figura 6, es importante que se haga un análisis constante de la situación del equipo para poder ir evaluando el progreso de la implantación de las prácticas seleccionadas y de esa forma ir incorporándolas acordes al equipo, guiados por la estrategia basada en objetivos de mejora continua.

Un primer análisis para generar un *roadmap* es dado por únicamente una selección de prácticas a priori, es decir, manual, solo guiados por el contexto del equipo. Una segunda selección puede ser más objetiva, tomando en cuenta los objetivos que se quieren alcanzar y las prácticas con las que tiene correspondencia. Esto llevará a realizar un análisis más concreto y con sentido, este proceso se puede repetir cuantas veces sea necesario.

La finalidad es ir incorporando las prácticas progresivamente, hasta llegar a que el equipo las conozca a fondo y se familiarice poco a poco con ellas, sin tener que depender de seguir una metodología ágil al pie de la letra, sino orientarlo a ir tomando lo mejor de cada práctica, considerando que esté al alcance del equipo. Este modelo incorporado a un constructo mejorado podría tener grandes beneficios para apoyar la gestión de transformación ágil, ya que, daría un contexto de la situación general de un equipo de trabajo, también permitiría ir midiendo el progreso de incorporación de

prácticas, considerando que este análisis (Evaluación de agilidad) podría hacerse por objetivos y áreas.

5. Conclusiones y trabajos futuros

El presente trabajo presenta un panorama general de la transformación ágil, abordando lo que se consideran anomalías (sección 2), que se definen como aquellos obstáculos para realizar un proceso de cambio entre un enfoque tradicional y uno ágil, ya que, si bien el enfoque ágil ha tenido éxito en los últimos años, existen barreras actualmente que impiden que se lleve a cabo una implantación del enfoque acorde. Lo anterior hace cuestionar que modelos métricas o herramientas existen dentro del mercado que puedan soportar dicha implantación ágil y apoyen a contrarrestar dichas anomalías. Dentro de un estudio realizado en [9], se constata que actualmente no existe una herramienta lo suficientemente completa, sin embargo si se encuentran un modelo (*Agileroadmap*) y dos herramientas (*Agileroadmap++ y Formulario de evaluación de agilidad*) que se complementan para proveer un apoyo para la gestión de la transformación ágil y que tienen como finalidad apoyar a un equipo de trabajo a alcanzar el objetivo de emprender la agilidad, tomando en cuenta las mejores prácticas acorde a la estrategia planteada.

Dentro de la sección 3, como parte de la estrategia para la gestión de la transformación ágil, se muestra y detalla la propuesta del modelo *Agileroadmap*, el cual pretende ayudar en el emprendimiento del camino hacia el enfoque ágil, llegando a la conclusión de que existe un área de oportunidad para poder solventar las deficiencias que tiene el modelo y las herramientas que los soportan, como, por ejemplo, la realización de una actualización tecnológica, mejoras en su fórmula del cálculo del nivel de agilidad, etc. Resumiendo, en la sección 4, como parte de aportes de investigación, la estrategia planteada que incorpora el modelo de *Agileroadmap*, derivada de la investigación.

Como parte de un trabajo a futuro se tiene la intención de desarrollar una herramienta/constructo que considere el modelo de *Agileroadmap* y que considere aspectos no tomados en cuenta en el momento de su elaboración. Con esto también se podría contar con el conocimiento, por ejemplo, de que prácticas son las utilizadas en los equipos de trabajo, lo que dejaría apertura a otra área de investigación y sería de gran aporte a la Ingeniería de Software (IS), en específico al conocimiento y estudio del enfoque ágil. También se podría evaluar el impacto de esta nueva herramienta/constructo en un caso de estudio aplicado, con esta evaluación se podría determinar su éxito y funcionalidad dentro de un equipo de trabajo.

Es claro el avance que se tiene hoy en día dentro de campo de la IS, en especial el agilismo, que ha sido una corriente que ha tomado mucho interés por la filosofía que maneja, misma que pretende apoyar a las organizaciones en su mejora de procesos de desarrollo de software, aunque actualmente se emplea además en otros ámbitos como la elaboración de productos, servicios etc.

La agilidad siempre está en constante evolución, por lo que este tema de investigación ha tenido bastante relevancia en los últimos años por varios investigadores que intentan mejorar los procesos para desarrollar ágilmente. Sin embargo, aún al día de hoy existe mucha incertidumbre en adoptar un paradigma ágil.

La estrategia planteada en este documento esboza una forma que pretende apoyar en la adopción de un enfoque ágil, tomando en cuenta todas las características generales de los considerados los métodos ágiles más empleados de los que existen actualmente, dando la facilidad de determinar cuáles son las prácticas adecuadas para comenzar a implantar y cuáles son los desafíos que se enfrentarían al tomarla, apoyándose para la selección de éstas en objetivos de mejora que se quieran alcanzar y su relación con ellas.

Referencias

1. VersionOne: 14th Annual State Of Agile Report, pp. 19 (2020)
2. Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., et. al.: Manifesto for agile software development. The Agile Alliance (2001)
3. Letelier-Torres, P.: Algunas anomalías en implantación de prácticas ágiles (2015)
4. Fowler, M.: FlaccidScrum (2020)
5. Sahota, M.: Una guía de supervivencia a la adopción y transformación ágil: trabajando con cultura organizacional. Gazafatonarioit (2012)
6. CAST: Agile transformation: Understanding what it means to be agile. Airfocus (2019)
7. Baskerville, R., Pries-Heje, J., Madsen, S.: Post-agility: What follows a decade of agility?. Information and Software Technology, 53(5), pp. 543–555 (2011)
8. Letelier-Torres, P.: Agility at work (2013)
9. Pérez-Castillo, Y.J., Orantes-Jiménez, S.D., Letelier-Torres, P.: Estudio de herramientas para determinar el nivel de agilidad en empresas de desarrollo de software (2019)
10. Amancio, N., Isaías, F.: AGILE Roadmap: Diagnóstico y evaluación de prácticas ágiles para ser implementadas en equipos de trabajo. Semantic Scholar (2014)
11. Pérez, M., Letelier-Torres, P.: Diseño de un modelo y construcción de una herramienta de apoyo para evaluar prácticas ágiles aplicables en equipos de trabajo (2014)
12. Palacio-León, A.: Evaluación del grado de agilismo basado en los objetivos y necesidades de los equipos de trabajo. Tesis de Universitat Politecnica de Valencia, pp. 52 (2015)

Propuesta de una tasa de desarrollo económico que contempla variables subjetivas para el caso de México

Miguel Ángel Sánchez García[1], Abraham Ramírez García[1], Agustín Ignacio Cabrera Llanos[2], Ana Lorena Jiménez Preciado[1]

[1] Instituto Politécnico Nacional,
Escuela Superior de Economía,
México

[2] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria de Biotecnología,
México

ajimenezp@ipn.mx

Resumen. En la presente investigación se propone una medición alternativa de bienestar económico para México que permita medir los niveles subjetivos de bienestar sin dejar de lado la coyuntura económica actual. Para tales fines se diseñó un modelo basado en lógica difusa que fue construido a partir del Índice de Desarrollo Humano (IDH), el indicador del consumo privado en el mercado interior y el indicador de la percepción sobre seguridad pública. El elemento distintivo de este indicador con respecto a otros ya existentes es que la inclusión de lógica difusa permite modelar funciones no lineales e incorporar elementos de percepción o subjetividad.

Palabras clave: Lógica difusa, indicadores subjetivos, desarrollo humano.

Proposal of an Economic Development Rate that Contemplates Subjective Variables for the Case of Mexico

Abstract. In this research, an alternative measurement of economic well-being is proposed for Mexico that allows measuring subjective levels of well-being without neglecting the current economic situation. For these purposes, a model based on fuzzy logic was designed that was built from the Human Development Index (HDI), the indicator of private consumption in the domestic market and the indicator of perception of public security. The distinctive element of this indicator with respect to other existing ones is that the inclusion of fuzzy logic allows modeling non-linear functions and incorporating elements of perception or subjectivity.

Keywords: Fuzzy logic, subjective indicators, human development.

1. Introducción

Con el paso del tiempo han surgido diferentes escuelas que sostienen que el crecimiento económico no es una métrica adecuada para medir el desarrollo económico. Dentro de estas escuelas se encuentra la Economía Social de Mercado (ESM) la cual afirma que es posible armonizar el libre mercado con ciertas medidas compensatorias que coadyuven a lograr una mayor justicia social (Schaeffler, 2004). Esto se ha intensificado a raíz de la cada vez más amplia brecha de desigualdad y riqueza que se registra en los países, en donde un mayor crecimiento económico no necesariamente implica mayor desarrollo. Derivado de ello, la necesidad de un índice de bienestar que incorpore variables subjetivas además de las clásicas ha tomado mayor relevancia con el paso del tiempo.

Previo a la crisis hipotecaria de 2008-2009, ya se contaba con una gran cantidad de propuestas de diversas metodologías que permitan medir de mejor manera el desarrollo económico y que actualmente se han modificado, mejorado y siguen prevaleciendo. Una de las medidas de desarrollo más utilizadas a nivel mundial es el Índice de Desarrollo Humano (IDH) diseñado por el Programa de las Naciones Unidas (PNUD), dicho indicador busca enfatizar que las personas y sus capacidades sean el criterio fundamental para evaluar el desarrollo de un país y no solo el crecimiento económico (United Nations Development Programme (UNDP), 2019).

El IDH incorpora tres aspectos fundamentales para la construcción de su indicador: salud, educación y riqueza. Si bien, incorpora elementos que permiten medir más allá de su ingreso (con salud y educación), el presente estudio incluye la percepción sobre inseguridad y el nivel de consumo para el caso mexicano como un primer acercamiento o una primera propuesta para una medición alternativa del bienestar. En la siguiente sección se realiza la revisión de literatura correspondiente, para posteriormente revisar la evolución de las variables y el modelo que se propone para medir el desarrollo, finalmente exponemos las conclusiones del estudio como también algunas sugerencias para futuros trabajos.

2. Breve revisión de literatura

En esencia la percepción sobre la inseguridad repercute en la calidad de vida de los ciudadanos; de acuerdo con (Jasso, 2013) un bajo nivel en la percepción de la seguridad resulta en una limitación del esparcimiento social lo cual a su vez reduce el nivel de actividad económica y bienestar en el sentido de que los agentes económicos por el miedo que les genera la inseguridad prefieren consumir lo menos posible. En ese sentido, el tema de la inseguridad ha tomado aún mayor relevancia en los últimos años, sobre todo en América Latina y países en vías de desarrollo. De acuerdo con (Wesley, 1986) la percepción de inseguridad puede deteriorar las condiciones de producción local debido a que ese factor psicológico afecta de manera negativa la productividad y el consumo lo cual se ve plasmado por los datos según (Quezada, Santillan, Hinojosa, & Rada, 2019) ya que en su estudio Percepción de inseguridad versus tasa delictiva, argumentan que un aumento del 10% en la percepción sobre inseguridad reduce en promedio 11.3 puntos porcentuales del Producto Interno Bruto (PIB) de México.

Este último estudio es preocupante ya que aun cuando no existe una clara causalidad entre crecimiento y desarrollo la correlación existente es irrefutable según (Ranis & Stewart, 2000).

Por su parte, (Muratori & Zubieta, 2016) demostraron que, para el caso de Argentina, el riesgo que se percibe por la inseguridad es un importante predictor de la actualización social y, por ende, una mayor percepción de la inseguridad afecta de manera negativa el potencial crecimiento y desarrollo de la sociedad, en este sentido se exhibe que el presente trabajo aun cuando se enfoca para el caso de México no es excluyente para otras economías.

De la breve revisión de literatura que se expuso anteriormente, se destaca que la percepción sobre la inseguridad no solo afecta de manera psicológica a la población, sino que también afecta de manera negativa a los niveles de productividad presionando los salarios a la baja y que finalmente se traduce en una reducción del consumo y por ende del bienestar individual. A continuación, se justifica el uso del nivel de consumo como una variable que repercute en el bienestar.

Existe una basta cantidad de modelos económicos que se basan en el consumo de los hogares para explicar el nivel de bienestar o utilidad, este tipo de modelo se conocen como modelos de crecimiento con optimización de consumo y se basan en el modelo de Ramsey[1]. Prácticamente toda la teoría microeconómica a través del equilibrio general busca explicar el equilibrio de mercado mediante la utilidad de los individuos y un sistema de precios.

El que es quizás la variante más popular del modelo de Ramsey es el modelo de generaciones traslapadas desarrollado por (Samuelson, 1958), en el cual se asume que cada persona vive por dos periodos, en el primer periodo es joven y trabaja y en el segundo es viejo y se retira, en este sentido, el comportamiento de los hogares se basa en la optimización de la utilidad de ambos periodos y esta proviene del consumo.

Por el lado de la teoría microeconómica, (Varian, 1992) centra el análisis de bienestar en una función que incorpora la distribución del análisis costo beneficio en la que supone una función de utilidad lineal ponderada por los pesos de bienestar que cada agente económico decida incorporar. En este sentido, tanto la teoría microeconómica como la macroeconómica centra el estudio del bienestar (desarrollo) en funciones que se encargan de optimizar la utilidad mediante el consumo.

Por su parte, en el documento que hizo ganador a Angus Deaton del premio nobel de economía del 2015, se hace mención de que el consumo de bienes y servicios es un determinante fundamental del bienestar humano (The Committee for the Prize in Economic Sciences, 2015), sin embargo, no considera que el consumo sea la única variable que es necesario incluir en el análisis del desarrollo económico ya que (Deaton, 2010) menciona que es necesario incluir en el análisis del desarrollo variables de carácter subjetivo además del ingreso y salud.

A pesar de que gran parte de la teoría económica considera que uno de los determinantes principales del bienestar proviene del consumo, en algunos casos el incremento del consumo no necesariamente implica en sí mejoras en el nivel de bienestar, principalmente cuando la capacidad del consumo ya es alta (Witt, 2016). En este sentido, consideramos que para el caso de México esta última afirmación no es necesariamente cierta ya que la capacidad del consumo en el país es baja si

[1] Véase: Barro, R. (2004) Economic Growth, The MIT press.

consideramos que el 41.9% de la población se encontraba en situación de pobreza en el 2018 según (CONEVAL, 2019).

Para demostrar en qué medida el consumo y la percepción sobre la inseguridad afectan el bienestar de México, se propone el uso de lógica difusa para modelar una tasa de desarrollo económico mediante la incorporación de elementos subjetivos e indicadores de coyuntura, en este sentido, se reitera que el modelo contempla la subjetividad de la percepción sobre inseguridad, el nivel de consumo en México, así como la evolución del IDH (que ya contempla riqueza, salud y educación).

La lógica difusa ha sido ampliamente usada en las ciencias sociales, especialmente en el área de economía y finanzas. Algunos estudios que aplican lógica difusa en este campo se pueden encontrar en (Cabrera-Llanos, Ortiz-Arango, & Cruz-Aranda, 2019) quienes desarrollaron un algoritmo basado en lógica difusa que modela un plan de gestión de mantenimiento de equipos médicos mediante un indicador que mide su importancia de acuerdo a su nivel de importancia y costo, sin embargo, no se limita a la generación de indicadores, por ejemplo en el estudio de (Abbassi, Abbassi, Heidari, & Mirjalili, 2019) se utiliza esta metodología para pronosticar precios del petróleo. En el mismo sentido, se destaca que en modelos económicos, la lógica difusa se ha implementado realizando una relación entre producto y consumo con el fin de obtener información más amplia y real que la resultante de la aplicación de modelos clásicos (Ferrer-Comalat, Corominas-Coll, & Linares-Mustaro, 2020), en este sentido, los autores incorporaron la metodología de lógica difusa asumiendo que los parámetros establecen un grado de dependencia.

Grosso modo, la metodología antes planteada permite modelar funciones no lineales, e incorporar elementos de percepción o subjetividad como el caso del equipo de mantenimiento y además de mantener las relaciones de dependencia es útil en la construcción de indicadores, de aquí la razón de utilizar esta metodología en la construcción del indicador del presente estudio.

En el siguiente apartado se realiza un análisis de las variables antes planteadas para el caso de México además de detallarse con mayor precisión la metodología para la construcción del modelo.

3. Análisis de las variables

El modelo tiene como entradas a evaluar: el Índice de Desarrollo Humano (IDH) obtenido con datos del PNUD, la percepción de seguridad de la población obtenida del INEGI y el consumo privado o de hogares obtenido del SIE Banxico y como salida obtenemos una tasa de crecimiento que toma en cuenta y pondera las variables ya mencionadas.

La elección de estas variables se basa en que el IDH muestra una parte subjetiva del desarrollo humano y lo cuantifica en un índice, pero que en este artículo se considera como incompleta y más adelante se detallará el porqué. La percepción de seguridad refleja de una manera cuantitativa el impacto de desequilibrios en la economía a través de los ojos de la sociedad en el sentido que se plantea en la revisión de literatura. Finalmente, el consumo es una parte fundamental en los modelos de crecimiento económico ya que en esta se ve reflejada tanto la demanda como la oferta de bienes y

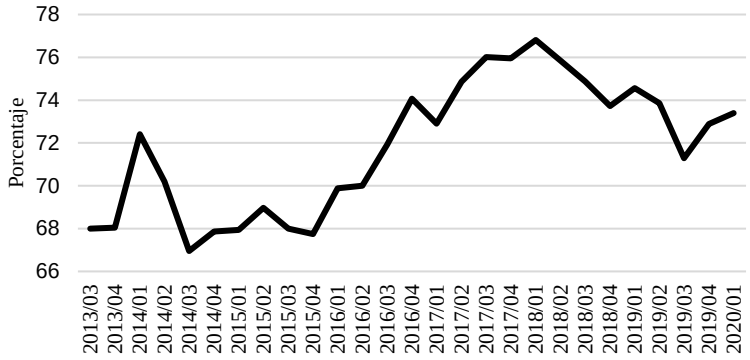


Fig. 1. Población de 18 años y más que considera insegura su ciudad (%). Fuente: Elaboración propia con datos de la Encuesta Nacional de Seguridad Pública Urbana (INEGI).

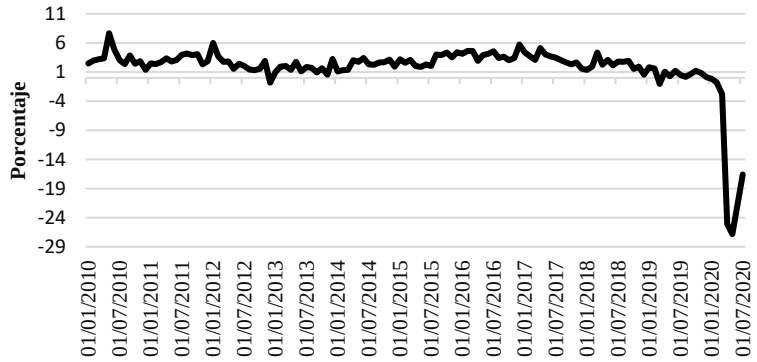


Fig. 2. Consumo privado Var (%) anual. Fuente: Elaboración propia con datos del Indicador Mensual del Consumo Privado en el Mercado Interior (INEGI).

servicios y que según la teoría microeconómica es la variable por optimizar en los modelos de equilibrio general.

Para el caso de México, la inseguridad se ha convertido en un tema fundamental ya que como se muestra en la figura 1, se puede apreciar que el porcentaje de población mayor de edad que considera que su ciudad es insegura es mayor al 65% desde el 2013. Además, en el primer trimestre del 2018 alcanzó su punto más alto al alcanzar un nivel de 76.8%, desde esa fecha el indicador cambió su tendencia hasta el penúltimo trimestre del 2019, para posteriormente presentar una tendencia creciente que probablemente continúe creciendo debido a la pérdida de empleo generada por la pandemia global de Coronavirus (Esquivel, 2020).

La segunda variable que se considera importante para la evaluación del desarrollo en México es el consumo, en la figura 2 se puede apreciar la tasa de crecimiento anual del índice mensual del consumo privado al interior del mercado para México.

Dicha variable crecía en promedio 2.9% hasta mediados de 2017, fecha desde la cual la tendencia del consumo ha cambiado, desde esa fecha, la tasa de crecimiento

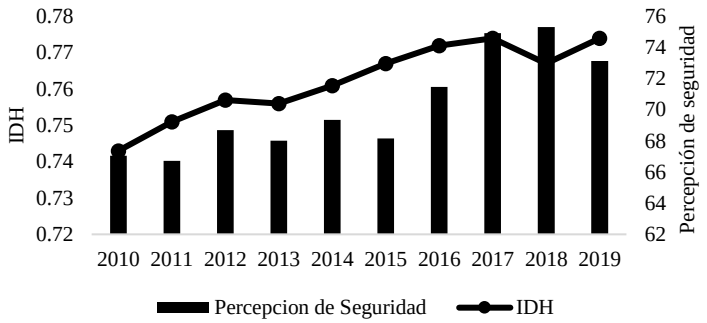


Fig. 3. Comparación entre el IDH y el Índice de Percepción de la Seguridad. Fuente: Elaboración propia con datos de la Encuesta Nacional de Seguridad Pública Urbana (INEGI) y datos de la PNUD.

Tabla 1. Grados de pertenencia.

Niveles de pertenencia	
Valor	**Grados de pertenencia**
0	Sin pertenencia
0.2	Pertenencia débil
0.5	Media Pertenencia
0.8	Pertenencia fuerte
1	Total pertenencia

promedio (hasta 2020) ha sido 1.8% mientras que en el 2020 (al menos hasta julio) la tasa de crecimiento promedio ha sido de -12.04%.

Finalmente, en la figura 3 se expone la correlación existente entre la percepción sobre la seguridad y el IDH. Por una parte, el índice de desarrollo humano presenta una tendencia creciente desde el año 2010, de acuerdo con la metodología planteada por el PNUD, México se encuentra en la categoría de desarrollo humano alto desde el año 2011, ya que se encuentra entre 0.7 y 0.8, sin embargo ¿cómo es esto posible si los datos del Coneval muestran que el 41.9% de la población en México se encuentra en alguna situación de pobreza? De aquí que en el presente artículo se considere que el IDH es un indicador que no muestra de manera adecuada el nivel de desarrollo del país y que se haya decidido incorporar nuevas variables.

Una vez que se han presentado las variables del modelo que se propone, en el siguiente apartado se desarrolla de manera más detallada la metodología que se utilizó en el presente estudio.

4. Metodología

La metodología se basa en la lógica difusa, una técnica computacional que permite llevar capacidades de razonamiento a sistemas computacionales a partir de información.

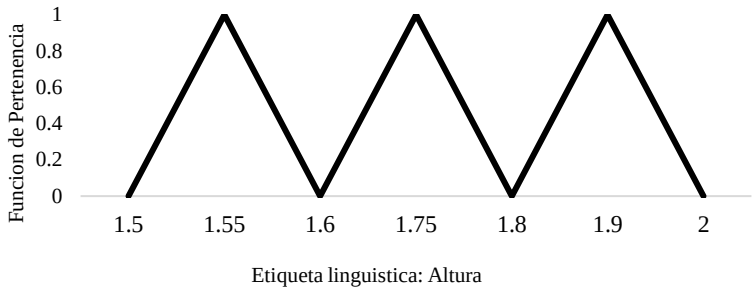


Fig. 4. Función de pertenencia. Fuente: Elaboración propia.

Tabla 2. Parámetros de los conjuntos difusos.

IDH		**Percepción de la seguridad**		**Consumo**	
Valor	**Etiqueta Lingüística**	**Valor**	**Etiqueta Lingüística**	**Valor**	**Etiqueta Lingüística**
< 0	Negativo	< 0	Negativo	< 0	Negativo
0~0.42	Bajo	0~2.2	Bajo	0~1.5	Bajo
0.25~0.74	Medio	1.5~3.5	Medio	1~3	Medio
0.58~1	Alto	2.8~4.8	Alto	2.5~4	Alto
< 1	Extraordinario	< 4.8	Extraordinario	< 4	Extraordinario

Esta técnica consta de funciones de pertenencia asociada a etiquetas lingüísticas que determinan la pertenencia de un objeto a la etiqueta lingüística, donde el cero es nula pertenencia del objeto y uno cuando el objeto pertenece en su totalidad a la etiqueta lingüística.

A ejemplo se considera la siguiente situación: ¿Qué tan alto es un objeto?, lo que determina si algo es alto o no, podría ser su medida cuantitativa en centímetros, pero esto parece diferir entre razonamientos, ya que el razonamiento en base a sus experiencias determina si algo es alto o bajo y pondera eso en centímetros para dar una calificación final. Este razonamiento se lleva a cabo en la lógica difusa mediante los elementos ya mencionados, que para el caso de este ejemplo las etiquetas lingüísticas son: Alto, mediano y bajo y la función de pertenencia va de cero a uno, midiendo que tanto pertenecen al grupo de la etiqueta lingüística asociada y finalmente el rango es la medida en centímetros que nosotros asociamos a la etiqueta lingüística.

Para el caso de este modelo se usan datos de las variables antes mencionadas en el desarrollo, obteniendo de ellas una tasa de variación que en economía se considera una tasa de crecimiento.

Estas variables serán las entradas del modelo, por lo que antes se hace un análisis estadístico de las tasas de crecimiento con el fin de crear las etiquetas lingüísticas a emplear en este modelo. Se obtienen los siguientes parámetros de los conjuntos difusos que se muestran con mayor detalle en la tabla 1.

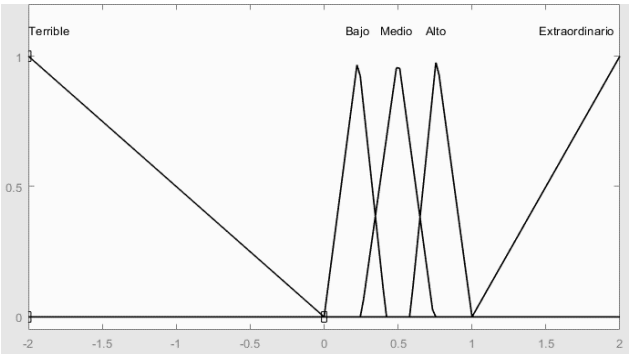


Fig. 5. Función del Índice de Desarrollo Humano.

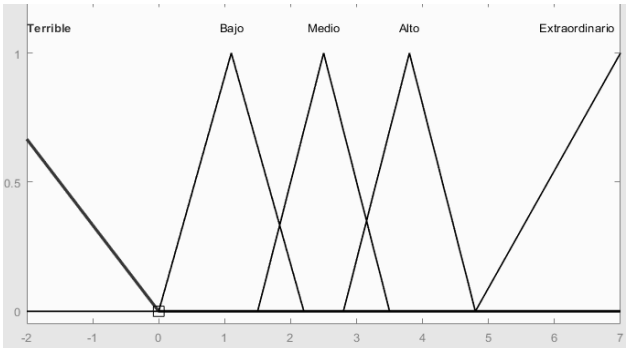


Fig. 6. Función de la percepción sobre la seguridad.

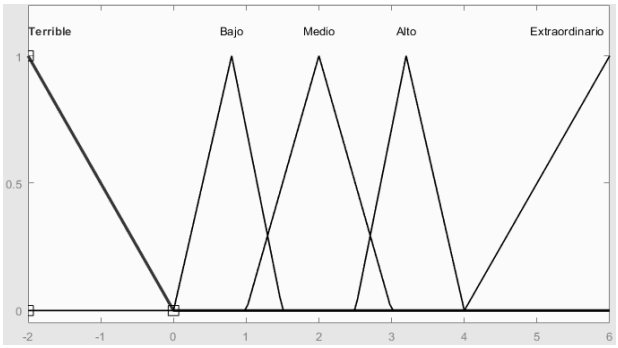


Fig. 7. Función del consumo.

Para las variables se formulan las siguientes funciones de pertenencia, basándonos en los parámetros anteriores. Son funciones triangulares ya que lo que buscamos es simplicidad en el modelo y definir bien el rango de la etiqueta lingüística.

El razonamiento económico es ingresado al modelo de manera abstracta mediante condicionales en la lógica difusa. La finalidad es ponderar las variables para que cuando

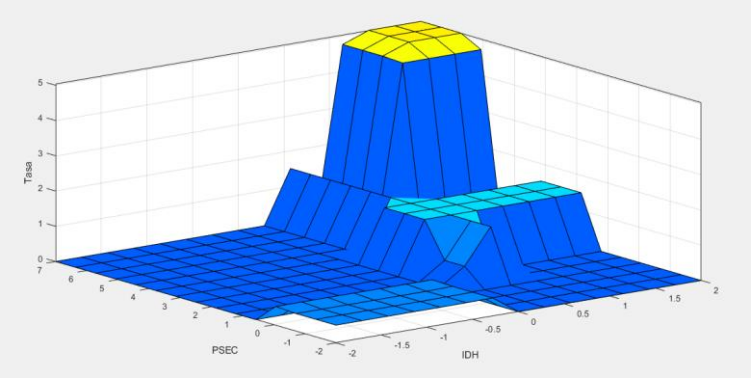


Fig. 8. Mapeo de las condicionales en el modelo: Seguridad vs IDH.

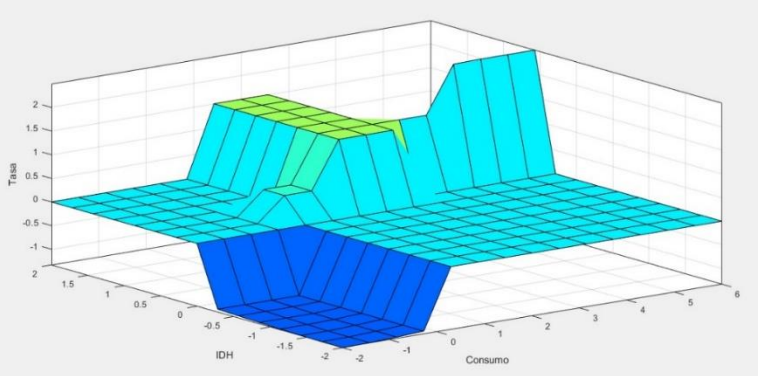


Fig. 9. Mapeo de las condicionales en el modelo: IDH vs Consumo.

exista un crecimiento en conjunto de las variables obtengamos una tasa que lo refleje tomando en cuenta un 60% será reflejo de la variable del consumo y un 40% de las variables que se consideran subjetivas en el IDH y la percepción de seguridad, con el fin de que a pesar de que exista un crecimiento económico sin crecimiento en las variables subjetivas la salida sea una tasa menor que la del consumo como tal en el mismo periodo, esto también en la situación inversa de que exista un crecimiento de las variables sin crecimiento en la tasa del consumo, el resultado sea menor que las tasas de las variables subjetivas para el mismo periodo. Con las condicionales se obtiene un mapeo en la figura 4 y 5:

$$Tasa < Cuantitativo_{OR}\ si\ Cualitativo_{OR} \geq Etiqueta\ "medio",$$

$$Tasa < Cualitativo_{OR}\ si\ Cuantitativo_{OR} \geq Etiqueta\ "medio".$$

Como lo que se busca es una tasa que muestre un crecimiento ponderado entre las variables cuantitativas y cualitativas, se escriben entre las condicionales algunas reglas que evitan que el resultado sea mayor a su observado real de la parte cuantitativa si la cualitativa no ha crecido lo suficiente para categorizarse como “medio” en las etiquetas lingüísticas antes mencionadas, esta condicional tiene una inversa que evita que la parte

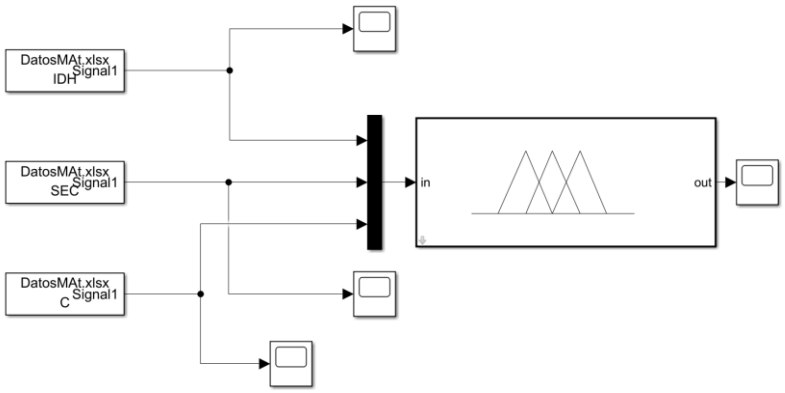


Fig. 10. Simulación del modelo.

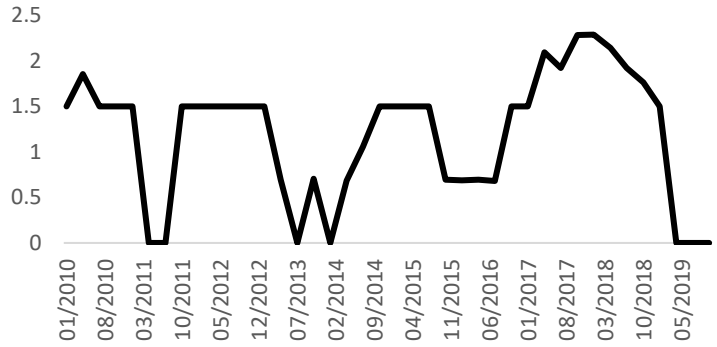


Fig. 11. Tasa obtenida a partir del modelo.

cualitativa sea menor a su observado real si la cuantitativa no se categoriza como "medio" en las etiquetas lingüísticas. Puede describirse de la manera presentada en Fig. 11, donde:

Tasa: Tasa subjetiva de desarrollo,

Cualitativo_{OR}: Dato real de la variable cualitativa observado para el periodo,

Cuantitativo_{OR}: Datos reales de las variables cuantitativas observadas para el periodo.

Finalmente, se plantea la simulación con la información de entrada para obtener una tasa de crecimiento ponderada mediante lógica difusa que refleje el crecimiento económico y subjetivo.

5. Resultados

Como resultado se obtiene una tasa de crecimiento que toma en cuenta parte de la subjetividad económica y aunado a una variable meramente económica como lo es el

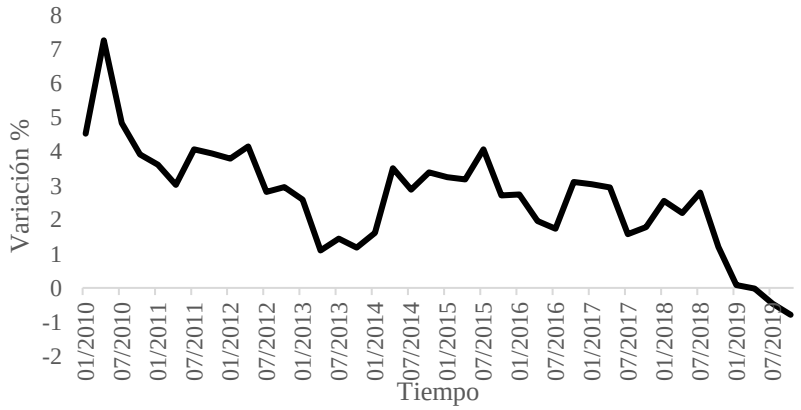


Fig. 12. Crecimiento medido con el PIB.

consumo. Con los datos del 2009 al 2019 de las variables ya antes mencionadas introducen al modelo con lo que obtenemos una tasa del 2010 al 2019, los datos son mostrados en la figura 11.

En la figura 11 se presentan periodos que marcan una variación del cero por ciento, eso significa que el modelo no observa crecimiento de las variables subjetivas o económicas razonables para poder obtener un verdadero crecimiento bajo la lógica antes mencionada, mientras que, en algunos puntos, como lo es el mes de octubre de 2017 se obtiene una tasa mayor, debido al comportamiento favorable de las variables cualitativas y cuantitativas. La tasa subjetiva de desarrollo presenta su mejor desempeño cuando el índice de inseguridad cae y el IDH aumenta, y a su vez, se presentan mayores niveles de consumo en la economía.

Esto se puede contrastar contra el indicador económico de crecimiento por excelencia que es la variación del PIB, variable que determina si hay crecimiento o no.

Como se observa en la figura 12, el crecimiento ha disminuido con el paso del tiempo.

El crecimiento medio de México en 2010 oscilaba en el 7% y posteriormente se presentan cambios en el crecimiento cada vez menores, llegando a estacarse entre 4% y 2% previo a 2018; en la etapa final de la gráfica, el crecimiento medio observado es negativo.

Cabe resaltar que esto no constituye una medida de desarrollo de la sociedad o de su calidad de vida, sin embargo, y como se ha señalado en la figura 11, un bajo crecimiento figura como un factor negativo para la tasa de desarrollo.

6. Conclusiones

Bajo la premisa de que un mayor nivel de crecimiento económico no necesariamente se ve reflejado en un mayor desarrollo, el objetivo de la presente investigación fue proponer un indicador de bienestar a través de una tasa de desarrollo económico que contempla variables subjetivas para el caso de México. La novedad de nuestro trabajo

con respecto a otras propuestas de medición de bienestar reside en incluir al IDH de México, pero, integrando la subjetividad de la percepción de inseguridad del país, así como el nivel de consumo implementando lógica difusa. Esto permite crear índices con las encuestas ya disponibles sin la necesidad de crear nuevas Este último punto es vital puesto que la mayor parte de las propuestas están orientadas a la generación e incorporación de nueva información para medir el bienestar cuando se pueden utilizar las percepciones y subjetividades de los agentes económicos como una métrica que permita distinguir la evolución del bienestar de dichos agentes.

La implementación de este tipo de metodologías no se limita al modelo expuesto en este estudio, las variables que se incluyan podrían cambiar de acuerdo con las condiciones y necesidades económicas de cada país, lo cual lo hace una herramienta flexible ante las coyunturas que afecten al país

Una variante al modelo que se presenta, podría ser la inclusión de variables de mayor frecuencia, esto permitiría dar un seguimiento más cercano a la evolución del desarrollo en las diferentes economías. En este mismo sentido, destacamos que el presente artículo es solo una propuesta de tasa de crecimiento del desarrollo sin demeritar la metodología propuesta por el PNUD.

Referencias

1. Abbassi, R., Abbassi, A., Heidari, A., Mirjalili, S.: Improving adaptive neuro-fuzzy inference system based on a modified salp swarm algorithm using genetic algorithm to forecast crude oil price. Energy Conversion and Management, 179, pp. 362–372 (2019)
2. Cabrera-Llanos, A.I., Ortiz-Arango, F., Cruz-Aranda, F.: Un modelo de minimización de costos de mantenimiento de equipo médico mediante lógica difusa. Revista Mexicana de Economía y Finanzas, pp. 379–396 (2019)
3. CONEVAL: Diez años de medición de pobreza multidimensional en México: avances y desafíos en política social (2019)
4. Deaton, A.: Price indexes, inequality, and the measurement of world poverty. American Economic Review, pp. 5–34 (2010)
5. Esquivel, G.: Los impactos económicosde la pandemia en México (2020)
6. Ferrer-Comalat, J.C., Corominas-Coll, D., Linares-Mustarós, S.: Fuzzy logic in economic models. Journal of Intelligent & Fuzzy Systems, pp. 5333–5342 (2020)
7. Jasso-López, C.: Percepción de inseguridad en México. Revista Mexicana de Opinión Pública, pp. 12–29 (2013)
8. Muratori, M., Zubieta, E.: La inseguridad subjetiva como mediadora del bienestar social y clima emocional. Psicodebate, pp. 95–120 (2016)
9. Quezada, P., Santillan, M., Hinojosa, R., Rada, J.: Percepción de inseguridad versus tasa delictiva: ¿Qué afecta más la economía mexicana?. Ensayos, pp. 205–226 (2019)
10. Ranis, G., Stewart, F.: Economic growth and human development. World Development, pp. 197–219 (2000)
11. Samuelson, P.: An exact consumption-loan model of interest with or without the social contrivance. Journal of Political Economy, pp. 467–482 (1958)
12. Schaeffler, K.: Economía de mercado con responsabilidad social. En ITESO (Ed.), Cátedra Konrad Adenauer (2004)
13. The Committee for the Prize in Economic Sciences: Angus Deaton: Consumption, poverty, and welfare. Sciences, T.C. (Ed.) (2015)
14. UNDP: Human development report 2019. United Nations Development Programme (2019)
15. Varian, H.R.: Microeconomic analysis. W.W. Norton & Company (1992)

16. Wesley, S.: Fear of crime and neighborhood change. Crime and Justice, pp. 203–229 (1986)
17. Witt, U.: The evolution of consumption and its welfare effects. Journal of Evolutionary Economics, pp. 274–293 (2016)

Diseño de biorreactor difuso

Diego Antonio Flores Solorzano[1], Gilberto Silos Chincoya[1], Gonzalo Guillermo Martínez Oliva[1], Francisco Javier García Camacho[1], Jesús Alberto Vázquez Santacruz[1], María Guadalupe Ramírez Sotelo[2], Agustín Ignacio Cabrera Llanos[1]

[1] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria de Biotecnología,
Departamento de Bioprocesos,
México

[2] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria de Biotecnología,
Departamento de Bioingeniería,
México

chinco.sg12@gmail.com

Resumen. Este trabajo presenta el diseño y desarrollo del control de un biorreactor de flujo ascendente mediante algoritmos de lógica difusa aplicados mediante la programación e implementación de código en Python. El trabajo se dividió en cuatro etapas: Diseño de los controles difusos para cada variable de control en un biorreactor e implementación del programa, desde la creación de funciones para facilitar el análisis de la lógica difusa y el uso de funciones predefinidas de librerías de Python. El procedimiento realizado se describe como: primeramente, se diseñaron los controles difusos para las variables temperatura del medio, velocidad de agitación y pH, el algoritmo difuso utilizado fue el método de Mamdani, proceso que se realiza en cuatro pasos: La fuzzificación de las variables de entrada, definiendo primeramente las funciones de membresía para después ser agregadas las variables lingüísticas; la evaluación de las reglas de inferencia mediante la creación de la función "cut" para evaluación; agregación de las salidas de las reglas mediante la creación de la función "unión", y la defuzzificación para poder obtener una respuesta de salida al sistema.

Palabras clave: Biorreactor, temperatura, ph, velocidad de agitación, lógica difusa, Python.

Fuzzy Bioreactor Design

Abstract. This work presents the design and development of the control of an upflow bioreactor through fuzzy logic algorithms applied through the programming and implementation of code in Python. The work was divided into four stages: Design of fuzzy controls for each control variable in a bioreactor and implementation of the program, from the creation of functions to facilitate the analysis of fuzzy logic and the use of predefined functions from Python libraries. The procedure performed is described as: first, the fuzzy controls were designed

for the variables temperature of the medium, stirring speed and pH, the fuzzy algorithm used was the Mamdani method, a process that is carried out in four steps: The fuzzification of the variables input, first defining the membership functions and then adding the linguistic variables; the evaluation of the rules of inference by creating the "cut" function for evaluation; aggregation of the outputs of the rules by creating the “union” function, and defuzzification in order to obtain an output response to the system.

Keywords: Bioreactor, temperature, ph, agitation speed, fuzzy logic, Python.

1. Introducción

Un biorreactor es un equipo capaz de simular medios de cultivo en estado sólido o líquido. En este dispositivo se debe garantizar la máxima conversión de materia prima en el producto de conversión, por lo que su participación es de vital importancia para este tipo de bioprocesos.

Debido a la incertidumbre que puede existir en un biorreactor por las variables que se deben de controlar como el pH, temperatura, saturación de oxígeno, velocidad de agitación, la naturaleza compleja del crecimiento de microorganismos y la formación de productos en cultivo. El uso de modelos matemáticos o sistemas de control por retroalimentación suelen ser imprecisos, siendo una opción viable la lógica difusa teniendo como principal ventaja la facilidad de implementación en un sistema, ya que no es necesario conocer el modelo matemático o emplear matemática avanza para describir el comportamiento que rige a dicho sistema para poder controlarlo, añadiendo el hecho de que en trabajos pasados donde se implementa este tipo de control se ha encontrado que han sido más eficientes en cuestión de la metodología para poder desarrollarlo de forma experimental; se puede dar la manipulación de varias variables del controlador como conjuntos difusos, factores de escalamiento, etc., junto con las simulaciones y el método de prueba y error.

La lógica difusa es utilizada en sistemas de control modernos siendo capaz de reaccionar a cambios continuos en el sistema y otorgar valores diferentes a la lógica booleana que es la lógica que utilizan los sistemas de control clásicos. El diseño de la lógica difusa permite interpretar y expresar el lenguaje ambiguo humano mediante funciones de pertenencia; en las cuales - como su nombre indica -, cada elemento de un universo dado pertenece a un conjunto específico.

Las funciones de pertenencia (o membresías), son asignadas con base en el rango en el cual se ejecutará dicho valor correspondiente a los valores de entrada de un control difuso, el cual puede ser de tipo SISO, MIMO (Single Input – Single Output, Multiple Input- Multiple Output), entre otros.

A partir de los valores de entrada, se obtienen diferentes valores de pertenencia para cada uno, a esto se le llama fuzzificación. De acuerdo con estos valores, se declaran las funciones de inferencia que, siguiendo condiciones si-entonces, dan como resultado otro valor difuso a la salida. Esta, por último, se defuzzifica mediante el método del centroide, el cual se basa en determinar el centro de gravedad del conjunto de salida.

Una función de membresía puede definirse como todos los conjuntos difusos que permiten medir el grado en que los objetos pertenecen a dichos conjuntos y que

satisfacen las propiedades definidas imprecisamente. Una función de membresía se define por tres propiedades:

- Núcleo: Es la región del universo caracterizado por un valor de membresía igual a 1 (membresía completa), en el conjunto.
- Soporte: Es la región del universo en el que la membresía tiene un valor distinto de 0 en el conjunto.
- Límites: Es la región del universo en donde una membresía contiene valores mayores que 0 pero menores de 1.

Python es el lenguaje de programación utilizado para programar la lógica difusa de los controles del biorreactor debido a varias razones. Una de ellas es porque es un lenguaje Open Source, lo cual significa que posee código abierto, librerías de funciones y brinda la facilidad para ejecutar los programas realizados en cualquier software reconozca este lenguaje, el cual es libre y gratuito para su descarga en múltiples plataformas tecnológicas; es útil para proyectos de programación orientada a objetos y de tarjetas de adquisición, lo cual, es bastante útil para implementar sistemas de embebido en objetos de interés como en el presente trabajo.

Python permite el desarrollo de páginas web, programación orientada a objetos y programación de tarjetas de adquisición. Es bastante útil para implementar sistemas de embebido en objetos de interés como un biorreactor.

2. Metodología

2.1. Conjuntos difusos

Los conjuntos difusos surgieron como una nueva forma de representar la imprecisión y la incertidumbre, permiten formalizar expresiones lingüísticas que típicamente contiene algún grado de ambigüedad. La teoría clásica de conjuntos se define un conjunto crisp A sobre X mediante la función característica de A como:

$$\mu_A(x) = \begin{cases} 1, & si\ x \in A, \\ 0, & si\ x \notin A. \end{cases} \quad (1)$$

Si la función de pertenencia para un valor dado de $\mu_A(x)$ toma el valor de 1, ese valor es un elemento conjunto A; por el contrario, si $\mu_A(x)$ toma el valor de cero, no pertenece al conjunto A (González-Morcillo, C,2011).

Para un conjunto difuso, sin embargo, la cuestión de pertenencia de un elemento al conjunto no es cuestión de todo o nada, existen diferentes grados de pertenencia, La función de pertenencia puede tomar cualquier valor dentro del intervalo de [0,1].

La función de pertenencia $\mu_A(x)$ de un conjunto difuso A es una función:

$$\mu_A: X \rightarrow [0,1]\,. \quad (2)$$

El conjunto difuso A se define como:

$$A = \{(x, \mu_A(x)) : x \in U\,, \mu_A(x) \in [0,1]\}\,. \quad (3)$$

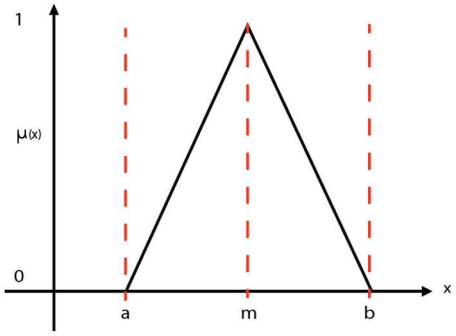


Fig. 1. Función Triangular.

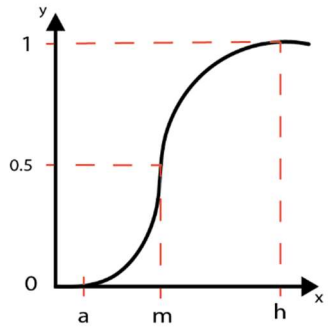


Fig. 2. Función Sigmoidal.

En cualquier elemento x en tiene un grado de pertenencia $\mu_A(x) \in [0,1]$.

2.2. Variables lingüísticas

Mientras que una variable algebraica toma números como valores, una variable lingüística toma palabras u oraciones como valores (Zadeh en Zimmermann 1993). El bloque fundamental de los sistemas basados en lógica difusa es la variable lingüística. La variable lingüística describe el razonamiento subjetivo que describen el contexto, es un medio de trasladar conceptos descripciones lingüísticas a descripciones numéricas. La forma de representar una variable lingüística esta descrita por cinco elementos:

$$(X, U, T(X), G, M),$$

donde X es el nombre de la variable, U el dominio subyacente, T(X) es el conjunto de términos o etiquetas que puede tomar X, G es una gramática para generar las etiquetas T(X): "Muy", "No muy", "Extremadamente", "Bajo", "Normal" y los conjuntos conectivos lógicos: operadores lógicos NOT, AND y OR y M es un regla semántica que asocia cada elemento de T(X) con un conjunto difuso en U de entre todos los posibles.

2.3. Funciones de membresía

Las funciones de pertenencia que se utilizaron en este trabajo son las siguientes:

A. Función triangular

La función triangular (4) se describe de la siguiente manera:

$$\mu_{\text{Triangular}}(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ \dfrac{c-x}{c-b}, & b \le x \le c. \\ 0, & x > c. \end{cases} \tag{4}$$

La representación gráfica se muestra en la Fig. 1.

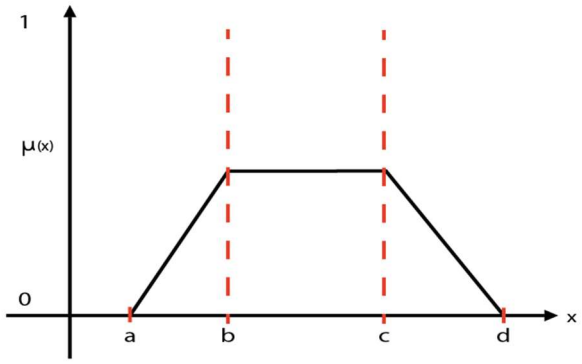


Fig. 3. Función Trapezoidal.

B. Función sigmoidal

La función sigmoidal (5) es descrita por:

$$\mu_{Sigmoidal}(x) = \begin{cases} 0, & x < a, \\ 2\left(\frac{x-a}{m-a}\right)^2, & a \le x \le m, \\ 1 - 2\left(\frac{x-a}{m-a}\right)^2, & m \le x \le h, \\ 1, & x > h. \end{cases} \tag{5}$$

La representación gráfica se muestra en la Fig. 2.

C. Función Zeta

La función zeta está definida (6) opuesta de la función sigmoidal:

$$\mu_{Zeta}(x) = 1 - \mu_{Sigmoidal}(x). \tag{6}$$

D. Función Trapezoidal

La función triangular (7) se describe de la siguiente manera:

$$\mu_{Trapezoidal}(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & b \le x \le c, \\ \frac{d-x}{d-c}, & c \le x \le d, \\ 0, & x < d. \end{cases} \tag{7}$$

La representación gráfica se muestra en la Fig. 3.

E. Función Singleton

La función singleton **¡Error! No se encuentra el origen de la referencia.** tiene un valor único cuando x=a, se describe de la siguiente manera:

$$\mu_{Singleton}(x) = \begin{cases} 0, & x = a, \\ 1, & x \neq a. \end{cases} \quad (8)$$

2.4. Diseño de control difuso

Utilizando los algoritmos de lógica difusa se puede diseñar un sistema de control inteligente que permita decidir entre las diferentes variables lingüísticas de la señal de entrada y proporcione una señal de salida que permita el control de esta variable. El desarrollo de la estructura de un sistema de control difuso se realiza mediante tres etapas.

2.5. Etapa de fuzzificación

Esta etapa calcula el grado de pertenencia de la entrada del sistema dentro de cada una de las funciones de membresía definidas para las variables de entrada.

Una función de membresía indica el grado de pertenencia de cada elemento dentro de un conjunto delimitado de valores llamado universo.

2.6. Etapa de reglas de inferencia

El control difuso usa reglas lingüísticas sobre los resultados que fueron generado en la etapa de fuzzificación, las reglas difusas son sentencia SI-ENTONCES (IF-THEN) que describen la acción a ser tomada en respuesta a varias entradas difusas. Las reglas de inferencia especifican conclusiones extraídas de afirmaciones conocidas o asumidas como verdaderas, para hacer las reglas de inferencia se parte de lo siguiente: A y B son conjuntos difusos definidos en el universo X y Y respectivamente. Esto es una implicación donde el antecedente es “x es A” y el consecuente es “y es B”., la premisa A’ es ligeramente diferente de A y por lo tanto la conclusión B’ es ligeramente diferente de B. Se representa de la siguiente forma:

$$\frac{A'}{\underline{A \rightarrow B}}.$$
$$B'$$

El modus ponens generalizado. Sean A y A’ conjuntos difusos definidos en X, y sean B sea un conjunto difuso definido en Y. Entonces el conjunto difuso B’, inducido por “x es A” de la regla difusa:

$$\frac{x\ es\ A'}{\underline{Si\ x\ es\ A\ entonces\ y\ es\ B}}. \quad (9)$$
$$y\ es\ B'$$

2.7. Etapa de defuzzificación

Esta etapa consiste en encontrar la salida final, las salidas que fueron encontradas en la etapa de reglas de evaluación modificaran a su respectiva función de pertenencia de salida. El método de defuzzificación utilizado es el Método de centro de Masa

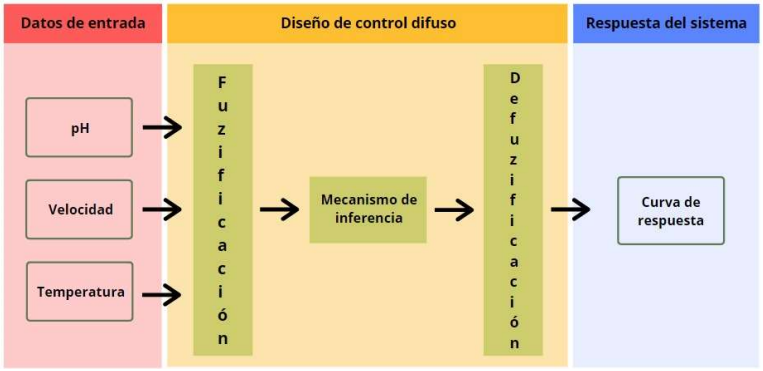


Fig. 4. Diagrama a bloques del algoritmo de programación.

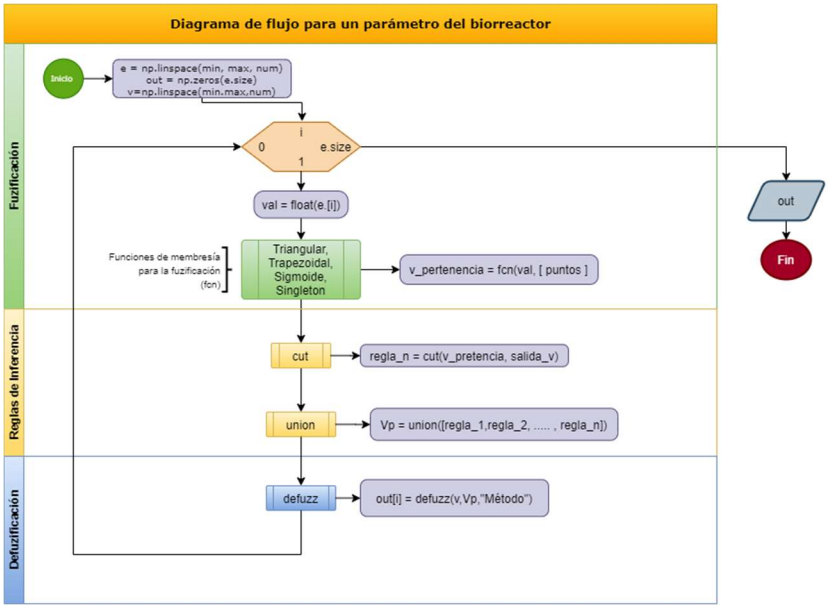


Fig. 5. Diagrama de flujo del control difuso.

(Centroide) (Silos et al., 2020), en el cual es determinado el centro de gravedad del conjunto de salida – se opta por este método debido a los resultados obtenidos en trabajos previos. Este método expresa el resultado final mediante un valor crítico o un promedio:

$$y_0 = \frac{\sum y\mu(y)}{\sum \mu(y)}. \tag{10}$$

Al valor obtenido con la expresión anterior permite el control de las variables temperatura, pH y velocidad de agitación, cada variable está determinado con su respectivo rango de valores.

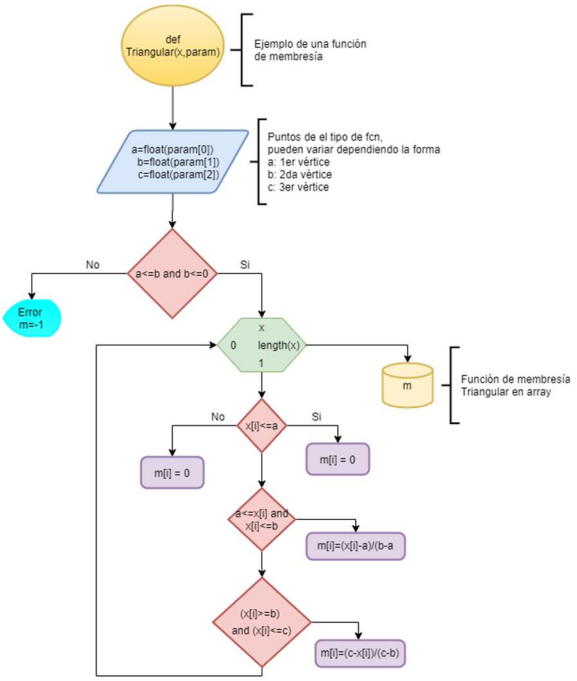


Fig. 6. Diagrama de flujo de la programación de una función de membresía.

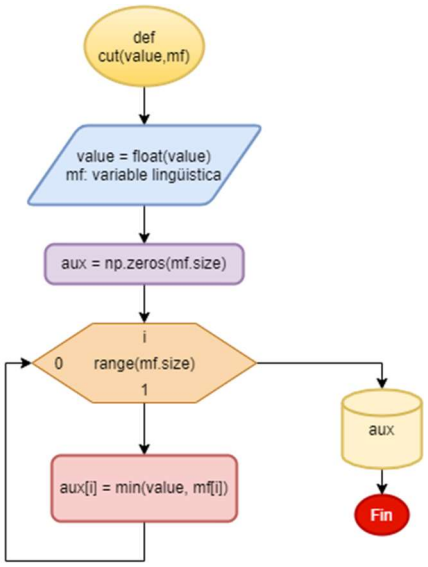


Fig. 7. Diagrama de flujo de la función "cut".

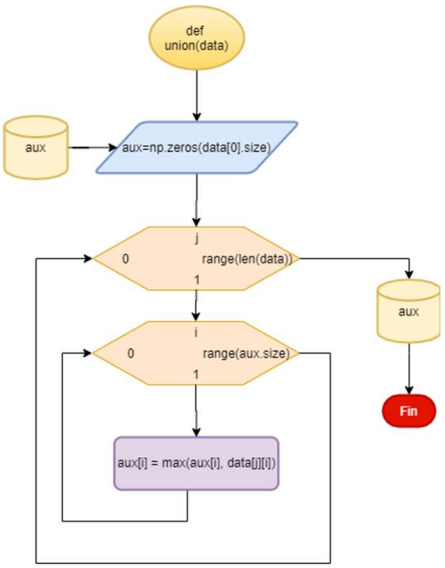


Fig. 8. Diagrama de flujo de la función "unión".

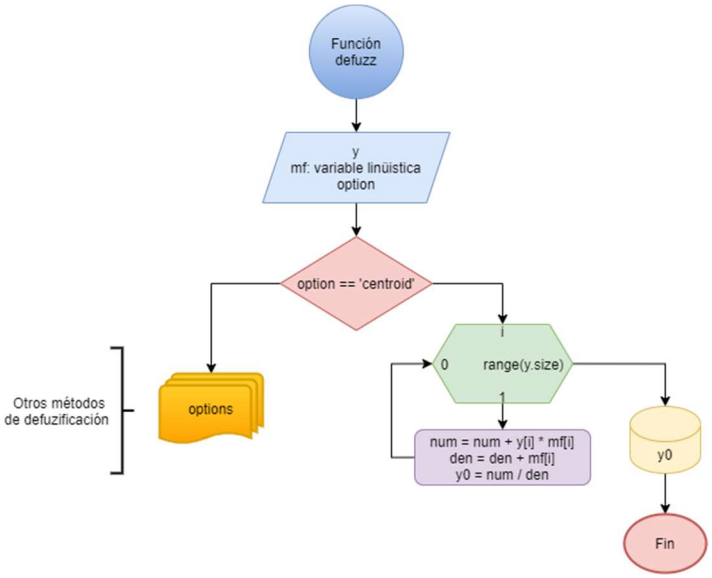


Fig. 9. Diagrama de flujo de la función “defuzz”.

2.8. Algoritmo de programación

El algoritmo que se presenta a continuación se desarrolló mediante el lenguaje de programación Python, se estructuró en cuatro etapas: Diseño de funciones de membresía, fuzzificación, reglas de inferencia y defuzzificación.

El código para el control difuso se representa en el diagrama de flujo de la Fig. 5. donde se crea un ciclo de lectura de las variables de entrada y se incia el proceso de control difuso mediante las funciones de fuzzificación con las funciones de membresía, las funciones cut y unión para las reglas de inferencia y finalmente la función defuzz que tiene la labor de defuzzificar los valores de pertenencia con base en las reglas de inferencia y mediante los métodos de centroide, bisector, MOM, SOMO o LOM. Finalmente, el programa entrega un vector con los valores de salida del control difuso.

A partir de las ecuaciones mencionadas anteriormente, se describen las funciones de membresía para el proceso de fuzzificación. Para la implementación de la función triangular se programó mediante el siguiente diagrama de flujo en la Fig. 6., así mismo, se programaron las funciones de membresía restantes vistas en el punto 2.3.

Las reglas de inferencia se definieron mediante las funciones “cut”, que relaciona los valores de pertencia con su valor fuzzificado, y “union” que se describe posteriormente. La función “cut” recibe los valores de las funciones de membresía de salida y un dato que recibe el nombre de valor de fuzzificación, este último delimita la función de membresía de salida, obteniendo como salida una función de membresía de salida cortada.

Se definió la función auxiliar “union”, los datos de entrada de la función fueron: las funciones de membresía de salida cortadas que se generaron en la etapa anterior.

A partir de la unión de todas las funciones de membresías de salidas cortadas, se obtiene un vector con los datos fuzzificados que son contrapuestos contra los valores del espacio de las funciones de membresía de salida. El método de defuzzificación es seleccionado a manera de string y la función fue programada con los procesos mencionados al principio de esta sección.

La salida de esta función son los valores para la curva de respuesta que representa el conjunto difuso de salida. Por último, para obtener un único dato de salida, se utiliza las formula del centroide y debido a que es un control de un biorreactor difuso y se han obtenido resultados óptimos bajo este método. (Silos et al., 2020).

A continuación, se muestra el programa realizado en Python para el diseño del control de velocidad de agitación, siguiendo este procedimiento manera se realiza el diseño de las variables temperatura y pH.

El pseudocódigo del programa de funciones de membresía velocidad de agitación:

Velocidad de agitación

```
        #ENTRADA
x_rpm = np.linspace(-100,100,1001)
rpm_cero = Sigmoide(x_rpm,[-0.4,-50])
rpm_alto_n = Triangular(x_rpm,[-60,-40,-10])
rpm_medio_n = Triangular(x_rpm,[-40,-20,0])
rpm_bajo_n = Triangular(x_rpm,[-10,0,10])
rpm_bajo_p = Triangular(x_rpm,[0,20,40])
rpm_medio_p = Triangular(x_rpm,[10,40,60])
rpm_alto_p = Sigmoide(x_rpm,[0.4,50])
        #SALIDA
x_corr = np.linspace(-0.05,0.05,1001)
corr_sin = Sigmoide(x_corr,[-600,-0.035])
corr_alta_n = Triangular(x_corr,[-0.04,-0.03,-0.01])
corr_media_n = Triangular(x_corr,[-0.03,-0.015,-0.005])
corr_baja_n = Triangular(x_corr,[-0.005,0,0.005])
corr_baja_p = Triangular(x_corr,[0.005,0.015,0.03])
corr_media_p = Triangular(x_corr,[0.01,0.03,0.04])
corr_alta_p = Sigmoide(x_corr,[800,0.035])
```

El código del Programa de todas las etapas:

```
e = np.linspace(-100, 100, 1001)
out = np.zeros(e.size)
v = np.linspace(-0.05,0.05,1001)
for i in range(e.size):
    val = float(e[i])
    # Fuzzyficando
    rpm_1 = Sigmoide(val, [-0.4, -50])
    rpm_2 = Triangular(val, [-60, -40, -10])
    rpm_3 = Triangular(val, [-40, -20, 0])
    rpm_4 = Triangular(val, [-10, 0, 10])
    rpm_5 = Triangular(val, [0, 20, 40])
    rpm_6 = Triangular(val, [10, 40, 60])
    rpm_7 = Sigmoide(val, [0.4, 50])
    ## Inferencia
    c_1 = cut(rpm_1, corr_alta_p)
```

```
c_2 = cut(rpm_2, corr_media_p)
c_3 = cut(rpm_3, corr_baja_p)
c_4 = cut(rpm_4, corr_sin)
c_5 = cut(rpm_5, corr_baja_n)
c_6 = cut(rpm_6, corr_media_n)
c_7 = cut(rpm_7, corr_alta_n)
# Se unen
Vp = union([c_1,c_2,c_3,c_4,c_5,c_6,c_7])
# Defuzzyficar
out[i] = defuzz(v, Vp, "centroid")
```

Tabla 1. Clasificaciones difusas para la señal de error de temperatura.

Etiqueta	Descripción	Parámetros
Negativo	Error muy negativo <0	Trapezoidal (-10,-10,-1,1.75)
Mínimo	Error mínimo cercano con un promedio 0	Triangular (0,1.75,3)
Intermedio Positivo	Error positivo con un promedio de 5	Triangular (0.5,5,7)
Máximo Positivo	Error muy positivo >7.5	Sigmoidal (2,6.5)

Tabla 2. Clasificaciones difusas para la señal de tiempo de encendido.

Etiqueta	Descripción	Parámetros
Apagado	No hay tiempo de encendido	Singleton (0)
Alto	tiempo de encendido con un promedio de 1 seg	Triangular (0,2,4)
Alto Medio	tiempo de encendido con un promedio de 2.5 seg	Triangular (1,3,5)
Medio	tiempo de encendido con un promedio de 5 seg	Triangular (3,5,7)
Bajo Medio	tiempo de encendido con un promedio de 7 seg	Triangular (5,7,9)
Bajo	tiempo de encendido con un promedio >8 seg	Sigmoidal (3.4,8.1)

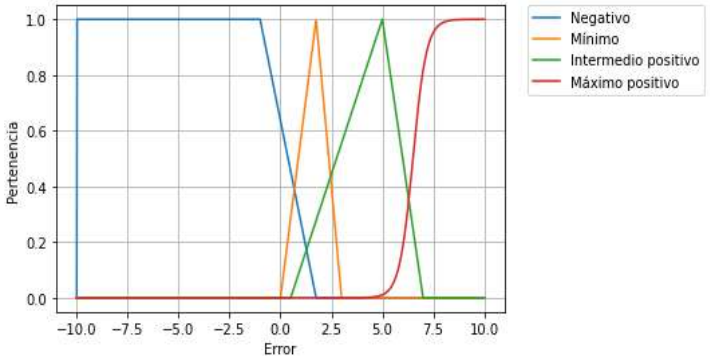


Fig. 10. Funciones de membresía de la variable de entrada “Error” del sistema difuso para el control de temperatura.

Tabla 4. Clasificaciones difusas para la señal de Voltaje PWM.

Etiqueta	Descripción	Parámetros
Vol Bajo	PWM con un promedio <80	Zeta (-0.6,85)
Vol Medio Bajo	PWM con un promedio de 100	Triangular (80,100,120)
Vol Regular	PWM con un promedio de 130	Triangular (110,130,160)
Vol Medio	PWM con un promedio de 160	Triangular (145,160,190)
Vol Medio Alto	PWM con un promedio de 210	Triangular (180,210,240)
Vol Muy Alto	PWM con un promedio 230	Sigmoidal (0.5,235)

Tabla 5. Clasificaciones difusas para la señal de error de la velocidad de agitación.

Etiqueta	Descripción	Parámetros
Alto Negativo	error con un promedio < -80	Zeta (-0.4,-50)
Medio Negativo	error con un promedio de -35	Triangular (-60,-40,-10)
Bajo Negativo	error con un promedio -20	Triangular (-40,-20,0)
Cero	error con un promedio 0	Triangular (-10,0,10)
Bajo Positivo	error con un promedio 25	Triangular (0,20,40)
Medio Positivo	error con un promedio 40	Triangular (10,40,60)
Alto Positivo	error con un promedio > 60	Sigmoidal (0.4,50)

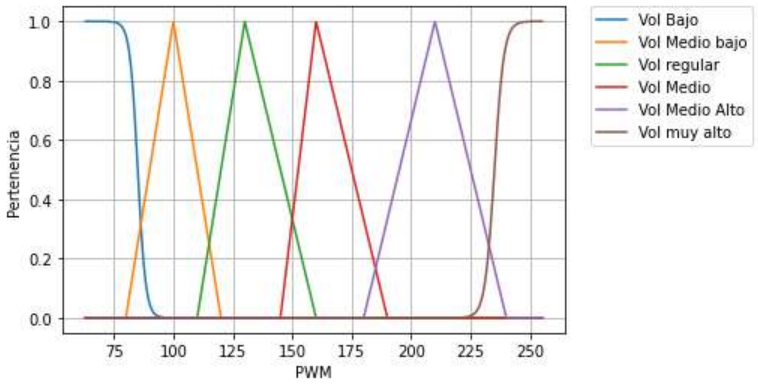


Fig. 13. Funciones de membresía de la variable de salida "Voltaje PWM" del sistema difuso para el control de pH.

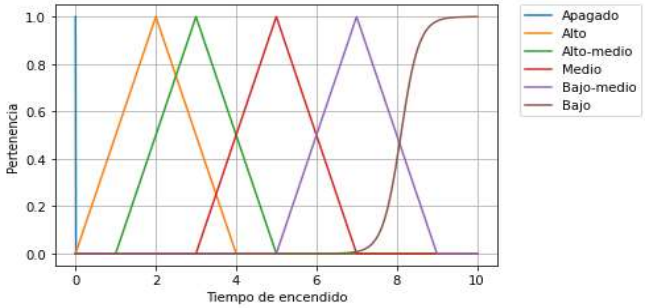


Fig. 11. Funciones de membresía de la variable de salida "Tiempo de encendido" del sistema difuso para el control de temperatura.

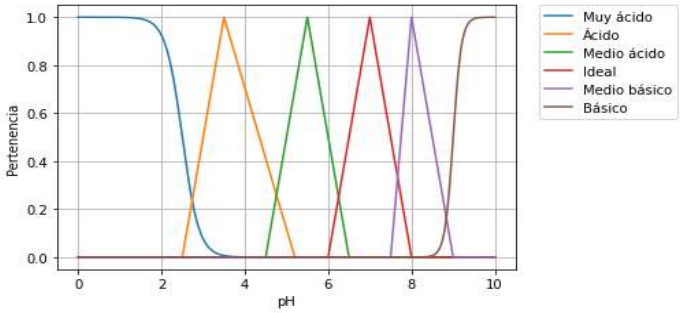


Fig. 12. Funciones de membresía de la variable de entrada "pH" del sistema difuso para el control de pH.

Tabla 3. Clasificaciones difusas para la señal de pH en el medio.

Etiqueta	Descripción	Parámetros
Muy Ácido	pH con un promedio < 2	Zeta (-5,2.5)
Ácido	pH con un promedio de 3	Triangular (2.5,3.5,5.2)
Medio Ácido	pH con un promedio de 5	Triangular (4.5,5.5,6.5)
Ideal	pH con un promedio de 7	Triangular (6,7,8)
Medio Básico	pH con un promedio de 8	Triangular (7.5,8,9)
Básico	pH con un promedio > 9	Sigmoidal (10,9)

2.9. Diseño de control difuso de temperatura

Se estableció un punto de referencia indicando la temperatura deseada ingresada por el usuario. La señal de entrada al sistema es el error calculado mediante la diferencia entre la señal deseada y la señal medida, usando este parámetro se diseñaron las

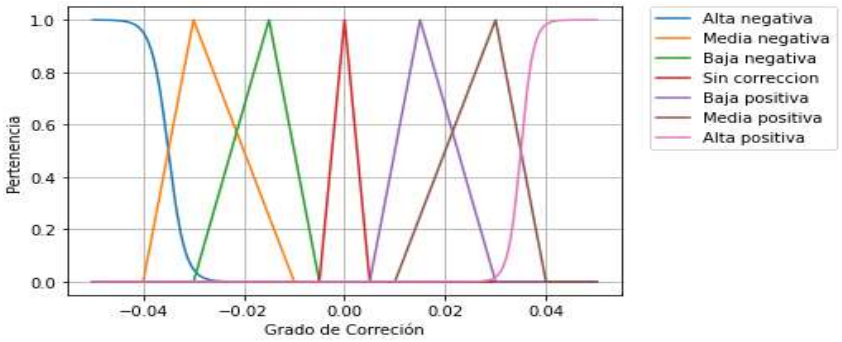


Fig. 15. Funciones de membresía de la variable de salida "Grado de corrección" del sistema difuso para el control de la velocidad de agitación.

Tabla 7. Reglas de inferencia de la variable temperatura.

Reglas
Reglas 1: IF "Error" IS "Negativo" THEN "Tiempo de Encendido" IS "Apagado"
Reglas 2: IF "Error" IS "Mínimo" THEN "Tiempo de Encendido" IS "Alto"

Tabla 8. Reglas de inferencia de la variable pH.

Reglas
Reglas 1: IF "pH" IS "Ideal" THEN "Voltaje PWM" IS "Voltaje-Medio"
Reglas 2: IF "pH" IS "Básico" THEN "Voltaje PWM" IS "Voltaje-Muy-Alto"

Tabla 9. Reglas de inferencia de la variable velocidad de agitación.

Reglas
Reglas 1: IF "Error" IS "Cero" THEN "Tiempo de Encendido" IS "Apagado"
Reglas 2: IF "Error" IS "Bajo-Negativo" THEN "Grado de corrección" IS "Baja-Positiva"

funciones de membresía de entrada para el algoritmo de control difuso cuyo rango de valores se encuentra en intervalo de -10 a 10, los parámetros de diseño se muestran en las Tabla 1.

El diseño de la señal de salida del sistema se contempla el tiempo de encendido de la resistencia para un intervalo de diez segundos, con este parámetro se diseñaron las funciones de membresía de salida para el control difuso, el rango de valores de salida se encuentra en el intervalo de 0 a 10 segundos. los parámetros de diseño se muestran en las Tabla 2.

2.10. Diseño de control difuso para pH

Las funciones de membresía de entrada se diseñaron a partir de la adquisición de la señal de pH del medio que se encuentra en el biorreactor usando este parámetro se

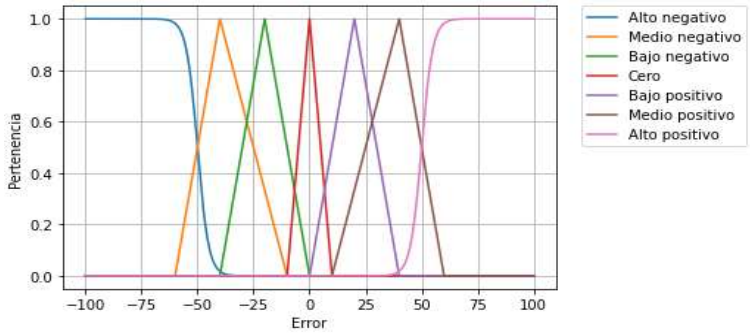


Fig. 14. Funciones de membresía de la variable de entrada "Error" del sistema difuso para el control de la velocidad de agitación.

Tabla 6. Clasificaciones difusas para la señal de grado de corrección.

Etiqueta	Descripción	Parámetros
Alto Negativo	error con un promedio < -80	Zeta (-0.4,-50)
Medio Negativo	error con un promedio de -35	Triangular (-60,-40,-10)
Bajo Negativo	error con un promedio -20	Triangular (-40,-20,0)
Cero	error con un promedio 0	Triangular (-10,0,10)
Bajo Negativo	error con un promedio 25	Triangular (0,20,40)
Medio Positivo	error con un promedio 40	Triangular (10,40,60)
Alto Positivo	error con un promedio > 60	Sigmoidal (0.4,50)

diseñaron las funciones clasificando la señal de pH cuyo rango de entrada se encuentra en el intervalo de 0 a 10 pH en el medio. los parámetros de diseño se muestran en las Tabla 3.

El control para la señal de salida es determinado por una bomba peristáltica regulando la Modulación de Ancho de Pulso (PWM), con este parámetro se diseña las funciones de membresía de salida del sistema difuso, clasificando la señal de modulación de ancho de pulso en seis variables lingüísticas, el rango de valores de salida se encuentra en el intervalo de 60 a 255 voltaje de modulación de ancho de pulso. Los parámetros de diseño se muestran en las Tabla 4.

2.11. Diseño de control difuso para velocidad de agitación

Se indica el punto de referencia indicando las revoluciones por minuto deseadas, se calcula el error mediante la diferencia entre la señal deseada y la señal adquirida. Se utilizó este parámetro para el diseño de las funciones de membresía de entrada clasificándolo en siete etiquetas, cuyo rango de entrada se encuentra en el intervalo de -100 a 100. Los parámetros de diseño se muestran en las Tabla 5.

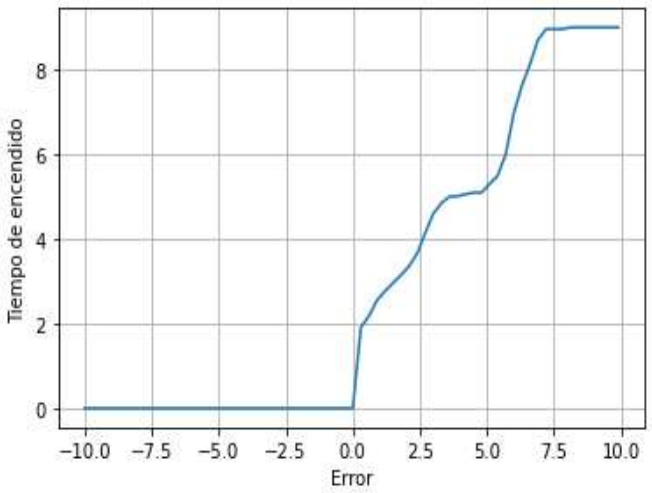


Fig. 16. Curva de respuesta del control difuso para la variable temperatura.

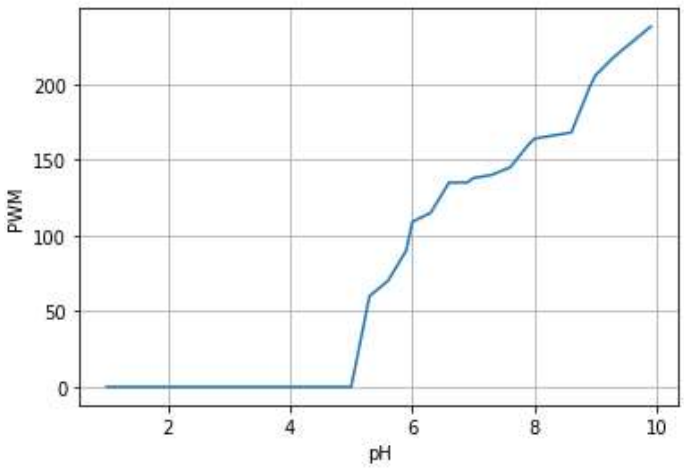


Fig. 17. Curva de respuesta del control difuso para el pH.

El diseño de las funciones de membresía para la señal de salida se estableció un grado de corrección de la Modulación del Ancho de Pulso (PWM), a partir de este parámetro de clasifica la señal de salida en siete variables lingüísticas, el rango de valores de salida se encuentra en el intervalo de -0.05 a 0.05 de grado de corrección. Los parámetros de diseño se muestran en las Tabla 6.

2.12. Diseño de reglas de inferencia

A continuación, se presenta las tablas de algunas reglas de inferencia desarrolladas.

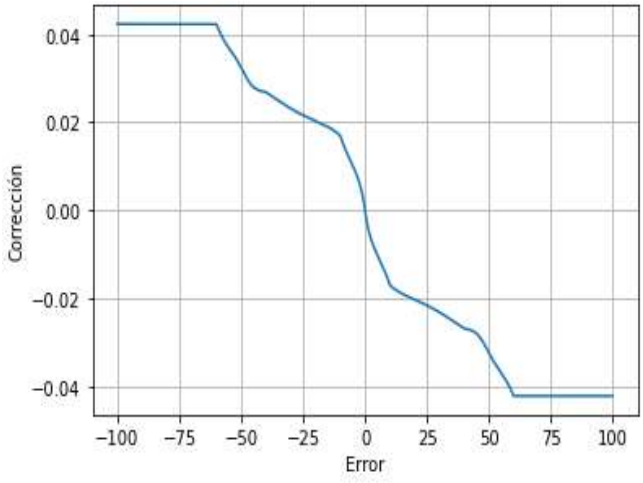


Fig. 18. Curva de respuesta del control difuso para velocidad de agitación.

3. Resultados

La respuesta obtenida para cada uno de los sistemas difusos se obtuvo mediante la relación establecida por las reglas de inferencia, se observa la curva de repuesta para la variable Temperatura en la Fig. 16.

En la Fig. 16. se muestra el funcionamiento del sistema difuso para el control de temperatura. Si el error de temperatura es negativo, implica que la medición de la temperatura sobrepasa el valor de referencia, por lo cual, la resistencia que permite el cambio de temperatura en el medio se mantiene apagada.

En el momento en el que el error se vuelve positivo, el sistema determina un tiempo de encendido de la resistencia que aumenta conforme aumenta el error para compensar el valor de la temperatura.

La respuesta del control difuso para el pH se observa en la curva de respuesta en la Fig.17.

En la Fig. 17. se muestra el comportamiento del sistema difuso para el control del pH en el medio. Si la señal de entrada permanece con un valor menor a 5, el ancho de modulación de pulso no genera una respuesta en el mecanismo de control de pH. En cambio, cuando se sobrepasa este umbral, aumenta en ciclo de trabajo conforme aumenta la entrada (ver Tabla 8).

En la Fig. 18. se muestra la respuesta del sistema difuso para el control de la velocidad de agitación. Si el error de velocidad es negativo, implica que el punto de referencia sobrepasa al valor medido, por lo cual, el grado de corrección del mecanismo es mayor. En caso contrario, si el error es positivo, el grado de corrección es bajo (ver Tabla 9).

4. Conclusiones

Se aplicó la teoría de lógica difusa a través del lenguaje de programación Python para diseñar un algoritmo que contenga las funciones necesarias para realizar los procesos de fuzzificación, inferencia y desfuzzificación para el diseño de tres sistemas difusos con la finalidad de controlar y monitorizar las variables de temperatura, pH y velocidad de agitación de un biorreactor. Se muestran las curvas de respuesta de los sistemas para mostrar las respuestas a los parámetros analizados en un biorreactor y demostrando que por este medio se pueden obtener las respuestas deseadas para el funcionamiento de un sistema control para un biorreactor. En este trabajo, se muestra la versatilidad del lenguaje Python para modelar sistemas de control difusos y proporciona las bases para desarrollar el modelado de manera física a través de tarjetas de adquisición, microprocesadores, microcontroladores que sean compatibles con el lenguaje Python.

Referencias

1. DisGonzález-Morcillo, C.: Lógica Difusa, una introducción práctica. Técnicas de Softcomputing (2011)
2. Fullér, R., Zimmermann, H.J.: On Zadeh's compositional rule of inference. In: Fuzzy Logic, pp. 193–200 (1993)
3. Kallestinova, E.: How to write your first research paper. Yale Journal of Biology and Medicine, 84(3), pp. 181–190 (2011)
4. Lee, K.D.: Introduction. Python programming fundamentals. Topics in Computer Science. Springer, London (2014)
5. Ruíz-Leza, H.A., Rodríguez-Jasso, R.M., Rodíguez-Herrera, R., Contreras-Esquivel, J.C., Aguilar, C.N.: Diseño de biorreactores para fermentación en medio sólido. Revista Mexicana de Ingenieria Química, pp. 33–40 (2007)
6. Jantzen, J.: Foundations of fuzzy control. John Wiley & Sons (2007)
7. Raich, V., Hooda, D.: Fuzzy logic models and fuzzy control: An introduction. Alpha Science International Ltd., pp. 2.1–2.3 (2017)
8. Hunt, J.: A beginners guide to python 3 programming. Springer Publishing (2019).
9. Silos-Chincoya, G., García-Estrada, H., Ramírez-Sotelo, M.G., Cabrera-Llanos, A.I.: Comparación de la variable de temperatura en un biorreactor: PID y lógica difusa. XVII Encuentro, Participación de La Mujer En La Ciencia (2020)

Grado de priorización para mantenimiento de equipo médico en un hospital por medio de lógica difusa y disposición aleatoria por el método de Montecarlo

Gonzalo Guillermo Martínez Oliva[1], Gilberto Silos Chincoya[1], Diego Antonio Flores Solorzano[1], Francisco Javier García Camacho[1], Jesús Alberto Vázquez Santacruz[1], Agustín Ignacio Cabrera Llanos[1], María Guadalupe Ramírez Sotelo[2]

[1] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria de Biotecnología,
Departamento de Bioprocesos,
México

[2] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria de Biotecnología,
Departamento de Bioingeniería,
México

gmtzoliva29@gmail.com

Resumen. Los dispositivos médicos tienen un papel importante en cada una de las etapas de atención dentro de las unidades de salud. Por esta razón, es importante mantener una buena gestión del equipo médico con la finalidad de asegurar la calidad del servicio que estas unidades otorgan. En este trabajo se presenta un sistema difuso capaz de priorizar la necesidad de mantenimiento de equipo médico considerando cinco factores: la función del equipo, el área en donde se encuentra, la carga del hospital, el tiempo desde la última solicitud de mantenimiento y la distancia relativa hacia otra unidad médica. Estas variables se utilizan como entrada para el sistema difuso, diseñado por medio de las herramientas proporcionadas por LabVIEW. La salida del sistema es el grado de priorización de equipo médico, clasificado como: muy baja, baja, media y alta. Para estimar el comportamiento del sistema se realizó una simulación por medio de valores aleatorios de entrada para calcular la salida. Los datos obtenidos se utilizaron para un análisis estadístico de Montecarlo. Los resultados mostraron que el sistema tiende hacia valores medios de priorización. Este estudio demuestra la utilidad de los sistemas difusos para resolver problemas de gestión administrativa en los hospitales.

Palabras clave: Hospital, lógica difusa, LabVIEW, gestión de equipo médico, administración hospitalaria, Monte Carlo.

Degree of Prioritization for Maintenance of Medical Equipment in a Hospital by means of Fuzzy Logic and Random Arrangement by the Monte Carlo Method

Abstract. Medical devices play an important role in each of the stages of care within health units. For this reason, it is important to maintain good management of the medical team in order to ensure the quality of service that these units provide. This paper presents a fuzzy system capable of prioritizing the need for maintenance of medical equipment considering five factors: the function of the equipment, the area where it is located, the load of the hospital, the time since the last maintenance request and the distance relative to another medical unit. These variables are used as input to the fuzzy system, designed using the tools provided by LabVIEW. The output of the system is the degree of prioritization of medical equipment, classified as: very low, low, medium and high. To estimate the behavior of the system, a simulation was carried out using random input values to calculate the output. The data obtained were used for a Monte Carlo statistical analysis. The results showed that the system tends towards average prioritization values. This study demonstrates the usefulness of fuzzy systems to solve administrative management problems in hospitals.

Keywords: Hospital, fuzzy logic, LabVIEW, medical team management, hospital administration, Monte Carlo.

1. Introducción

Dentro de un hospital, los equipos médicos tienen un papel de vital importancia para el diagnóstico, el tratamiento y la atención de los pacientes. Cada vez se diseñan dispositivos más completos y especializados, por la cual se deben de contar con un mantenimiento de calidad y así asegurar un excelente desempeño durante su interacción con el paciente.

Para lograr este objetivo se debe implementar una mejor planificación del programa de mantenimiento dentro de los hospitales para brindar una atención de calidad. La implementación de este sistema está a cargo del departamento de ingeniería de cada unidad de salud, de donde surge la necesidad de sistemas que faciliten esta tarea.

De acuerdo con la Organización Mundial de la Salud [5], existen dos maneras principales de clasificar al mantenimiento de equipo médico: inspección y mantenimiento preventivo (IMP) y mantenimiento correctivo (MC). Dentro del mantenimiento preventivo se consideran todas aquellas actividades que tienen como objetivo asegurar la funcionalidad de los equipos, así como prevenir posibles fallas y de esta manera prolongar la vida útil del dispositivo. La OMS también recomienda que cada una de las instituciones de salud cuente con el diseño de un programa de mantenimiento de equipo médicos de acuerdo con sus necesidades.

Uno de los métodos más utilizados para el diseño de sistemas de gestión de equipo médico es el de Fennigkoh-Smith, que permite relacionar tres factores principales: la función del equipo, riesgo físico para el paciente y los requisitos de mantenimiento [4]. Este sistema propone un valor de prioridad llamada Gestión de Equipo (GE) que se calcula como la suma de la calificación asignada al dispositivo médico:

Tabla 1. Equipment Function. Estos valores están divididos en la importancia que tiene el equipo médico de acuerdo con la función que se desempeña en el hospital [2].

EF	Numeric Value
Therapeutic – Life support	10
Therapeutic – Surgical or Intensive Care	9
Therapeutic – Physical Therapy or Treatment	8
Diagnostic – Surgical or Intensive Care Monitoring	7
Diagnostic – Other physiological monitoring	6
Analytical – Laboratory	5
Analytical – Computer and related	3
Miscellaneous – Patient related	2
Miscellaneous – Non patient related	1

Tabla 2. Location of equipment. Valores asignados al lugar de uso dentro del hospital [2].

L	Numeric Value
Anesthetizing	5
Critical Care Areas, Operational Rooms	4
Wet Locations/Labs/Exam Areas	3
General Patient Care Areas	2
Non-Patient Care Areas	1

Tabla 3. Hospital Load. Valores asignados basado en la carga del hospital (número de camas) [2]

HL [beds]	Numeric Value	HL [beds]	Numeric Value
>550	12	251-300	6
501-550	11	201-250	5
451-500	10	151-200	4
401-450	9	101-150	3
351-400	8	51-100	2
301-350	7	0-50	1

GE= Función del equipo + Riesgo + Requisitos de Mantenimiento.

Otro modelo utilizado para la priorización del mantenimiento es el de Wang-Levenson quienes reinterpretaron el criterio de la función del equipo, implementado anteriormente por Fennigkoh-Smith, como la misión crítica [6]. De igual manera,

Tabla 4. Time. Valores asignados desde la última solicitud de mantenimiento. El tiempo en el que la solicitud de mantenimiento fue emitida también considera el lapso en que el equipo está fuera de servicio [2].

T [days]	Numeric Value	T [days]	Numeric Value
>10	22	5	10
10	20	4	8
9	18	3	6
8	16	2	4
7	14	1	2
6	12		

Tabla 5. Distance to nearest alternative. Valores asignados a la distancia más cercana a un hospital que cuenta con el mismo equipo médico [2].

D [km]	Numeric Value	D [km]	Numeric Value
>90	26	20.1-30	12
80.1-90	24	10.1-20	10
70.1-80	22	5.1-10	8
60.1-70	20	2.1-5	6
50.1-60	18	1.1-2	4
40.1-50	16	0-1	2
30.1-40	14		

renombraron el valor de prioridad como el índice de valor de equipo proponiendo la siguiente ecuación:

#GE = Misión crítica + 2*Riesgo + 2*Requisitos de Mantenimiento.

Wang y Levenson también introdujeron el valor de tasa de uso para representar la urgencia de reparación del equipo, por lo cual se ajusta la ecuación de la siguiente manera:

#GE = (Misión crítica + 2*Requisitos de Mantenimiento) *Tasa de uso + 2*Riesgo.

Actualmente, existen muchas variantes del algoritmo de Wang-Levenson lo que permite seguir desarrollando sistemas que mejoren el desempeño de la administración de equipo médico dentro de las unidades de salud.

2. Metodología

La metodología desarrollada en este trabajo permite obtener un grado de priorización para cualquier solicitud de mantenimiento de equipo médico e indica la importancia relativa de esta solicitud, el sistema permite determinar el grado de priorización más

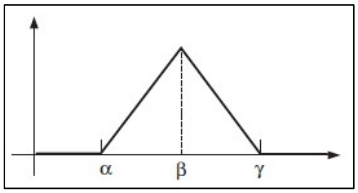


Fig. 1. Función Triangular.

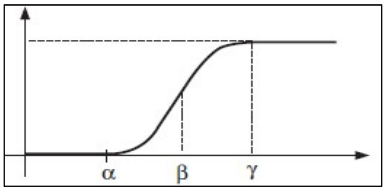


Fig. 2. Función Sigmoidal.

importante basado en las necesidades médicas del hospital y la seguridad del paciente. Los parámetros de diseño que permiten la importancia de la solicitud son basados en los siguientes criterios:

El número de prioridad de cada solicitud se calcula mediante el algoritmo de control difuso cuyo rango de valores se encuentra dentro del intervalo de 1 a 10 que se desarrolla en este trabajo, al número de prioridad más alto se le da servicio primero.

2.1. Etapas del control difuso

2.1.1. Fusificación

Esta etapa pondera a cada una de las entradas del sistema, con sus funciones de membresía definidas, en un rango de 0 a 1. Donde 0 representa la mínima pertenencia y 1 la máxima. Los conjuntos clásicos se representan como funciones de membresía o funciones de pertenencia esta descrita como $\mu_A(x)$: $U \rightarrow \{0,1\}$ es definido como 0 si x no se encuentra en A y si x toma el valor de 1 se encuentra en el elemento A. En los conjuntos difusos a x se le asigna un grado de pertenencia dentro del intervalo de [0,1].

Las funciones de membresía utilizadas en este trabajo son las siguientes:

Función Triangular

La representación de la función triangular se muestra a continuación.

$$\mu_T(x) = \begin{cases} 0 & x < \alpha, \\ \dfrac{x-\alpha}{\beta-\alpha} & \alpha \le x \le \beta, \\ \dfrac{x-\gamma}{\beta-\gamma} & \beta \le x \le \gamma, \\ 0 & x > \gamma. \end{cases}$$

Función Sigmoide

La representación de la función triangular esta descrita como:

$$\mu_S(x) = \begin{cases} 0, & x < \alpha, \\ 2\left(\frac{x-\alpha}{m-\alpha}\right)^2, & \alpha \le x \le \beta, \\ 1 - 2\left(\frac{x-\alpha}{m-\alpha}\right)^2, & \beta \le x \le \gamma, \\ 1, & x > \gamma. \end{cases}$$

2.1.2. Reglas de inferencia

Las reglas de inferencia toman como fundamente el concepto de variable lingüística para describir las entradas y salidas del sistema. Una variable convencional es numérica y precisa. No es capaz de soportar la vaguedad.

En los sistemas difusos, una variable lingüística está formada por palabras, frases o lenguaje artificial que son menos precisos que los números. Proporciona los medios de caracterización aproximada de fenómenos complejos o mal definidos, por mencionar algunas variables lingüísticas "Muy", "Mucho", "Poco", "Menos".

La base de reglas es esencialmente la estrategia de control del sistema. Generalmente se obtiene de conocimiento de reglas o heurística y expresado como un conjunto de reglas SI – ENTONCES.

Las reglas son basadas en el concepto de inferencia difusa y los antecedentes y consecuentes están asociados con variables lingüísticas. Se considera A y B como conjuntos difusos de entrada y salida, x es A haciendo referencia a los antecedentes, así como y es B como las consecuencias:

$$R: (x = A) \rightarrow (y = B)\ \ IF\ x\ is\ A\ THEN\ y\ e\ B.$$

2.1.3. Defusificación

Al diseñar todas las reglas difusas en la etapa anterior se obtiene una conclusión difusa, una variable lingüística cuyos valores han sido asignados por los diferentes grados de pertenencia. La conclusión es un conjunto difuso, sin embargo, se necesita un valor escalar que corresponda a estos grados de pertenencia.

El método de defusificación se realizó utilizado el método del centroide, conocido como Centro de Área, el cual calcula el centro de gravedad del polígono (curva de respuesta) que se generó mediante la etapa de inferencia, se obtiene mediante la siguiente expresión:

$$f(y) = \frac{\sum \mu(y) \cdot y}{\sum \mu(y)}.$$

2.2. Algoritmos de programación

El grado de priorización fue calculado con lógica difusa por medio del software LabVIEW con la herramienta Fuzzy System Designer de la sección Control and Simulation. Para realizar un análisis de priorización se tomaron en consideración las entradas que teóricamente dan funcionalidad óptima a un hospital: Funcionamiento del equipo, área de uso, la carga de trabajo en el hospital (número de camas), tiempo desde el último mantenimiento, la distancia al hospital más cercano.

Estas variables fueron usadas como valores de entrada difusas para otorgar una salida pertinente al estado en que se encuentre un hospital.

Tabla 6. Parámetros de diseño de funciones de membresía de entrada de función del equipo.

Función de Membresía	Forma	Puntos / Intervalo
Diverso	Sigmoide	[0, 0, 1, 3.5]
Analítico	Triangular	[2, 4, 6]
Diagnóstico	Triangular	[4.5, 6.5, 8]
Terapéutico	Sigmoide	[7, 9, 10, 10]

Tabla 7. Parámetros de diseño de funciones de membresía de lugar de uso.

Función de Membresía	Forma	Puntos / Intervalo
Ninguno	Sigmoide	[1, 1, 1.25, 2.25]
General	Triangular	[1.5, 2.5, 3.25]
Áreas de exámenes / Húmedas	Triangular	[2.75, 3.5, 4.5]
Critico	Sigmoide	[3.75, 4.75, 5, 5]

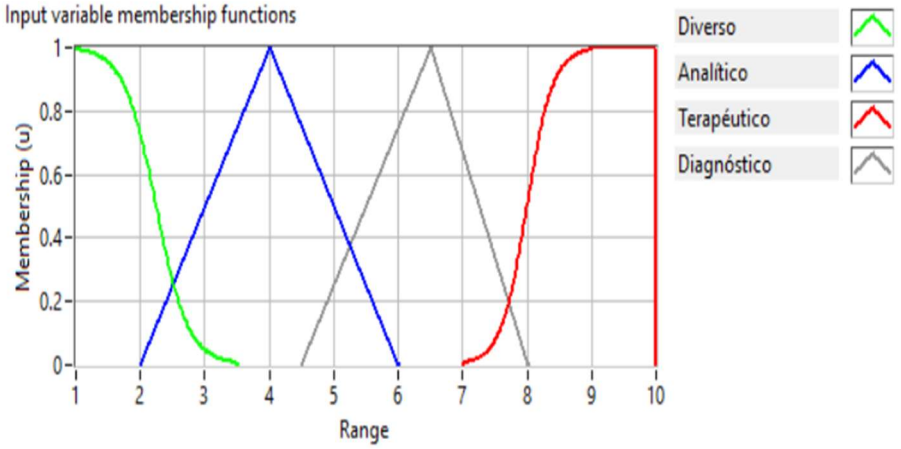


Fig. 3. Funciones de membresía de función de equipo.

2.3. Diseño de las funciones de membresía de entrada y salida

Función del equipo

Esta variable se refiere al papel que cumple el equipo. La función diversa se refiere a la relación con el paciente; analítica desempeñando su función en el laboratorio clínico o como equipo de cómputo; diagnóstico para supervisión fisiológica presente o no en cirugía y, finalmente de función terapéutica como un equipo en el que su funcionamiento dependa la vida del paciente en cirugía o sala de cuidados intensivos y terapia o un tratamiento. Esta variable tiene gran peso dentro del sistema difuso.

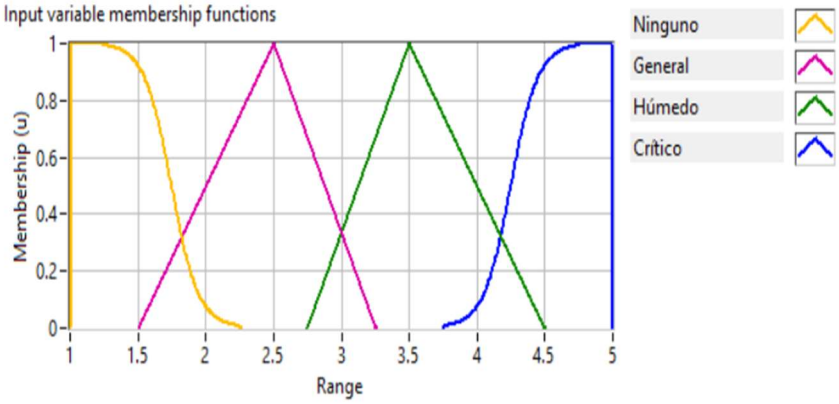


Fig. 4. Funciones de membresía de lugar de uso.

Tabla 8. Parámetros de diseño de funciones de membresía de carga del hospital.

Función de Membresía	Forma	Puntos / Intervalo
Muy poca	Sigmoide	[1, 1, 2, 4.5]
Poca	Triangular	[2.75, 5, 7.25]
Promedio	Triangular	[6.25, 8, 10.25]
Mucha	Sigmoide	[8.5, 10.5, 12, 12]

Área

Esta variable hace referencia al área en la que se necesita el equipo, empezando donde no hay necesidad del cuidado del paciente, áreas de cuidado general, laboratorios o áreas donde se realizan exámenes; áreas de cuidado crítico o salas de operación y, por último, las áreas críticas citando como las áreas de anestesia o en Qx. Esta variable tiene gran peso dentro del sistema difuso.

Carga de trabajo

Se refiere al número de camas designadas para dicho equipo donde muy pocas son de 0 a 200 camas, poca de 201 a 350, promedio de 351 a 500 y mucha de 501 a mayor de 550.

Tiempo desde el último mantenimiento

Es el tiempo desde el último mantenimiento dado al equipo médico, dando un intervalo de d a 7 días para reciente, de 5 a 12 días para hace poco, de 10 a 18 días para bastante y de 15 a 22 días para mucho.

Distancia

Se refiere a la distancia en kilómetros con respecto al hospital más cercano que cuente con el mismo servicio, siendo muy cerca de 2 a 8 km, cerca de 6 a 15 km, retirado de 12 a 22 km, lejos de 19 a mayor a 26 km.

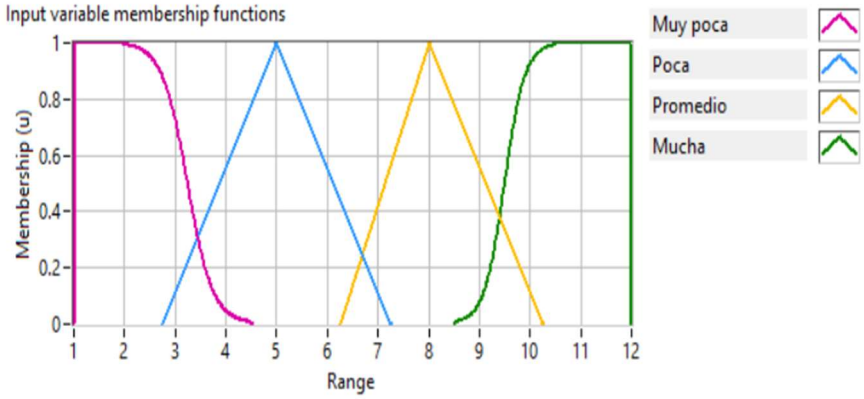


Fig. 5. Funciones de membresía de carga del hospital.

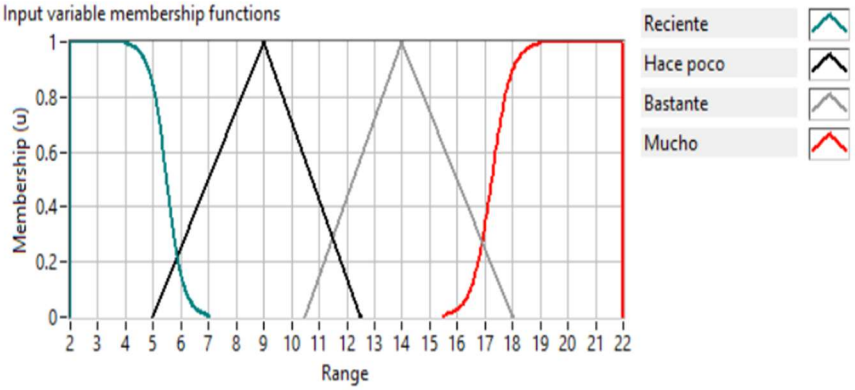


Fig. 6. Funciones de membresía del tiempo desde el último mantenimiento.

Tabla 9. Parámetros de diseño de funciones de membresía del tiempo desde el último mantenimiento.

Función de Membresía	Forma	Puntos / Intervalo
Reciente	Sigmoide	[2, 2, 4, 7]
Hace poco	Triangular	[5, 9, 12.5]
Bastante	Triangular	[10.5, 14, 18]
Mucho	Sigmoide	[15.5, 19, 22, 22]

Prioridad

Por último, tenemos la salida que es el grado de priorización de mantenimiento que recibirá el equipo médico teniendo los siguientes intervalos:

Tabla 10. Parámetros de diseño de funciones de membresía de la distancia entre hospitales cercanos.

Función de Membresía	Forma	Puntos / Intervalo
Muy cerca	Sigmoide	[2, 2, 4, 8]
Cerca	Triangular	[6, 11, 15]
Retirado	Triangular	[12, 17, 22]
Lejos	Sigmoide	[19, 24, 26, 26]

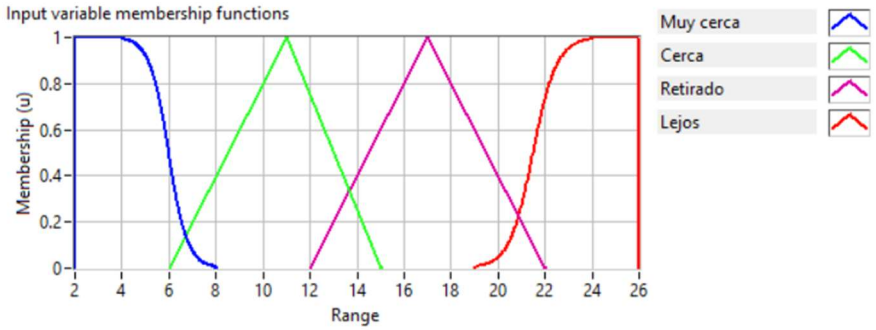


Fig. 7. Funciones de membresía de la distancia entre hospitales cercanos.

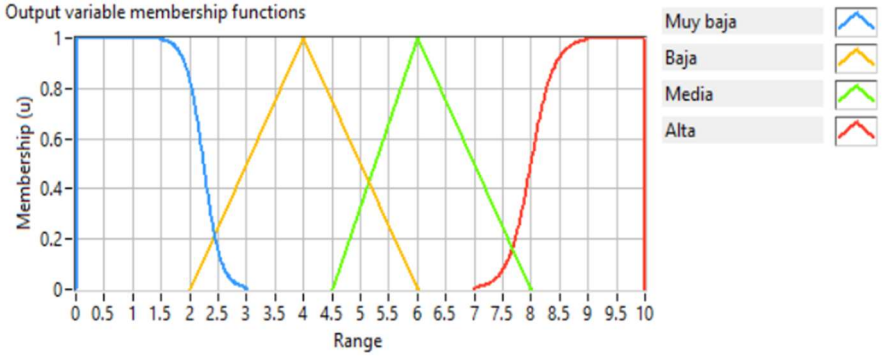


Fig. 8. Funciones de membresía del grado de priorización.

Tabla 11. Parámetros de diseño de funciones de membresía de grado de priorización de equipo médico.

Función de Membresía	Forma	Puntos / Intervalo
Muy baja	Sigmoide	[0, 0, 1.5, 3]
Baja	Triangular	[2, 4, 6]
Media	Triangular	[4.5, 6, 8]
Alta	Sigmoide	[7, 9, 10, 10]

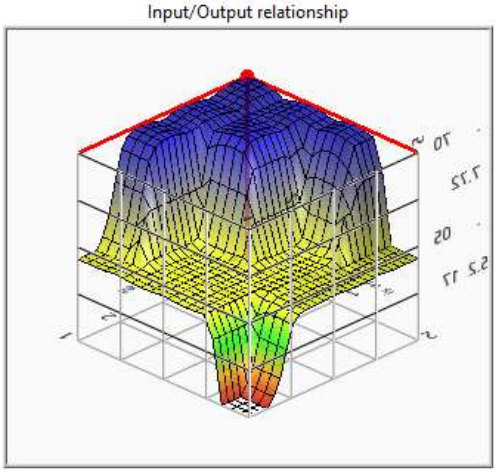


Fig. 9. Curva de respuesta del sistema.

Tabla 12. Diseño de reglas de inferencia.

1. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'reciente' AND 'Distancia' IS 'muy cerca' THEN 'Prioridad' IS 'muy baja'
2. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'reciente' AND 'Distancia' IS 'cerca' THEN 'Prioridad' IS 'muy baja'
3. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'reciente' AND 'Distancia' IS 'retirado' THEN 'Prioridad' IS 'baja'
4. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'reciente' AND 'Distancia' IS 'lejos' THEN 'Prioridad' IS 'baja'
5. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'hace poco' AND 'Distancia' IS 'muy cerca' THEN 'Prioridad' IS 'muy baja'
6. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'hace poco' AND 'Distancia' IS 'cerca' THEN 'Prioridad' IS 'muy baja'
7. IF 'Función del equipo' IS 'Diverso' AND 'Área ' IS 'Ninguno' AND 'Carga de trabajo' IS 'Muy poca' AND 'Tiempo desde el mantenimiento' IS 'hace poco' AND 'Distancia' IS 'retirado' THEN 'Prioridad' IS 'baja'

2.4. Reglas de inferencia

Para determinar el grado de priorización se utilizan reglas para comparar las variables lingüísticas de entrada fusificadas, y posteriormente, evaluar las

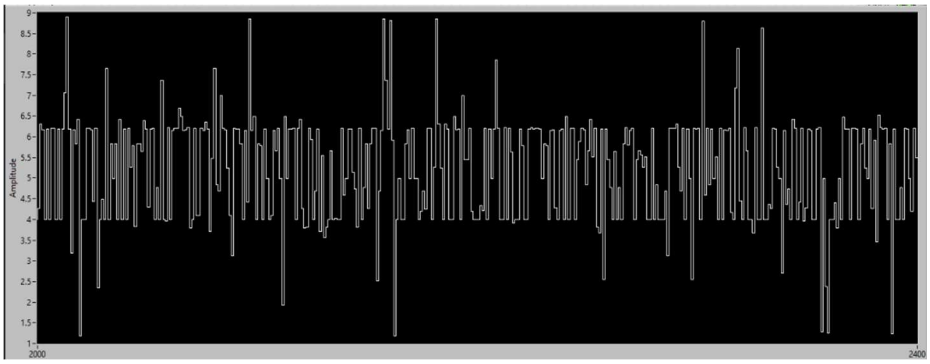


Fig. 10. Interfaz de usuario del programa en operación.

Tabla 13. Relación del rango con el número de intervalo.

Intervalo	Rango
1	[1.2, 1.9]
2	[1.91, 2.62]
3	[2.63, 3.33]
4	[3.34, 4.05]
5	[4.06, 4.76]
6	[4.77, 5.47]
7	[5.48, 6.19]
8	[6.2, 6.9]
9	[6.91, 7.62]
10	[7.63, 8.33]
11	[8.34, 9.05]

combinaciones por medio de sentencias SI-ENTONCES (IF-THEN), donde cada regla se va etiquetando en función a los valores de pertenencia otorgados.

2.5. Defusificación

Esta es la etapa final que consiste en obtener una salida ponderada en los rangos de priorización elegidos de 1 a 10, por medio de una superficie generada en comparación de dos entradas en los ejes *x* y *y* y la salida en el eje z. Un ejemplo de las superficies donde se observa el comportamiento se muestra en la Fig. 9.

2.6. Método Monte Carlo

La simulación del sistema que se presenta en el desarrollo de este trabajo se realizó mediante el método Monte Carlo. La simulación Monte Carlo permite la creación de

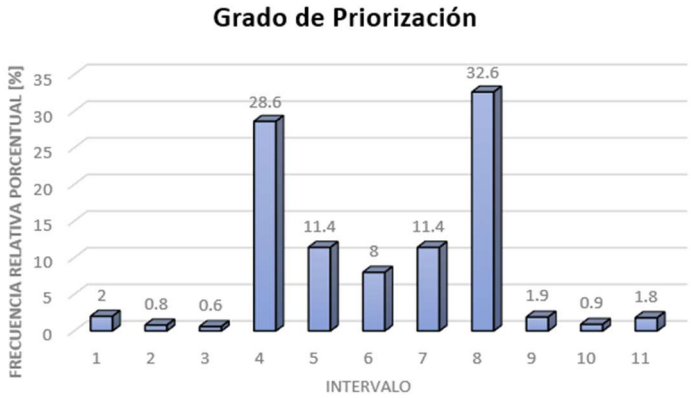


Fig. 11. Resultados de grado de priorización de mantenimiento de equipo médico, primera simulación.

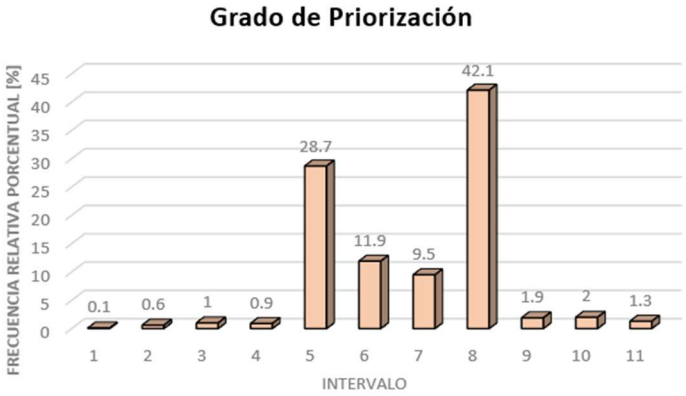


Fig. 12. Resultados de grado de priorización de mantenimiento de equipo médico, segunda simulación.

un modelo de estimación matemática mediante un análisis frecuencial utilizando una muestra de números aleatorios divididos por rangos. Se generaron 1000 diferentes números aleatorios para cada parámetro, los cuales se introducen al sistema difuso para generar una salida de la misma magnitud.

3. Resultados

3.1. Simulación en LabVIEW

Se realizaron tres simulaciones diferentes de 1000 iteraciones aleatorias para evaluar la eficiencia del sistema y poder analizar los resultados. Los números aleatorios fueron

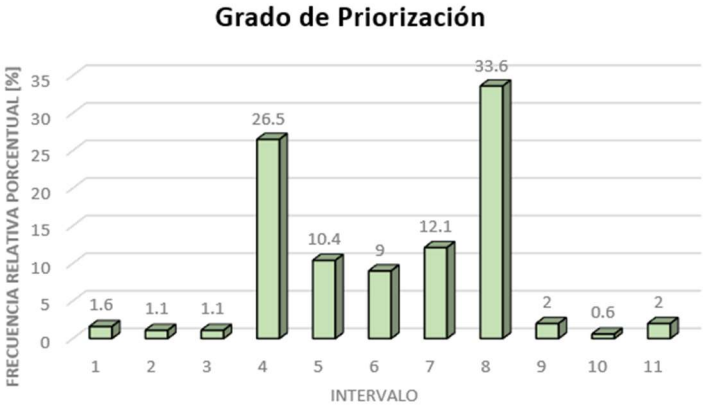


Fig. 13. Resultados de grado de priorización de mantenimiento de equipo médico, tercera simulación.

generados para los rangos de cada entrada. Dichas iteraciones fueron capturadas por medio del ploteo de LabVIEW.

3.2. Análisis estadístico

Se generó una hoja de datos en Excel a partir de la simulación en LabVIEW con los cuales se realizó el análisis de Monte Carlo. Se establecieron 11 intervalos definidos por el número de datos, los cuales se utilizaron para clasificar cada una de las salidas.

Se realizaron tres gráficas, una por cada simulación, de la frecuencia relativa porcentual ($\%fr$) para cada uno de los intervalos. Con este análisis se puede realizar una comparación para determinar la respuesta probable del sistema.

Las tres simulaciones concentraron el mayor porcentaje de valores en el intervalo 8, correspondiente a un rango de [6.2, 6.9]. También se obtuvo una ligera tendencia hacia el intervalo 4 y 5, dentro del rango [3.34, 4.76].

4. Conclusiones

El sistema difuso muestra un comportamiento concentrado a valores bajos y medios de priorización de mantenimiento de equipo médico, esto debido al peso dado a las variables de entrada: función de equipo y área.

En este estudio, se logró demostrar la eficiencia de la lógica difusa ante la lógica tradicional, plantando las bases para la implementación del método en futuras plataformas de lectura de priorización en hospitales de la zona metropolitana.

El uso de lógica difusa tiene aplicaciones eficientes en el campo de la administración para resolver problemas convencionales con principios de inteligencia artificial.

Referencias

1. Cabrera-Llanos, A.I., Ortiz-Arango, F., Cruz-Aranda, F.: Un modelo de minimización de costos de mantenimiento de equipo médico mediante lógica difusa. Revista Mexicana de Economía y Finanzas Nueva Época, 14(3), pp. 379–396 (2019)
2. Hamdi, N., Oweis, R., Zraiq, H., Sammour, D.: An intelligent healthcare management system: A new approach in workorder prioritization for medical equipment maintenance requests. Journal of Medical Systems, 36(2), pp. 557–567 (2010)
3. Jantzen, J.: Foundations of fuzzy control. John Wiley & Sons (2007)
4. Masmoudi, M., Houria, Z., Hanbali, A., Masmoudi, F.: Decision support procedure for medical equipment maintenance management. Journal of Clinical Engineering, 41(1), pp. 19–29 (2016)
5. Organización Mundial de la Salud: Introducción al programa de mantenimiento de equipos médicos. In: Serie de Documentos Técnicos de la OMS sobre Dispositivos Médicos (2012)
6. Tawfik, B., Ouda, B., El Samad, Y.: A fuzzy logic model for medical risk clasification. Journal of Clinical Engineering, 38(4), pp. 185–190 (2020)

Identificación biométrica vascular del dorso de la mano mediante imágenes infrarrojas

Marco A. Mayén García[1], Daniela Rodríguez García[1],
Benjamín Luna Benoso[1], Uriel Corona Bermúdez[2]

[1] Instituto Politécnico Nacional,
Escuela Superior de Cómputo,
México

[2] Instituto Politécnico Nacional,
Centro de Investigación en Computación
México

{antoniomayen1997, mobius_95}@hotmail.com,
{danierodga, urielcoro}@gmail.com

Resumen. Actualmente uno de los mayores problemas de seguridad en las organizaciones es la falta de sistemas computacionales que integren tecnologías que permitan la identificación confiable de miembros y visitantes. En este trabajo se propone una arquitectura que permite la identificación de un individuo por medio de la red de venas en el dorso de la mano haciendo uso de imágenes infrarrojas obtenidas en un ambiente controlado a través de una cámara digital, aplicando técnicas de análisis de imágenes y reconocimiento de patrones.

Palabras clave: Biometría, patrón vascular, procesamiento de imágenes, reconocimiento de patrones, identificación biométrica.

Vascular bBometric Identification of the Back of the Hand using Infrared Images

Abstract. Currently one of the biggest security problems in organizations is the lack of computer systems that integrate technologies that allow the reliable identification of members and visitors. In this work, an architecture is proposed that allows the identification of an individual through the network of veins on the back of the hand using infrared images obtained in a controlled environment through a digital camera, applying image analysis techniques and pattern recognition.

Keywords: Biometrics, vascular pattern, image processing, pattern recognition, biometric identification.

1. Introducción

En México, el delito de robo de identidad va en aumento día con día, según datos del Banco de México, el país ocupa el octavo lugar a nivel mundial en este delito. Confirmar la identidad de un individuo se ha convertido en la pieza clave para reducir el riesgo de fraudes por robo de identidad en las organizaciones alrededor del mundo [1, 2].

Los identificadores biométricos son mediciones del cuerpo humano vivo. El reconocimiento biométrico, o simplemente biometría, se refiere al uso de características o identificadores anatómicos y conductuales distintivos (por ejemplo, huellas dactilares, cara, iris, voz, geometría de la mano, etc.) para reconocer automáticamente a una persona [3].

Los identificadores biométricos no se pueden extraviar, falsificar o compartir fácilmente, se consideran más confiables para el reconocimiento de personas que los métodos tradicionales de tarjeta (tarjetas de identificación) o basados en conocimientos como contraseñas o número de identificación personal (PIN, por sus siglas en inglés).

Los objetivos del reconocimiento biométrico son los beneficios del usuario (retiro de dinero en un cajero automático sin tarjeta o PIN), mejor seguridad (solo la persona autorizada puede entrar en una instalación), una mayor responsabilidad (difícil de negar haber accedido a registros confidenciales), y una mayor eficiencia (menor sobrecarga que el mantenimiento de contraseñas en computadora) [3].

En este artículo se presenta una arquitectura que permite la identificación de una persona por medio del patrón biométrico vascular de las venas del dorso de la mano. La tecnología de patrón de venas utiliza la red vascular subcutánea en el dorso de la mano para verificar la identidad de los individuos en las aplicaciones biométricas. El principio de esta tecnología se basa en el hecho de que el patrón de los vasos sanguíneos es único para cada individuo, incluso entre gemelos idénticos [4].

2. Materiales y métodos

2.1. Ambiente controlado

Un ambiente controlado, es un ambiente en el que los factores de luz, temperatura, humedad relativa, etc., se establecen artificialmente [5]. En el presente trabajo, se ha implementado un ambiente que permite la entrada de luz natural o artificial únicamente por un cuadrado de lado de 11.5cm. En su interior contiene un aparato localizador de venas con un regulador manual de potencia que utiliza leds infrarrojos para transmitir luz infrarroja a través de la palma de la mano hasta el dorso, develando así la red vascular dorsal, como se puede visualizar en la fig. 1. También, contiene tres espejos en las caras inferior, posterior y una lateral para mayor reflexión de la luz infrarroja transmitida por el localizador de venas.

Fig. 1. Imagen ilustrativa de uno de los usos del localizador de venas [6].

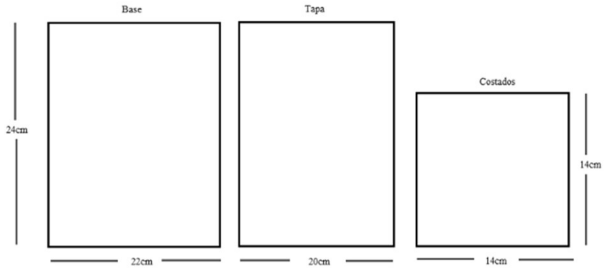


Fig. 2. Planos del prototipo de ambiente controlado.

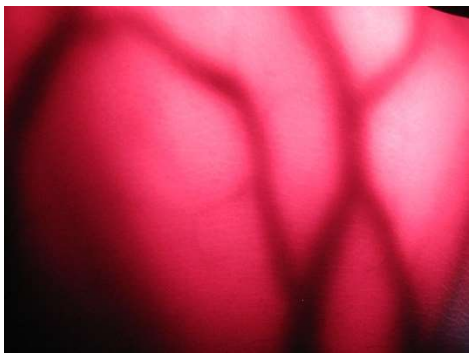


Fig. 3. Imagen capturada en el ambiente controlado.

El ambiente controlado provee un espacio lo suficientemente amplio al usuario para introducir la palma de una mano y cerrar el puño tomando el localizador. Además, permite la entrada del lente de una cámara digital, por la tapa, para tomar fotografías de las venas visualizadas. En la fig. 2 se pueden apreciar las dimensiones del ambiente.

Posteriormente, se capturaron un conjunto de 10 imágenes, que correspondían al dorso de una mano de un sujeto, de un total de 18 sujetos. Dichas imágenes fueron tomadas con una cámara digital Canon PowerShot A470 de 7.2 megapíxeles, utilizando una configuración de ISO 200, que permitía que el sensor fuese menos sensible a la luz,

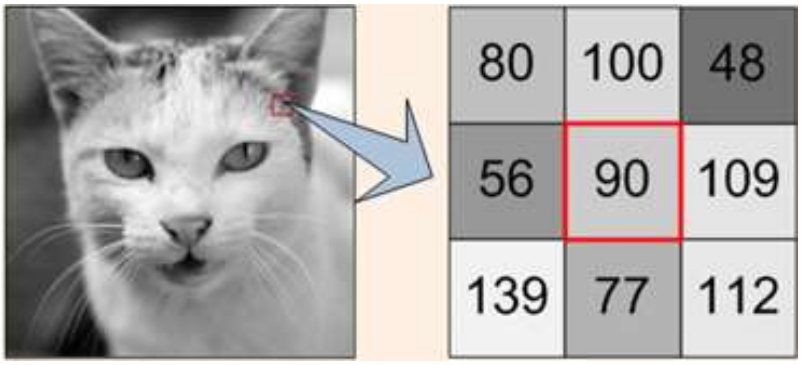


Fig. 4. Imagen en escala de grises [8].

esto con el fin de evitar obtener reflejos rojizos o blanquecinos en la imagen final, y el modo macro, que aseguraba el acercamiento a un objeto relativamente pequeño como lo era la superficie plana en el dorso de una mano. La intensidad del localizador de venas se colocó en la máxima potencia. La fig. 3 es una imagen tomada dentro del ambiente controlado.

Con el fin de incrementar el tamaño del banco de imágenes para el proceso de entrenamiento y prueba del clasificador, se realizaron 24 transformaciones geométricas que incluyen rotaciones en diferentes ángulos, traslaciones, sesgamientos, zoom y reflejos.

2.2. Mejoramiento de la imagen

El procesamiento digital de imágenes, incluye un conjunto de técnicas que operan sobre la representación digital de una imagen, a objeto de destacar algunos de los elementos que conforman la escena, de modo que se facilite su posterior análisis por parte de un sistema de visión artificial [7].

La técnica de suavizado, eliminación de ruido y la detección y realce de bordes se realizó haciendo uso de diferentes técnicas, descritas a continuación:

2.2.1. Modo de imagen escala de grises

En este modo de color la imagen está constituida por píxeles que pueden adoptar distintas tonalidades de un mismo color, por ejemplo, desde el blanco (0% de negro) hasta el negro (100% de negro). Este espectro de tono se gradúa, normalmente, en una escala que tiene 256 niveles [8]. La fig. 4 muestra un ejemplo de una imagen en escala de grises y su representación de niveles de intensidad luminosa.

2.2.2. Convolución

La convolución es un filtro de uso general que proporciona una forma de realizar un producto de dos matrices numéricas, normalmente de diferente tamaño, pero con el mismo número de dimensiones, para producir una tercera matriz numérica de la misma dimensión [9].

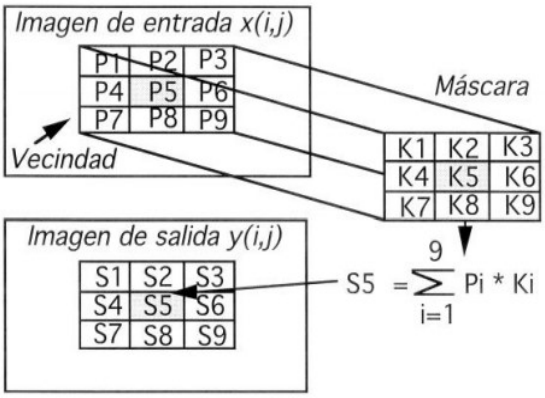


Fig. 5. Ilustración del proceso de convolución con una máscara [9].

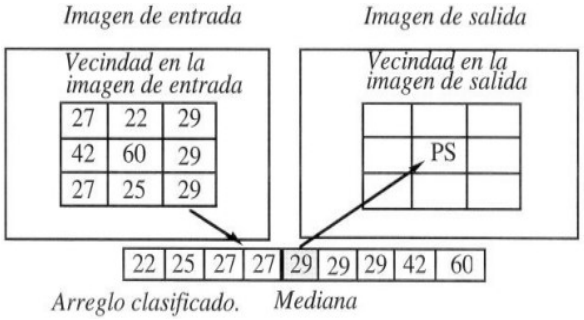


Fig. 6. Ilustración del procedimiento para implantar el filtro mediana [9].

Si consideramos una imagen como un arreglo bidimensional denotado por $x_{(i,j)}$ y el filtro (núcleo o máscara de convolución) con respuesta impulsiva $h_{(i,j)}$, su convolución produce una imagen de salida $y_{(i,j)}$, de acuerdo a la siguiente ecuación, en donde m y n definen la vecindad a considerar de acuerdo al tamaño del núcleo de convolución $h_{(i,j)}$ [9]:

$$y(i,j) = \sum_{m=-K1}^{K2} \sum_{n=-L1}^{L2} h(m,n)x(i-m, j-n)\,. \tag{1}$$

La implementación de esta ecuación de convolución se hace de manera directa cuando el tamaño del filtro o máscara de convolución es pequeño (usualmente menor que 9x9 píxeles). Para la implementación directa de la ecuación de convolución es una matriz de tamaño (N_1xN_2), usualmente de 3x3 píxeles, la cual adicionalmente define el tamaño de la vecindad dentro de la imagen, de manera que sea del mismo tamaño de la máscara [9].

Como se puede observar en la fig. 5, cada píxel en la imagen de salida es el resultado de la suma de los productos entre los píxeles de la máscara y los píxeles incluidos en la vecindad correspondiente en la imagen de entrada.

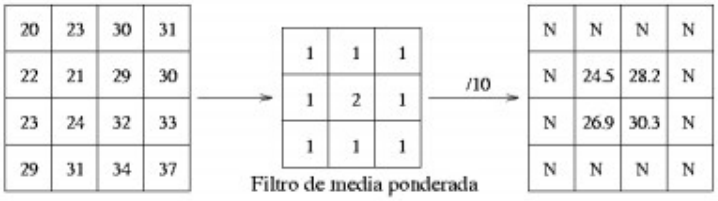


Fig. 7. Filtro media ponderada [10].

2.2.3. Filtro mediana

El filtraje mediana, es un procedimiento no lineal, útil para reducir el ruido impulsivo y del tipo "sal y pimienta", presente en las imágenes. El filtro mediana utiliza los valores de los píxeles contenidos en una vecindad de tamaño impar, para determinar el nuevo valor de píxel de interés. El procedimiento para ello, consiste en ordenar todos los píxeles incluidos en la ventana en orden creciente y sustituir el píxel ubicado en el centro de la vecindad por el píxel mediano luego de la clasificación, es decir, si tenemos una secuencia discreta de tamaño *N* impar, entonces la mediana de tal secuencia, es aquel miembro de la secuencia, para el cual, (*N-1*)/2 elementos son más pequeños o a lo sumo iguales y (*N-1*)/2 elementos son más grandes. En la fig. 6 se muestra un ejemplo de la implantación del filtro mediana [9].

2.2.4. Filtro media ponderada

El filtro media ponderada es similar al filtro media, donde se le asigna a un píxel central la media de todos los píxeles incluidos en la ventana. Sin embargo, la matriz de filtrado está compuesta no únicamente por unos, sino que se le otorga más peso a uno de ellos (el central, habitualmente), para obtener un resultado similar a la imagen original y evitar que parezca borrosa [10]. La fig. 7 muestra una máscara media ponderada.

2.2.5. Modificaciones al brillo (aclarado)

El brillo aumenta la luminosidad total de una imagen, es decir, modificarlo permite aclarar los píxeles oscuros de una imagen y blanquear totalmente los claros [11].

2.2.6. Operación morfológica de dilatación

Sea A y B conjuntos de Z^2, la dilatación de A y B, denotada como $A \oplus B$, está definida como:

$$A \oplus B = \left\{ z \middle| (\hat{B})_z \cap A = \emptyset \right\}. \quad (2)$$

Esta ecuación se basa en reflejar B sobre su origen, y moviendo esta reflexión de z. La dilatación de A y B es el conjunto de todos los desplazamientos, z, tal que B y A se superponen por al menos un elemento. Basado en esta interpretación, puede ser escrita una equivalencia como:

$$A \oplus B = \left\{ z \middle| \left[(\hat{B})_z \cap A \right] \subseteq A \right\}. \quad (3)$$

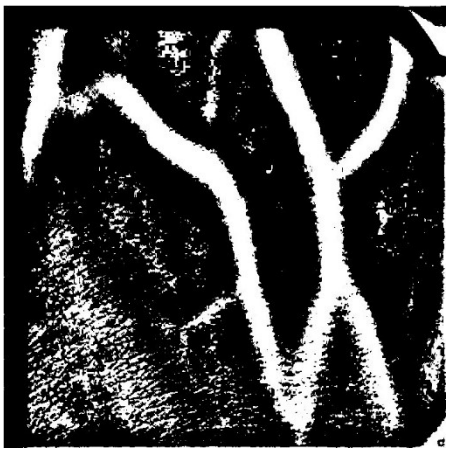


Fig. 8. Umbralado por método de Niblack.

Como antes, asumimos que B es un elemento estructural y A es el conjunto (objeto de la imagen) a ser dilatado [12].

2.3. Segmentación de la imagen

2.3.1. Segmentación por umbralado local

Este tipo de segmentación permite separar un objeto dentro de la imagen del fondo que lo circunda. La técnica se basa en comparar alguna propiedad de una imagen con un umbral fijo o variable, realizando tal comparación para cada uno de los píxeles que conforman la imagen, si el valor de la profundidad del píxel supera el valor del umbral, entonces el píxel pertenece al objeto, en caso contrario, el píxel pertenece al fondo [13].

El algoritmo de Niblack determina un valor umbral para cada píxel en sentido deslizando una ventana rectangular sobre la imagen en niveles de gris. El tamaño de la ventana rectangular puede diferir. El umbral se calcula en función de la media local *m* y la desviación estándar *S* de todos los píxeles en la ventana como se muestra en la ecuación 5:

$$T_{Niblack} = m + k * s\,, \tag{4}$$

$$T_{Niblack} = m + k\sqrt{\frac{1}{NP}\sum(p_i - m)^2} = m + k\sqrt{\frac{\sum {p_i}^2}{NP} - m^2} = m + k\sqrt{B}\,, \tag{5}$$

donde *NP* es el número total de píxeles presentes en la imagen en escala de grises, $T_{Niblack}$ representa el valor de umbral, *m* es el valor promedio de los píxeles p_i, y *k* es fijado dependiendo del ruido que aún existe en el fondo, pudiendo ser -0.1 o -0.2 [14]. En la fig. 8 puede verse el algoritmo aplicado a la imagen capturada en el ambiente controlado.

Fig. 9. Imagen aplicando Niblack, reversión y componentes conexos.

Fig. 10. Imagen original contrastada con su esqueleto.

2.3.2. Componentes conexos

Sea G=(*V, E*) un grafo no conexo. Se denomina componente conexo de G a un subgrafo conexo maximal de G [15].

Aplicar componentes conexos permite el etiquetado de diferentes conjuntos de información conectados, y su diferenciación por el área que abarcan. Se colocó una restricción para que cualquier componente conexo menor de una dimensión de 200x200, fuese eliminado de la imagen, para únicamente quedarse con la mayor cantidad de información de la red, como se aprecia en la fig. 9.

2.3.3. Segmentación basada en esqueletizado

El esqueletizado es un proceso para reducir las regiones de primer plano en una imagen binaria a un remanente esquelético que conserva en gran medida el alcance y la conectividad de la región original mientras se descarta la mayoría de los píxeles de primer plano originales [16].

El esqueleto se puede producir de dos formas principales. El primero es usar algún tipo de adelgazamiento morfológico que erosione sucesivamente los píxeles del límite (al tiempo que conserva los puntos finales de los segmentos de línea) hasta que no sea

posible un adelgazamiento, en cuyo punto lo que queda se aproxima al esqueleto. El método alternativo es calcular primero la transformación de distancia de la imagen.

El esqueleto se encuentra a lo largo de las singularidades (es decir, pliegues o discontinuidades de curvatura) en la transformación de distancia [16]. A continuación, la fig. 10 muestra una comparación entre la imagen original capturada en el ambiente controlado y la misma imagen esqueletizada.

2.4. Extracción de características para el análisis de imágenes

La forma característica de un objeto puede cuantificarse mediante momentos, los cuales describen la manera en que se distribuyen los píxeles de un objeto sobre el plano de la imagen. Los momentos deben ser invariantes (es decir, valores similares para objetos del mismo tipo) a las transformaciones geométricas (traslación, rotación y escala) que pueden sufrir los objetos y al mismo tiempo deben ser discriminantes (es decir, valores distintos para objetos de diferente tipo). Estas características son deseables para poder reconocer los objetos con mayor facilidad [17].

Los momentos de Hu son un conjunto de siete descriptores invariantes que expresan numéricamente la forma de un objeto, donde los dos primeros momentos se computan como:

$$\phi_1 = \eta_{20} + \eta_{02}, \tag{6}$$

$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4{\eta_{11}}^2, \tag{7}$$

donde los momentos centrales normalizados de segundo orden se calculan como:

$$\eta_{20} = \frac{m_{20} - \frac{{m_{10}}^2}{m_{00}}}{{m_{00}}^2}, \tag{8}$$

$$\eta_{02} = \frac{m_{02} - \frac{{m_{01}}^2}{m_{00}}}{{m_{00}}^2}, \tag{9}$$

$$\eta_{11} = \frac{m_{11} - \frac{m_{10}m_{01}}{m_{00}}}{{m_{00}}^2}, \tag{10}$$

y los momentos geométricos de orden $p + q$ se calculan como:

$$m_{pq} = \sum_x \sum_y x^p y^q I(x, y), \tag{11}$$

donde $I(x, y)$ es un píxel del objeto con las coordenadas (x, y) [17].

2.5. Entrenamiento del clasificador de patrones

Un patrón es un conjunto de características, mientras que una clase de patrones es un conjunto de patrones similares. El objetivo del reconocimiento de patrones es asignar un patrón a la clase a la que pertenece [18].

Random Forest es un método versátil de aprendizaje automático capacitado de efectuar tanto tareas de regresión como clasificación. Es un tipo de método de

Tabla 1. Comparativa de exactitud de algoritmos clasificadores.

Algoritmo	Exactitud	Imágenes clasificadas correctamente
SVM	83.00%	747/900
MLP	≈84.44%	760/900
kNN	≈84.44%	760/900
Árbol de decisión	≈84.44%	760/900
Random Forest	≈90.22%	812/900

Tabla 2. Comparativa de exactitud para diferentes valores de *k* utilizando validación cruzada.

K	Exactitud promedio
10	≈94.51%
30	≈97.15%
50	≈97.62%
100	≈98.19%

aprendizaje por conjuntos, donde un grupo de modelos débiles se combinan para formar un modelo robusto. Se generan múltiples árboles y cada árbol otorga una clasificación, es decir, vota por una clase y el resultado es la clase con mayor número de votos en todo el bosque [19].

Cada árbol se construye así:

a) Dado que el número de casos en el conjunto de entrenamiento es *N*. Una muestra de esos *N* casos se toma aleatoriamente, pero con reemplazo. Esta muestra será el conjunto de entrenamiento para construir el árbol i.

b) Si existen *M* variables de entrada, un número $m<M$ se especifica tal que, para cada nodo, m variables se seleccionan aleatoriamente de *M*. La mejor división de estos m atributos es usado para ramificar el árbol. El valor *m* se mantiene constante durante la generación de todo el bosque.

c) Cada árbol crece hasta su máxima extensión posible y no hay proceso de poda [20].

3. Los resultados

Después de aplicar 24 transformaciones geométricas, el banco de imágenes final se incrementó en un 2,500%, obteniendo 4,500 imágenes de cada una de las cuales fue sustraído un vector característico de tamaño 7.

Se destinó un 80% de estos vectores obtenidos para el conjunto de entrenamiento y el 20% restante para el conjunto de prueba, es decir, que se usaron 3,600 vectores característicos para el primer conjunto y 900 para el segundo.

Se realizaron pruebas con otros tipos de clasificadores de patrones, como Máquina de Soporte Vectorial (SVM), Red Neuronal Multicapa (MLP), kNN, y árbol de decisión, para comparar el desempeño de éstos con respecto al clasificador utilizado en este trabajo, Random Forest. Enseguida, en la tabla 1, se muestran los porcentajes de exactitud y el número de imágenes clasificadas correctamente alcanzados por cada uno de los algoritmos implementados haciendo uso de librerías del lenguaje de alto nivel Python en su versión 3.7.

Sin embargo, dada la naturaleza estocástica de la asignación de porcentajes a los conjuntos de entrenamiento y de prueba, no es posible asegurar que se obtendrán los mismos resultados y, en consecuencia, las mismas precisiones y errores en cada corrida del algoritmo. Por lo cual, se implementó un algoritmo de validación para observar el desempeño del clasificador de patrones en diferentes escenarios.

El método de validación cruzada consiste en tomar los datos originales, en este caso los 4,500 vectores característicos generados, para elaborar a partir de ellos los conjuntos separados de entrenamiento y prueba, para luego dividir en k subconjuntos el conjunto de entrenamiento, y en el momento del entrenamiento del algoritmo, tomará cada *k* subconjunto como conjunto de prueba del modelo, mientras que el resto será utilizado como conjunto de entrenamiento. El proceso está pensado para repetirse *k* veces, y en cada repetición, se elegirá un conjunto de prueba diferente y nuevamente los restantes serán parte del conjunto de entrenamiento. Cuando se realicen en su totalidad las *k* iteraciones, se calcula la precisión y el error para cada uno de los modelos producidos y se calcula el promedio de los *k* modelos para obtener la precisión y error finales [21].

Para diferentes valores de *k*, se muestra a continuación la exactitud promedio de los modelos generados en la tabla 2.

4. Conclusiones y trabajos futuros

Se ha presentado una arquitectura que permite la identificación de un sujeto, con una exactitud promedio mayor al 90%. Esto nos hace capaces de afirmar que, será capaz de identificar correctamente una imagen correspondiente al dorso de una mano de una persona, 9 de cada 10 veces que esa imagen sea ingresada.

Se logró implementar un ambiente controlado, que permite la captura de imágenes infrarrojas a través de la técnica de transiluminación.

También, se consiguió segmentar la red vascular subcutánea del dorso de una mano, manteniendo una abstracción de la información original.

Con todo esto dicho, las implicaciones de esta arquitectura podrían llevar a implementar la arquitectura en el ámbito público o privado, permitiendo la identificación de individuos de organizaciones pequeñas y medianas, dados los tiempos de procesamiento.

Como trabajo futuro debería considerarse la reducción de tiempos de procesamiento y optimización de los algoritmos implementados para realizar la identificación, así como la automatización de los procesos de captura de imágenes infrarrojas y transformar una sola toma del dorso de una mano, en un barrido fotográfico del dorso de una mano.

Referencias

1. Barrera-Rubio, P.: Realidad y prevención: Robo de identidad en México. El Economista (2020)
2. Cero Papel.: El 26% de las empresas en el mundo utilizan biometría para combatir el fraude (2019)
3. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of fingerprint recognition. Springer-Verlag, pp. 2–3 (2009)
4. Barrales-López, A.L.: Identificación biométrica por medio de la detección de venas de la mano en imágenes infrarrojas. Tesis de maestría CIDETEC (2015)
5. Real Academia de Ingeniería: Ambiente controlado. Diccionario de la Real Academia de Ingeniería (2015)
6. Sonomedical: Localizador de venas (2019)
7. Passariello, G., Mora, F.: Imágenes médicas: Adquisición, análisis, procesamiento e interpretación. Ediciones de la Universidad Simón Bolívar, pp. 61–98 (1995)
8. Passariello, G., Mora, F.: Imágenes médicas: Adquisición, análisis, procesamiento e interpretación. Ediciones de la Universidad Simón Bolívar, pp. 36– 38 (1995)
9. González, R., Woods, R.: Image segmentation. Digital Image Processing (pp. 778-780). Ed. Perason-Prentice Hall (2007)
10. Alonso, F.: Tema 6 técnicas de filtrado. https://um.es/geograf/sigmur/teledet/tema06.pdf (2005)
11. Arcgis.com: Función de brillo y contraste. ArcGIS for Desktop https://desktop.arcgis.com/es/arcmap/10.3/manage-data/raster-and-images/contrast-and-brightness-function.htm (2020)
12. González, R., Woods, R.: Morphological image processing. Digital Image Processing. Pearson-Prentice Hall, pp. 653 (2007)
13. Gómez, W.: Análisis de imágenes digitales, segmentación por umbralado. Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), pp. 35 (2015)
14. Ramudu, K., Krishna-Reddy, V., Abdul-Rahim, B.: Niblack method based segmentation for microscopic imagery. International Journal of Electrical and Electronics Engineers, 7 (1), pp. 26−27 (2015)
15. UDEA: 1.6. Conectividad, componentes conexas, vértice y aristas de corte. Universidad de Antioquia. http://docencia.udea.edu.co/regionalizacion/teoriaderedes/informaci%F3n/C1_Conectividad.pdf (2005)
16. Homepages: Morphology-skeletonization/medial axis transform. https://homepages.inf.ed.ac.uk/rbf/HIPR2/skeleton.htm (2020)
17. Gómez, W.: Reconocimiento de objetos en fotografías. Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), https://tamps.cinvestav.mx/~wgomez/toptamps/presentacion.pdf (2020)
18. Grupo de Topología Computacional y Matemática Aplicada: Introducción al reconocimiento de objetos. http://grupo.us.es/gtocoma/pid/tema7.pdf (2020)
19. González, L.: Aprendizaje supervisado: Random forest classification - Ligdi González https://ligdigonzalez.com/aprendizaje-supervisado-random-forest-classification/ (2020)
20. Orellana, J.: Árboles de decisión y Random Forest. https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html (2020)
21. Delgado, R.: Introducción a la validación cruzada (k-fold cross validation) en R. http://rstudio-pubs-static.s3.amazonaws.com/405322_6d94d05e54b24ba99438f49a6f8662a9.html (2018)

Propuesta metodológica para la predicción a corto plazo de contagios de COVID-19

María del Carmen Santiago Díaz, Ana Claudia Zenteno Vázquez,
Yeiny Romero Hernández, Judith Pérez Marcial,
Gustavo T. Rubín Linares,
Antonio Eduardo Álvarez Núñez

Benemérita Universidad Autónoma De Puebla,
México

{marycarmen.santiago, ana.zenteno, yeiny.romero, judith.perez, gustavo.rubin}@correo.buap.mx, eduard-alvarez@live.com.mx

Resumen. El coronavirus covid-19 es una pandemia muy grande y se requieren modelos matemáticos para simular escenarios y proyecciones que brinden información más precisa basada en las variables de comportamiento actuales. En este trabajo partimos del análisis y procesamiento de los reportes oficiales diarios del número de casos positivos confirmados N, cada reporte se divide por estados y se calcula el orden "n" de una regresión polinomial P(x) que maximice R^2, a los coeficientes A_i de ésta también se les aplican regresiones polinomiales de diversos órdenes para maximizar R^2 para cada estado y cada reporte, enseguida se aplica la derivada a estas regresiones polinomiales de los coeficientes y con ella se determina el orden más adecuado para el polinomio de los coeficientes el cual se utiliza para generar el polinomio de cada estado nuevamente. Con esta metodología se encontró un mecanismo para predecir a corto plazo el escenario estatal y nacional en días posteriores al último reporte oficial.

Palabras clave: Covid-19, modelo, predicción, contagios.

Methodological Proposal for the Short-Term Prediction of COVID-19 Infections

Abstract: The covid-19 coronavirus is a very large pandemic and mathematical models are required to simulate scenarios and projections that provide more accurate information based on current behavioral variables. In this work, we start from the analysis and processing of the daily official reports of the number of confirmed positive cases N, each report is divided by states and the order "n" of a polynomial regression P (x) is calculated that maximizes R2, to the coefficients Ai of this, polynomial regressions of various orders are also applied to maximize R2 for each state and each report, then the derivative is applied to these polynomial regressions of the coefficients and with it the most suitable order for the polynomial of the coefficients is determined, which is used to generate the polynomial of each state again. With this methodology, a mechanism was found

to predict in the short term the state and national scenario in days after the last official report.

Keywords: Covid-19, model, prediction, contagion.

1. Introducción

La sociedad a lo largo de la historia ha tenido que combatir diversos tipos de enfermedades que han ocasionado altas tasas de mortalidad y graves consecuencias económicas y de salud. Los virus que causan enfermedades siempre han existido, se han mutado y han aparecido en diferentes momentos del mundo. En diciembre de 2019, surgen casos de virus en China, particularmente en Wuhan (Hubei). Estos casos están vinculados a un mercado mayorista de mariscos, pescado y animales vivos, y han propiciado una ola de investigaciones del desarrollo de la pandemia en todos los países. La Organización Mundial de la Salud (OMS) recibió reportes de varios casos de neumonía de etiología desconocida. El 3 de enero de 2020, las autoridades de China notificaron a la OMS que existían 44 pacientes con neumonía de etiología desconocida, de entre los cuales, 11 pacientes estaban gravemente enfermos. Según informaciones difundidas en los medios de comunicación, el mercado implicado en Wuhan se cerró el 1 de enero de 2020 para realizar acciones de saneamiento y desinfección ambiental [1]. El día 8 de enero en Tailandia detectó un primer caso (fuera de China), siendo el 10 de enero el día que se presenta el primer fallecimiento causado por el virus. El incremento de los casos que aparecen en China y en otros países pone en evidencia la gravedad de la situación y la OMS el 10 de enero publica orientaciones técnicas y recomendaciones para todos los países sobre el modo de detectar y gestionar casos y para realizar pruebas de laboratorio. Para el 30 de enero la OMS señala la existencia de un total de 7818 casos confirmados en todo el mundo, la mayoría de ellos en China y 82 en otros 18 países. La OMS evalúa el riesgo en China como muy alto y el riesgo mundial como alto [2]. Para el mes de septiembre, a nivel mundial se contabilizan más de 34 millones de contagios y aproximadamente un millón de decesos. El 1 de octubre, en México se confirman 748,315 casos y 78,078 decesos [3].

2. Estado del arte

Diferentes investigaciones se están llevando con el propósito de modelar el número de casos sospechosos, confirmados, decesos, ocupación de instalaciones hospitalarias, formas y patrones de contagio, entre otras variables. En los modelos epidemiológicos se parte del supuesto de que los individuos se encuentran en uno de varios estados posibles, Susceptible (S), Infectado (I) o Recuperado (R).

Los modelos matemáticos SI, SIS, SIR y sus derivados se emplean para predecir el impacto de las pandemias en las sociedades, en los que la interacción de los individuos es crucial para la propagación del virus y por eso es recomendación el confinamiento para minimizar los estragos en las sociedades. Son usados para tratar enfermedades que afectan a poblaciones grandes y a menudo surgen representados a través de ecuaciones diferenciales[4].

En el Centro de Investigación y Docencia Económicas (CIDE) el equipo del profesor Escudero desarrolló un modelo matemático de proyecciones sobre los efectos de las distintas prácticas de mitigación sobre COVID-19. Otro modelo, denominado SC-COSMO por las siglas Stanford-CIDE COronavirus Simulation MOdel se encarga de analizar cómo evoluciona la enfermedad y modela también los mecanismos en los que los individuos interactúan entre sí [5]. Este modelo realiza análisis demográficos para considerar a los individuos susceptibles, a los expuestos, los infectados y los recuperados, conforme a los patrones de contacto que fueron clave para la transmisión de covid-19.

Los casos confirmados, sospechosos y decesos se miden diariamente bridando información a los gobiernos sobre la capacidad de atención a la población [6]. En este punto, el tiempo es un factor que permite observar las tendencias de los casos cuando hay mejora, los decesos y la evolución en duplicación en los casos a nivel local y global. El "Covid-19 Modelo numérico de casos de infección y estimaciones epidémicas modelo asimétrico - Gompertz" [6] obtiene una estimación de la demanda hospitalaria para casos graves que necesitan terapia intensiva en España y muestra los beneficios de mantenerse en casa, para que, en caso de ser contagiado se cuente con una mejor atención médica.

En México, la información del aumento de casos fluía lentamente y no permitía generar modelos que brindaran información concreta del desarrollo de la pandemia a nivel estatal y nacional. Cuando los casos confirmados comenzaron a incrementar, a partir de 500 fue posible aplicar diversas estrategias metodológicas para extraer de la información reportada diariamente, modelos y tendencias de la información. Aunque el número de casos acumulados aumentaba diariamente se observaron comportamientos característicos en los reportes presentados, además se analizó el porcentaje de casos distribuidos con respecto a las fechas de inicio de síntomas, las de ingreso y defunciones, y se encuentra que un gran porcentaje de los casos confirmados fallecen pocos días después de su ingreso a los servicios hospitalarios y más alarmante aún es que de estas lamentables defunciones la mayoría presentó síntomas mucho tiempo antes de su ingreso.

Esta información es muy importante para nosotros porque reafirma el hecho de que el número de casos acumulados se alimenta con casos ocurridos varios días antes del reporte. Por eso, este modelo utiliza los reportes oficiales publicados diariamente y genera los modelos matemáticos para cada uno, los cuales mediante un proceso de análisis similar se descomponen en sus coeficientes o variables numéricas para optimizarse y a partir de ellos mejorar la estimación del comportamiento del número de casos confirmados positivos reportados diariamente y generar un modelo que describa mejor los escenarios futuros.

3. Metodología

La información reportada por la Dirección General de Epidemiología del gobierno de México y publicados en el portal oficial brindan un panorama muy completo de la evolución de la pandemia, sin embargo la interpretación no es trivial en el sentido de que la información requiere ser procesada con extremo cuidado. Utilizamos para este trabajo sólo la información correspondiente al número de casos confirmados positivos

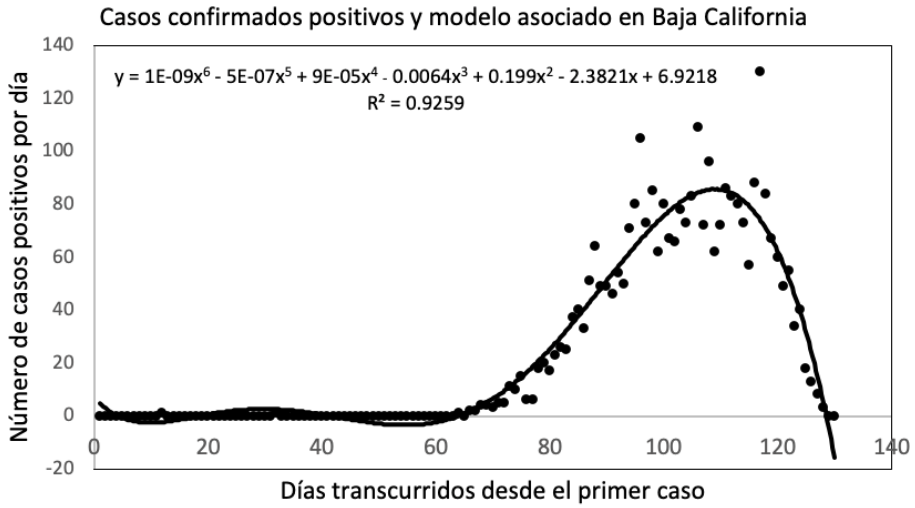


Fig. 1. Gráfica que muestra los casos confirmados positivos por día en Baja California, el polinomio de ajuste y el coeficiente de correlación R^2.

para la prueba de SARS-COv2 por día, por estado y por municipio. En un trabajo relacionado procesamos a nivel nacional una metodología basada en los coeficientes de los polinomios de regresión aplicados a cada día de reporte nacional [7]. En este trabajo mostramos una variante a esta metodología, ya que el llevar el modelo polinomial a la base de datos de casos confirmados por estado, implica estados donde hay una alta dispersión en los datos, es decir, no siguen una tendencia totalmente homogénea. El modelo polinomial se justifica de acuerdo a las siguientes evidencias obtenidas al procesar la información:

1. Brinda un Coeficiente de Correlación R^2 cercano a 1 al utilizarlo para describir el número de casos diarios reportados.
2. Genera en el número acumulado de casos una aproximación menor al 5%.
3. A partir del orden 6 los cambios en el Coeficiente de Correlación son menores a 0.01, sin importar el día que se analice.
4. El pico máximo del número de casos diarios sigue el mismo comportamiento en el modelo polinomial que en el reportado.

Entonces realizamos un filtrado e inspección de la información por estado y aplicamos regresiones polinomiales encontrando que los coeficientes de correlación para cada estado presentan el valor más próximo a 1 para polinomios de orden 6, con lo cual utilizamos los 7 coeficientes por cada estado en la información reportada por día y realizamos una regresión polinomial a cada familia de coeficientes. Posteriormente aplicamos el coeficiente de correlación para analizar cual polinomio nos brinda una mejor aproximación y en este punto el sistema nos muestra lo obvio, que el orden más alto nos va a generar el coeficiente R^2 más alto, sin embargo una inspección nos revela que este hecho nos lleva a que una vez que el polinomio genera una curva muy similar a la información de los coeficientes, al pasar por el último dato tiene un cambio significativo en su comportamiento, por lo cual, nos surge de forma

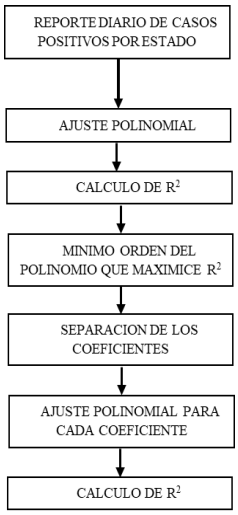


Fig. 2. Diagrama a bloques que muestra la metodología de análisis y procesamiento para la información de casos positivos por día para cada estado.

natural la necesidad de aplicar la derivada a la curva posterior a su entorno conocido y así quedarnos con el que tenga la derivada más suave y un coeficiente de correlación en un rango aceptable. En seguida mostraremos las etapas discutidas previamente.

3.1. Modelo matemático

A la información filtrada por estados le aplicamos el modelo polinomial para obtener una representación analítica del comportamiento de los casos confirmados positivos con lo cual determinamos el mínimo orden con el mayor coeficiente R^2, este modelo es del mismo orden para todos los estados, tanto los que concentran el 70% de los casos a nivel nacional como aquellos con solo el 1%, como se muestra en la figura 1.

Se llevó a cabo el análisis de todos los estados desde el 15 de mayo hasta el 24 de junio y como puede mostrarse en la figura 1 en el caso de Baja California, hay una gran dispersión en la información, sin embargo, el modelo tiene un buen coeficiente R2 para orden 6. Partiendo de estos hechos, se establece una ecuación que determina el comportamiento de cada uno de los 7 coeficientes del polinomio de orden 6, para cada estado. Es decir, para un estado tenemos en este rango de fechas 41 ecuaciones de orden 6 de las cuales obtenemos 41 elementos para cada coeficiente y que enseguida aplicamos regresiones de orden 2, 3, 4, 5 y 6 y sus respectivos coeficientes de correlación. Como se muestra en el siguiente diagrama a bloques.

Una vez que realizamos el procesamiento hasta la separación de los coeficientes de la figura 2, encontramos que cada uno de estos coeficientes del polinomio de orden 6 de la ecuación 1 tienen características como las que se muestran en la figura 3 para el caso de Aguascalientes, donde ya se concentra la información del coeficiente A6 para los 41 reportes:

$$y = A_6x^6 + A_5x^5 + A_4x^4 + A_3x^3 + A_2x^2 + A_1x^1 + A_0\ . \qquad (1)$$

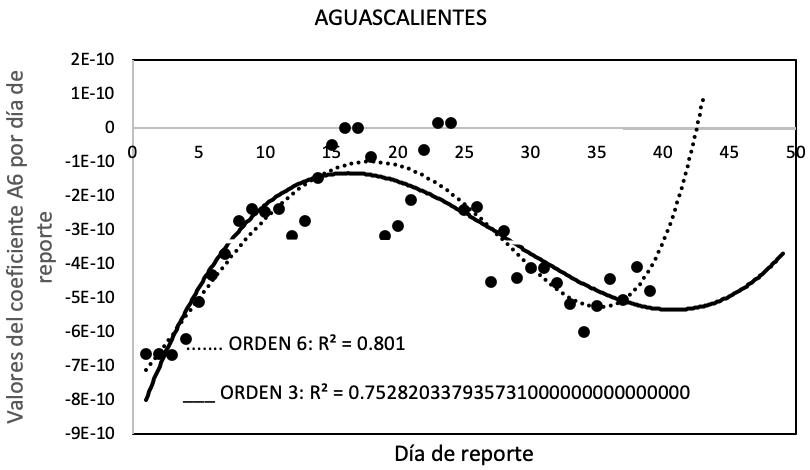


Fig. 3. Gráfica del coeficiente A_6 obtenido del polinomio de orden 6 del reporte diario para el número de casos positivos en Aguascalientes y las curvas polinomiales de orden 6 y 3, respectivamente.

La figura 3 muestra un resultado muy frecuente que se presenta al llevar a cabo la determinación del modelo de un conjunto de información, el aumentar el orden de ajuste polinomial nos lleva a mejorar el coeficiente R^2, sin embargo la curva de orden mayor puede tener un comportamiento asintótico o demasiado brusco comparado con el comportamiento de la información que está modelando, por ello aplicamos la derivada a la curva y con esto obtenemos el orden que posea un balance entre el coeficiente R^2 y la suavidad de cambio que mide la derivada.

Enseguida aplicamos las regresiones polinomiales de orden 2, 3, 4, 5 y 6 a los 7 coeficientes de cada estado para los 41 días reportados. Y con ayuda de la derivada que implementamos de forma discreta, es decir, a partir de las diferencias entre pares de puntos, disminuimos el orden del R^2 máximo a un orden con $R^2>0.7$, este hecho disminuye el orden a 5 y cuatro y en algunos casos hasta 3, esta decisión se lleva a cabo de forma automática por el sistema, obteniendo fluctuaciones de 2 y 3, a diferencia de lo que se obtendría únicamente con el coeficiente R^2, en cuyo caso obtenemos que predominan los órdenes 5 y 4, y de forma poco significativa 2 y 3.

Estos ordenes se promedian para cada estado, es decir en cada estado se promedian los órdenes de los 7 coeficientes y este promedio se establece como el orden de todos los coeficientes para ese estado.

4. Resultados

Una vez que se determina el orden del polinomio para cada coeficiente de cada estado se utiliza este polinomio para generar el nuevo polinomio de orden 6 del estado, pero ahora con el ajuste de los coeficientes.

En la figura 4 se muestra la información reportada los días 3 y 16 de junio, es notoria la dispersión en la información sin embargo el ajuste después del tratamiento

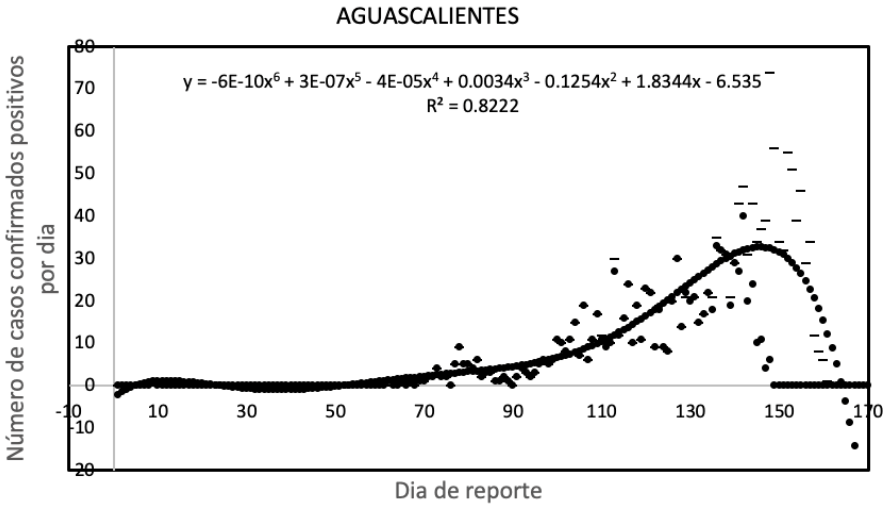


Fig. 4. Gráfica de casos positivos para los días 3 y 16 de junio (- - , . . , respectivamente) y la simulación polinomial de orden 6 con los coeficientes ajustados en orden 4.

Tabla 1. Validación y proyecciones a 2 días de las simulaciones.

	No. de casos reportados día 41	**No. de casos simulados día 41**	**No. de casos simulados día 42**	**No. de casos simulados día 43**
Nacional	129184	131801.65	134085.74	138567.29
Baja California	6436	6343.87	6475.17	6616.65
Puebla	4742	4700.12	4878.48	5091.51

mencionado anteriormente obtiene una curva de orden 6 para el estado de Aguascalientes con un coeficiente R^2=0.82, utilizando un coeficiente de orden 4 para los coeficientes, siendo que el coeficiente sugerido sin la derivada es de orden 5.

Esta información obtenida para todos los estados nos permite determinar el número de casos confirmados positivos a nivel nacional y estatal, para los días reportados y días posteriores, como se observa en la siguiente tabla.

5. Conclusiones

El sistema presenta errores con respecto a los reportes oficiales en un rango promedio del 4.91%, aunque se obtuvieron errores cercanos a cero en general, también se tienen estados donde la gran dispersión origina errores del 30%, esto es debido a que el sistema polinomial ajusta buscando el mejor polinomio de los casos positivos y posteriormente hacemos otro ajuste polinomial en los coeficientes del polinomio

anterior, pero los dos conjuntos de información presentan una alta dispersión, es decir, el costo por buscar una mayor precisión en el número de casos resulta en algunos estados en una excelente aproximación, sin embargo hay otros en los cuales se debe considerar un criterio distinto debido a que internamente la información en cada estado presenta diferentes velocidades de contagio por día y en la información estatal solo se observa como una alta dispersión. La estrategia que ya estamos desarrollando es el análisis por municipio, el cual tendrá mucho más detalle de la velocidad de propagación de la pandemia, pero sin duda el problema será que las regresiones polinomiales deberán aplicarse de forma más cuidadosa, lo cual en el caso por estados ya se utilizaron algunas consideraciones, como restringir el dominio de la información útil que modela el polinomio, es decir, no considerando en algunos estados el mismo inicio, ya que no todos los estados tienen el primer caso el mismo día y esto afecta considerablemente el modelo, ya que al ser de orden alto es natural que posea diversas oscilaciones aun cuando la información se mantenga constante.

6. Trabajo futuro

Los resultados obtenidos han demostrado que la metodología es adecuada ya que brinda resultados cercanos a los casos reportados, así que ahora se debe integrar la información de municipios, estados y país, además de la información de casos negativos, sospechosos y defunciones, lo cual, junto con las características de la población que ha sido reportada nos permitirá conocer más detalles del comportamiento matemático. Aunque no se tiene ninguna referencia de alguna metodología parecida si se está comparando con aquellas que utilizan el número de casos acumulados, a fin de encontrar mejoras en el modelo.

Referencias

1. OMS: Neumonía de Causa desconocida China (2020)
2. Novel: Coronavirus (2019-nCoV) Situation Report-10 (2020)
3. Gobierno de México: Datos Coronavirus (2020)
4. Sánchez-Villegas, P., Daponte-Codina, A.: Modelos predictivos de la epidemia de COVID-19 en España con curvas de Gompertz. Gaceta Sanitaria (2020)
5. CIDE: CIDE y Stanford desarrollan modelo matemático de proyecciones sobre COVID-19 (2020)
6. Borja, A., Grasso, D., Llaneras, K., Galindo, J.: Así evoluciona la curva del coronavirus en México, Colombia, Chile, Argentina y el resto de Latinoamérica. El País (2020)
7. Zenteno, A.C., Santiago, M.C., Romero, Y., Pérez, J., Rubín, G.T., Álvarez, A.E.: Optimización de los coeficientes del modelo predictivo del número de casos diarios de coronavirus Covid-19 en México. In: Simposio Nacional de Inteligencia Artificial e Industria 4.0 (2020)
8. Li, L., Yang, Z., Dang, Z., Meng, C., Huang, J., Meng, H., Wang, D., Chen, G., Zhang, J., Peng, H., Shao, Y.: Propagation analysis and prediction of the COVID-19. Infectious Disease Modelling, 5, pp. 282−292 (2020)
9. Ketema-Mamo, D.: Model the transmission dynamics of COVID-19 propagation with public health intervention. Applied Mathematics (2020)

10. Sameni, R.: Mathematical modeling of epidemic diseases: A case study of the COVID-19 (2020)

Exploración y dimensionamiento de espacios desconocidos utilizando un robot terrestre

A. Bello-Germán[1], A. Gumeta-López[1], O. Villegas-Olguín[1], P.J. Escamilla-Ambrosio[2]

[1] Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Zacatenco,
México

[2] Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

pescamilla@cic.ipn.mx

Resumen. En este artículo se presenta el trabajo llevado a cabo para realizar una prueba de concepto (PoC) en un ambiente controlado para demostrar la implementación de un robot móvil terrestre en situaciones donde debe de evitar obstáculos. El robot cuenta con una cámara RGBD, una cámara de seguimiento (tracking camera), y un sensor ultrasónico, que tiene como finalidad realizar la construcción de un mapa en 3D, y el registro de la trayectoria del robot terrestre demostrando que es viable su utilización. Además, con la realización de la fusión de sensores mediante la implementación del algoritmo conocido como filtro Kalman, para poder predecir el estado del sistema. Los sensores y demás periféricos son integrados en una tarjeta de desarrollo NVIDIA Jetson Nano, la cual controla los parámetros recolectados, además de procesar los datos (distancias, velocidad, puntos de referencia, etc.) que posteriormente son de utilidad para su uso de la estrategia de navegación desarrollado. Además, se demuestra que la nube de puntos es capaz de mostrar la forma de diferentes objetos haciéndolos de fácil identificación en el entorno, y haciendo posible su uso en situaciones de difícil acceso para el ser humano.

Palabras clave: Fusión de sensores, RGBD, seguimiento, Mapa 3D, robot móvil, exploración, espacios desconocidos.

Exploration and Sizing of Unknown Spaces Using a Ground Robot

Abstract. This article presents the work carried out to carry out a proof of concept (PoC) in a controlled environment to demonstrate the implementation of a land mobile robot in situations where it must avoid obstacles. The robot has an RGBD camera, a tracking camera, and an ultrasonic sensor, whose purpose is to construct a 3D map, and record the trajectory of the terrestrial robot, demonstrating that its use is feasible. . In addition, with the realization of the

fusion of sensors through the implementation of the algorithm known as the Kalman filter, to be able to predict the state of the system. The sensors and other peripherals are integrated into an NVIDIA Jetson Nano development card, which controls the collected parameters, in addition to processing the data (distances, speed, reference points, etc.) that are later useful for the use of the device, navigation strategy is developed. In addition, it is shown that the point cloud is capable of showing the shape of different objects, making them easy to identify in the environment, and making it possible to use them in situations that are difficult to access for humans.

Keywords: Sensor fusion, RGBD, tracking, 3D map, mobile robot, exploration, unknown spaces.

1. Introducción

La robótica móvil ha tenido un gran desarrollo a través de estos años, permitiendo la utilización de diversos robots en muchos campos de la industria con un alto grado de confiabilidad. Desde la década de los 70's, se han planteado las bases para poder llevar a cabo estos avances, las cuales se han plasmado en el libro titulado "Manipulators: Mathematics, Programming and Control", escrito por Richard Paul del MIT [1].

Actualmente, diferentes robots móviles han tenido participación en la navegación de lugares hostiles, como lo es el caso de la luna o el desierto de Mojave [2]. Esto ha causado que su utilización para el reconocimiento de lugares desconocidos sea altamente demandado, facilitando la tarea al ser humano, y no solo eso, sino también se vuelven productos para la realización de tareas complejas, tediosas o de alto riesgo.

Para la contribución de este trabajo, se planteó la utilización de dos cámaras desarrolladas por Intel para la creación de la trayectoria que sigue un robot y la generación del mapa del área que recorre, haciendo la utilización de técnicas como la fusión de sensores para poder tener una mayor confiabilidad en los datos recabados por las cámaras, y así tener una predicción del movimiento del robot, facilitando además la estrategia de navegación.

Este trabajo se estructura en ocho partes las cuales seis (de la sección 2 a la 7) son las que explican el desarrollo del proyecto. Comenzando la lectura sobre el proyecto en la sección número dos, donde se describe la integración de los componentes que se utilizan para la construcción y desarrollo del robot móvil, en la sección tres se tocan las consideraciones de software que se utilizó para la construcción del mapa en 3D.

Para la sección cuatro, se describe la estrategia de desplazamiento que se desarrolló para la navegación en el entorno. En la sección cinco se habla sobre la alineación espacial que se debe de tomar en cuenta para el uso de las cámaras y el sensor ultrasónico, además de mostrar los pormenores matemáticos que se realizaron para la fusión de sensores.

Para la sección seis, se habla sobre las pruebas que se realizaron para comprobar la navegación del robot evitando obstáculos, en donde la nube de puntos es de utilidad para la creación del mapa y la identificación de objetos, así como que es posible conocer su trayectoria. En la séptima sección se presentan las conclusiones y trabajo futuro de esta investigación.

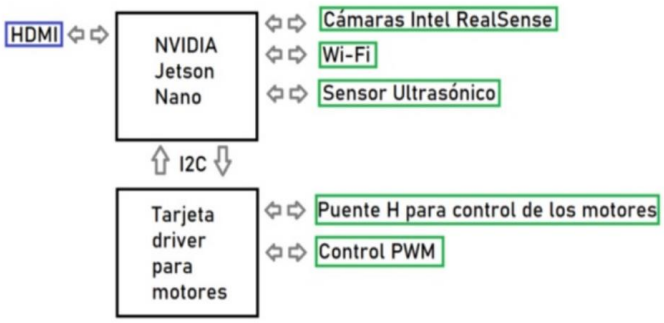


Fig. 1. Diagrama a bloques de la integración de los componentes.

2. Integración de componentes

El robot está conformado por una tarjeta de desarrollo NVIDIA Jetson Nano (SBC, Single Board Computer), la cual tiene como sistema operativo Ubuntu 18.04 LTS. Se tiene un sensor ultrasónico HC-SR04, el cual será de utilidad para la navegación del robot. Cuenta con dos cámaras desarrolladas por INTEL de la gama RealSense: Tracking Camera T265 y Depth D435i [3].

La tarjeta Jetson Nano es utilizada para el procesamiento de la información, control e integración de los componentes periféricos. Además, ejecuta el framework de ROS para la integración de todos los códigos necesarios para poder llevar a cabo en conjunto la navegación del robot junto con la creación del mapa.

La función del sensor HC-SR04 es orientar al robot en espacios estrechos, así como facilitar el algoritmo de navegación de éste.

La implementación de ambas cámaras tiene como fin realizar una nube de puntos y conocer la trayectoria que realiza el robot. La cámara T265 es la encargada de marcar la trayectoria que recorre el robot a través del sensor interno IMU (Inertial Measurement Unit), el cual cuenta con sensores de aceleración y velocidad angular de 3 ejes, que facilita la medición de la posición en la que se encuentra la cámara, otorgando datos en los tres ejes (X, Y, Z) del plano cartesiano.

La cámara D435i realiza la nube de puntos, el cual ayuda con la generación del mapa en 3D, así como la recolección de las distancias que mide con ayuda de un láser interno. Las distancias medidas son entre el robot y un objeto interpuesto en la trayectoria.

En conjunto con las cámaras, el sensor ultrasónico, los periféricos (motores dc, baterías, etc.), se realiza la interconexión de los elementos en la tarjeta Jetson Nano para el funcionamiento apropiado del robot como se muestra en la Figura 1 [4].

Los componentes que se pueden observar en la Figura 1, muestran cómo están interconectados a la tarjeta Jetson Nano y a un puente H, haciendo que sea de fácil entendimiento la conexión a las tarjetas.

Fig. 2. Diagrama a bloques de la transformación de los datos recolectados por las cámaras.

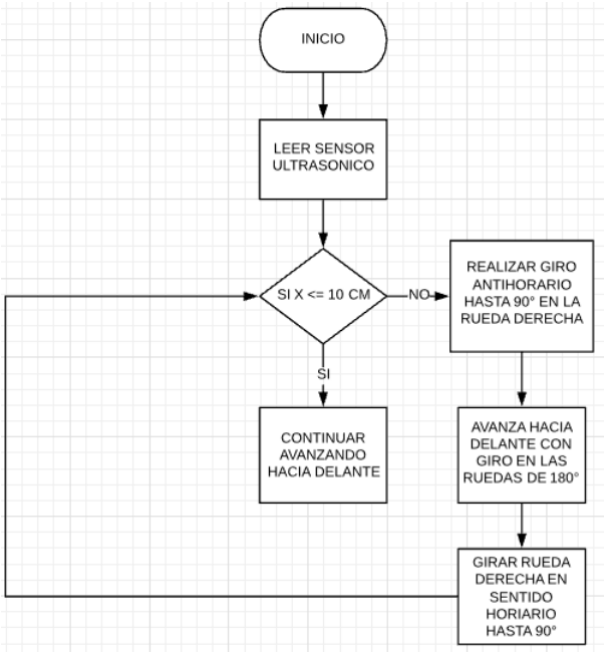


Fig. 3. Diagrama de estrategia de navegación con el sensor ultrasónico.

3. Software

Se utilizó ROS (Robotic Operating System) para la implementación del software encargado del manejo del robot, se emplean los programas Rviz y RealSense Viewer, para la observación de la nube de puntos, y la construcción del mapa en 3D, además de la utilización de los lenguajes de programación en Python y C++, para desarrollar la estrategia de navegación con los datos recolectados por los sensores.

La programación del robot es desarrollada e integrada por el framework llamado ROS Melodic, integrando los códigos desarrollados en Python y en C++ para el control del sensor ultrasónico junto con los motores DC, que tiene como finalidad la corrección del movimiento del robot móvil conforme a la estrategia de desplazamiento en el entorno. El framework es el encargado de abstraer la información y utilizarla para que

el robot pueda moverse libremente, para evitar obstáculos, para la construcción del mapa y la obtención de datos de la posición del entorno [5].

El proceso del framework para la transformación de los datos recolectados por las cámaras se observa en la Figura 2, pero es importante realizar el énfasis que, para poder realizar el rastreo y mapeo con las cámaras, se debe de tener en consideración como será el flujo de la información, que es como se muestra en la figura ya mencionada. Además, es de importancia remarcar que la cámara D435i es la encargada de realizar la construcción de la nube de puntos, y la cámara T265 es la que realiza el posicionamiento del robot.

4. Estrategia de desplazamiento

La estrategia que se muestra en la Figura 3, se encuentra apoyado en el uso de un sensor ultrasónico el cual hace que el robot móvil siempre conduzca cercano a alguna pared u objeto que se encuentre al lado derecho del robot. Lo anterior con el fin de facilitar la navegación del robot, sin la necesidad de desarrollar algún otro algoritmo que requiera de algún tipo de entrenamiento como lo sería con la utilización de machine learning.

Con el diagrama anterior, se asegura que el robot siempre guarde una distancia de 10 cm con respecto a alguna pared u objeto que se encuentre en su lado izquierdo, haciendo que su desplazamiento sea siempre predecible y seguro.

Para realizar la lectura de los datos recabados por el sensor, se implementa un código realizado en Python, el cual recaba las mediciones y realiza las comparaciones pertinentes para poder aproximar al robot a la pared u objeto, además de realizar las correcciones necesarias con la activación de los motores DC para siempre mantener el margen deseado.

5. Alineación espacial y fusión de sensores

Las cámaras se enlazan entre sí para realizar el entendimiento del entorno en 3D que se genera, de igual manera para conocer la posición y la orientación del robot. A la técnica empleada para que este enlace se lleve a cabo se le conoce como Alineación Espacial [3], lo cual implica marcar un punto de referencia estático en el entorno y contar con un cuadro de referencia utilizando las coordenadas 0,0,0 (x,y,z). La Alineación Espacial también tiene como fin que los puntos obtenidos sean transformados a cuadros de profundidad (Depth Frame), y así tener una referencia real del entorno.

Los parámetros para tener en consideración son la traslación (t_x, t_y, t_z) y la rotación, definida como una matriz 3x3, como se muestra en la ecuación 1.

$$a_p = \begin{bmatrix} r_{11} & r_{21} & r_{31} \\ r_{12} & r_{22} & r_{32} \\ r_{13} & r_{23} & r_{33} \end{bmatrix} b_p + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}. \quad (1)$$

Los programas que se utilizan arrojan las coordenadas r_{ij}, t_x, t_y y t_z, siempre con unidades en metros [3].

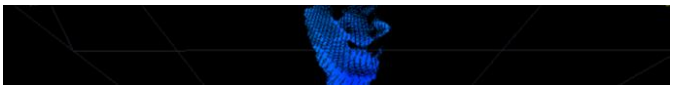


Fig. 4. Identificación de una persona con la generación de la nube de puntos.

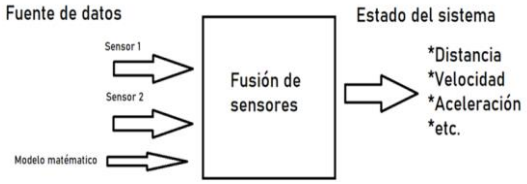


Fig.5. Descripción gráfica de la fusión de sensores.

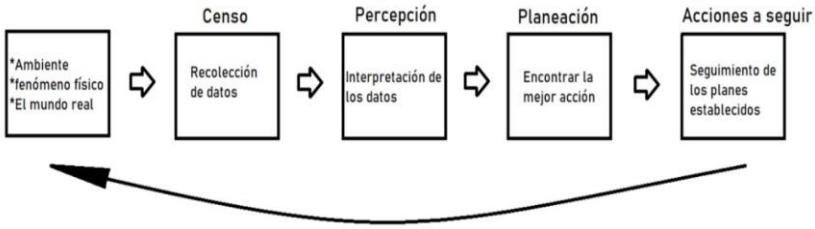


Fig. 6. Diagrama a bloques de los pasos a seguir para la fusión de sensores.

Con la ayuda del láser de la cámara D435i, se escanea el ambiente para construir el mapa en 3D, con esto se mide de manera automática las distancias y la profundidad, así como la distinción a los objetos. En la Figura 4 se observa la distinción de una persona (al centro en azul) con la ayuda de la nube de puntos [6, 7].

Al tener diferentes fuentes de datos la medición de un fenómeno físico se vuelve más complicada la comprensión de los datos, es por esto, por lo que con la fusión de sensores se puede tener un dato confiable y de gran precisión [8] En la Figura 5, se muestra un diagrama a bloques de cómo opera la fusión de sensores.

Al combinar múltiples sensores para la medición de un fenómeno físico como la aceleración o la posición de un objeto, se llevan a cabo diferentes procesos para poder determinar qué acción siguiente será la mejor, además de obtener información más precisa. Los pasos que seguir para llevar a cabo la fusión de sensores se observan en la Figura 6 [9, 10].

Para llevar a cabo la fusión de sensores se necesitan las distribuciones normales de probabilidad de las diferentes medidas que se obtienen al momento de caracterizar los sensores, ya que ayuda a conocer el promedio y la varianza entre los diferentes sensores, además que al utilizar un método como el filtro Kalman para fusionar los sensores, haciendo que se tenga la estimación del sistema de forma recursiva. Para este proyecto se realizaron 100 mediciones en el rango de 20 cm a 1.13 m, (a menos de 20 cm la cámara D435i no detecta eficientemente los objetos). En la Figura 7 se muestra la caracterización de la cámara D435i, junto con el sensor HC-SR04.

Al hacer la comparación de las medidas entre la cámara y el sensor ultrasónico se observó que la cámara es mucho más precisa debido a la precisión del láser para obtener

Fig. 7. caracterización de los sensores.

Tabla 1. Promedio de sensores.

Promedio distancia	Promedio cámara RGB	Promedio sensor ultrasónico
66.6606	65.5757	64.3030

la distribución normal, como se muestra en la Tabla 1. Además, el sensor ultrasónico cuenta con el inconveniente de que se encuentra sujeto al su ángulo de medición, y de la superficie en la que reboten las ondas, ya que la onda debe de incidir directamente sin contar con ningún ángulo ya que medición pude ser errónea.

Con los datos obtenidos de la Tabla 1 se obtuvo la distribución normal con la ecuación 2, como también se observa en las distribuciones normales de la Figura 8.

$$G_{x,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (2)$$

Observando las gráficas que se encuentran en la Figura 8, se notan que son muy similares, indicando que no existe gran variación en cuanto a los datos obtenidos por los sensores respectos con las medidas realizadas con un flexómetro. Entonces se puede decir que se cuenta con una resolución bastante buena en los sensores.

Esta ligera diferencia también puede ser debido a que cuentan con un funcionamiento diferente, también de que el ruido que produce el sensor ultrasónico es mayor al de la cámara RGBD.

El ruido que producen los sensores es de gran importancia en su análisis, ya que pueden producir errores en las mediciones que, en este caso, son las distancias medidas por los sensores.

A partir de la ecuación 2 se obtienen los pesos (w) para cada sensor, para conocer cuál será el sensor que tendrá mayor relevancia al momento de obtener los datos.

Primero se obtiene la varianza de cada sensor teniendo como resultado los que se muestran en la Tabla 2.

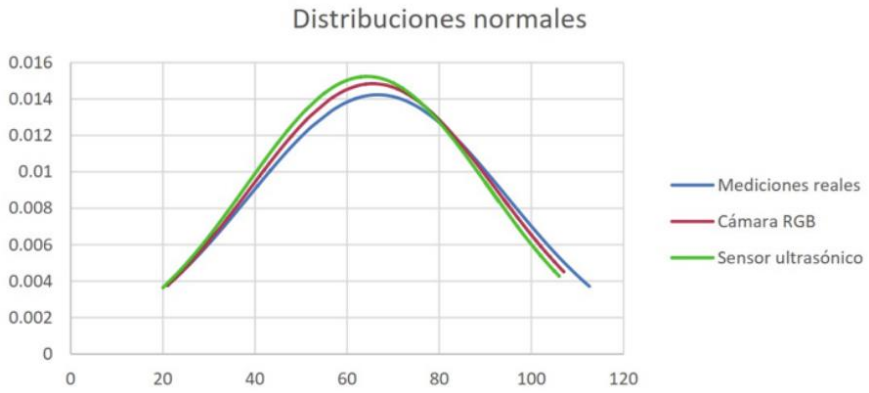


Fig. 8. Mapa en 3D generado con la nube de puntos.

Tabla 2. Promedio de sensores.

Varianza cámara RGB	Varianza sensor ultrasónico
722.593692	685.8460111

Tabla 3. Promedio de sensores.

w_1 cámara RGB	w_2 sensor ultrasónico
0.486954	0.513045

Con la siguiente ecuación, se obtienen los pesos para cada sensor:

$$x = \left[\frac{\sigma^2{}_{z2}}{\sigma^2{}_{z1} + \sigma^2{}_{z2}}\right] Z_1 + \left[\frac{\sigma^2{}_{z2}}{\sigma^2{}_{z1} + \sigma^2{}_{z2}}\right] Z_2 \,, \tag{3}$$

donde Z_1 y Z_2 son las correcciones de término, que sirven para que los datos finales no permanezcan con ruido innecesario debido a las mediciones [10].

En la Tabla 3 se muestran los pesos para cada sensor. La media o promedio de la fusión de sensores es $X_2 = 64.9558$. Al utilizar esta forma de fusionar los sensores a partir de los datos recabados, se tiene como fin el solo reconocer los objetos que pudieran estar enfrente del robot.

6. Prueba de funcionamiento

Para comprobar la construcción del mapa en 3D se planteó que el robot realice un recorrido a través de 3 obstáculos, donde sea capaz de evitar los objetos que encuentre durante el recorrido, que se puedan identificar la forma de los objetos, y finalmente

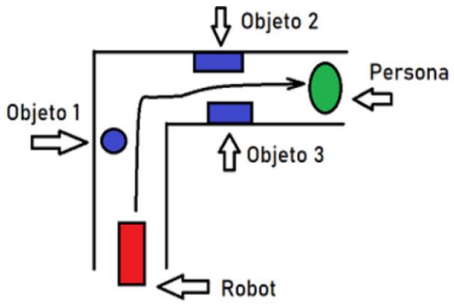


Fig. 9. Ruta de desplazamiento realizado por el robot.

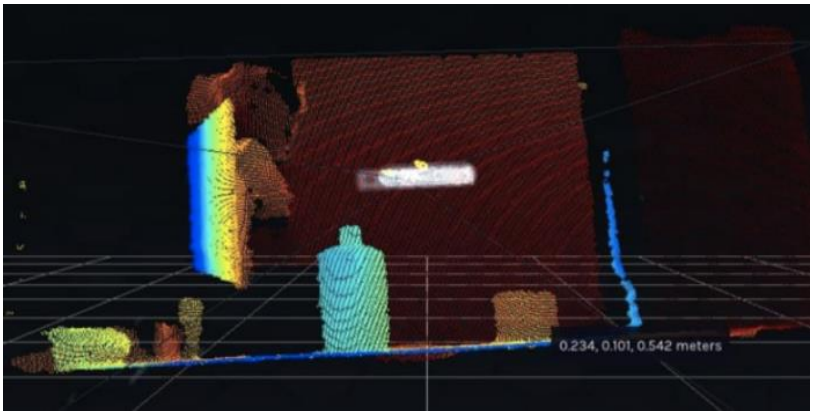


Fig. 10. Posición inicial del robot, mostrando un bote enfrente.

encontrar al final del recorrido a una persona recostada en el suelo. El recorrido planteado se observa en la Figura 9. Este recorrido debe de cumplir con los requerimientos planteados en la estrategia de navegación del cual ya se ha hablado.

Para el recorrido del robot, se planteó la identificación de tres objetos, los cuales son un bote y dos cajas. En la posición inicial colocamos el robot en la posición (0,0,0) donde se muestra en la Figura 10, y en una primera instancia al objeto 1, el cual es un bote que se puede distinguir con vasta claridad con la nube de puntos, además de proporcionarnos la distancia en la que se encuentra en relación con el objeto.

De acuerdo con la estrategia desarrollada, se inicia acercando hacia el lado izquierdo, ya que se encuentra colocado el sensor ultrasónico de ese lado, manteniendo una distancia con la pared de 10 cm, y activando ambas ruedas para avanzar hacia adelante, realizando solamente las correcciones necesarias para mantener la distancia fijada. A continuación, se realiza un giro antihorario en la rueda derecha para dar un giro hacia la derecha, que es como se muestra en la Figura 9, y así continuar su desplazamiento hacia adelante. Como se observa en la Figura 11 el robot avanza generando una línea de seguimiento, la cual muestra con precisión el andar del robot, además se muestra como la cámara RGBD genera la nube de puntos identificando al objeto 2.

Para la continuación del recorrido, se tiene que pasar por en medio de dos objetos (objeto 2 y 3), las cuales son unas cajas. Por las condiciones del recorrido, se permite

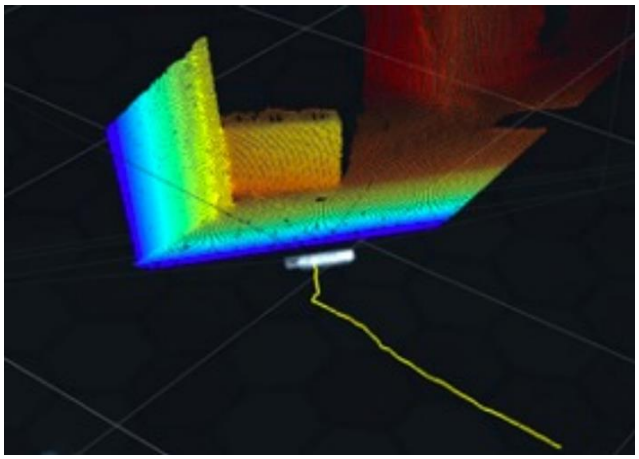


Fig. 11. Desplazamiento del robot, mostrando su trayectoria y un segundo objeto enfrente del robot móvil.

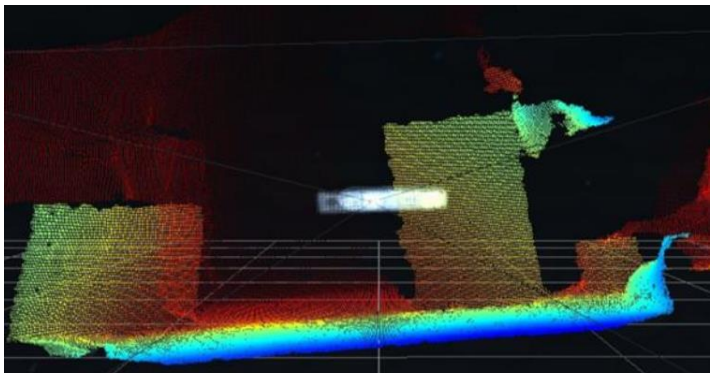


Fig. 12. Después de realizar el giro hacia la derecha, se pueden observar los objetos 2 y 3.

que el robot pueda pasar con titubeos entre los dos objetos, ya que tiende a alejarse del objeto 2, acercándose demasiado al objeto 3.

Después de cruzar por los objetos 2 y 3, los cuales se observan en la figura 12, el recorrido prosigue con ajustes menores hacia delante, y regresando a la tendencia de mantener la distancia con la pared. En la instancia final del recorrido, se encuentra a una persona recostada en el suelo, simulando que se encuentra inconsciente. Se puede verificar que con la nube de puntos y la línea de trayectoria es posible identificar objetos con vasta claridad, además de conocer su posición como se observa en la Figura 13.

7. Conclusiones y trabajo a futuro

El robot móvil terrestre que se desarrolló a través de la implementación de un sensor ultrasónico y cámaras (cámara tracking y cámara RGBD), que hacen posible conocer

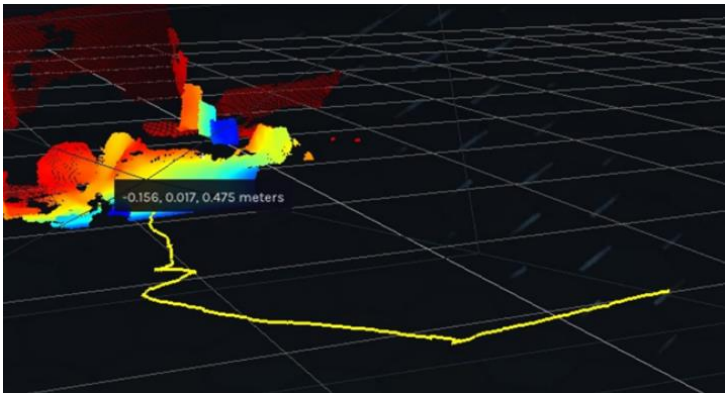


Fig. 13. Desplazamiento del robot, mostrando su trayectoria.

el entorno de forma gráfica, y poder conocer una ruta libre de obstáculos para desplazarse. La cámara RGBD demostró que es capaz de crear un mapa en 3D mediante la nube de puntos, y con la cámara tracking se puede seguir la línea de desplazamiento, siendo posible conocer la localización exacta de algún objeto de interés, además de arrojar una imagen con objetos de fácil reconocimiento para el ser humano.

Un reto constante en la realización del robot móvil es conocer la ubicación del robot, y tener la certeza de que su desplazamiento es realizado conforme al plan, es por eso por lo que con la implementación de la fusión de sensores a través del filtro Kalman, el robot tiene una mejor captación de la información que es obtenida del entorno donde se encuentra desplazándose, permitiendo un mejor entendimiento de los datos. El robot tiene un desplazamiento mucho más fluido, ya que las señales se encuentran en constante recursividad, haciendo que sus mediciones sean procesadas en tiempo real, teniendo la capacidad de hacer las correcciones necesarias en el momento, y prediciendo un posible movimiento con los datos obtenidos.

En cuanto a la programación, la tarjeta Jetson Nano solo soporta la versión de 18.04 de Ubuntu, reduciendo la utilización de librerías que se han desarrollado anteriormente, siendo así retrasado el desarrollo del robot móvil, cuando el tiempo con el que se cuenta es muy poco. Las diferentes herramientas que se han desarrollado para versiones anteriores no funcionan para esta versión del sistema operativo de Linux, haciendo que muchas de las cosas ya existentes se deban de comenzar desde cero.

Una mejora a futuro es el de otorgar al robot la posibilidad de tomar alguna decisión con base a los datos que está recabando (Inteligencia artificial o machine learning), para mejorar la navegación del robot, ya que, con la utilización del sensor ultrasónico, el desplazamiento se ve entorpecido por los múltiples ajustes que se deben de realizar para evitar los obstáculos que se encuentran en el entorno. Siendo clave algún tipo de entrenamiento para el robot para mejorar su navegación.

También se podría tomar la elección de implementar un Lidar (Light Detection and Ranging), para complementar la creación de un mapa en 3D, sin la necesidad de recorrer todo el entorno para conocer los por menores del lugar. Ya que, con el solo uso de las cámaras, si se desea conocer un entorno grande, tomaría mucho tiempo el

recorrido del lugar, ya que el robot tendría que pasar por todos los rincones del lugar, haciendo ineficiente la captura del mapa.

Referencias

1. Brady, M., Paul, R.: Robotics research. In: First International Symposium (1984)
2. Dempsey, C.: Robotics in cartography. Gislounge (2008)
3. Schmidt, P., Scaife, J., Harville, M., Liman, S., Ahmed, A.: Intel® RealSense™ Tracking Camera T265 and Intel® RealSense™ Depth Camera D435 - Tracking and Depth. Real Sense (2019)
4. Braunl, T.: EyeBot: A family of autonomous mobile robots. In: 6th International Conference on Neural Information Processing. Proceedings (ICONIP'99, ANZIIS'99, ANNES'99 & ACNN'99), 2, pp. 645−649 (1999)
5. Lentin, J.: ROS robotics projects. Packt Publishing Ltd (2017)
6. Intel: Intel® RealSense™ tracking camera T265 and depth cameras D400 series – Better together (2020)
7. Mujica, D.: Navegación autónoma de un robot móvil dentro de entornos real/virtual. Tesis Instituto Politécnico Nacional (2008)
8. MathWorks: Fusión de datos de sensores para sistemas autónomos (2019)
9. Marković, I., Petrović, I.: Bayesian sensor fusion methods for dynamic object tracking. A Comparative Study. Automatika Journal for Control, Measurement, Electronics, Computing and Communications, 55(4), pp. 386−398 (2017)
10. Escamilla-Ambrosio, P.J., Mort, N.: A hybrid Kalman filter-fuzzy logic architecture for multisensor data fusion. In: Proceeding of the IEEE International Symposium on Intelligent Control ISIC'01, pp. 364−369 (2001)

Análisis geoespacial del COVID-19 en Ciudad de México y Estado de México

Catherine Montiel Porcayo, Carlos Alonso Medina Cortes,
Ana María Magdalena Saldaña Pérez

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
México

{a190460, a190757}@sagitario.cic.ipn.mx, amsaldanap@ipn.mx

Resumen. El virus coronavirus (SARS-CoV-2), responsable de la enfermedad COVID-19 es hasta el día de hoy, la causante de una pandemia no vista desde la causada por la gripe española en 1918. Desde su aparición a finales de diciembre de 2019 en Wuhan China, se han reportado aproximadamente 22 675 000 casos en todo el mundo [1] y 543,806 casos en México [2]. Debido a la gravedad de esta enfermedad, se han llevado a cabo grandes esfuerzos desde diferentes frentes como el médico, el gubernamental, el social, y también el científico. Gracias al avance de la tecnología y a las herramientas computacionales con las que se cuenta, además de los análisis de casos de personas infectadas por el COVID-19, se pueden hacer aproximaciones del esparcimiento de contagios, conocer la ocupación hospitalaria, y generar estrategias para controlar el fenómeno. El área geoespacial no es la excepción y cobra amplia relevancia al proporcionar técnicas que permiten modelar el esparcimiento de los casos de personas infectadas en un área geográfica determinada, hacer predicciones del comportamiento de contagios, y conocer las relaciones geográficas que coadyuvan al esparcimiento del virus. En este trabajo se analiza desde una perspectiva geoespacial cómo es que ha evolucionado la epidemia en México. Aplicando técnicas y modelado geoespacial se realiza un análisis desde que surgió el primer caso en territorio mexicano hasta la tercera semana de agosto del 2020, las áreas geográficas son la ciudad de México y del estado de México, las dos urbes con más población por metro cuadrado en el territorio mexicano. En el artículo se presentan mapas generados por técnicas de geoprocesamiento, con datos obtenidos de fuentes oficiales del gobierno mexicano.

Palabras clave: COVID-19, coronavirus, pandemia, geoprocesamiento, análisis geoespacial.

Geospatial Analysis of COVID-19 in Mexico City and the State of Mexico

Abstract. Until today, the coronavirus virus (SARS-CoV-2), responsible for the COVID-19 disease, is the cause of a pandemic not seen since the one caused by

the Spanish flu in 1918. Since its appearance at the end of December 2019 in Wuhan China, approximately 22,675,000 cases have been reported worldwide [1] and 543,806 cases in Mexico [2]. Due to the seriousness of this disease, great efforts have been made from different fronts such as the medical, the government, the social, and also the scientific. Thanks to the advancement of technology and the computational tools that are available, in addition to the analysis of cases of people infected by COVID-19, it is possible to make approximations of the spread of infections, know the hospital occupation, and generate strategies to control the phenomenon. The geospatial area is no exception and is widely relevant by providing techniques that allow modeling the spread of cases of infected people in a given geographical area, making predictions of the behavior of contagions, and knowing the geographical relationships that contribute to the spread of the virus. This paper analyzes from a geospatial perspective how the epidemic has evolved in Mexico. Applying techniques and geospatial modeling, an analysis is carried out from the moment the first case arose in Mexican territory until the third week of August 2020, the geographical areas are Mexico City and the state of Mexico, the two cities with the most population per square meter. in the Mexican territory. The article presents maps generated by geoprocessing techniques, with data obtained from official sources of the Mexican government.

Keywords: COVID-19, coronavirus, pandemic, geoprocessing, geospatial analysis.

1. Introducción

En México el COVID-19 impactó drásticamente a la población, tomando la vida de miles de personas y cambiando la vida de toda la nación. A pesar de los esfuerzos del gobierno para reducir la tasa de contagios, en algunas localidades sobrepobladas fue imposible reducir el número de personas infectadas diariamente, muchas de ellas con comorbilidades relacionadas a la obesidad, problemas respiratorios, entre otros.

Otro aspecto importante que propició el avance del virus en la sociedad mexicana es la falta de compromiso para evitar el número de contagios, si bien las autoridades iniciaron varias campañas para promover las medidas de prevención como es el lavado de manos, el estornudo de etiqueta, evitar los saludos de mano, mantener la distancia entre personas y en lo posible trabajar desde casa no en todos los casos se siguieron adecuadamente [3], estas actividades requieren un cambio en el comportamiento de las personas y por lo tanto, un ritmo de aprendizaje para poder tomar las medidas de prevención con la seriedad necesaria.

En este trabajo se busca analizar el avance del virus a través de las dos grandes urbes de México, la Ciudad de México y del Estado de México, ambas presentan altos índices de contagios y defunciones, otro objetivo del trabajo es encontrar las zonas de dichas entidades federativas con los números de contagios más altos.

Ya que el avance del virus no es instantáneo es necesario realizar un estudio a lo largo del tiempo del comportamiento de su propagación, para lo cual se toma una ventana de tiempo que comprende desde la segunda semana de mayo hasta la tercera semana de agosto de 2020.

Dentro de los hallazgos que se pueden percibir en la presente investigación es el comportamiento espacial que presentan los datos, mostrado en visualizaciones espaciales; algunos de los análisis más importantes son la relación entre el área de las zonas de mayor concentración de casos con respecto al número de infectados, observándose que las dos grandes urbes fueron seriamente afectadas, al igual que sus periferias; es decir los focos de infección se concentraron en las ciudades y posteriormente el virus se esparcía hacia los municipios vecinos.

Este comportamiento se observa de manera puntual en la Ciudad de México y en sus municipios envolventes, llamados comúnmente zona metropolitana; de igual forma el fenómeno es fácilmente observable en la relación de contagios ocurridos en la ciudad de Toluca y sus municipios vecinos.

2. Estado del Arte

De acuerdo con la Organización Mundial de la Salud (OMS), la primera aparición del COVID-19 fue en Wuhan, China, a finales del 2019 y fue hasta marzo 11 del 2020 que OMS declaró pandemia a la enfermedad del COVID-19 [4]. Después de este evento varios países europeos reportaron un incremento en el número de casos a través del continente, resaltando a Italia, Francia y España. En México el primer caso de COVID-19 se reportó el 27 de febrero de 2020 por el titular de la Subsecretaría de Prevención y Promoción de la Salud, Hugo López-Gatell Ramírez [5]. El 19 de marzo, los 32 estados de la república mexicana reportaban casos positivos de COVID-19. Hasta el 23 agosto se reportaron 595,144 casos positivos de COVID-19, así como 62,241 defunciones a través de todo el territorio mexicano.

El análisis de una pandemia requiere que varias ramas de las ciencias trabajen de manera interdisciplinaria para estudiar y resolver la problemática, dentro de estas la geografía juega un papel clave, ya que se relacionan las actividades humanas con las biofísicas [6], esto nos permite estudiar cómo es que un virus se transporta a través de un espacio geográfico y el proceso que implica [7]. Si bien la geografía nos ayuda a conocer el desarrollo del virus, hay otras ramas como las ciencias de la salud que buscan encontrar métodos para sesgar el crecimiento del virus realizando análisis y pruebas clínicas con fármacos conocidos [8], así como soporte para realizar estudios en otras ramas como es la inteligencia artificial, rama en la cual se han realizado estudios para conocer el diagnóstico de pacientes para así de manera no invasiva saber si una persona es diagnosticada o no con COVID-19 [9, 10].

Los Sistemas de Información Geográfica (GIS) han resultado ser una herramienta fundamental en el análisis de la distribución de enfermedades a través del mundo, para así determinar su alcance, así como para la toma de medidas de prevención sanitarias para reducir el nivel de mortalidad. Desde el surgimiento del virus se han desarrollado varias aplicaciones utilizando GIS como son los paneles geográficos del Centro de Sistemas e Ingeniería de la universidad Johns Hopkins (JHU CSSE) [2], o el de la Organización Mundial de la Salud [11], para así tener una representación de los países con un mayor número de casos positivos por COVID-19, el número de defunciones o el número de personas recuperadas de la enfermedad.

La densidad de población de cada ciudad es particular, en un estudio por Gibson y Rush. los autores realizando un análisis geoespacial que consistía en la comparación

entre dos poblados usando un buffer de dos metros para simular el distanciamiento social, y con el uso del algoritmo de aprendizaje automático k vecinos cercanos (knn, k nearest neighbours) pudieron determinar qué tan cercanas están las personas una de las otras cuando ocurre un contagio, apoyando ideas de prevención de contagio como el uso de la sana distancia. En la ciudad de Masiphumelele, en Cabo Sudáfrica fue posible llevar a cabo las prácticas de distanciamiento social debido a que hay una menor densidad de población, sin embargo, en la ciudad de Klipfontein Glebe fue difícil mantener 2 metros entre persona y persona [12].

3. Metodología

En este apartado se describe la metodología por medio de la cual fue posible procesar, modelar, y analizar el comportamiento que el virus COVID-19 ha tenido sobre varias ciudades de México; en el presente artículo el área de estudio se limita específicamente a la Ciudad de México y al Estado de México. La metodología aplicada consta de cuatro etapas, a través de las cuales se lleva a cabo la recuperación y procesamiento de los datos, para su posterior análisis geoespacial.

La metodología propuesta cuenta con 2 características principales:

- Hace uso de datos provenientes de un repositorio de datos abiertos proporcionados por el gobierno mexicano, para obtener información sobre el esparcimiento del virus COVID-19 en la ciudad de México y en el Estado de México.

- Modela geoespacialmente los datos consultados y los resultados obtenidos del análisis del esparcimiento del virus en el área de interés.

Etapa 1.-Extracción de los datos: en esta etapa se obtuvieron los datos provenientes del programa de datos abiertos del gobierno de México, este corpus era actualizado todos los días por el mismo gobierno reportando todas las pruebas negativas y positivas que se realizan a pacientes que presentaban síntomas del virus COVID-19.

Etapa 2.- Tratamiento de datos: los datos son procesados para limpiarlos, filtrarlos por áreas deseadas (Estado de México y Ciudad de México), y obtener los atributos de interés, que para la presente investigación son los pacientes que han salido positivos a la prueba en cada división política de las áreas de interés.

Etapa 3.- Modelado espacial, en el que se analizan los datos obtenidos y se mapean en el área de estudio. Se aplican técnicas geoespaciales para permitir un mayor análisis de los datos procesados, y obtener el modelado geoespacial adecuado.

Etapa 4.- Análisis de las variables generadas a partir de los datos previamente procesados de manera geoespacial, para determinar patrones y características en el comportamiento que tiene el virus sobre el área de estudio.

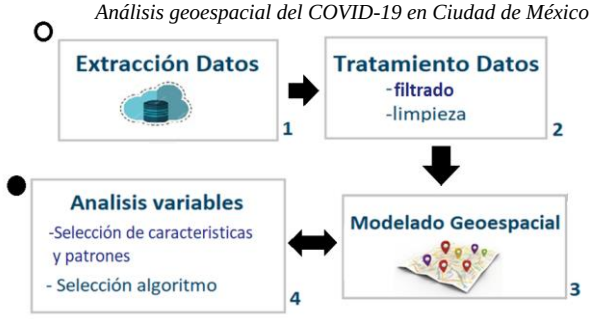


Fig. 1. Diagrama de actividades que muestra la secuencia de las etapas.

Tabla 1. Distribución de las variables a utilizar.

ETIDAD_UM	MUNICIPIO_RES	FECHA_DEF	RESULTADO
1	1	dd/mm/aaaa (fallecido)	1 (Positivo)
2	2	9999-99-99 (vivo)	2 (No Positivo)
3	3		3 (Pendiente)
…	…		
31	570		
32	999 (No especifícado)		

La metodología planteada permite llevar a cabo el análisis espacio temporal del comportamiento del COVID-19 al procesar y modelar los datos en periodos de tiempo específicos planteados con base en la variación significativa del número de infectados. En la Figura 1, se muestra un diagrama de actividades de la secuencia de estas etapas, cabe destacar que la etapa 3 y 4 se retroalimentan una de la otra ya que algunos de los patrones y tendencias que muestra el virus en su esparcimiento se pueden visualizar de una mejor manera con un modelo geoespacial reiterativo.

La implementación de la metodología se llevó a cabo de la siguiente manera, primero se realizó una extracción continua de los datos de COVID-19 semanalmente, ya que en el trabajo realizado por el laboratorio en el transcurso de tres meses se mantuvo la actualización semanal de los mismos.

Como segundo paso e indagando en la etapa dos, se realizó un filtrado de los datos ya que se trabajaron tres áreas geográficas específicas, la República Mexicana dividida por estados, la Ciudad de México dividida por alcaldías, y el Estado de México dividido por municipios; posterior a esto, se realizó una limpieza de los datos ya que solo se trabajó con el número positivos de infectados por cada una de estas zonas y también con el número de fallecidos.

Para llegar a este filtrado se trabajó con las siguientes características dentro del corpus de datos abiertos de COVID-19 proveniente del gobierno: ENTIDAD_UM, que representa el código numérico de cada estado de la república, cabe destacar que cada tupla dentro del corpus representa una prueba de covid aplicada a un paciente con síntomas sospechosos de COVID-19, durante el desarrollo del proyecto se observó que el número de tuplas iba en aumento con el paso de los días empezando

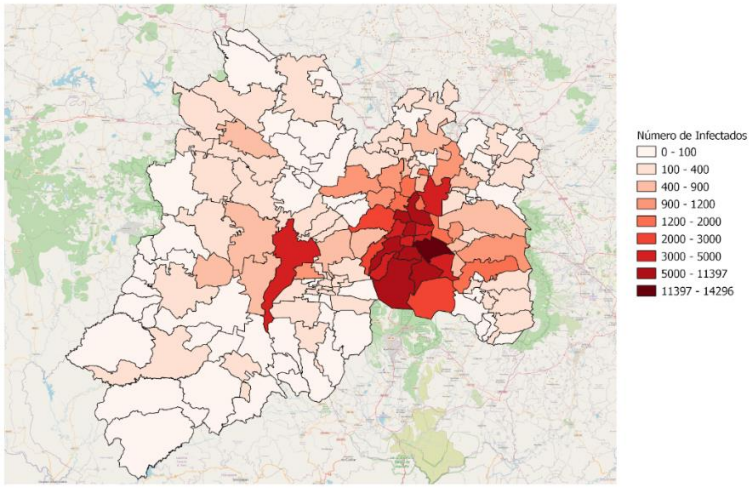


Fig. 2. Mapa de la ciudad de México y el Estado de México que muestra el número de infectados con fecha hasta el 20 agosto 2020.

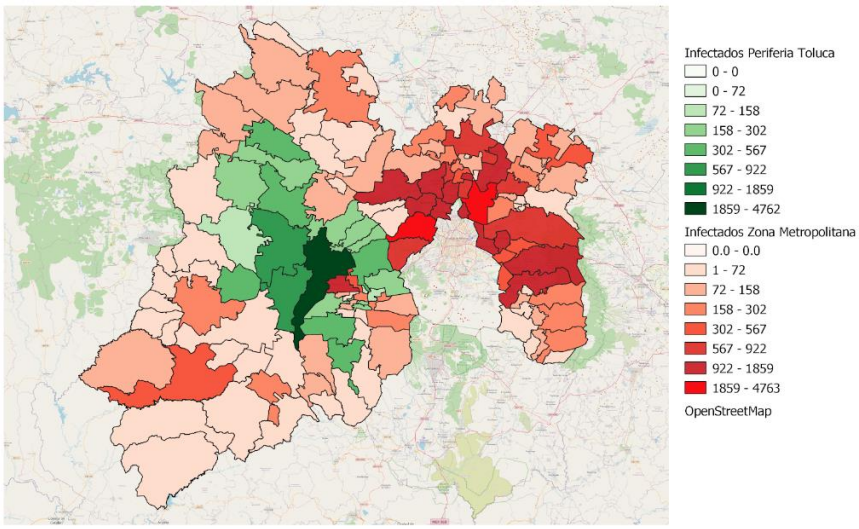


Fig. 3. Mapa del Estado de México que muestra el número de infectados por municipio hasta el 20 agosto de 2020.

con un corpus inicial de 110,995 tuplas con fecha del 6 Mayo del 2020, hasta un total de tuplas de 1,048,576 hasta el 20 de Agosto 2020.

Otra de las características a utilizar fue MUNICIPIO_RES que representa el código de municipio por cada estado de la república o el código de alcaldía para el caso de la ciudad de México, FECHA_DEF fue otra característica utilizada ya que esta nos indicaba si el paciente había muerto o no, y por último RESULTADO en donde sólo contabilizábamos los casos positivos, la siguiente tabla nos muestra las características

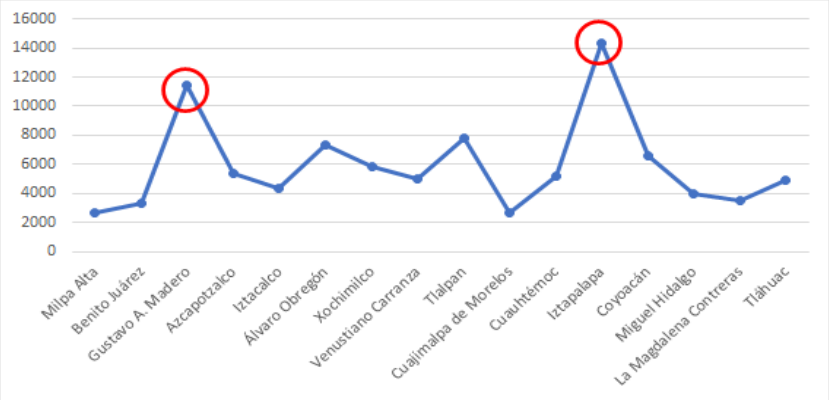


Fig. 4. Gráfica que muestra el número de pacientes positivos en la Ciudad de México en cada una de sus alcaldías.

de las variables que se utilizaron y qué valores podrían tomar cada una, los valores fueron procesados para ser representados como numéricos.

Posterior a la selección de características se realizó el conteo de datos, utilizando herramientas QGIS, consultas con PGADMIN (gestor de base de datos espacial para manejo de datos geoespaciales), y código en python.

En la tercera etapa se procedió a realizar el modelado Geoespacial en QGIS con periodicidad semanal, del número de infectados y el número de muertos en las tres zonas señaladas anteriormente.

Para el desarrollo del presente artículo, se decidió poner mayor interés en el Estado de México y en la Ciudad de México. La Figura 2 muestra un mapa de la ciudad de México y el Estado de México tomando como referencia el número de infectados con fecha última de 20 Agosto 2020, en el mapa se puede notar un principal foco de interés sobre la ciudad de México en donde algunas de sus alcaldías llegan hasta un pico arriba de 14,000 infectados, en contraste con los demás municipios del Estado de México donde los puntos más altos llegan aproximadamente a 4,700 infectados, uno de estos ejemplos es la ciudad de Toluca, esto se puede visualizar de manera rápida en el mapa de la Figura 2, basándonos en estas evidencias se decidió indagar en la hipótesis inicial, la expansión del virus tomando focos de infección para su posterior esparcimiento por sus municipios vecinos.

Al notar el mapa anterior y darnos cuenta que la diferencia de infectados en la Ciudad de México y el Estado de México era considerable decidimos crear un mapa aparte con coropletas guiándonos solo con los datos en el estado de México, con lo que llegamos al mapa de la Figura 3, de esta manera pudimos apreciar de manera espacial los principales focos de infección llegando a dos grupos y para fines explicativos decidimos mostrar de color verde para la zona de Toluca y roja para la zona metropolitana, en el mapa es claro apreciar estas zonas de contagio por su color más profundo.

Tomando en cuenta esta distinción de zonas infectadas y partiendo por los puntos de infección que en el mapa se muestran, las Ciudades de Toluca, Ecatepec y Naucalpan además de dos delegaciones en específico en la ciudad de México,

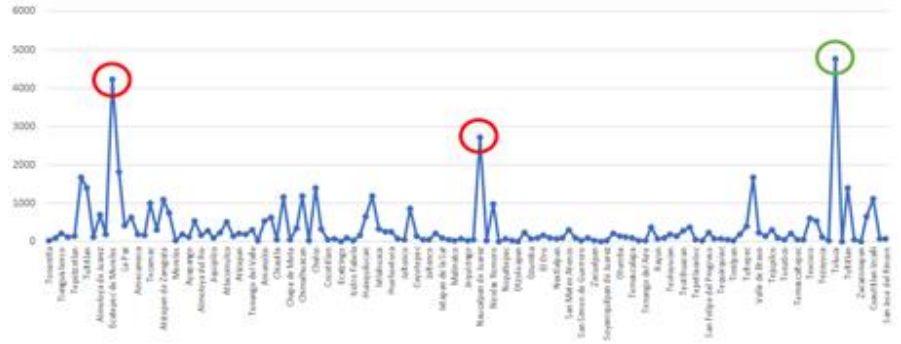


Fig. 5. Gráfica que muestra el número de pacientes positivos en el estado de México en cada uno de sus municipios.

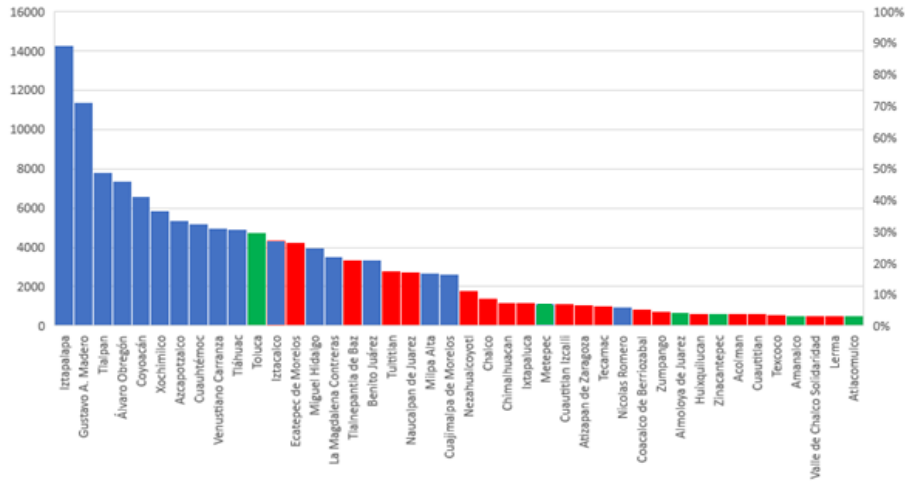


Fig. 6. Histograma que muestra el número de infectados ordenados de mayor a menor por municipio en el Estado de México y Ciudad de México.

Gustavo A Madero e Iztapalapa, decidimos corroborar los datos con gráficas e histogramas que nos mostraran los resultados esperados.

4. Discusiones

Analizando los datos de casos positivos en la Ciudad de México y el Estado de México se encontró que hay cierto patrón en el número de contagiados, agrupando el mayor número de contagiados en la alcaldía de Iztapalapa y Gustavo A. Madero en la Ciudad de México como se muestra en la figura 4, llegando a los 14,000 y 12,000 casos positivos respectivamente. Con lo que respecta al Estado de México, los

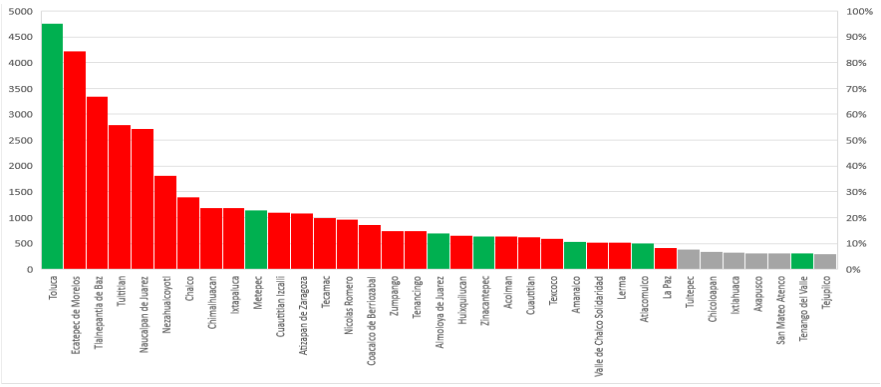


Fig. 7. Histograma que muestra el número de infectados ordenados de mayor a menor por municipio en el Estado de México, considerando solo los 35 municipios más afectados.

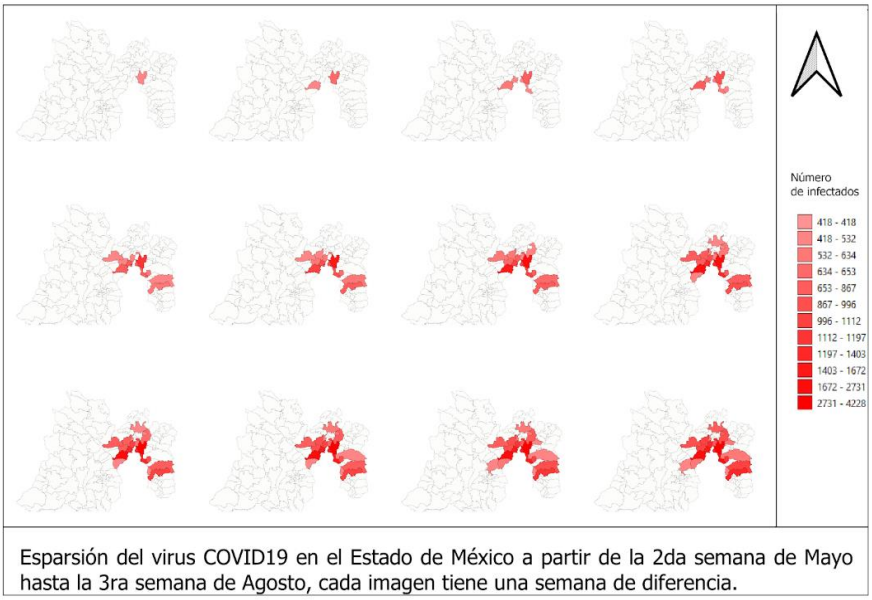


Fig. 8. Expansión del virus Covid 19 en el Estado de México con foco en la ciudad de México a partir de la 2da semana de mayo hasta 3ra semana de agosto, cada imagen tiene una semana de diferencia.

municipios de Toluca, Ecatepec y Naucalpan fueron los municipios con mayor número de casos positivos, como se muestra en la figura 5, alcanzando los 5,000,4,500 y 2,800 casos respectivamente. Las zonas analizadas representan los focos de infección y es a partir de estos que el número de casos positivos en las alcaldías y municipios colindantes fue aumentando de manera exponencial.

Se observa que los picos de infección se localizan en Iztapalapa y en la alcaldía Gustavo A. Madero, la gráfica muestra cómo las alcaldías próximas a las que

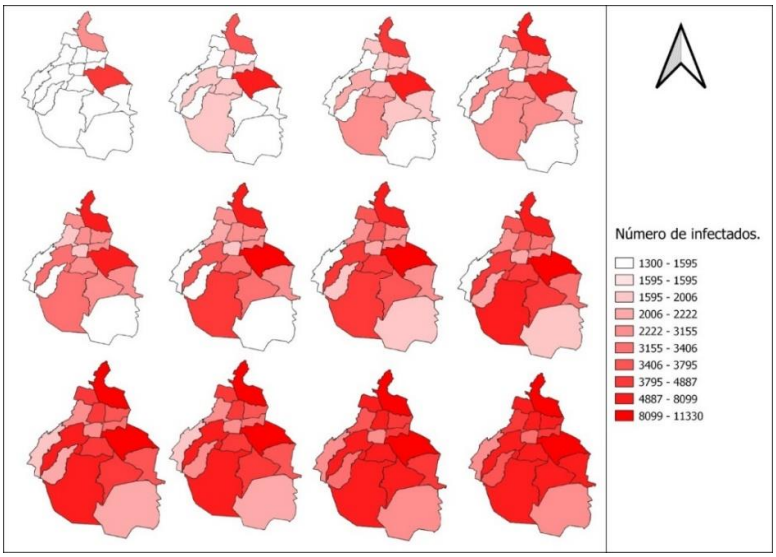


Fig. 9. Expansión del virus Covid 19 en la Ciudad de México a partir de la segunda semana de mayo hasta 3ra semana de agosto, cada imagen tiene una semana de diferencia.

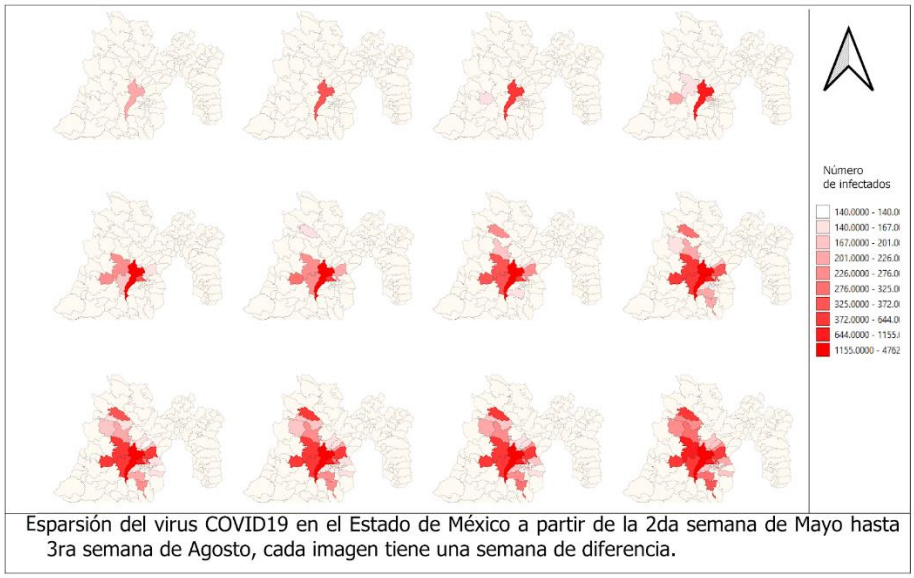


Fig. 10. Expansión del virus COVID-19 en el Estado de México con foco en la ciudad de Toluca a partir de la segunda semana de mayo hasta la tercera semana de agosto, cada imagen tiene una semana de diferencia.

presentan los mayores índices, también poseen un marcador elevado, a diferencia de las que se ubican más retiradas de los focos geográficos de infección.

Tomando en cuenta las gráficas anteriores se esperaría que los siguientes puntos en magnitud fuesen los municipios circundantes a los principales focos de infección, la figura 6 muestra un histograma ordenado de mayor a menor por número de infectados en las alcaldías de la ciudad de México (color azul) y los municipios del estado de México (color verde para la zona de Toluca, y color rojo para la zona metropolitana).

Basándonos en los resultados de la gráfica es posible percatarse de que las alcaldías de la Ciudad de México ocupan los primeros 10 lugares por arriba de Toluca y Ecatepec, analizando y esperando que la hipótesis planteada se cumpla en esta zona se esperaría que los dos primeros puestos fueran ocupados por Iztapalapa y la Gustavo A. Madero y que las alcaldías circundantes fuesen los siguientes puestos, viendo los resultados de la gráfica nos damos cuenta que se cumple la hipótesis los municipios en las posiciones 3 hasta 10 son las alcaldías que circundan a los principales focos, en la Figura 8 se puede ver este tipo de expansión en el mapa de manera semanal durante 3 meses.

Por otra parte, debido a que como en el mapa de la Figura 2 las alcaldías de la ciudad de México eclipsan los resultados de las dos zonas de Toluca y Zona metropolitana decidimos presentar otro histograma, solo tomando en cuenta los datos del Estado de México esperando que la hipótesis se cumpla también en estas, la Figura 7 muestra un histograma ordenado de mayor a menor por número de infectados en el estado de México rojo para la zona Metropolitana, Verde para la zona de circundante a Toluca y gris para los municipios restantes que no se encuentran involucrados en las zonas de esparcimiento.

Al analizar el comportamiento de la gráfica nos damos cuenta que las tres primeras posiciones corresponden a los principales 3 puntos que se mostraban en la gráfica 2, las siguientes barras muestran el comportamiento del virus por los municipios vecinos ya que podemos ver que las siguientes barras desde la posición 4 hasta la posición 29 corresponden a todos los municipios vecinos, también podemos notar que el segundo foco de infección más fuerte después de la Ciudad de México se encontró dentro de la zona metropolitana ya que la gráfica está pintada en su mayoría de color rojo.

Como tercer foco localizamos Toluca, es importante hacer notar que la diferencia de casos positivos a COVID-19 entre los municipios vecinos de Toluca y esta ciudad es grande, ello puede ser atribuido a la cantidad de hospitales que se encuentran en la ciudad a comparación de sus municipios vecinos; el conteo se hizo a partir de la característica **ENTIDAD_UM** que se refiere a la entidad donde se encuentra el hospital o la unidad de medicina que asistió al paciente.

A lo largo de los tres meses que duró este estudio se puede concluir que los focos de infección jugaron un importante lugar en la tasa de contagios en México, esto se puede visualizar en los mapas de coropletas de las Figuras 8, 9 y 10, cada uno representando las tres zonas antes mencionadas, donde se puede ver cómo es que a partir de las zonas con mayor número de contagios el virus se empieza a esparcir entre las demás vecindades, cambiando el color de blanco a los diferentes tonos de rojo, mostrando así la severidad de la situación y la expansión del virus en el territorio nacional.

Estos mapas muestran claramente los resultados ya antes discutidos en las gráficas analizadas, de tal manera que el lector pueda ver a través del tiempo la expansión del virus.

5. Conclusiones

El virus proveniente de Wuhan, China ha avanzado miles de kilómetros a través de la tierra debido a distintos factores. En México, el Covid se ha logrado contener en ciertos epicentros de las urbes. Gracias a este trabajo podemos localizar geográficamente los territorios más afectados, así como las zonas donde se reportan un pequeño número de infectados. El número de infectados varía de acuerdo con qué tan cerca están los poblados de los focos de la enfermedad.

La investigación proveniente de datos gubernamentales resulta muy útil para modelar el ritmo de infección del virus en cualquiera de las regiones del territorio nacional a través del tiempo. Con esta investigación se pudo modelar la situación actual del impacto de la epidemia del COVID-19 en el centro del territorio nacional, en trabajos futuros podrían analizarse los datos en todo el territorio mexicano para conocer qué estados o regiones fueron las más afectadas por el virus.

Los mapas generados semanalmente pueden ser consultados en el tablero del Laboratorio de Procesamiento Inteligente de Información Geoespacial. https://blup.com.mx/piig/. Como continuación de este trabajo de investigación se realizará un sitio web con un tablero en tiempo real del avance del COVID-19 a lo largo del territorio mexicano.

Referencias

1. Gobierno de México: Comunicado técnico 20 de Agosto (2020)
2. University Johns Hopkins University: Mapa global Johns Hopkins (2020)
3. Acuña-Zegarra, M.A., Santana-Cibrian, M., Velasco-Hernandez, J.X.: Modeling behavioral change and COVID-19 containment in Mexico: A trade-off between lockdown and compliance. Mathematical Biosciences, 325, pp. 108370 (2020)
4. Organización Mundial de la Salud: COVID-19: Cronología de la actuación de la OMS (2020)
5. Wikipedia: Pandemia de COVID-19 en México (2020)
6. Turner, B.L.: Contested identities: Human-environment geography and disciplinary implications in a restructuring academy. In: Annals of the Association of American Geographers, 92(1), pp. 52−74 (2002)
7. Franch-Pardo, I., Napoletano, B.M., Rosete-Verges, F., Billa, L.: Spatial analysis and GIS in the study of COVID-19. Science of The Total Environment, 140033 (2020)
8. Tobaiqy, M., Qashqary, M., Al-Dahery, S., Mujallad, A., Hershan, A.A., Kamal, M.A., Helmi, N.: Therapeutic management of COVID-19 patients: A systematic review. Infection Prevention in Practice, 100061 (2020)
9. Laguarta, J., Hueto, F., Subirana, B.: COVID-19 artificial intelligence diagnosis using only cough recordings. In: IEEE Open Journal of Engineering in Medicine and Biology, pp. 275−281 (2020)
10. Mei, X., Lee, H.C., Diao, K.Y., Huang, M., Lin, B., Liu, C., Bernheim, A.: Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature Medicine, 26, pp. 1−5 (2020)
11. Organización Mundial de la Salud: Who coronavirus disease dashboard (2020)
12. Gibson, L., Rush, D.: Novel coronavirus in Cape Town informal settlements: feasibility of using informal dwelling outlines to identify high risk areas for COVID-19 transmission from a social distancing perspective. JMIR Public Health and Surveillance, 6(2), pp. 9 (2020)

Programación políglota con la máquina virtual Graal

José Antonio Romero Ventura, Ulises Juárez Martínez,
Lisbeth Rodríguez Mazahua, María Antonieta Abud Figueroa,
S. Gustavo Peláez Camarena

Tecnológico Nacional de México,
Instituto Tecnológico de Orizaba,
México

antonioromerov@gmail.com,
{ujuarezm, mabudf}@orizaba.tecnm.mx,
lrodriguez@itorizaba.edu.mx,
gpelaez@ito-depi.edu.mx

Resumen. Hoy en día, el desarrollo de aplicaciones hace uso de entornos separados con el propósito de que los lenguajes de programación usados sean capaces de ejecutarse en ambientes con características específicas, pero esto implica que sea difícil, tardado y costoso el desarrollo y mantenimiento de dichas aplicaciones. Como solución, la programación políglota con GraalVM permite el desarrollo de aplicaciones usando más de un lenguaje de programación a la vez, en un mismo entorno de desarrollo y permitiendo la comunicación entre diferentes lenguajes de programación. En este artículo se emplea la programación políglota en dos escenarios, uno para explorar las capacidades nativas de GraalVM y otro para explorar sus capacidades de interoperabilidad usando Node.js, para observar el potencial de la programación políglota y cómo facilita el desarrollo de aplicaciones al momento de combinar lenguajes de programación que de manera cotidiana se encuentran en entornos separados y se comunican por medio de API's.

Palabras clave: GraalVM, políglota, interoperabilidad, imagen nativa, Node.js.

Polyglot Programming with the Virtual Machine Graal

Abstract. Nowadays, application development makes use of separate environments so that the programming languages used are capable of running in environments with specific characteristics, but this implies that the development and maintenance of said applications is difficult, time-consuming and expensive. As a solution, polyglot programming with GraalVM allows the development of applications using more than one programming language at the same time, in the same development environment and allowing communication between different programming languages. This article uses polyglot programming in two scenarios, one to explore the native capabilities of GraalVM and the other to explore its interoperability capabilities using Node.js. To see the potential of polyglot programming and how it makes it easier to develop applications on the

fly. to combine programming languages that are found in separate environments on a daily basis and communicate through APIs.

Keywords: GraalVM, polyglot, interoperability, native image, Node.js.

1. Introducción

Hoy en día, el desarrollo de aplicaciones se lleva a cabo por medio de diferentes lenguajes de programación [1], ya sea en el desarrollo web [2], aplicaciones de escritorio, aplicaciones de dispositivos móviles o aplicaciones en la nube. Cada uno de los lenguajes requeridos necesita un entorno específico para ejecutarse, ya sea un navegador web, un servidor de aplicaciones, una máquina virtual o un conjunto de intérpretes, así como el uso de entornos separados por sus características de ejecución o compilación. En consecuencia, se tiene el uso de diferentes servidores, aplicaciones en equipos separados debido a ciertas características del programa y el uso de *API's (Application Programming Interface*, Interfaz de programación de aplicaciones*)* que permitan la comunicación entre programas para el envío de mensajes o ejecución de tareas, éstas *API's* en ocasiones son desarrolladas por terceros.

GraalVM [3] *(Graal Virtual Machine*, Máquina Virtual de Graal*)* cuenta con un amplio soporte de lenguajes de programación e interoperabilidad entre estos, así como la ejecución de lenguajes de programación a nivel nativo. *GraalVM* es compatible con sistemas operativos de *Linux, MacOS* y recientemente con *Windows*, aunque la versión de este último aún es versión beta. Estas características permiten ubicar a *GraalVM* como una plataforma universal multilenguaje de alto desempeño para el desarrollo aplicaciones compiladas.

En este artículo se presentan dos escenarios de programación políglota usando *GraalVM*. El primer escenario explora las capacidades nativas y demuestra el potencial de *GraalVM* con diferentes lenguajes de programación; el segundo escenario, con base en un caso de estudio, demuestra las capacidades de interoperabilidad de *GraalVM* usando *Node.js* [4] como *framework*, explorando las capacidades políglotas con diferentes lenguajes de programación, como lo es *Java, Ruby, Python* [5] *y C*. Se analiza y se observa cómo es el desempeño de *GraalVM* en cada uno de los escenarios y cómo es que se comunican los diferentes lenguajes de programación entre sí, lenguajes que comúnmente se encuentran en entornos separados.

El presente artículo está estructurado de la siguiente forma: la sección 2 comprende estado del arte; la sección 3 menciona qué es la programación políglota y presenta algunos ejemplos; la sección 4 habla del desarrollo y cómo funciona la máquina virtual Graal; la sección 5 habla acerca de un caso de estudio y cómo la programación políglota con *GraalVM* ayudó en su mejora; y por último en la sección 6 se mencionan las respectivas conclusiones.

2. Estado del arte

Esta sección presenta los trabajos relacionados a la programación políglota considerando su evolución y aplicación.

En [6], se menciona acerca de los adaptadores que se usan en cuanto al envío de datos entre diferentes lenguajes usando *GraalVM* como herramienta de ejecución de código. Sin embargo, también existen problemas en dichas ejecuciones, un problema común es el paso de datos entre lenguajes. Se implementó un prototipo de adaptadores políglotas usados en *Python 3* y se demostró cómo trabajan en combinación con el *shell* de *GraalVM*.

En [7] *Niephaus et al.* se han enfocado en la interoperabilidad de los lenguajes y el diseño y la implementación de ejecuciones políglotas rápidas. Para ello se utilizó *GraalSqueak*, una implementación de una máquina virtual de *Squeak/SmallTalk* para *GraalVM*. Como conclusión, se tiene que *GraalSqueak* tiene limitaciones en cuanto a la integración de lenguajes que serán investigadas más a futuro. El objetivo principal fue encontrar una manera apropiada de manejar *Squeak/SmallTalk* en el uso de interrupciones y en las grandes cantidades de objetos políglotas.

En [8], *Niephaus* et al. continuaron con la programación políglota, haciendo uso de *bytecodes* con *Truffle*. *Truffle* es un *framework* de implementación de lenguajes, que está diseñado para crear *AST* [9] *(Abstrac Syntax Tree,* Árbol de sintaxis abstracta*)* como intérpretes, el proceso para llevar a cabo la implementación está muy bien documentado. Los *AST* serán los encargados de generar los intérpretes de *bytecodes*, sin embargo, el implementar *bytecodes* en *Truffle* no es algo intuitivo, por lo cual, se requiere la creación de nodos *AST*. Se implementaron todos los *bytecodes* y primitivas necesarias para ejecutar *tinyBenchmarks* de *Squeak* y los resultados de *OpenSmallTalkVM* (*OpenSmallTalk Virtual Machine,* Máquina Virtual de *OpenSmallTalk*) se tratan como base de una máquina virtual para *Squeak/SmallTalk*.

Würthinger et al. [10] describe la construcción de una nueva máquina virtual que aminora el esfuerzo inicial al momento de usar nuevos lenguajes de programación. Comúnmente, las implementaciones de estos nuevos lenguajes se crean en lenguaje *C* o *C++*, que las hace poco seguras, con un grado de complejidad alto, y trabajan comúnmente con interfaces de tipo *bytecode*. El compilador explora la estructura del intérprete y realiza una evaluación parcial del mismo cuando el código se genera. *Würthinger et al.* [10] obtuvieron un alto rendimiento de la combinación de las siguientes técnicas: reescritura de nodos usando *AST*, y des-optimización desde el código máquina de vuelta a los intérpretes *AST*.

Šipek et al. [11] mencionan que la interoperabilidad entre lenguajes de programación puede provocar una baja considerable en el rendimiento del software. Uno de los proyectos que solucionan el problema mencionado y soporta una gran cantidad de lenguajes de programación, así como la interoperabilidad entre estos en la *JVM* (*Java Virtual Machine*, máquina virtual de Java), es el proyecto de Graal *OpenJDK (Open Java Development Kit,* Kit de herramientas de Java *Open)*, que evolucionó del proyecto *Maxine VM* [12] (*Virtual Machine,* máquina virtual).

Salim et al. [13] trabajaron con *WebAssembly* [14], que es un compilador de formato binario para lenguajes como *C/C++*, *Rust* y *Go*. Además, habilita la ejecución en navegadores web y programas de tipo *Standalone* (programa de carácter único, sin dependencias). Los módulos compilados interactúan con otros lenguajes de programación, como *JavaScript*. La compilación *WebAssembly* usa la infraestructura *LLVM* [15] *(Low Level Virtual Machine*, Máquina Virtual de Bajo Nivel*)*, para producir binarios de *WebAssembly* sin hacer uso de una *API* en específico.

Tabla 1. Tabla comparativa de aspectos principales con *GraalVM.*

Artículo	Aspectos			
	Multiplataforma	**Web**	**Múltiple soporte de lenguajes**	**Ambiente GUI**
Niephaus et al. [6]	X		X	
Niephaus et al. [7]	X		X	
Niephaus et al. [8]	X		X	
Würthinger et al. [10]	X		X	
Šipek et al. [11]	X		X	
Salim et al. [13]	X	X	X	
Niephaus et al. [16]	X	X		X
Programación políglota con la máquina virtual Graal	**X**	**X**	**X**	**X**

Niephaus et al. [16] evaluaron *PolyJus*, por medio de una demostración de un ambiente políglota y discutiendo las ventajas y desventajas que se encontraron. Desde que la comunidad científica hace uso de una gran cantidad de lenguajes de programación, especialmente en el análisis de datos y *machine learning* (aprendizaje automático), pueden seleccionar cualquier lenguaje de programación para su uso en los ambientes. Pero esta libertad es un poco limitada, desde que solo un lenguaje se usa por ambiente. Para resolver este problema, hicieron uso del proyecto *Jupyter*, los ambientes *Jupyter* evolucionaron de *IPython*. En un futuro planean crear más ambientes para analizar conjuntos de información más grandes.

En la Tabla 1, se presenta una comparación de características principales entre los artículos del estado del arte y el trabajo del presente artículo.

Como se observa en la Tabla 1, y considerando el estado actual que presenta *GraalVM*, es posible considerar todos los aspectos políglotas para el desarrollo de aplicaciones. Cabe mencionar que, aunque el entorno web ha sido naturalmente políglota, *GraalVM* lo supera al generar código compilado y nativo para cualquier plataforma y combinación de lenguajes utilizados.

3. Programación políglota

La programación políglota es el desarrollo de una solución de software usando más de un lenguaje de programación a la vez en un mismo ambiente de desarrollo. Por ejemplo, en el desarrollo de aplicaciones web, es necesario: el manejo de *HTML* [17] *(HyperText Markup Language*, Lenguaje de Marcas de HiperTexto*)* para la presentación de datos al usuario desde un navegador web; *SQL* [18] (*Structured Query Language*, Lenguaje Estructurado de Consulta) para la obtención y manejo de datos desde un sistema gestor de bases de datos; y por último, por lo menos uno o dos

```
1    const express = require('express');
2    const app = express();
3    app.listen(3000);
4    app.get('/', function(req,res){
5        var text = 'Hello World!';
6        const BigInteger = Java.type('java.math.BigInteger');
7        text += BigInteger.valueOf(2).pow(100).toString(16);
8        text += Polyglot.eval('R','runif(100)')[0];
9        res.send(text);
10   });
```

Fig. 1. Código políglota *Node.js* con *Java y R.*

```
13  try(Context context = Context.newBuilder()
14    .allowAllAccess(true).allowIO(true).build()){
```

Fig. 2. Creación de objeto context.

```
20  for(String arg : args){ texto += arg + " ";}
21      texto += "- JAVA";
```

Fig. 3. Almacenamiento de la cadena de texto en una variable de *Java*.

```
24  codigoRuby += "out_file = File.new('" + filename + "', 'w+') \n";
25  codigoRuby += "out_file.puts('"+ texto + " - RUBY') \n";
26  codigoRuby += "out_file.close \n";
27  context.eval("ruby", codigoRuby);
```

Fig. 4. Creación del archivo en *Ruby*.

lenguajes de *scripting* para el procesamiento de datos desde el servidor. Todos estos componentes combinados en un solo ambiente, sin el uso de software de terceros o alguna *API* intermedia que logre las comunicaciones entre estos, hace que la programación políglota se aplique en una misma solución de software. Desde el punto de vista metodológico, la programación políglota hace uso de las metodologías convencionales (pesadas y ágiles) de desarrollo de software debido a que los lenguajes que se utilizan son orientados a objetos, independientemente de la parte híbrida que actualmente se tiene con enfoques funcionales.

En los siguientes ejemplos se presenta cómo se realiza la programación políglota por medio de la plataforma *GraalVM*, usando diferentes lenguajes de programación y combinándolos en un solo código. En la Fig. 1 se observa un ejemplo de código políglota combinando *JavaScript* [19] potenciado por *Node.js* como lenguaje base, con *Java* [20] y *R* como lenguaje de uso interno o complementario. Como se observa en la Fig. 1, en la línea 5 se tiene el concatenado de un texto para mostrar directamente en *JavaScript*, en la línea 6 se observa el manejo de clases de Java, en este caso el manejo de "*java.math.BigInteger*" que crea un numero a partir de dicha clase, en la línea 8 se genera un número aleatoriamente usando lenguaje *R.* Todos estos resultados se concatenan en una cadena de texto que se mostrará en pantalla desde un navegador *web* por medio de un *request GET* de *Node.js* desde el servidor local en el puerto 3000.

A continuación, se presenta un ejemplo políglota de un programa con *Java* como lenguaje base (cabe mencionar que se pueden tener los demás lenguajes soportados por *GraalVM* como lenguajes base por igual) que obtiene una cadena de texto desde una terminal. Posteriormente hace la creación de un archivo escribiendo la cadena de texto introducida desde la terminal en lenguaje *Ruby*, después hace la lectura de dicho archivo

en lenguaje *Python* y por último hace la impresión en pantalla del contenido del archivo en lenguaje *C*. El código es más complejo ya que se tiene la combinación de cuatro lenguajes de programación en un mismo programa, debido a esto es un programa más extenso, por lo tanto, solo se mostrarán los bloques principales donde se presenta la combinación de lenguajes y la sección de comunicación entre estos.

En la Fig. 2 se observa la creación del objeto "*context*", este objeto permite la creación del ambiente políglota de *GraalVM* donde se ejecutan los diferentes lenguajes de programación. En la línea 14 se especifican los permisos generales a *GraalVM*, y el permiso a la creación, lectura y escritura de archivos dentro del equipo de cómputo. En la fig. 3 se realiza la obtención de la cadena de texto que se introdujo desde una terminal y se almacena en una variable de *Java*, en el concatenado de la línea 21 es solamente para mostrar el nombre del lenguaje que lo está ejecutando, solo para cuestiones de flujo de código y ver cómo la cadena de texto va pasando por los diferentes lenguajes de programación.

En la Fig. 4 se observa la creación del archivo en lenguaje *Ruby*, se genera una cadena de texto con el script a ejecutar de *Ruby* y se almacena en una variable en *Java*, después en la línea 27 se ejecuta con el objeto "*context*" usando la función "*eval()*", se especifica primero el lenguaje de programación y después el *script* o líneas de código a ejecutar.

En la Fig. 5 se realiza la ejecución de la lectura del archivo. Este bloque es muy parecido al código de la Fig. 4 ya que se concatena una cadena de texto con el código a ejecutar, en este caso es en lenguaje *Python.* Por último, en la línea 35 se realiza la ejecución de dicho código con el objeto "*context*".

Por último, en la Fig. 6 se realiza la impresión en pantalla de la cadena de texto introducida desde la terminal junto con los nombres de los lenguajes en los cuales se ejecutó. En *GraalVM*, el lenguaje *C* [21] también es compilado, por lo tanto se tienen acceso a este como un programa objeto y se ejecuta desde un objeto "*context*", posteriormente se almacena en un objeto de tipo "*Value*", como se observa en la línea 39. Con esto se tiene acceso a las funciones internas del programa generadas por el desarrollador, incluso se tiene acceso a funciones propias del lenguaje *C*. En la línea 40 se ejecuta "*printMensajeArray*", que es una función desarrollada en *C* que imprime en pantalla una cadena de texto que se obtiene como parámetro de función. La línea 42 hace la ejecución final de la impresión en pantalla desde *C*.

En el siguiente ejemplo se presenta lo que es la implementación del programa anterior, pero desde una imagen nativa, el código es el mismo, solo varia una sección que se explicará a continuación. En la fig. 7 se observa lo que es la declaración de las propiedades "*option*" en el objeto "*context*", estas opciones se declaran para permitir a la imagen nativa localizar las ubicaciones de los lenguajes a usar en ella, por ello se observan palabras como "*ruby*" o "*python*" en las líneas 15 a 20, que son los lenguajes de programación que se usaron en dicha prueba.

Al ejecutar la imagen nativa, esta se ejecuta como un objeto *bash* de *Linux*, en este ejemplo se pasa como parámetro de ejecución el valor "*-Dllvm.home*", este parámetro es necesario ya que en el programa de la imagen nativa se ejecuta código en lenguaje *C* usando *LLVM*, esta "*option*" se envía de esta manera debido a la versión de *GraalVM* que se está manejando, la versión 20.1.

Para efectos de rendimiento se hizo un conjunto de pruebas de desempeño de programas, se realizó la ejecución del programa de creación y lectura de archivos

```
30  codigoPython += "cadena = '' \n";
31  codigoPython += "for line in open('" + filename + "'): \n";
32  codigoPython += "       cadena = cadena + line \n\n";
33  codigoPython += "cadena.strip()+' - PYTHON'";
34  Value filecontent = context.eval("python", codigoPython);
35  String textodesdePython = filecontent.asString();
```

Fig. 5. Lectura del archivo en *Python*.

```
38  Source s = Source.newBuilder("llvm", new File("imprimeTexto.o")).build();
39  Value lib = context.eval(s);
40  Value printMensajeArray = lib.getMember("printMensajeArray");
41  Value msgarray = getvaluyeArrayCharASCII(textodesdePython,context);
42  printMensajeArray.executeVoid(msgarray); //Imprime en pantalla desde C
```

Fig. 6. Impresión en pantalla desde *C*.

```
13  try(Context context = Context.newBuilder().allowNativeAccess(true)
14  .allowHostAccess(HostAccess.ALL).allowAllAccess(true).allowIO(true)
15  .option("ruby.home","GRAALVM/jre/languages/ruby")
16  .option("python.SysPrefix","GRAALVM/jre/languages/python")
17  .option("python.CoreHome","GRAALVM/jre/languages/python/lib-graalpython")
18  .option("python.StdLibHome","GRAALVM/jre/languages/python/lib-python/3")
19  .option("python.Executable","GRAALVM/jre/languages/python/bin/graalpython")
20  .option("python.CAPI","GRAALVM/jre/languages/python/lib-graalpython")
```

Fig. 7. Declaración de valores "*option*" en context.

explicados anteriormente, con y sin imagen nativa; ambas versiones se ejecutaron un total de 100 veces en un equipo *Lenovo Legion Y530*, con procesador *Core i5 8300H* a 60 *Hz* y 16 *GB* de *RAM* en Ubuntu 18.04 64 *bits*. Los resultados fueron los siguientes:

Como se observa en la Tabla 2, los resultados oscilan de 5 a 7 incluso 9 segundos, dando un promedio de 5.840 segundos de tiempo de ejecución. En la Tabla 3 se observa lo que son los resultados en segundos del tiempo de ejecución del programa, pero desde una imagen nativa, por lo que se observa un tiempo que oscila de los 0.424 segundos a los 1.529 segundos, dando un promedio de todos los datos de 0.43997 segundos de tiempo de ejecución, por lo que se aprecia el gran potencial que tiene la imagen nativa sobre el programa de Java.

4. Componentes de la programación políglota

En la Fig. 8, se presenta un diagrama que muestra los componentes principales que interactúan con *GraalVM*, algunos componentes internos y los productos que se obtienen de procesar códigos políglotas. Como se observa en la Fig. 8, GraalVM funciona con los siguientes lenguajes de programación: *Java*, *JavaScript*, *Python*, *Ruby*, *R*, *C/C++*, *WebAssembly (Wasm)*. Dichos lenguajes son interpretados o compilados según sea su naturaleza por *GraalVM (*en el caso de *C* y *C++* que son lenguajes compilados). Sus componentes principales que llevan a cabo dicha tarea son: (1) los *AST*, estos permiten la interpretación de los diferentes lenguajes de programación por medio de árboles de sintaxis abstracta; (2) *Truffle* [22] es un *framework* de *GraalVM* que permite implementar estos lenguajes de programación con ayuda de los *AST*; (3) el compilador *JIT (Just In Time*, Justo en tiempo*)*, este componente permite la ejecución o interpretación de los programas políglotas en

Tabla 2. Resultados del tiempo de ejecución del programa sin imagen nativa.

5.699	5.563	5.791	5.565	5.899	5.681	5.666	5.766	5.616	5.573
5.897	5.700	5.593	5.770	5.717	5.843	5.770	5.762	5.767	5.802
5.766	5.876	5.761	5.554	5.971	5.670	5.943	5.834	5.671	5.752
5.749	5.962	5.605	5.651	5.607	7.302	5.734	5.774	5.870	5.733
9.375	5.589	5.855	5.790	5.641	5.624	6.064	5.845	5.967	5.709
5.762	5.623	5.830	5.811	5.697	5.902	5.735	5.745	6.376	5.880
7.522	5.743	5.665	5.722	5.879	5.616	5.872	5.607	6.665	5.631
7.029	5.634	5.675	5.753	5.771	5.940	5.745	5.966	5.952	5.715
5.803	5.649	5.748	5.676	5.773	5.745	5.683	5.684	5.696	5.574
5.674	5.671	5.977	5.618	5.581	5.770	5.560	5.704	5.726	5.585

Tabla 3. Resultados del tiempo de ejecución del programa con imagen nativa.

1.529	0.428	0.427	0.425	0.427	0.427	0.435	0.426	0.433	0.433
0.433	0.426	0.437	0.427	0.429	0.428	0.425	0.431	0.427	0.426
0.427	0.426	0.431	0.426	0.426	0.432	0.428	0.433	0.429	0.438
0.432	0.432	0.433	0.425	0.427	0.425	0.427	0.430	0.426	0.432
0.430	0.429	0.429	0.429	0.426	0.438	0.427	0.427	0.430	0.434
0.426	0.437	0.427	0.426	0.433	0.429	0.438	0.427	0.430	0.428
0.426	0.424	0.432	0.430	0.432	0.425	0.429	0.432	0.427	0.433
0.435	0.426	0.432	0.426	0.426	0.435	0.427	0.425	0.426	0.425
0.426	0.426	0.425	0.426	0.426	0.431	0.430	0.432	0.428	0.430
0.428	0.425	0.433	0.425	0.431	0.433	0.427	0.426	0.426	0.426

tiempo de ejecución permitiendo detectar los puntos de fuga en las excepciones o manejo de errores de sintaxis que le permiten al usuario corregir sus códigos o encontrar las excepciones de manera más rápida entre los enlaces de los lenguajes políglotas; por último, (4) el compilador *AOT (Ahead Of Time,* Antes de tiempo*)*, esta propiedad es principalmente aplicable a las imágenes nativas generadas por *GraalVM*, ya que gracias a ello la ejecución de estas son más rápidas, eficientes, y mantienen su interoperabilidad de lenguajes.

5. Caso de estudio

El caso de estudio consiste en una aplicación web de una empresa de desarrollo de software en la ciudad de Monterrey, Nuevo León, México, y debido a derechos de autor y temas de confidencialidad, solo se mostrarán diagramas representativos del sistema estudiado de la compañía.

En dicho sistema se tiene la separación de lenguajes de programación debido a que un lenguaje es para desarrollo propio del entorno web, como lo es *HTML, CSS* [23]

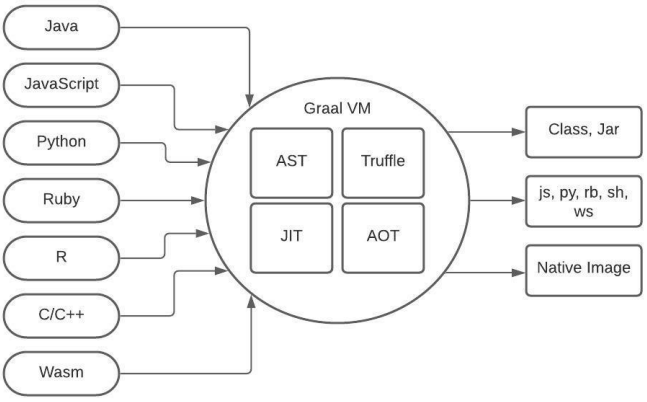


Fig. 8. Diagrama que muestra los componentes dentro de *GraalVM.*

Fig. 9. Diagrama de Arquitectura SOA en sistema del caso de estudio.

(*Cascading Style Sheets*, Hojas de Estilo en Cascada), *JavaScript*; los lenguajes de *scripting*, que en este caso es *ASP.NET (Active Server Pages .NET,* páginas activas del servidor .NET*)* y *C#* como lenguaje de programación; y el lenguaje *SQL*. Cada uno de estos ambientes se encuentran en entornos separados, aplicando procesos robustos de la ingeniería de software como lo es la separación de intereses para atender los requerimientos no funcionales.

Como se observa en la Fig. 9, se aplica la arquitectura *SOA (Service Oriented Architecture,* Arquitectura Orientada a Servicios*)*, esto permite que la aplicación se consulte desde canales *web* y dispositivos móviles. En dicho diagrama se observa de forma general la separación de intereses por medio de una capa de "Lógica de negocio" que realiza todas las operaciones y algoritmos referentes a las funciones de la compañía; en la capa de "Acceso a datos" es propiamente para el manejo de consultas a bases de datos para su posterior procesamiento en la capa de "Lógica de negocio"; la capa de "Cliente sistemas externos" se usa para la obtención de datos desde un sistema externo; se tiene como agentes externos lo que es la "base de datos" y el "sistema externo" que interactúan con los servicios por medio de *API* o software de terceros; por último la implementación es una solución de servicios SOAP/REST (*Representational State Transfer*, transferencia de estado representacional).

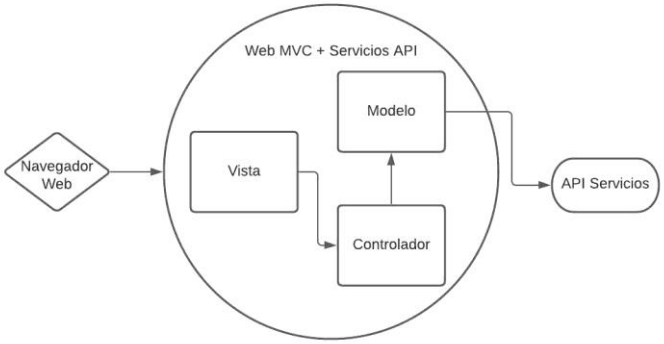


Fig. 10. Diagrama de aplicación del modelo MVC con consumo de una API de servicios.

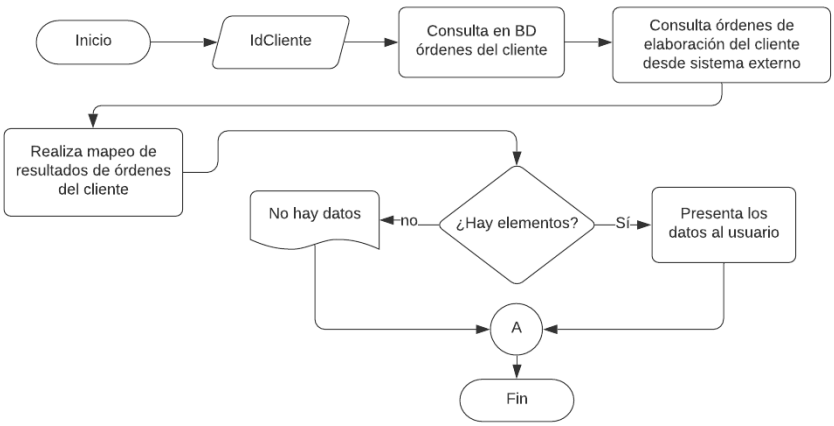


Fig. 11. Diagrama de flujo de programa políglota con base en el caso de estudio.

Tal y como se observa en la Fig.10, por medio del canal web, se maneja el modelo *MVC (Model View Controller,* Modelo Vista Controlador*)* junto con el consumo de una *API* pare el consumo de servicios explicado en la Fig. 9, también se tiene la vista que se presenta desde un navegador web al usuario, el controlador que realiza la administración de llamadas y obtención de datos desde diferentes acciones del usuario y el modelo que lleva a cabo el procesamiento de datos y obtención de los mismos desde el uso de una *API* de servicios. Como se aprecia, la programación políglota fortalece la separación en capas requeridas por el patrón arquitectónico MVC, al homogenizar en un solo estilo de programación el trabajo con varios lenguajes.

Cada capa representa a un conjunto de asuntos en específico, incluyendo los intereses como capacidad de mantenimiento, de documentación y en este caso de implementación homogénea. La programación políglota facilita el desarrollo de sistemas complejos al permitir un lenguaje base para interactuar con los otros lenguajes requeridos.

También se hizo un rediseño e implementación de cómo sería su funcionamiento si este se aplicara por medio de un entorno políglota con *GraalVM*, por ello, se realizó el

```
70  function getWorkOrdersByClientFromExternalSystem(idclient) {
71      return new Promise(resolve => {
72          var WorkOrderJava = Java.type('WorkOrder');
73          var list = WorkOrderJava.GetWorkOrdersByClientArrayObj(idclient);
74          var jsonordersfromextsys = JSON.stringify(list);
75          resolve(jsonordersfromextsys);
76  }); }
```

Fig. 12. Consulta ordenes de elaboración del cliente desde sistema externo.

```
80  function getOrdersOnProcessOrFinished(orders, workorders) {
81  return new Promise(resolve => {
82    resolve(Polyglot.eval('python', "set("+orders+") & set("+workorders+")"));
83  });  }
```

Fig. 13. Comparación de conjuntos de ordenes en *Python*.

```
20  app.get('/', function(req, res) {
21      var text = 'Welcome!<br> '
22      text += "Client = " + clientId + '<br>'
23      if (ordersonprocessorfinished.length <= 0) {
24          text += span("No orders in process".red) + " <br>"
25      } else {
26          text += span("Orders in process".green) + " <br><ul>"
27          for (var i = 0; i < ordersonprocessorfinished.length; i++) {
28              text += "<li> " + ordersonprocessorfinished[i] + " </li>"
29          }
30          text += "</ul>"
31      }
32      res.send(text)
33  })
```

Fig. 14. Presenta en pantalla resultados por medio de un *request GET* de *Node.js.*

diagrama de la Fig. 11. En la Fig. 11 se tiene el diagrama de flujo del funcionamiento del programa políglota: se introduce el identificador del cliente "IdCliente", con este valor se busca en la base de datos las órdenes o pedidos que ha realizado el cliente, después se consulta desde un sistema externo por medio de una clase en Java, se obtienen los datos de órdenes de elaboración del cliente, sin hacer uso de alguna API o software de terceros. Después se realiza un mapeo de dichas órdenes para ver si hay una intersección de ambos conjuntos de elementos donde el conjunto A serían las órdenes en la base de datos y el conjunto B serían las órdenes de elaboración desde el sistema externo obteniendo $C = A \cap B$ donde C es el conjunto de órdenes en proceso de elaboración pertenecientes al id del cliente. Después solo se analiza si el conjunto C tiene elementos, si no, presenta en pantalla un mensaje de "No hay datos", y si los hay, presenta la lista de las órdenes de elaboración.

El código se realizó en *JavaScript* con *Node.js* [24], y debido a que es un código largo se explicarán los bloques principales siguiendo la lógica del diagrama de la Fig. 11. En la Fig. 12 se observa el bloque de código que realiza la ejecución de un programa en *Java*, obteniendo las órdenes de elaboración desde un sistema externo. Anteriormente se consultaban las ordenes en proceso desde un sistema externo consumiendo una API/REST, sin embargo, ahora se usa la clase de Java "WorkOrder" desde Node.js, por medio de propiedad políglota "Java.type()" como se ve en la línea 72. Después se obtiene el listado de órdenes por el identificador del cliente ejecutando la función "GetWorkOrdersByClientArrayObj" en la línea 73, por último, se hace un

formateo a *JSON (JavaScript Object Notation*, Notación de Objetos de JavaScript*)* en la línea 74.

En la Fig. 13, el código realiza la comparativa de conjuntos de órdenes y órdenes de elaboración para obtener la intersección de ambos conjuntos y presentarlos al usuario, esta comparativa se realiza en lenguaje Python. Se ejecuta el elemento políglota "*Polyglot.eval()*" en la línea 82, donde se especifica como primer parámetro el lenguaje de programación a ejecutar seguido de la línea de código, en este caso se especifica el lenguaje *Python*.

En la fig. 14, se encuentra el código que representa la lógica del diagrama de la Fig. 11 donde hace la condición si el conjunto de órdenes de elaboración en proceso tiene o no elementos, si no los tiene presenta un mensaje en color rojo como se ve en la línea 26, y si los tiene presenta un mensaje en color verde como se ve en la línea 29, seguido de las órdenes en proceso.

El programa con Node.js funcionó de acuerdo con el funcionamiento esperado como se manejaba en el sistema de .NET. La gran diferencia es que se realizó una disminución de capas de desarrollo debido al diferente manejo de lenguajes y sistemas externos. Dicho cambio a código políglota permitió una ejecución más transparente y rápida. La parte políglota del programa permite de manera homogénea trabajar con menos capas, ya sea de manera compilada o como una imagen nativa. En general el caso de estudio presenta un considerable cambio de estructura del programa original debido a la disminución de capas.

6. Conclusiones

Con base en los resultados y comparaciones realizadas, la programación políglota con *GraalVM* muestra un amplio soporte de lenguajes e interoperabilidad entre los mismos, las capacidades nativas favorecen el desempeño de las aplicaciones al tener archivos ejecutables compilados y no se requiere de metodologías nuevas para soportar el desarrollo de aplicaciones políglotas. Desde el punto de vista arquitectónico, el modelo MVC, así como los estilos de programación, no se ven afectados por el contexto políglota, solo se favorece la implementación al usar cada parte de la solución en el lenguaje que mejores propiedades ofrece. También es remarcable mencionar la disminución de código observado en el caso de estudio, así como la simplificación de las capas requeridas en su arquitectura. En el mismo sentido, la comunicación entre componentes se mejora al tener de forma transparente un solo medio de programación.

Adicionalmemte, a través del framework Truffle, es posible incorporar nuevos lenguajes de programación que actualmente no se soportan en *GraalVM*, lo cual trae como consecuencia mejorar las capacidades de implementar sistemas que requieran necesidades muy específicas de diversas áreas, por ejemplo, la programación lógica para favorecer implementaciones de inteligencia artificial.

Como trabajo futuro se tiene considerado hacer pruebas con casos de mayor complejidad, evaluar de forma exhaustiva el desempeño nativo y no nativo de *GraalVM*, incorporar al menos un lenguaje no soportado actualmente en *GraalVM* para evaluar la facilidad de extensibilidad nata de la plataforma, y finalmente, se considera utilizar la naturaleza políglota de *GraalVM* para reimplementar aplicaciones que trabajan en área transversales como es el caso de la bioinformática.

Referencias

1. Juganaru-Mathieu, M.: Introducción a la programación. Grupo Editorial Patria (2014)
2. Mateu, C.: Desarrollo de aplicaciones web. Eureca Media, SL, pp. 39−43 (2004)
3. Oracle and/or its affiliates: GraalVM. https://graalvm.org/ (2020)
4. Nodejs: Acerca de Node.js. https://nodejs.org/es/about/ (2020)
5. Wiki.python: Beginner's Guide to Python. https://wiki.python.org/moin/ BeginnersGuide (2020)
6. Niephaus, F., Felgentreff, T., Hirschfeld, R.: Towards polyglot adapters for the GraalVM. In: Proceedings of the Conference Companion of the 3rd International Conference on Art, Science, and Engineering of Programming, pp. 1–3 (2019)
7. Niephaus, F., Felgentreff, T., Hirschfeld, R.: GraalSqueak: Toward a smalltalk-based tooling platform for polyglot programming. In: Proceedings of the 16th ACM (SIGPLAN) International Conference on Managed Programming Languages and Runtimes, pp. 14–26 (2019)
8. Niephaus, F., Felgentreff, T., Hirschfeld, R.: GraalSqueak: A fast smalltalk bytecode interpreter written in an AST interpreter framework. In: Proceedings of the 13th Workshop on Implementation, Compilation, Optimization of Object-Oriented Languages, Programs and Systems, pp. 30–35 (2018)
9. Eclipse: AST. https://eclipse.org/jdt/ui/astview/index.php (2020)
10. Würthinger, T., Wimmer, C., Wöß, A., Stadler, L., Duboscq, G., Humer, C., Richards, G., Simon, D., Wolczko, M.: Wolczko One VM to rule them all. In: Proceedings of the ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming & Software, pp. 187–204 (2013)
11. Šipek, M., Mihaljević, B., Radovan, A.: Exploring aspects of polyglot high-performance virtual machine graalVM. In: International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1671-1676 (2019)
12. Maxine-VM: Welcome to the Maxine VM project. https://maxine-vm.readthedocs.io/en/stable/ (2020)
13. Salim, S.S., Nisbet, A., Luján, M.: Towards a WebAssembly standalone runtime on GraalVM. In: Proceedings Companion of the ACM SIGPLAN International Conference on Systems, Programming, Languages, and Applications: Software for Humanity, Athens, Greece, pp. 15–16 (2019)
14. Webassembly.org: https://webassembly.org/ (2020)
15. Llvm.org: The LLVM Compiler Infrastructure. https://llvm.org/ (2020)
16. Niephaus, F., Krebs, E., Flach, C., Lincke, J., Hirschfeld, R.: PolyJuS: A Squeak/Smalltalk-based polyglot notebook system for the GraalVM. In: Proceedings of the Conference Companion of the 3rd International Conference on Art, Science, and Engineering of Programming, pp. 1–6 (2019)
17. Mozilla.org: HTML. https://developer.mozilla.org/es/docs/Web/HTML (2020)
18. Oppel, A., Sheldon, R.: Fundamentos de SQL. McGraw-Hill Interamericana Editores (2009)
19. Mozilla.org: JavaScript. https://developer.mozilla.org/es/docs/Web/JavaScript (2020)
20. Java: Java. https://java.com/en/download/faq/whatis_java.xml (2020)
21. Maas, A.J., Nazaré, H., Liblit, B.: Array length inference for C library bindings. In: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, pp. 461–471 (2016)
22. Github: Truffle. https://github.com/oracle/graal/tree/master/truffle (2020)
23. Mozilla.org: CSS. https://developer.mozilla.org/es/docs/Web/CSS (2020)
24. Sun, H., Bonetta, D., Humer, C., Binder, W.: Efficient dynamic analysis for Node.js. In: Proceedings of the 27th International Conference on Compiler Construction, pp. 196–206 (2018)

Identificador de movimientos mediante el análisis estadístico de la señal electromiográfica

Marco Antonio Franco-Rivas, Alfredo Ramírez-García

Universidad Autónoma de Aguascalientes,
Centro de Ciencias de la Ingeniería,
México

marcoafrancor@gmail.com, alfredo.ramirez@edu.uaa.mx

Resumen. Este artículo presenta el diseño de un sistema capaz de identificar movimientos realizados mediante el análisis de la señal de electromiografía superficial (sEMG) utilizando distintas métricas de similitud estadística. En el sistema propuesto, la señal de sEMG multicanal es procesada en Matlab, utilizando 3 métricas para determinar la similitud de elementos: el coeficiente de Bhattacharyya, la similitud del coseno y la diferencia de energía, el cual compara la señal sEMG de interés con registros previos en una base de datos para determinar el movimiento que ha generado dicha señal. Las pruebas experimentales desarrolladas, demuestran que el sistema propuesto es capaz de identificar los movimientos realizados por el usuario con una precisión de hasta 87 % obteniendo resultados competitivos con el estado del arte a un costo bajo.

Palabras clave: Señal EMG, coeficiente de Bhattacharyya, similitud del coseno, diferencia de energía, reconocimiento de patrones.

Movement Identifier through Statistical Analysis of Electromyographic Signal

Abstract. This article presents the design of a system capable of identifying movements made by analyzing the surface electromyographic (sEMG) signal using different statistical similarity metrics. In the proposed system, the multichannel sEMG signal is processed in Matlab, using 3 metrics to determine the similarity of elements: the Bhattacharyya coefficient, the cosine similarity and the energy difference, which compares the sEMG signal of interest with records in the previous data in a database to determine the movement that has generated the signal. The experimental tests developed show that the proposed system is capable of identifying the movements made by the user with an accuracy of up to 87 %, obtaining competitive results with the state of the art at a low cost.

Keywords: EMG signal, Bhattacharyya coefficient, cosine similarity, energy difference, pattern recognition.

1. Introducción

La señal electromiográfica (EMG) registrada sobre la superficie de la piel ha sido utilizada en diversas aplicaciones de diagnóstico, tratamiento y rehabilitación. En el campo de la rehabilitación los esfuerzos están orientados a procesar la señal EMG como señal de entrada para la manipulación de dispositivos protésicos [1,2]. En este sentido, en la práctica los puntos de registro de señal son limitados, por lo que es necesario desarrollar propuestas donde el número reducido de canales de registro de señal proporcionen resultados adecuados para a la identificación de movimientos.

En [3], se presenta un método basado en máquinas de vectores de soporte que puede detectar la apertura y cierre del pulgar, y los otros dedos a través de la señal de EMG superficial, mediante la colocación de 10 electrodos en diversos músculos del antebrazo, les fue posible clasificar la flexión y extensión del pulgar, el índice, y el resto de los dedos en conjunto con un porcentaje de acierto de entre 89 y 97 %, obteniendo resultados precisos en diferentes sesiones e independientemente de la posición del brazo.

De manera similar, en [4] se buscó implementar el EMG superficial para controlar dispositivos que pudieran asistir a personas con masa muscular reducida; se desarrolló un sistema de reconocimiento de patrones que permitía estimar el torque aplicado por la muñeca, mediante el análisis de la información de EMG superficial obtenida de 4 canales conectados a distintos músculos del antebrazo (flexor cubital del carpo, palmar largo, extensor común de los dedos y extensor radial corto del carpo). Los autores analizaron la flexión y extensión de la muñeca, así como la desviación radial y cubital, logrando un porcentaje de acierto de hasta 88 % al usar 19 clases y hasta 96 % al usar 13 clases.

Algunas otras técnicas de procesamiento que ha sido utilizadas para lograr reconocer movimientos mediante señales de sEMG incluyen el modelo oculto de Márkov [5] y la extracción de distintas características a partir de segmentos de la señal de sEMG [6,7].

En el presente trabajo, se presenta una propuesta de diseño de un sistema capaz de identificar movimientos de la mano realizados por un usuario mediante el análisis de una señal de electromiografía superficial (sEMG) de 2 canales. La intención del proyecto es conseguir identificar, con la mayor precisión posible, los cinco movimientos básicos de la mano: flexión, extensión, pronación, supinación y cierre; a la vez que se procura mantener el costo de dicho sistema lo más bajo posible.

Este trabajo está estructurado de la siguiente manera, en la primer sección se presenta la metodología de trabajo, se explican las distintas métrica utilizadas y la manera en que fueron implementadas en el sistema, además de las pruebas desarrolladas para determinar la posición optima de los electrodos. Luego, se presentan los resultados obtenidos en dichas pruebas, así como las discusiones de estos. Finalmente, se desarrollan las conclusiones a las que se ha llegado en este trabajo y el trabajo futuro a desarrollar.

2. Metodología

2.1. Implementación de métricas

Como previamente fue mencionado, el método propuesto consiste en utilizar 3 métricas distintas que permiten determinar la similitud existente entre 2 elementos: el coeficiente de Bhattacharyya, la similitud del coseno y la diferencia de energía. El coeficiente de Bhattacharyya es usado ampliamente en la investigación de extracción y selección de características, procesamiento de imágenes, reconocimiento de locutores y agrupación de teléfonos. Por otro lado, la similitud del coseno se utiliza en procesos de minería de datos, recuperación de información y coincidencia de texto. Tomando en cuenta las aplicaciones en las que son utilizadas, se consideró que podrían ser de utilidad para el objetivo deseado.

Es importante señalar como se obtiene la información: un registro está formado por la información de los 5 movimientos, la información de cada movimiento está compuesta por la señal proveniente de 2 canales, para cada uno de los canales se obtienen 9000 muestras. Con base en las características del hardware y software del sistema, y para garantizar el cumplimiento del teorema de muestreo de Nyquist, se decidió trabajar con la frecuencia de muestreo máxima que se pudo obtener del sistema: 1.7 kHz por canal. La señal se obtuvo de un circuito de diseño propio que delimita la información captada en el rango de 5-500 Hz. Puesto que la señal de sEMG es una señal bastante aleatoria y puede llegar a cambiar bastante con gran velocidad, se optó por trabajar las señales de EMG utilizando el valor promedio cada cierto número de muestras. A partir de los registros obtenidos, se decidió utilizar el promedio cada 50 muestras ya que se consideró un valor adecuado para no perder tanta información. De esta forma, los registros pasan de tener una longitud 9000 a tan solo 180 muestras.

El primer paso requerido para utilizar las métricas propuestas para procesar los datos es la normalización de los éstos, tanto de la matriz correspondiente a la base de datos como de la matriz con la señal de interés, para convertirlos en funciones de distribución de probabilidad. La normalización se realiza de tal forma que la suma de la información normalizada de ambos canales de EMG de cada movimiento en cada registro sume 1, Fig. 1.

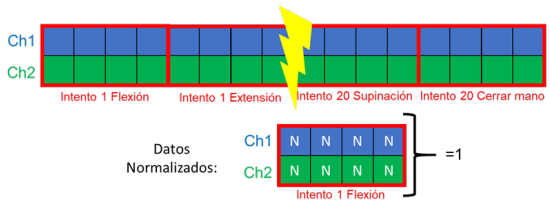


Fig. 1. Esquema de normalización de matrices. La suma de la información normalizada de ambos canales de cada movimiento de cada registro debe ser igual a 1.

El cálculo del coeficiente de Bhattacharyya implica una forma rudimentaria de integración de la superposición de las dos muestras, dicha operación está descrita en 1:

$$BC(p,q) = \sum_{i=1}^{n} \sqrt{p_i q_i}, \tag{1}$$

donde, considerando las muestras p y q , n es el número de particiones, p_i y q_i son los números de miembros de las muestras p y q en la i-ésima partición. Se diseño una función en Matlab para llevar a cabo el cálculo de los coeficientes. Utilizando las matrices normalizadas, se realiza el cálculo del coeficiente de Bhattacharyya entre la matriz del registro de interés (de tamaño 2x180) y cada una de las N matrices correspondientes a cada uno de los distintos movimientos de cada registro de la base de datos (2x180xN). De esta operación, se obtiene un vector de N elementos el cual indica el coeficiente de Bhattacharyya entre el registro de intereses y los de la base de datos. Para conocer el registro de la base de datos que posee una mayor similitud con el registro de intereses, se obtiene el valor máximo del vector obtenido previamente, y a partir de la posición de éste, se puede determinar a qué movimiento pertenece con ayuda de la operación módulo.

Para utilizar la similitud del coseno y la diferencia de energía es necesario modificar la forma de los datos al momento de ser procesados ya que dichas métricas trabajan solamente con vectores. Para esto, se optó por mover el canal 2 al final del canal 1 para cada movimiento de cada registro de las matrices normalizadas. En la fig. 2 se ejemplifica lo anterior de una forma visual.

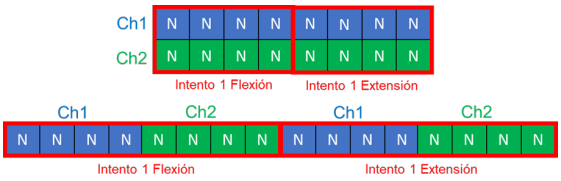


Fig. 2. Esquema de reacomodo de datos. Los datos pasan de estar representados en una matriz a un vector.

La similitud de coseno mide el coseno del ángulo entre 2 vectores distintos de 0 de un espacio de producto interno. Es decir, dos vectores coseno que están alineados en la misma orientación tendrán una medida de similitud de 1, mientras que dos vectores alineados perpendicularmente tendrán una similitud de 0. La similitud del coseno entre 2 vectores A y B se describe como se muestra en 2:

$$Similitud = cos(\theta) = \frac{(A \cdot B)}{\|A\| \, \|B\|}. \tag{2}$$

Se diseño una función en Matlab para llevar a cabo dicho cálculo. Utilizando el vector de la variable de interés y el arreglo de vectores normalizados, se

realiza el cálculo entre el vector normalizado del registro de interés (de tamaño 1x360) y cada uno de los N vectores correspondientes a cada uno de los distintos movimientos de cada registro de la base de datos (1x360xN). De esta operación, se obtiene un vector de N elementos el cual indica la similitud del coseno entre el registro de intereses y los de la base de datos. De manera similar al caso anterior, se determina el movimiento utilizando el valor máximo del vector y la operación módulo.

La diferencia de energía, como el nombre lo indica, permite conocer la diferencia en la energía de 2 vectores. Entre más similares sean ambos vectores, la diferencia de energía será menor. La diferencia de energía entre 2 vectores A y B está dada por 3:

$$EN(i) = \|A - B\|^2 . \tag{3}$$

Para realizar esta operación se creó una función en Matlab que realiza el cálculo de la diferencia de energía entre el registro de interés y cada uno de los vectores de la base de datos. Se obtiene un vector de N elementos el cual indica la diferencia de energía entre el vector del registro de intereses y los de la base de datos. Para conocer el registro de la base de datos que posee una mayor similitud con el registro de intereses, se obtiene el valor mínimo del vector obtenido previamente.

Para observar la respuesta que otorga cada una de las técnicas antes mencionadas, así como realizar otras pruebas de funcionamiento, se diseño un programa sencillo en Matlab el cual toma un registro de sEMG, de dimensiones iguales a los anteriores, lo compara con la base de datos previamente generada utilizando cada una de las métricas propuestas y finalmente muestra en pantalla que movimiento fue identificado en cada caso.

2.2. Determinación de posición de electrodos

Al realizar cada uno de los movimientos de la mano, distintos grupos de músculos se activan con diferentes intensidades para llevar a cabo dicho movimiento. Son estas diferencias en el nivel de activación de los grupos musculares lo que permite identificar el tipo de movimiento que ha originado dicha activación muscular. Puesto que el sistema propuesto solamente cuenta con 2 canales, es importante determinar la mejor combinación de aquellos músculos en los que es conveniente colocar los electrodos con el fin de obtener la mayor cantidad de información útil que permita diferenciar los movimientos de interés.

Mediante una investigación bibliográfica de las características de los músculos del antebrazo, como su localización y profundidad, así como las funciones que desempeñan y como son activados [8], se seleccionaron 4 músculos: flexor cubital del carpo, palmar largo, extensor común de los dedos y extensor radial del carpo, puesto que dichos músculos son superficiales e influyen en gran medida en las realización de los movimientos de interés. Con base en la información recabada acerca de la función de cada uno de los músculos candidatos, se optó por utilizar el músculo cuya función principal es la flexión de la muñeca, flexor cubital del carpo (FCU), en combinación con alguno de los otros músculos, ya sea alguno

de los encargados de la extensión, el extensor común de los dedos (ED) y el extensor radial del carpo (ECR), o el músculo palmar largo (PL); quedando de esta forma 3 posibles combinaciones: 1. FCU y ED, 2. FCU y PL, y 3. FCU y ECR. En las figuras 3, 4 y 5 se muestra la posición en la que fueron colocados los electrodos para cada una de las 3 combinaciones respectivamente.

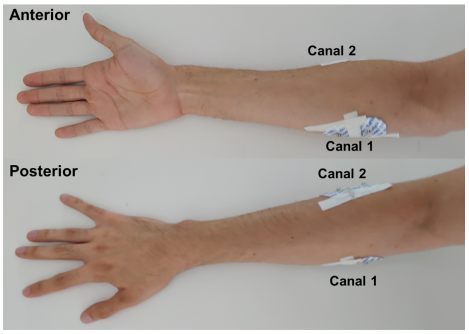


Fig. 3. Posición de electrodos en FCU (canal 1) y ED (canal 2).

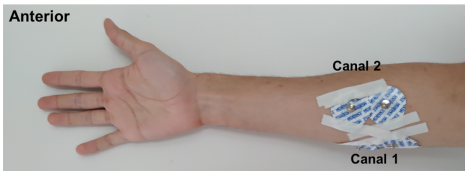


Fig. 4. Posición de electrodos en FCU (canal 1) y PL (canal 2).

Para conocer el porcentaje de acierto de cada combinación de músculos se diseñó la siguiente prueba. Para cada una de las 3 combinaciones, se generó una base de datos con 20 registros de los 5 movimientos de interés. Luego, utilizando el programa de prueba se realizaron 20 intentos para identificar cada uno de los 5 movimientos, registrando el número de casos por movimiento y por métrica que eran identificados correctamente. A partir de los resultados obtenidos con cada una de las 3 métricas utilizadas, se calculó el porcentaje de acierto para determinar que combinación de músculos proporcionaban mejores resultados.

3. Resultados y discusión

En la tabla 1 se presenta el porcentaje de acierto por movimiento usando cada una de las métricas, así como los porcentajes de acierto promedio por

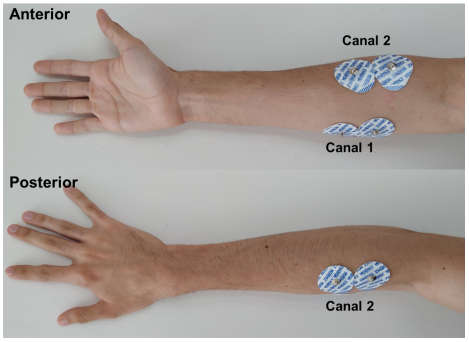


Fig. 5. Posición de electrodos en FCU (canal 1) y ECR (canal 2).

movimiento, por métrica y el promedio correspondiente a la combinación de los músculos FCU y ED.

Tabla 1. Porcentajes de acierto promedio usando músculos FCU y ED.

Movimiento	***Porcentaje de Acierto (%)***			
	C. Bhattacharyya	***S. Coseno***	***D. Energía***	***Promedio***
Flexión	100	100	100	100
Extensión	100	100	100	100
Pronación	75	75	65	72
Supinación	100	100	100	100
Cerrar Puño	80	55	50	62
Promedio	91	86	83	87

De la tabla 1 se observa que, utilizando esta combinación de músculos es posible identificar los 5 movimientos de interés con un porcentaje de acierto global de 87 %. La respuesta promedio obtenida utilizando las distintas métricas presenta una variación considerable, obteniendo una mejor respuesta al usar el coeficiente de Bhattacharyya.

Por otro lado, en la tabla 2 se presenta el porcentaje de acierto por movimiento usando cada una de las métricas, así como los porcentajes de acierto promedio correspondiente a la combinación de los músculos FCU y PL. Estos resultados muestran que también es posible identificar los 5 movimientos propuestos en este trabajo con un porcentaje de acierto global del 73 %. La respuesta promedio obtenida utilizando las distintas métricas no presenta una variación considerable, teniendo una respuesta ligeramente mejor al utilizar el coeficiente de Bhattacharyya.

Finalmente, en la tabla 3 se presenta el porcentaje de acierto por movimiento usando cada una de las métricas, así como los porcentajes de acierto promedio correspondiente a la combinación de los músculos FCU y ECR. De esta última

Tabla 2. Porcentajes de acierto promedio usando músculos FCU y PL.

Movimiento	***Porcentaje de Acierto (%)***			
	C. Bhattacharyya	***S. Coseno***	***D. Energía***	***Promedio***
Flexión	90	95	95	93
Extensión	60	45	45	50
Pronación	40	40	40	40
Supinación	100	100	100	100
Cerrar Puño	80	80	80	80
Promedio	74	72	72	73

Tabla 3. Porcentajes de acierto promedio usando músculos FCU y ECR.

Movimiento	***Porcentaje de Acierto (%)***			
	C. Bhattacharyya	***S. Coseno***	***D. Energía***	***Promedio***
Flexión	90	95	80	88
Extensión	0	0	0	0
Pronación	0	5	5	3
Supinación	100	100	100	100
Cerrar Puño	60	55	55	57
Promedio	50	51	48	50

prueba se observa que, a diferencia de con las combinaciones anteriores, no es posible identificar los 5 movimientos de interés. El porcentaje de acierto promedio es de 50 %. En este caso, la respuesta promedio obtenida utilizando las distintas métricas tampoco presenta una variación considerable; la respuesta obtenida al utilizar la similitud del coseno fue ligeramente superior a la de las otras 2 métricas.

Se decidió trabajar utilizando la combinación de músculos 1 por encima de la combinación de 2, ya que, a pesar de que con ambas combinaciones es posible identificar los 5 movimientos, la combinación 1 obtuvo un mayor porcentaje de acierto en cada una de las 3 métricas. Se optó por descartar la combinación 3, ya que con dicha combinación no fue posible identificar la extensión ni la pronación. Al seleccionar utilizar la combinación 1 y mediante el método propuesto es posible identificar con un porcentaje de acierto de hasta el 87 % los movimientos de la mano hechos por el usuario.

Una posible limitación del método es que es necesario que los registros de la base de datos correspondan al usuario del sistema, ya que la forma de llevar a cabo los movimientos llega a variar bastante de una persona a otra, sin embargo, esto podría solventarse en el futuro aplicando otros criterios de normalización de datos. Asimismo, es recomendable establecer una cierta forma en que se llevarán a cabo cada uno de los movimientos, y procurar realizarlos lo más similar posible. En contrapeso, el sistema de procesamiento propuesto presenta la ventaja de que su rendimiento va a ser mejor si se añade un mayor número de canales. La respuesta observada en este trabajo demuestra que el sistema es capaz de

identificar correctamente, en la mayoría de los casos, aquellos movimientos en los que los músculos censados brindan información suficiente como sucedió con la flexión y la extensión, los cuales están estrechamente relacionados con el músculo FCU y el ED respectivamente.

4. Conclusiones y trabajos futuros

El sistema de procesamiento propuesto es una alternativa de bajo costo y fácil implementación. Ofrece una precisión bastante buena, siempre y cuando la información con la que se alimente sea relevante para el tipo de movimientos que se desean identificar. Se ha demostrado que la combinación de los músculos FCU y ED ofrecen la mayor cantidad de información útil para identificar los movimientos de la mano con un sistema de 2 canales.

El sistema propuesto ofrece resultados similares a lo presentado en [3] y [4], con la ventaja de que usando este sistema solo se requiere de 2 canales para identificar un número cercano de movimientos, lo que representa una reducción significativa en el costo del sistema.

Como parte del trabajo futuro, y una posible aplicación para el sistema, se planea diseñar un sistema de entrenamiento que otorgue retroalimentación al usuario con la intención de facilitarle el aprendizaje de la forma de realizar los movimientos para posteriormente implementar el mismo sistema de procesamiento en el control de una prótesis activa sencilla. Asimismo, como trabajo a futuro se plantea, utilizando la combinación de los músculos aquí encontrada, desarrollar un procesamiento que permita examinar las señales de sEMG como una unidad, y no realizando una comparación elemento a elemento de los componentes de la señal.

Referencias

1. Contreras, D., Ramírez-García, A., Gallegos, F., Bazán, I.: Prototipo de una prótesis mioeléctrica para la emulación de una articulación de codo. Revista Mexicana de Ingeniería Biomédica, 36(1), pp. 6580 (2015)
2. de la Cruz, H., López, C.E., Rodríguez, E.E., Sandoval, L.M., Ramírez-García, A.: Propuesta de un entrenador mioeléctrico basado en una aplicación móvil. Pistas Educativas, 39(128), pp. 395–411 (2018)
3. Bitzer, S., van der Smagt, P.: Learning EMG control of a robotic hand: Towards active prostheses. In: Proceedings IEEE International Conference on Robotics and Automation, ICRA, pp. 2819–2823 (2006)
4. Khokhar, Z.O., Xiao, Z.G., Menon, C.: Surface EMG pattern recognition for real-time control of a wrist exoskeleton. BioMed Eng OnLine, 9, pp. 41 (2010)
5. Lu Zhiyuan, Xiang Chen, Li Qiang, Zhang Xu, Zhou Ping: A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. IEEE Transactions on Human-Machine Systems, 44, pp. 293–299 (2014)
6. Smith, L., Hargrove, L., Lock, B., Kuiken, T.: Determining the optimal window length for pattern recognition-based myoelectric control: Balancing the competing effects of classification error and controller delay. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 19, pp. 186–92 (2010)

7. Hakonen, M., Piitulainen, H., Visala, A.: Current state of digital signal processing in myoelectric interfaces and related applications. Biomedical Signal Processing and Control, 18, pp. 334–359 (2015)
8. Pease, W.S., Lew, H.L., Johnson, E.W.: Johnson's practical electromyography. Wolters Kluwer Health (2015)

Diseño de un módulo web automatizado para la recuperación de metadatos de referencias de artículos

Alma Delia Apale Zitzihua, Ignacio López Martínez,
Giner Alor Hernández, José Luis Sánchez Cervantes,
Luis Ángel Reyes Hernández

Tecnológico Nacional de México,
Maestría en Sistemas Computacionales
México

almazitzihua@gmail.com,
{ilopez, galor}@ito-depi.edu.mx,
jlsanchez@conacyt.mx, lreyesh@orizaba.tecnm.mx

Resumen. En la actualidad surge la necesidad por parte de estudiantes e investigadores de extraer e integrar la información bibliográfica de los artículos científicos que han sido publicados en distintas revistas o libros de divulgación científica que se concentran en internet con el fin de visualizar, extraer, almacenar y gestionar estos resultados en los repositorios de sus instituciones que servirán para llevar un control así como el seguimiento del trabajo y producción de trabajos de investigación, los metadatos son indispensables al momento de realizar una investigación de la localización de un artículo digital. Entre los metadatos que contienen las referencias de un artículo uno de los más relevantes que poseen es el Identificador de Objeto Digital (DOI); el cual facilita la localización dentro de las bases de datos en las que son almacenados, ya que se compone de un código alfanumérico que es único para cada objeto digital. La cantidad de información concentrada en Internet es enorme y conforme pasa el tiempo se van desarrollando nuevas y mejores herramientas así también métodos de recuperación de esta información. En este trabajo se presenta el diseño de un módulo Web automatizado para la recuperación de metadatos de los artículos científicos de distintas bases de datos y repositorios de gran relevancia. Adicionalmente se presenta y se describe una arquitectura de los componentes y su funcionamiento.

Palabras clave: Internet, divulgación científica, repositorios, Metadatos, DOI.

Design of an Automated Web Module for Article Reference Metadata Retrieval

Abstract. Currently, there is a need for students and researchers to extract and integrate the bibliographic information of scientific articles that have been published in different popular science magazines or books that are concentrated on the Internet in order to visualize, extract, store and manage these results in

the repositories of their institutions that will serve to control and monitor the work and production of research work, metadata is essential when conducting an investigation of the location of a digital article. Among the metadata contained in the references of an article, one of the most relevant they have is the Digital Object Identifier (DOI), which facilitates the location within the databases in which they are stored, since it is made up of an alphanumeric code that is unique for each digital object. The amount of information concentrated on the Internet is enormous and as time goes by, new and better tools are developed, as well as methods for retrieving this information. This paper presents the design of an automated Web module for retrieving metadata of scientific articles from different highly relevant databases and repositories. Additionally, an architecture of the components and their operation is presented and described.

Keywords: Internet, popular science, repositories, Metadata, DOI.

1. Introducción

La innovación, así como los nuevos descubrimientos científicos requieren largos periodos de investigación que resultan en nuevas herramientas y métodos que ayudan a mejorar la vida diaria, este tipo de resultados en su mayoría se reflejan actualmente en productos digitales como lo son artículos. Es importante recordar que estos productos de investigación se clasifican e identifican mediante su referencia bibliográfica; la cual se compone de un conjunto de datos como: autor, título del artículo, fecha de publicación, título de la revista, así como un identificador único, estos datos sirven para buscar, clasificar y almacenar los artículos.

Existen herramientas como los gestores de referencias [1] que se encargan de gestionar grandes cantidades de productos digitales de investigación, además permiten buscar y almacenar de manera organizada dichos productos. Mendeley [2], Zotero [3], EndNote [4] son tres de los gestores más conocidos y utilizados, entre sus características se destacan que Mendeley provee 2 GB de almacenamiento gratuito y funciona mediante una aplicación de escritorio que cuenta con cinco maneras de extracción de artículos los cuales están incorporados a las bases de datos pertenecientes a la editorial Holandesa Elsevier [5], estas herramientas utilizan *plugins* (programa informático de inserción) [6] en páginas Web y bases de datos pero para la búsqueda se requiere ingresar la mayor parte de los datos del trabajo de investigación en algunos casos no se presentan problemas de compatibilidad.

Por otro lado, Zotero lleva a cabo la misma función, pero a diferencia del gestor anterior, solo provee de forma gratuita 300 MB de almacenamiento y presenta problemas de compatibilidad con algunas versiones de Word. Finalmente, el gestor EndNote permite almacenar distintos datos de los trabajos de investigación como imágenes y referencias, pero es muy poco conocido por la comunidad investigadora. Los tres gestores descritos anteriormente son de gran utilidad al momento de realizar una investigación, pero existen varios motivos que hacen de ellos muy complejos al momento de utilizarlos como dificultades al momento de la instalación en los dispositivos que en ocasiones presentan incompatibilidad en las versiones de software, el uso de complementos en los distintos navegadores, entre otros. A causa de esto es

necesario el desarrollo de una herramienta más intuitiva, sencilla y fácil de utilizar para la extracción de referencias con base en las necesidades del usuario final.

Tomando en cuenta lo anterior este artículo presenta una propuesta de solución que consiste en el desarrollo de un módulo Web automatizado que permita recuperar información de referencias de artículos a nivel de metadatos para asegurar una información más precisa y sin duplicidad, es decir, una búsqueda más simple ingresando la menor cantidad de datos del trabajo de investigación y así localizarlo en las principales bases de datos; la sección 2 presenta trabajos relacionados sobre métodos de extracción de metadatos con diferentes tipos de tecnologías; la sección 3 explica el diseño y funcionalidad del módulo Web; por último en la sección 4 se muestran las conclusiones del trabajo que se está realizando y los trabajos a futuro.

2. Trabajos relacionados

La producción en el ámbito de divulgación científica crece día a día y genera que los investigadores, así como las instituciones donde llevan a cabo este proceso se vean en la necesidad de llevar el control de manera organizada, así como de cada uno de sus productos generados; se han desarrollado herramientas que ponen a disposición para buscar, citar y almacenar estos trabajos de investigación. A continuación, se presenta una revisión del estado del arte de los trabajos existentes que han utilizado herramientas y/o métodos de búsqueda y extracción de datos de artículos científicos.

Klein y Van de Sompel [7] afirmaron que en las últimas dos décadas, la comunicación de la investigación ha pasado de ser un esfuerzo basado en el papel a una empresa digital basada en la Web. Con el paso del tiempo el proceso de investigación ha comenzado a evolucionar, ya que pasó a ser una actividad abierta y visible a nivel mundial. Con el fin de apoyar a los distintos grupos de investigadores surgió una gran variedad de repositorios, bases de datos, plataformas virtuales como una herramienta de divulgación científica. Sin embargo, algo que ha pasado rara vez pero que significa mucho para los usuarios es que las plataformas pueden desaparecer sin dejar rastro causando la pérdida de información importante.

Algunas de ellas no ofrecen garantías del contenido publicado, es por eso que se propuso un nuevo paradigma de archivo que se centra en el descubrimiento de entidades y artefactos Web en el ámbito del enfoque de archivo Web. Dado que el ORCID [8] surgió como una infraestructura Web académica de alto potencial que asigna identificadores únicas digitales a los académicos, permite listar identidades Web adicionales, así como artefactos. Sin embargo, también se descubrió que los crecimientos de ORCID tienen una tasa muy significativa que supera el crecimiento de los investigadores y eso ha causado también que varias plataformas últimamente opten por añadir esta función a sus tecnologías. Por lo tanto, existe una posibilidad real de que ORCID alcance un nivel de cobertura en un futuro próximo más adecuado para resolver las necesidades de los investigadores.

Hozlmann y Runnwerth [9] mencionaron que existen muchos tipos de identificadores para referirse a los trabajos académicos principalmente hicieron notar la existencia de DOI (*Digital Object Identifier*, Identificador de Objetos Digitales) Sistema que se inició en el año 1998[10] por Crossref una organización de miembros sin ánimos de lucro[11] , estos identificadores están constituidos por datos en conjunto,

referentes a la obra y que son llamados metadatos. Como problemática presentada en este artículo se argumentó que las referencias provenientes de sitios Web como lo son los blogs o programas informáticos, son muy vagas, aunque esta contenga la URL (*Uniform Resource Locator*, Identificador de recursos uniforme) [12] y la fecha en que se visitó, ya que, los sitios están en constante modificación de su contenido. Por lo anterior se presentaron los Micro Archivos que son colecciones microscópicas como su nombre lo dice, de los recursos archivados en la Web, utilizados para describir los objetos o entidades mediante el prototipo Micrawler (*Micro Crawler*, micro rastreador) que es una implementación y prueba de referencia encargado de la gestión de estos conjuntos de micro información. Micrawler tiene como objetivo principal que los archivos consultados en la Web archiven todos los recursos relacionados incluyendo el código fuente resultando en un micro Archivo para una consulta permanente mediante una URL corta o un mediante DOI.

A su vez Bangert y Frances [13] afirmaron que nos encontramos en un panorama en el que los productos académicos aumentan exponencialmente, debido a esto; los Identificadores Persistentes (PID) son una tecnología clave para permitir el acceso y la interoperabilidad entre los sistemas involucrados en la comunicación académica, estos identificadores permiten que un trabajo de investigación sea rastreado, visualizado, pero debido al surgimiento de nuevas plataformas que almacenan estos documentos de investigación surgió la necesidad de implementar mejoras en esta tecnología, es así, que varias organizaciones han trabajado para mejorar la integración en la infraestructura de investigación internacional y se propuso que sea a través de proyectos de colaboración con plataformas como ORCID que provee una forma de identificación al autor y la Red de Interoperabilidad de DataCite.

Este proyecto consistió en el diseño y la entrega de los servicios de PID en la Biblioteca de la UNSW que está guiado por las características de los identificadores como es el caso de DOI.

Los identificadores asignados a los resultados de la investigación son interoperables, basados en las fuentes verídicas, y contienen metadatos legibles por humanos y máquinas. DOI permite tener un acceso persistente al recurso está garantizado por la biblioteca como custodio de los identificadores y del contenido del repositorio asociado. Los resultados fueron que a medida que los identificadores pasan a formar parte de cada etapa del ciclo de vida de la investigación, el desafío para las instituciones será seguir rigiendo su asignación de manera efectiva, seguir las normas de la comunidad, y optimizar su uso para y por los investigadores.

Además Yang y Zhang [14] dedujeron que debido a las diversas plataformas que albergan miles de trabajos de investigación científica los investigadores tienen un acceso ilimitado a estas herramientas digitales, existen distintas formas de realizar la búsqueda de información y una de las más utilizadas es empleando las palabras claves para una recopilación precisa y necesaria de lo que se busca, esto se lleva a cabo por medio de la utilización de algoritmos de búsqueda o métodos de rastreo, en algunos casos estas formas muestran artículos más citados recientemente o su reputación dependiendo el autor, esto en algunos casos puede no ser lo que se requiere para el trabajo de investigación que se está desarrollando. Teniendo esta problemática se propuso la utilización del algoritmo de integración de texto llamado "TextRank" que es utilizado para ayudar a los investigadores en la realización de la revisión de la literatura. El objetivo principal de TextRank es resumir una obra determinada de documento que

Tabla 1. Comparativa de trabajos que muestran la falta de un sistema Web que se complemente con estas tecnologías, en el que se incluyan identificadores de autores, así como de artículos para recopilar y gestionar información precisa de obras de investigación que se encuentren almacenadas en distintas bases de datos y de diferentes editoriales.

Artículos	Extracción por ID	Extracción de Metadatos	Base de Datos	Integración de ORCID	Integración de Crossref
Klein y Van de Sompel [7]	✓	✓	✓	X	**X**
Holzmann y Runnwerth [9]	X	X	✓	✓	**X**
Bangert y Frances [13]	✓	**X**	✓	✓	**X**
Yang y Zhang [14]	**X**	✓	**X**	**X**	**X**
Módulo Web Propuesto	✓	✓	✓	✓	✓

Tabla 2. Tipos de metadatos a utilizar para la extracción de referencias de artículos.

Tipo de metadato	Contenido
Descriptivos	– Titulo – Autor
Estructurales	– Volumen de revista – Número de página

se encuentre contenido en alguna base de datos, por lo que se puede decir de acuerdo a su funcionamiento que este algoritmo se basa en la minería de texto y al ser compleja se requiere un gran esfuerzo humano.

A continuación, se presenta una tabla comparativa de trabajos relacionados y el alcance que cada uno tiene respecto a los métodos y tecnologías utilizadas.

3. Diseño y funcionalidad del módulo Web

La propuesta de solución surge con el fin de proveer a los investigadores y académicos una herramienta automatizada para recopilar las referencias de las obras publicadas en distintas bases de datos y que permita extraer metadatos para llevar el control de referencias de los artículos, así como almacenarlas además de monitorear el estatus de dicha información que se incluye en otras investigaciones en donde han sido citadas.

Se muestra una descripción general de la propuesta de solución y su arquitectura preliminar para el funcionamiento del módulo Web presentando las partes más importantes que son la búsqueda, recolección de información, selección y

Fig. 1 Diagrama del funcionamiento de las herramientas seleccionadas.

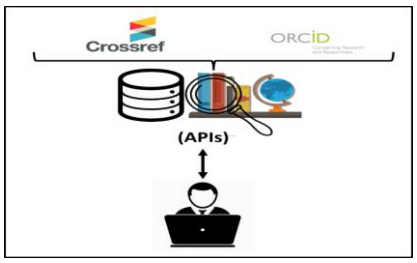


Fig. 2 Funcionamiento de un segmento de la arquitectura preliminar.

funcionamiento. Para llevar a cabo esta extracción es importante mencionar que se realizará la extracción de metadatos que son datos que describen otros datos [15] para el caso de las referencias de artículos se utilizaran los siguientes dos tipos de metadatos:

Búsqueda y recolección de metadatos. Se describe el módulo Web que interactúe con las bases de datos científicas que pertenezcan a trabajos de investigación en sus distintos formatos se logren recuperar metadatos de referencias bibliográficas de artículos utilizando principalmente los identificadores internacionales y el ingreso de la menor cantidad de datos para el rastreo de los trabajos de investigación, para la búsqueda y extracción.

Selección de herramientas. Para la selección de las herramientas a implementar en el módulo se basó en el siguiente diagrama que muestra el funcionamiento de recolección de productos de investigación:

Funcionamiento. En primer lugar, las bases de datos como: Scopus, Web of Science, entre otros; pertenecen a la mayor editorial de contenido científico que es Elsevier la cual posteriormente se encarga de depositar sus metadatos a la agencia de identificadores que en este caso es Crossref [10] ; de igual forma los investigadores depositan sus metadatos mediante el uso de ORCID. En base a este funcionamiento se eligió utilizar ambas herramientas para realizar la búsqueda y extracción requerida para el módulo tomando en cuenta la forma en que ellos recolectan la información que se necesita para alimentar el módulo y aprovechando la oportunidad de alcance a esta concentración de datos al público en general, así como desarrolladores y este caso resolviendo los problemas anteriormente planteados.

En el siguiente diagrama se muestra el funcionamiento entre usuario y las herramientas seleccionadas. En primera instancia el usuario realiza una petición de

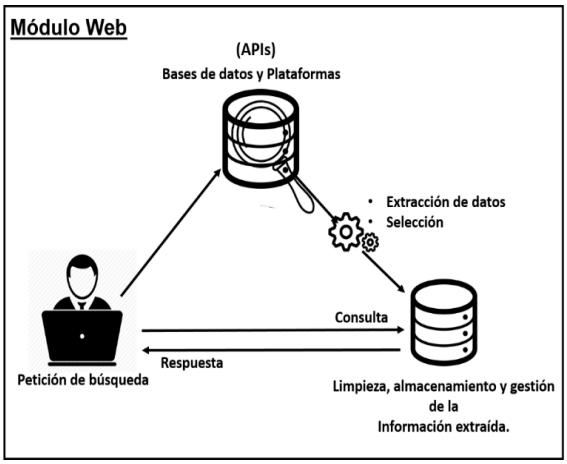


Fig. 3. Arquitectura preliminar basado en el modelo orientado a servicios.

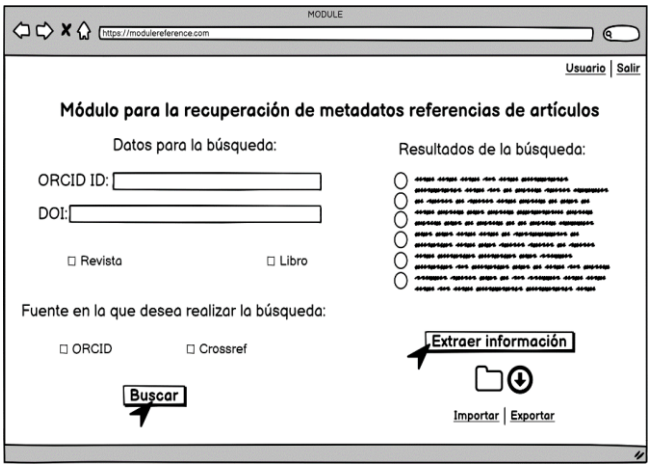


Fig. 4. Mockup de la interfaz del Módulo Web propuesto.

búsqueda de metadatos mediante el módulo que con ayuda de las API´s (Interfaces de Programación de Aplicaciones del Navegador, *Application Programming Interface*) [16] realizará la entrega de la respuesta de la información requerida.

Agregando la otra parte del funcionamiento de la arquitectura propuesta contempla la utilización de una base de datos local que se encargará de reunir la información para su posterior selección de la información requerida y envío para la concentración en algún repositorio dependiendo la decisión de uso del usuario.

La figura que se presenta a continuación muestra la interfaz que va interactuar con el usuario (estudiante o investigador) que tiene como objetivo realizar la búsqueda automatizada minimizando el ingreso de datos de manera que ingresando el Identificador de Objeto o Identificador de autor puede ser buscado también en libros y revistas.

Tabla 3. Requerimientos solicitados por ORCID para el cumplimiento de políticas de seguridad de datos y sus funciones principales a realizar dentro del módulo Web.

Requerimientos	**Función dentro del módulo**
Almacenar Identificadores	Cada usuario con cuenta en ORCID maneja un identificador único que ayudará a clasificar el contenido dentro del módulo de forma más segura.
Usar Tokens de acceso persistentes y actualizados	Los Tokens de acceso que funcionan como llaves digitales dentro del módulo serán utilizados para el tratamiento seguro de cada uno de los productos.
Registro de interacciones con la API	Se realizará el registro de todas las llamadas y respuestas recibidas.
Contacto de soporte	Se debe incluir una función que sirva como medio de contacto con soporte técnico cuando ocurra alguna interacción inesperada.

La búsqueda entre las bases de datos será opcional entre utilizar una en específico o ambas simultáneamente. Una vez finalizada la búsqueda el módulo mostrará los resultados obtenidos para la selección de interés por parte del usuario y con las opciones de importación y exportación de sus datos contemplando las políticas de seguridad que pongan en vigor ambas organizaciones que se incluyen en el módulo.

Las organizaciones en la elección de búsqueda ponen a disposición las siguientes Apis:

API REST Pública Crossref: La cual expone los metadatos que son depositados por los miembros en la plataforma de Crossref que son importantes para localizar publicaciones por autor, cabe mencionar que dichos datos públicos se pueden utilizar sin restricción [17].

Funcionalidades de los componentes de búsqueda:

- Trabajos utilizando el DOI,
- Miembros,
- Tipos de publicaciones,
- Revistas.

Clasificación:

- Relevancia,
- Fecha,
- Hora.

API RESTful Pública ORCID: API que proporciona el acceso a la base de datos de los datos de registros ORCID ID públicos función que es elección del miembro[18].

Funcionalidades de los componentes de búsqueda:

- Obtener ORCID ID Autenticado,
- Buscar,

Tabla 4. Propuesta de tecnologías para el desarrollo del módulo, se muestra la solución propuesta la cual considera la utilización del lenguaje de programación Node.js esto debido a que es adecuado para el desarrollo Web, como entorno de desarrollo se contempló el uso de NetBeans ya que provee soporte para las aplicaciones orientadas a servicios. En cuanto al sistema gestor de base de datos MariaDB para almacenar los datos extraídos, y finalmente la metodología Scrum la cual maneja un entorno colaborativo, organizado por iteraciones entres otras cualidades.

Lenguaje de programación	IDE	SGBD	Metodología
Node.js	NetBeans	MariaDB	Scrum

– Recuperar datos públicos.

Para una implementación óptima y segura de la API de ORCID es indispensable cumplir con los siguientes requerimientos como políticas de seguridad por parte de ORCID en su documentación de uso de su API [19]:

El desarrollo de este módulo contempla la utilización de tecnologías que permitan una interacción entre la base de datos con el servidor y el usuario con el módulo. Para ello se requiere de un entorno de desarrollo que soporte aplicaciones orientadas a servicios. Finalmente se propone una base de datos local para un mejor control y gestión de la información extraída.

4. Conclusiones y trabajos futuros

Existen gestores de referencias y componentes que ayudan a extraer información referente a trabajos de investigación en bases de datos científicas pero en particular con el módulo Web propuesto realizará esta función de manera más sencilla y eficaz con una búsqueda ingresando la menor cantidad de datos además de que sea compatible con los buscadores resolviendo los problemas que presentan los actuales gestores de referencias anteriormente mencionados para esto se utilizaran las normas y estándares de calidad para recopilar dicha información , esto se logrará utilizando los identificadores usados en los distintos formatos para mayor confiabilidad tomando en cuenta sus políticas de seguridad así como las recomendaciones de mejores prácticas.

El seguimiento, almacenamiento y control de productos científicos es relevante e importante al momento de llevar a cabo el desarrollo de una nueva tecnología, método o herramienta que resuelva una problemática actual, ya que, como en este caso se toman en cuenta los resultados obtenidos para tomarlos en cuenta en todo momento del proceso de desarrollo.

El objetivo principal y más importante de este trabajo es beneficiar no solo a personas que participen en algún tipo de investigación como lo hacen los estudiantes al momento de concluir un grado de estudio, sino también a instituciones que requieran llevar un control de las obras realizadas por sus académicos en las distintas áreas.

Como trabajo a futuro se desarrollará el módulo Web automatizado propuesto y se validará mediante un caso de estudio implementándolo en una institución para gestionar los productos sus estudiantes o investigadores autores además de agregar un apartado extra que obtenga el informe sobre el estatus de producción para encontrar los puntos de oportunidad en las cuales trabajar y mejorar. Finalmente se realizarán las pruebas

necesarias con el equipo de ORCID para la validación, aprobación y publicación del módulo desarrollado para llevar a cabo la transición a la API de producción.

Referencias

1. PoliScience: Gestores de referencias. https://poliscience.blogs.upv.es/investiga dores-2/mis-citas/gestores-de-citas/ (2020)
2. Mendeley: Software de gestión de referencias y red de investigadores. https://mendeley.com/?interaction_required=true (2020)
3. Zotero: Tu asistente de investigación personal. https://zotero.org/ (2020)
4. EndNote: Clarivate Analytics. https://endnote.com/ (2020)
5. Elsevier: Zona de Lectura. https://elsevier.es/es (2020)
6. Neo Wiki: ¿Qué es un Plugin y para que sirve?. https://neoattack.com/ neowiki/plugin/ (2020)
7. Klein, M., van de Sompel, H.: Discovering scholarly orphans using ORCID. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–10 (2017)
8. ORCID: Un sistema para identificar de manera única a los investigadores. HAAK, Learned Publishing (2012)
9. Holzmann, H., Runnwerth, M.: Micro archives as rich digital object representations. In: Proceedings of the 10th ACM Conference on Web Science, pp. 353–357 (2018)
10. Doi.org: Agencias de registro de DOI. https://doi.org/registration_agencies.html (2020)
11. Wilkinson, L.J.: You are Crossref, https://crossref.org/ (2020)
12. Significados.com: Significado de URL (Qué es, Concepto y Definición) Significados. https://significados.com/url/ (2020)
13. Bangert, D., Frances, M.: PIDs to support discovery and citation: Persistent identifier service design and delivery at UNSW library. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–2 (2017)
14. Yang, D., Zhang, A.N.: Performing literature review using text mining, Part III: Summarizing articles using TextRank. In: IEEE International Conference on Big Data (Big Data), Seattle, pp. 3186−3190 (2018)
15. PowerData: ¿Qué son los metadatos y cuál es su utilidad? https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-son-los-metadatos-y-cual-es-su-utilidad (2020)
16. Hipertextual.com: Qué es una API. https://hipertextual.com/archivo /2014/05/que-es-api/ (2020)
17. Crossref: Rest Api. https://crossref.org/education/retrieve-metadata/rest-api/ (2020)
18. Krznarich, L.: Orcid Api. https://orcid.org/organizations/integrators/API (2014)
19. Orcid Members: Getting started with your ORCID integration. https://members.orcid.org/api/getting-started (2020)

Wearable para monitoreo de ritmo cardíaco y actividad electrodérmica

Luis Brayan Zacatelco Barrios[1], Blanca Tovar Corona[2], Javier Pindter Medina[3]

[1] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas,
Sección de Estudios de Posgrado e Investigación,
México

[2] Instituto Politécnico Nacional
Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas,
Laboratorio de Instrumentación y Procesamiento de Señales,
México

[3]Investigación, Desarrollo e Innovación,
PindNET R&D, SA de CV,
México

lzacatelcob1400@alumno.ipn.mx, bltovar@ipn.mx,
idi@pindnet.com

Resumen. En nuestros días el uso de wearables que permitan monitorear algunas señales fisiológicas mientras realizamos actividades cotidianas es cada vez más común por las diferentes aplicaciones que se le pueden dar al análisis de estas señales en ámbitos como; salud, estilo de vida, vida fitness, vida industrial y entretenimiento. Las lecturas obtenidas de estas señales con el uso de este tipo de dispositivos tienen diversas aplicaciones, por ejemplo, el seguimiento a cambios en el estado de nuestra salud, monitoreo del comportamiento de personas con alguna enfermedad, detección de emociones, entre otras. En el presente trabajo se describe un prototipo, denominado "B1", capaz de registrar actividad electrodérmica (EDA por sus siglas en inglés) y ritmo cardíaco (HR por sus siglas en inglés) el cual fue evaluado al comparar con el sistema certificado E4 wristband de Empática Inc. Se obtuvo una diferencia de 0.4 % en mediciones de HR y de 0.18 % en lecturas de EDA.

Palabras clave: Ritmo cardíaco, fotopletismografía, actividad electrodérmica, respuesta galvánica de la piel.

Wearable for Monitoring Heart Rate and Electrodermal Activity

Abstract. In our days the use of wearables that allow us to monitor some physiological signals while we carry out daily activities is increasingly common

due to the different applications that can be given to the analysis of these signals in areas such as; health, lifestyle, fitness life, industrial life and entertainment. The readings obtained from these signals with the use of this type of device have various applications, for example, monitoring changes in the state of our health, monitoring the behavior of people with a disease, detection of emotions, among others. This paper describes a prototype, called "B1", capable of recording electrodermal activity (EDA) and heart rate (HR), which was evaluated when compared with the certified E4 wristband system. Empatica Inc. A difference of 0.4% in HR measurements and 0.18% in EDA readings was obtained.

Keywords: Heart rate, photoplethysmography, electrodermal activity, galvanic skin response.

1. Introducción

El monitoreo de señales fisiológicas ayuda a dar seguimiento a cambios en el estado de salud mediante un sistema medidor de señales biomédicas que permita observar los cambios. El sistema aquí descrito registra dos variables: HR y EDA.

El dispositivo B1, por su batería y su capacidad de transmisión inalámbrica, se posiciona en la categoría de wearables (Asociación de Centros Tecnológicos de Galicia, 2017). Consta de dos subsistemas para la toma de mediciones, ambos están integrados en una pulsera la cual deberá llevar el sujeto a monitorear mientras realiza sus actividades diarias o es sometido a algún tipo de estudio.

El desarrollo de este proyecto tiene como intención bajar los costos en comparación a sistemas ya existentes de medición de señales biomédicas, pero sin bajar la calidad de las mediciones. Para esto se analizaron y eligieron los sensores que mejor se adaptaron a estas características; esto es posible gracias al desarrollo que se tiene tanto tecnológico como científico en nuestros días donde de igual manera, se encuentran en el mercado de una forma amplia. Además de esto, se realizó una conexión inalámbrica con un servidor, el cual, recopila los diversos datos personales de cada usuario, así como los resultados del monitoreo, dando posibilidad a ser utilizados posteriormente al quedar almacenados en una base de datos.

1.1. Ritmo cardíaco

HR es el conteo de latidos del corazón por unidad de tiempo, generalmente se expresa en latidos por minuto (BPM por sus siglas en inglés) (Mandal, 2019). El HR es uno de los parámetros no-invasivos más utilizado en el análisis y en la valoración de la actividad cardíaca. En una persona sana, en reposo, los latidos se van produciendo con una frecuencia variable, es decir, el tiempo entre dos latidos va variando latido a latido. Este aspecto representa el concepto de variabilidad de la frecuencia cardíaca (HRV por sus siglas en inglés), que se define como la variación de la frecuencia del latido cardíaco durante un intervalo de tiempo definido con anterioridad (nunca superior a 24 horas) en un análisis de períodos circadianos consecutivos (Rodas, Pedret Carbadillo, Ramos, & Capdevila, 2008).

Para el correcto funcionamiento del organismo es necesario que el corazón actúe bombeando la sangre hacia todos los órganos, pero además lo debe hacer a una

determinada presión (presión arterial) y a una determinada frecuencia (Facultad de Medicina, 2019)

1.1.1. Fotopletismografía

La fotopletismografía es una técnica de pletismografía que consiste en registrar de manera no invasiva las variaciones de volumen sanguíneo en las diferentes partes del cuerpo de una persona, especialmente en sus extremidades (Celi, Rocha, & Yapur, 2015).

Su principio de funcionamiento parte de la emisión de un haz de luz infrarroja sobre la piel para iluminar los vasos subcutáneos, estos reflejan parte de dicho haz dependiendo la cantidad de hematíes que contienen. La luz reflejada incide en un fotosensor (usualmente de cadmio-selenio) que la convierte en un voltaje equivalente. Debido a que la piel absorbe más del 90 % de la luz, el par foto-diodo se acompaña de amplificadores y filtros que garantizan un voltaje adecuado. El ciclo cardíaco puede obtenerse midiendo el intervalo que existe entre cada pico de voltaje (Shelley & Shelley, 2011). Esta técnica permite la medición tanto de HR como de HRV.

1.2. Actividad electrodérmica

EDA consiste en la variación de las propiedades eléctricas de la piel al producirse sudor. Estas variaciones en la conductancia de la piel se pueden medir aplicando una corriente continua de baja intensidad de forma no invasiva. Dicha reacción fisiológica está relacionada con la activación del eje Hipotálamo-Hipofisario-Adrenal (HHA), que genera en última instancia la activación de las glándulas sudoríparas de la piel. Se trata de uno de los principales métodos psicofisiológicos para medir procesos psicológicos como la emoción, el arousal o la atención; y sus variaciones han sido relacionadas con cambios en el estado cognitivo o emocional en el individuo, especialmente con estados de estrés (Díaz Robledo & Sánchez, 2018).

1.2.1. Respuesta galvánica de la piel

Una forma de registrar de actividad electrodérmica es obteniendo los valores de la resistencia o impedancia de la piel ante el paso de una pequeña corriente que se aplica por medio de electrodos, uno de los métodos que se usan es la respuesta galvánica de la piel (GSR por sus siglas en inglés).

La GSR es la medida de las continuas variaciones en las características eléctricas de la piel, por ejemplo, la conductancia, causada por la variación de la sudoración del cuerpo humano. Para registrar la GSR son necesarios dos electrodos y la variación de una corriente aplicada de bajo voltaje (Sapienza Universita Di Roma, 2018).

1.3. Comunicaciones inalámbricas

La comunicación inalámbrica es aquella en la que ni el emisor ni el receptor se encuentran unidos de manera física y se comunican mediante el uso de ondas electromagnéticas y mecánicas (Consinfin, 2012).

La mayor efectividad y alcance logrado entre dos equipos inalámbricos es cuando no existen obstáculos entre sus antenas lo cual es conocido en el ambiente técnico como

Tabla 1. Wearables que registran EDA y HR.

Nombre del artículo o proyecto	Señales fisiológicas que se miden	Aplicación	Síntesis
Diseño de un prototipo de medición de señales fisiológicas utilizadas en Biofeedback (Nieto & Vega, 2017).	HR, EDA y ritmo respiratorio (RR por sus siglas en inglés).	El wearable puede ser utilizado para el tratamiento de arritmias o hipertensión, para investigación científica o como herramienta complementaria al diagnóstico.	Se desarrolló un wearable de tamaño pequeño, cómodo, portátil y autónomo que permitiera tomar mediciones de diferentes señales biomédicas. Las señales biomédicas que se miden son: HR, RR y EDA. El subsistema HR presenta gran sensibilidad a los movimientos del usuario, provocando error en las lecturas. En RR se presenta un problema de desfase en tiempo En EDA las lecturas se consideran adecuadas de acuerdo a los rangos propuestos por los autores.
Monitoreo del ritmo cardiaco a través de dispositivos móviles (Aveiga-Paini, Criollo Altamirano, & Cruz-Quijije, 2018).	HR	Auxiliar en el diagnóstico de problemas detectados por cambios en HR, principalmente en personas hipertensas. Envía alertas cuando los valores de HR salen del rango.	Genera registros personalizados de HR mediante el uso del dispositivo Zephyr HxM Bluetooth Wireless Heart Rate Sensor for Android and Windows (ZEPHYR, 2013), el cual, se coloca en el tórax del usuario. Genera registros que el médico utiliza en los historiales médicos utilizados para dar seguimiento.
Actividad electrodérmica aplicada a la psicología: análisis bibliométrico (Mojica-Ledoño, 2017).	EDA	Análisis de procesos psicológicos que incluyen EDA para: detección del engaño, neuromarketing, afectaciones en niños y adolescentes por violencia en videojuegos, estudios sobre emociones según el tipo de música que escuche una persona, análisis durante la toma de decisiones, tratamiento de pacientes con dolor lumbar crónico a través de biofeedback y neurofeedback, manifestación de agresividad y la disregulación afectiva en niños.	Se llevó a cabo una revisión teórica y empírica de 36 artículos publicados entre los años 2006 y 2016, y de 3 libros en relación con la utilidad de EDA en psicología. El análisis de este material permite concluir que EDA es la señal fisiológica más utilizada para dar sustento a procesos psicológicos en relación con la emoción, el arousal y la atención.

"línea de vista". Si no hay paredes, edificios o cerros la comunicación será más efectiva (Foroz, 2017).

1.3.1. Bluetooth de bajo consumo (BLE, Bluetooth Low Energy)

Bluetooth es una tecnología inalámbrica estándar para el intercambio de datos en distancias cortas de hasta 100 metros usando las ondas de radio de onda corta en las bandas industriales, científicas y médicas de 2,4 a 2,485 GHz (Hernández Aquino, 2008).

Tabla 1. Continuación.

Nombre del artículo o proyecto	Señales fisiológicas que se miden	Aplicación	Síntesis
A Review of Wearable Solutions for Physiological and Emotional Monitoring for Use by People with Autism Spectrum Disorder and Their Caregivers (Taj-Eldin, Ryan, O´Flynn, & Paul, 2018).	HR, HRV, Reactividad de ritmo cardiaco, RR, EDA, Temperatura corporal y de la piel, niveles de cortisol, presión sanguínea, volumen de flujo sanguíneo, saturación de oxígeno en la sangre, electromio-grafía (EMG), electroence-falografía (EEG).	Monitoreo de personas con autismo, así como personas con desordenes intelectuales. Detección de cambios en estados emocionales para evitar desordenes crónicos cardio-vasculares al rebasar los límites de seguridad de parámetros obtenidos de las diferentes señales a monitorear. Detección de factores estresantes que provoquen fobia social, ansiedad, agresividad e irritabilidad. Detección en cambios de niveles de estrés, desordenes postraumáticos y de calidad del sueño.	Se lleva a cabo una revisión de wearables presentes en el mercado con el objetivo de exponer que estos dispositivos son instrumentos que ayudan a dar soluciones potenciales a diferentes enfermedades, así como apoyar su diagnóstico. Este análisis cuenta con el soporte de validaciones clínicas, análisis de prototipos y perspectivas diferentes sobre discusiones clínicas que apoyan el uso de estos dispositivos, ya que en un futuro cercano ofrecerán más soluciones prometedoras a diferentes problemas como los mencionados.
Validation of Wireless Sensors for Psychophysiol ogical Studies (Silva Moreira, Chaves, Dias, Dias, & R. Almeida, 2019).	EDA, señal de fotopletis-mografía.	Detección de emociones, arousal, ansiedad, psicosis, desordenes de dependencia, epilepsia, síndrome de Tourette,	La validación del Sistema James One, basado en el uso del chip BLE nRF52832 de Nordic Semiconductor, contra el sistema BIOPAC MP36, se lleva a cabo mediante el monitoreo de 20 sujetos de estudio, donde los resultados de similitud arrojan que, la señal de EDA presenta un 95 % de similitud, mientras que los registro de BPM presentan una correlación mayor al 0.999. Con esto se concluye que el dispositivo James One se puede usar para las aplicaciones anteriormente mencionadas.
Oportunidade s Industria 4.0 en Galicia (Asociación de Centros Tecnológicos de Galicia, 2017).	Temperatura corporal y de la piel, EDA, HR, HRV, movimiento corporal, ocular, bióxido de carbono y concentra-ción de oxígeno en la sangre, EMG, EEG, presión arterial, glucosa y cualquier tipo de señal asociada al sistema nervioso autónomo.	Sistemas de control para exoesqueletos, aplicaciones industriales, militares y en la moda. También, para monitoreo de vida fitness, de salud laboral y diaria, y con fines de entretenimiento. Simulaciones en realidad virtual, aumentada y creación de sistemas de seguimiento. Desarrollo de wearables para protección de extremidades o de cuerpo entero.	Los wearables que se analizan abarcan desde los dispositivos para cabeza, dispositivos de muñeca hasta los que se llevan en otras partes del cuerpo, presentado sus ventajas en diferentes aplicaciones.

BLE se fundamenta en la reducción del consumo de energía y se está posicionando como el estándar clave para dar soporte a la nueva ola de dispositivos wearables, también es la base de aplicaciones centradas en el monitoreo de la salud. Un sensor con soporte BLE puede durar encendido hasta meses si es alimentado con una "pila de botón" (Velasco, 2013).

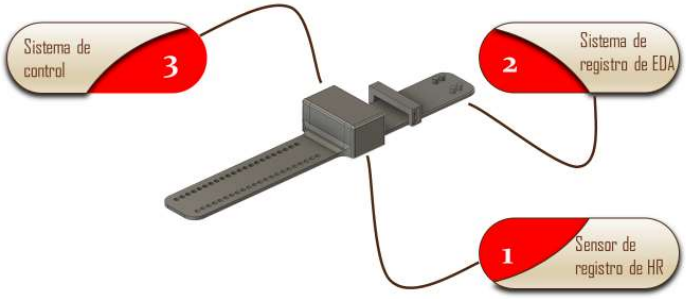


Fig. 1. Componentes del sistema B1.

Fig. 2. Intervalo de tiempo entre pulsos cardíacos (IBI).

2. Estado del arte

En la Tabla 1, se muestra el análisis de diferentes estudios en relación con el uso de wearables y las señales de EDA y de HR denotando para cada estudio o proyecto las señales biomédicas que se registran, las aplicaciones del mismo y una pequeña síntesis que engloba parte de la problemática que se atacó, así como parte de las conclusiones obtenidas. Estos dispositivos se tomaron como referencia para el desarrollo de la B1.

3. Metodología

El sistema B1 consta de 3 partes que describen su funcionamiento, sistema de control, medidor de HR y medidor de EDA, cuya localización en la pulsera se observa en la Fig. 1.

3.1. Sistema de registro de HR

Para llevar a cabo la medición de pulsos cardíacos se hace la diferencia de tiempos en que se da una primera pulsación y una segunda (IBI por sus siglas en inglés), tomando como referencia los niveles más altos en cada ciclo como se muestra en la Fig. 2; este dato se obtiene en milisegundos. Posteriormente se pasa a segundos para obtener como resultado el número total de BPM como se muestra en la ecuación (1 y 2):

$$1\ pulso = X[s], \tag{1}$$

$$BPM = \frac{60[s]}{(X[s])(1000)}, \tag{2}$$

donde X[s] está dado en milisegundos y el factor 1000 de la ecuación (2), resulta del paso de milisegundos a segundos.

Dentro del programa que controla el sistema medidor de HR se hacen consideraciones en las mediciones, para evitar contar como pulsaciones las señales provenientes de ruido debido al movimiento del usuario, tomando en cuenta los rangos de BPM conocidos para un adulto.

Para limitar el mínimo de pulsaciones que puede tener una persona, se toma en consideración que es deportista en estado de reposo y que sus pulsaciones varían entre 40 y 60 como se muestra en la Tabla 2, por lo cual como mínimo de BPM que puede tener una persona adulta se toma 40.

Tabla 2. BPM de personas con diferentes perfiles *(Vice Staff, 2017)*.

	Adulto Sedentario	**Adulto en forma**	**Deportista**
REPOSO (BPM)	Entre 60 y 90	Entre 60 y 80	Entre 40 y 60

Tabla 3. BPM de una mujer en relación con su edad *(Muy en Forma, 2015)*.

EDAD (AÑOS)	**INADECUADO (BPM)**	**NORMAL (BPM)**	**BUENO (BPM)**	**EXCELENTE (BPM)**
20 – 29	96 o más	78 – 94	72 – 76	70 o menos
30 – 39	98 o más	80 – 96	72 – 78	70 o menos
40 – 49	100 o más	80 – 98	74 – 78	72 o menos
50 o más	104 o más	84 – 102	76 – 82	74 o menos

Para el máximo de pulsaciones permitidas se tomó como límite las pulsaciones de una mujer (ya que presentan más pulsaciones por minuto a diferencia que los hombres) en rango de edad 50 o más, es decir, 104 BPM como se muestra en la Tabla 3.

Teniendo los límites definidos en los cuales trabaja el sistema para llevar a cabo las mediciones de pulsaciones cardíacas, se calcula el tiempo que se presenta un pulso cuando se tiene un BPM de 40 (límite inferior de BPM) y cuando se presenta un BPM de 104 (límite superior de BPM).

Para un BPM igual a 40 el valor de IBI queda definido por las ecuaciones (3 y 4):

$$\frac{40[pulsos]}{60[s]} = \frac{1[pulso]}{IBI}, \tag{3}$$

$$IBI = \frac{(60[s])(1[pulso])}{40[pulsos]} = \frac{60[s]}{40} = \frac{3}{2}[s] = 1.5\ [s]. \tag{4}$$

Para un BPM igual a 104, el valor de IBI queda definido por las ecuaciones (5 y 6):

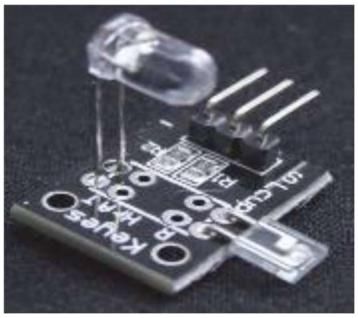


Fig. 3. Sensor de pulso de ritmo cardiaco para dedo.

$$\frac{104[pulsos]}{60[s]} = \frac{1[pulso]}{IBI}, \tag{5}$$

$$IBI = \frac{(60[s])(1[pulso])}{104[pulsos]} = \frac{60[s]}{104} = \frac{15}{26}\,[s]. \tag{6}$$

Si la lectura se encuentra fuera de los rangos inferior y superior, no se toma en cuenta. Las mediciones se llevan a cabo con un sensor comercial capaz de cumplir con las tareas del sistema medidor de ritmo cardiaco de manera satisfactoria, por lo cual, se llevó a cabo un análisis de diferentes sensores en el mercado, los cuales se describen a continuación.

- Sensor de pulso de ritmo cardiaco para dedo. Este sensor, que se muestra en la Fig. 3, consta de un diodo emisor de luz infrarroja y un fototransistor. Funciona emitiendo luz infrarroja mediante un diodo emisor de luz (LED por sus siglas en inglés) y detectando cuanta luz arriba al LED opuesto. Cuando hay un pulso cardiaco la densidad de sangre es mayor en el dedo por lo tanto existe una variación en la luz detectada.

  Especificaciones:

  - Temperatura de operación: 0 a 60ºC.
  - Voltaje de alimentación: 5V.
  - Receptor: Fototransistor
  - Emisor: LED IR.
  - Dimensiones: 24 x 21 x 18 mm
  - Peso: 05 g.
  - Marca: OEM.

  Observaciones: No cumple con los requerimientos para ser montado en una pulsera.

- Sensor de frecuencia cardiaca MAX30100. Este sensor óptico, presentado en la Fig. 4, que deriva sus lecturas de emisión de dos longitudes de onda de la luz a partir de dos dispositivos LED, uno rojo y un puerto de infrarrojos, midiendo la absorbancia de pulso de la sangre a través de un fotodetector. Esta combinación particular de colores LED está optimizado para la lectura de los datos a través

Fig. 4. Sensor de frecuencia cardiaca MAX30100.

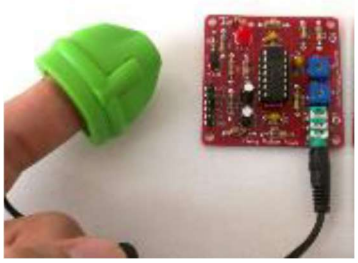


Fig. 5. Easy Pulse Sensor.

de la punta de un dedo. La señal es procesada por una unidad de procesamiento de señales analógicas de bajo ruido y comunicada al circuito integrado a través de la interfaz I2C. Se debe tener en cuenta que las lecturas pueden verse negativamente afectadas por el exceso de movimiento y los cambios de temperatura. También, el exceso de presión puede hacer variar el flujo de sangre y por lo tanto disminuir la fiabilidad de los datos.

Especificaciones:

- Integrado con el MAX30100.
- Interfaz I2C.
- Sensor eficaz.
- Listo para utilizarse.
- Utiliza la fuente de alimentación 3.3V.
- Marca: MikroElektronika.

Observaciones: Su funcionamiento se ve comprometido con el movimiento y cambios de temperatura.

– Easy Pulse Sensor. El sensor Easy Pulse, presentado en la Fig. 5, está diseñado para aplicaciones educativas y de pasatiempo para ilustrar el principio de la fotopletismografía como una técnica óptica no invasiva para detectar la onda cardiovascular del pulso desde la punta del dedo. Utiliza una fuente de luz infrarroja para iluminar el dedo en un lado, y un fotodetector colocado en el otro. Mide las pequeñas variaciones en la intensidad de la luz transmitida.

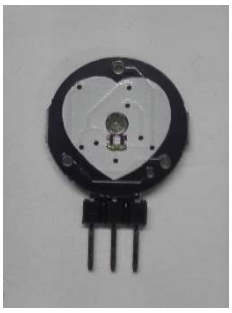


Fig. 6. Pulse sensor.

Especificaciones:

- Utiliza el sensor de transmisión PPG HRM-2511E para lecturas estables.
- MCP6004 Op-amp con capacidad de salida de riel a riel para una máxima oscilación de la señal.
- Salidas analógicas y digitales separadas.
- Potenciómetro de control de ganancia para la salida analógica.
- Control de ancho de pulso para la salida digital.
- Puntos de prueba adicionales a bordo para analizar señales en diferentes etapas de instrumentación.
- El módulo amplificador funciona bien, pero su ancho de banda fue limitado a sólo 3 Hz para eliminar interferencias en la frecuencia de (50 o 60 Hz) productos de cable no blindado.

Observaciones: El sistema solo viene adecuado para medir pulso en el dedo.

– Pulse Sensor. Este sensor, presentado en la Fig. 6, bien diseñado, que mide el ritmo cardíaco de las personas, puede ser utilizado por estudiantes, artistas, atletas, y desarrolladores que quieran incorporar fácilmente datos de frecuencia cardíaca en sus proyectos. Está basado en un LED emisor y un sensor receptor de intensidad, la cantidad de luz reflejada por el dedo cuando hay paso de corriente sanguínea define la salida del sensor. Por lo que es posible visualizar gráfica o numéricamente la información de las lecturas del mismo.

  Especificaciones:

  - Alimentación: 3.3V/5V.
  - Salida: Voltaje analógico.
  - Voltaje de Operación: 3 a 5v.
  - Corriente de Operación: 40ma con 5V.
  - Diámetro de la tarjeta: 10mm.
  - Fácil conexión.

  Observaciones: Fácil implementación y compatibilidad con Arduino. Su forma permite ser usado en el sistema para darle la ergonomía correcta al montaje del sistema.

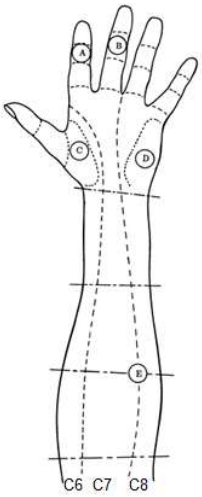


Fig. 7. Distribución de puntos de medición de EDA *(Wolfram, 2012)*.

Después del análisis llevado a cabo y con los resultados mostrados con anterioridad, se decidió usar el sensor llamado Pulse Sensor ya que se considera el más adecuado en cuanto a las características que debe de cumplir para llevar a cabo satisfactoriamente las tareas del sistema medidor de HR.

3.2. Sistema de registro de EDA

El sistema medidor de EDA consta de un circuito cuyo funcionamiento es el mismo que los sistemas detectores de mentiras (Montgomery, 2012) y de dos electrodos de grafeno que permiten medir la respuesta galvánica de la piel. Los electrodos tienen un diámetro de 1 cm.

La colocación de estos electrodos es cerca de la palma de la mano en el dermatoma C7 que se muestra en la Fig. 7. La toma de mediciones es en este punto por el gran número de glándulas sudoríparas en comparación a otras partes de la piel (Márquez, 2014).

El funcionamiento del sistema medidor de EDA inicia con una etapa de acoplamiento, esta se da cuando los electrodos toman mediciones mientras se adecuan a la temperatura de la piel, dichas mediciones se descartan.

3.3. Integración de los sistemas medidores de HR y de EDA en pulsera (Sistema de control)

Se construyó una pulsera la cual se coloca en la muñeca, que cuenta con el sistema medidor de HR, así como el sistema medidor de EDA. También integra el sistema mínimo Arduino Pro Mini 3.3 V a 8 MHz, (ATmega328), con una SRAM de 2 Kbytes (Arduino, s.f.), el cual procesa la información recibida por los sensores y envía esta información vía inalámbrica a un servidor, por lo cual al sistema mínimo se conecta un módulo de comunicación inalámbrica Bluetooth HM-11 (CC2541).

Fig. 8. Pulsera contenedora de sistema medidor de señales biomédicas, el cual se coloca en la caja central.

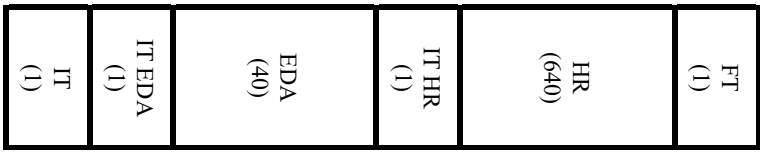


Fig. 9. Trama de envío de datos, donde los números en paréntesis indican la cantidad de datos que se tienen para cada casilla de la trama.

Por último, se tiene una batería de 3.7 V que alimenta al sistema. Todos estos componentes constituyen al sistema medidor de señales biomédicas, el cual, está colocado en una pulsera cuyo diseño fue realizado mediante técnicas de manufactura aditiva utilizando un elastómero de poliuretano termoplástico (TPU por sus siglas en inglés), dicho diseño se observa en la Fig. 8. Se utilizó la tecnología de montaje superficial para la construcción del sistema de control de la B1.

La parte que contiene al sistema de control en la pulsera tiene unas medidas de 3.1 cm x 5.3 cm x 3.3 cm y el sistema B1 tiene un peso de 120 g.

3.3.1. Transmisión de datos

El chip del Arduino Pro Mini por sus características en almacenamiento permite el envío de tramas cada 10 segundos. Las lecturas obtenidas por la B1 son enviadas al servidor, la trama de datos queda estructurada de la siguiente forma; inicia con un dato que se diferencia de todos los demás de la trama y que se denomina como el inicio de la trama (IT), seguido de este dato va otro que da inicio a las lecturas de EDA (IT EDA), posterior a esto se tienen 40 lecturas de EDA, al término de estos datos va el dato correspondiente a inicio de trama de HR (IT HR) seguido de 640 lecturas de HR, y la trama finaliza con el dato correspondiente al fin de trama (FT). Dicha distribución se observa en la Fig. 9.

4. Resultados

Se llevó a cabo el monitoreo de un hombre de 22 años mientras se mantenía en reposo, sentado, con la pulsera B1 colocada en la mano izquierda, y la pulsera E4

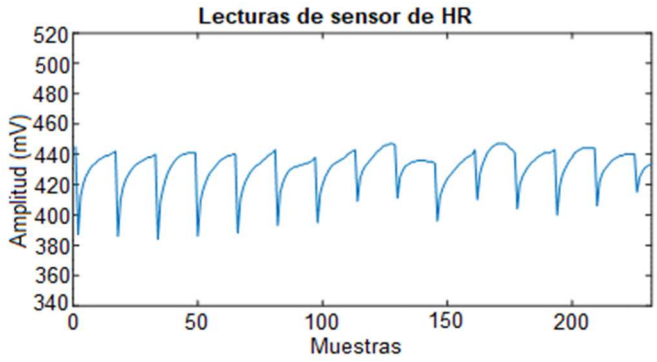


Fig. 10. Señal obtenida de monitoreo con sensor de ritmo cardíaco.

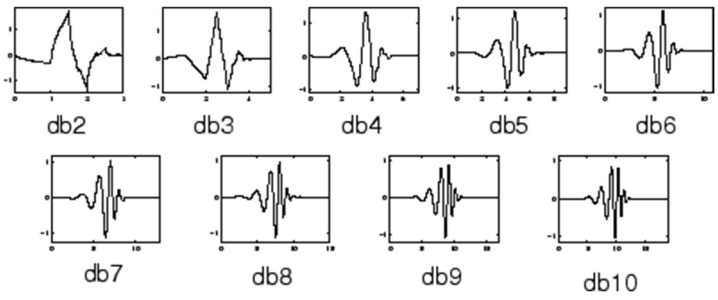


Fig. 11. Familias de wavelets *(MathWorks, s.f.)*.

wristband de Empatica (Empatica, s.f.) colocada en la mano derecha. La E4 es un wearable certificado de grado médico.

El usuario de ambas pulseras se encontraba en un rango menor a los 10 metros de separación con el servidor de recolección de datos, esto para tener lecturas claras de las mediciones de las señales fisiológicas de la B1, ya que al salir de esta zona de cobertura se presentan errores en el envío de las lecturas debido a la presencia de otras ondas de radiofrecuencia que interfieren en el canal de comunicación.

El monitoreo fue realizado con la intención de llevar a cabo un chequeo de HR y EDA del sujeto de estudio con ambas pulseras de forma simultánea para evaluar las mediciones de las señales de la B1 contra las obtenidas por un sistema comercial y certificado.

La señal obtenida del monitoreo de HR de la B1 que fue a una frecuencia de muestreo igual a 64 Hz (misma que usa el sistema de la E4 wristband), no presenta de forma clara los picos más altos de cada representación de un pulso cardíaco como se muestra en la Fig. 10, por lo que se filtró mediante el uso de wavelets.

Por la similitud de esta señal con la wavelet madre de la familia Daubechies, que se presenta en la Fig. 11, se decidió utilizar la tipo 2.

Utilizando la herramienta Wavelet Analyzer de MATLAB® (MathWorks, s.f.) para el análisis de la señal, se utilizaron los detalles arrojados en el primer nivel de descomposición, ya que muestra mejores resultados como se muestra en la Fig. 12 a comparación de los detalles y aproximaciones de un nivel 2.

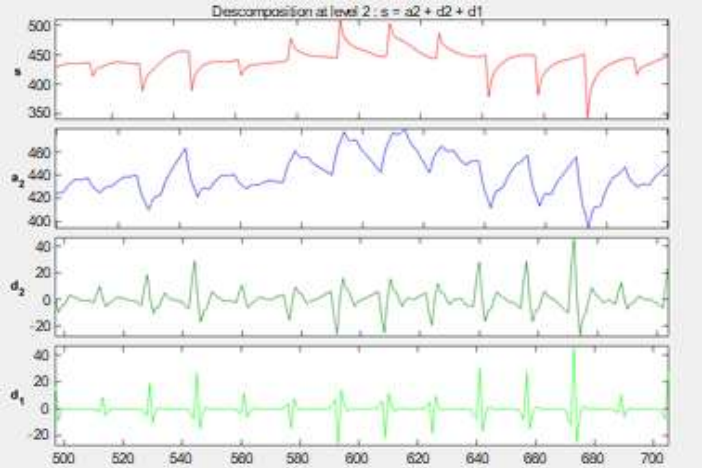


Fig. 12. Descomposición de señal mediante la herramienta Wavelet Analyzer.

Tabla 4. BPM obtenidas en registro simultáneo analizando 8 minutos.

Número de minuto	E4 wristband	B1
1	60.3778	65
2	60.6903	62
3	60.6746	60
4	60.5340	60
5	60.1121	60
6	60.4403	60
7	60.2215	60
8	60.5965	60

Al obtener los BPM durante 8 minutos se obtuvieron los resultados presentados en la Tabla 4, donde se observan minuto a minuto los cambios y el contraste en las mediciones obtenidas con la B1 y la pulsera E4.

En los primeros dos minutos del análisis, se observa un mayor conteo de pulsaciones por parte del sistema B1 debido a que requiere de un periodo de acoplamiento. Para el monitoreo del sujeto de estudios, la E4 ya había pasado su periodo de acoplamiento, mientras que la B1 inició su registro sin dicho periodo.

Descartando los dos primeros minutos de acoplamiento, los BPM se aproximan más en ambos sistemas teniendo diferencias de 2.579 BPM más en la E4 que en la B1, lo que equivale al 0.40769 %, las cuales se dan ya que el sistema de la B1 no está diseñado para dar lecturas decimales de conteos de pulsos, es decir, solo cuenta la presencia de pulsos completos. Tomando esto en cuenta no se tiene diferencias entre las mediciones obtenidas por ambos sistemas.

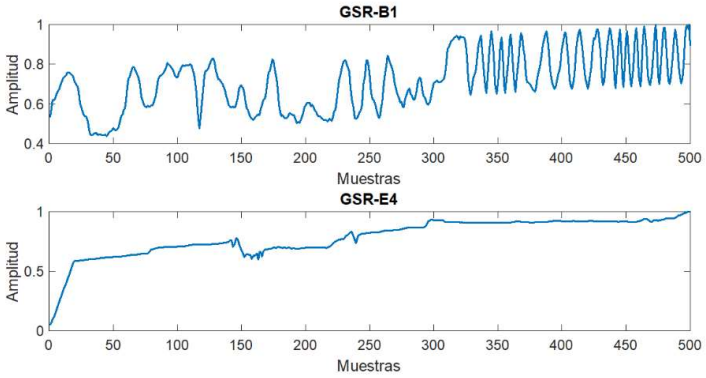


Fig. 13. Señales obtenidas de EDA en ambos sistemas de medición.

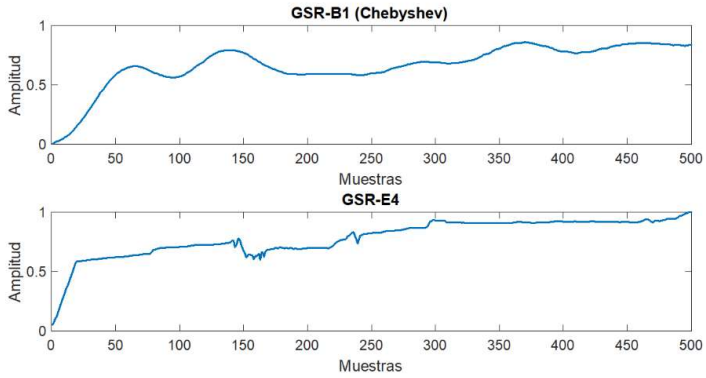


Fig. 14. Comparación de comportamientos de señal filtrada obtenida con la B1 y señal obtenida con la E4 wristband.

La señal resultante del monitoreo correspondiente a 500 muestras obtenidas a una frecuencia de muestreo igual a 4 Hz (misma que utiliza el sistema de la E4) por parte del sistema medidor de EDA de la B1 se muestra en la Fig. 13 en la parte superior, mientras que en la parte inferior se muestra la señal obtenida con la E4.

Para eliminar el ruido de la señal EDA obtenida con el sistema B1, se filtró utilizando un filtro Chebyshev tipo 2 pasa bajas, con frecuencia de corte en 0.05 Hz de quinto orden cuyo resultado se observa en la Fig. 14 en la parte superior, mientras que en la parte inferior se presenta la señal obtenida con el sistema de medición de la E4.

Para tener una comparación cuantitativa de estas señales, se segmentaron ambas señales en ventanas con una duración de 3 segundos (equivalente a 12 muestras por la frecuencia de muestreo igual a 4 Hz) ya que J Braithwaite en 2015 reportó que la ventana de la latencia va de 1 a 3 segundos que es donde se presentan variaciones abruptas en el nivel electrodérmico (Braithwaite, Watson, Jones, & Rowe, 2015). Estas ventanas son rectangulares y se desplazan cada segundo (4 muestras), de esta forma se tiene un traslape de dos segundos (8 muestras).

A cada segmento de datos se le calcula la energía implementando la ecuación (7):

$$E = \sum_{n=1}^{12} |x(n)|^2 , \quad (7)$$

donde x(n) son las muestras del segmento de la señal a analizar y *n* representa el número de muestras que se tiene por cada segmento.

De esta forma la diferencia que existe en todo el registro de EDA tomado con ambos sistemas es de 1.8549 Joules lo que equivale al 0.1823 %, esto quiere decir que la energía obtenida en el registro tomado con la E4 wristband presenta 1.8549 Joules más que la energía obtenida de la señal registrada con la B1.

5. Conclusiones

El desarrollo de la B1 por las señales que mide y por obtener el almacenamiento de las lecturas crudas de EDA y de la señal fotopletismográfica, permite aplicar diversos tratamientos dependiendo de la aplicación, como la medición de niveles de estrés, inicio de crisis epilépticas, monitoreo del comportamiento de personas con alguna enfermedad, entre otras que se mencionan en el estado del arte.

Los resultados obtenidos de las comparaciones para señales de EDA con el sistema certificado muestran una similitud de 99.8177 % en el registro de un sujeto durante 125 segundos, mientras que, el resultado reportado en "Validation of Wireless Sensors for Psycophyiological Studies" (Silva Moreira, Chaves, Dias, Dias, & R. Almeida, 2019), muestra un porcentaje de similitud de 95 %, utilizando mediciones de 20 sujetos, durante 15 segundos. Aunque el porcentaje de similitud obtenido con la B1 es mayor, cabe mencionar que solo se registró un sujeto bajo las mismas condiciones que las reportadas en la mencionada referencia.

Las diferencias obtenidas en el análisis del registro de HR muestra que en los primeros 2 minutos, los valores distintos de BPM se deben al periodo de acoplamiento del sensor, por lo cual, se toma como duración de 2 minutos el tiempo que tarda este proceso, ya que, a partir del minuto 3 se encuentran mediciones de BPM estables. Las diferencias que a partir del minuto 3 se presentan se deben a que el sistema de la B1 no fue diseñado para dar lecturas fraccionarias del conteo. Comparando resultados con la E4 se observó una similitud de 99.5923 % mientras que el estudio mencionado reporta una similitud de 99.9 %.

Como trabajo a futuro se harán modificaciones al diseño de la pulsera para mejorar su ergonomía. De igual forma se trabajará en la miniaturización de todo el sistema mediante la implementación de sistemas embebidos más adecuados.

También se trabajará en el diseño de un nuevo filtro para la señal de EDA que permita ver mayor resolución, tales como se presentan en la E4.

Referencias

1. Arduino: Arduino Store. https//store.arduino.cc/usa/arduino-pro-mini (2019)
2. Asociación de Centros Tecnológicos de Galicia: Oportunidades industria 4.0 en Galicia. Estudio IGAPE, http://igape.es/es/ser-mas-competitivo/galiciaindustria4-0/estudos-e-informes/item/download/71_7e5c2e1b8028489a94dce74bae90ec15 (2017)

3. Aveiga-Paini, C.E., Criollo Altamirano, B.G., Cruz-Quijije, A.M.: Monitoreo del ritmo cardiaco a través de dispositivos móviles. Ciencias de la salud, IV(2), pp. 3−19 (2018)
4. Braithwaite, J., Watson, D., Jones, R., Rowe, M.: A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRS) for psychological experiments. Birmingham (2015)
5. Celi, G., Rocha, M., Yapur, M.: Mediciones fotopletismográficas. Tesis de posgrado Escuela Superior Politécnica del Litoral (2015)
6. Consinfin: ¿Qué es la comunicación inalámbrica (wireless)?. http://consinfin.com/que-es-la-comunicacion-inalambrica-wireless/ (2012)
7. Díaz-Robledo, L., Sánchez, R.J.: La actividad electrodérmica de la piel como indicador de activación psicofisiológica en pilotos de caza españoles: Un estudio preliminar. Sanidad Militar, 74(1), pp. 7−12 (2018)
8. Empatica: Empatica. https://empatica.com/research/e4/ (2020)
9. Facultad de Medicina: El corazón como bomba: fases del ciclo cardíaco. Universidad Nacional Autónoma de México (2019)
10. Foroz: ¿Qué es la transmisión de datos vía bluetooth?. https://foroz.es/que-es-la-transmision-de-datos-via-bluetooth.html (2017)
11. Hernández-Aquino, R.: Diseño, simulación y construcción de antenas tipo parche para buetooth y WI-FI, bandas 2.4 ghz y 5.8 ghz. Tesis Universidad de las Américas Puebla (2008)
12. Mandal, A.: News Medical Life Sciences. In: Cashin-Garbutt, A. (Ed.), ¿Cuál es el ritmo cardiaco?. https://news-medical.net/health/What-is-Heart-Rate-(Spanish).as px (2019)
13. Márquez, F.J.: Diseño de un sistema de reconocimiento de estrés en seres humanos. Tesis Universidad Nacional Autónoma de México (2014)
14. MathWorks: Introduction to Wavelet Families. https://mathworks.com/help/ wavelet/gs/introduction-to-the-wavelet-families.html (2018)
15. MathWorks: Matlab. https://mathworks.com/products/matlab.html (2018)
16. Mojica-Ledoño, A.G.: Actividad electrodérmica aplicada a la psicología: Análisis bibliométrico. Revista Mexicana de Neurociencia, XVIII(4), pp. 46−56 (2017)
17. Montgomery, S.: Make: Community. The Truth Meter. https://makezine.com/ projects/the-truth-meter-2/ (2012)
18. Muy en forma: Pulsaciones en reposo. https://muyenforma.com/pulsaciones-en-reposo.html (2019)
19. Nieto, N., Vega, M.L.: Diseño de un prototipo de medición de señales fisiológicas utilizadas en Biofeedback. Proyecto Integrador, Universidad Nacional de Córdoba (2017)
20. Rodas, G., Pedret-Carbadillo, C., Ramos, J., Capdevila, L.: Variabilidad de la frecuencia cardiaca: Concepto, medidas y relación con aspectos clínicos (1). Archivos de medicina del deporte, XXV(123), pp. 41−47 (2008)
21. Sapienza Universita Di Roma: Brain Signs: Respuesta galvánica de la piel (GSR). https://brainsigns.com/es/science/s2/technologies/gsr (2018)
22. Shelley, K., Shelley, S.: Pulse oximeter waveform: Photoelectric plethysmography. Clinical Monitoring: Practical Applications for Anesthesia and Critical Care, pp. 420−423 (2011)
23. Silva-Moreira, P., Chaves, P., Dias, R., Dias, N., Almeida, P.: Validation of wireless sensors for psychophysiological studies. Sensors, XIX(4824), pp. 1−24 (2019)
24. Taj-Eldin, M., Ryan, C., O´Flynn, B., Paul, G.: A review of wearable solutions for physiological and emotional monitoring for use by people with autism spectrum disorder and their caregivers. Sensors, 18(4271), pp. 1−29 (2018)
25. Velasco, J.: Hipertextual. ¿En qué consiste Bluetooth LE?: https://hipertextual.com/2013/12/que-es-bluetooth-le (2013)
26. Vice Staff: Guía básica para ser mejor deportista: La frecuencia cardiaca. https:// vice.com/es/article/vv9pwa/guia-basica-para-ser-mejor-deportista-la-frecuencia-cardi aca (2017)
27. Wolfram, B.: Electrodermal Activity. Wuppertal, Springer (2012)

28. Zephyr: Medtronic. https://zephyranywhere.com/resources/hxm (2013)

Modelo computacional para el análisis de la calidad del aire en interiores

Christian Olvera García[1], José Juan Carbajal Hernández[2], Víctor Manuel Landassuri Moreno[1], Miguel Ángel Olvera García[2]

[1] Universidad Autónoma del Estado de México, Centro Universitario del Valle de México, México

[2] Instituto Politécnico Nacional, Centro de Investigación en Computación, México

{christian.olvera38, landassuri, ma.olvera58}@gmail.com, jcarbajalh@cic.ipn.mx

Resumen. El presente estudio propone la creación de un modelo computacional para el análisis de la calidad del aire en interiores. Mediante el monitoreo continuo de parámetros como PM_{10}, $PM_{2.5}$, TVOC, CH_2O, CO_2, temperatura y humedad relativa, se establecen criterios de evaluación de niveles perjudiciales para la salud humana. Asimismo, se establecen los criterios de calidad del aire en interiores que son definidos mediante la conjunción de las evaluaciones de cada parámetro y mismos que corresponden con 5 niveles: excelente, buena, regular, mala y peligrosa calidad.

Palabras clave: Calidad del aire, modelo, sistema, monitoreo.

Computational Model for the Analysis of Indoor Air Quality

Abstract. This study proposes the creation of a computational model for the analysis of indoor air quality. Through the continuous monitoring of parameters such as PM10, PM2.5, TVOC, CH2O, CO2, temperature and relative humidity, evaluation criteria are established for levels that are harmful to human health. Likewise, the indoor air quality criteria are established, which are defined by the conjunction of the evaluations of each parameter and which correspond to 5 levels: excellent, good, regular, poor and dangerous quality.

Keywords: Air quality, model, system, monitoring.

1. Introducción

Hoy en día es muy importante la evaluación de la calidad del aire en interiores, en virtud de que se han observado elevados niveles de contaminantes en edificios públicos

como oficinas, escuelas y hospitales [1]. La falta de ventilación en interiores causa una mala calidad del aire y provoca mayor incidencia del síndrome del edificio enfermo [2].

En la Ciudad de México, existen 1717 edificios, todos ellos cerrados y con poca ventilación donde el aire exterior influye dramáticamente sobre el aire interior [3,4].

La falta de mantenimiento a los ductos de ventilación y circulación del aire implica un gasto energético y económico [5]. Esto genera la acumulación de polvo y la creación de gérmenes, bacterias como el *staphylococcus aureus* y ácaros [6,7].

Incluso en los trenes subterráneos se han detectado niveles altos de contaminantes [8,9]. Sin embargo, en México no existe un índice de la calidad del aire interior normativo similar al IMECA, el cual es utilizado para la evaluación de la calidad del aire exterior. Esto hace necesario la implementación de un modelo de evaluación de la calidad del aire interior, que evalúe los niveles de contaminantes perjudiciales.

Sin embargo, debemos señalar que como no existe una norma mexicana (NOM) para los límites máximos permisibles, es necesario utilizar los límites propuestos por la Agencia de Protección Ambiental de Estados Unidos (USEPA) [10], la Sociedad Estadounidense de Ingenieros de Calefacción, Refrigeración y Aire Acondicionado (ASHRAE) [11] y Liderazgo en Energía y Diseño Ambiental (LEED) [12] para los principales contaminantes del aire interior como las partículas mayores a 2.5 y 10 micras ($PM_{2.5}$ y PM_{10}), los compuestos orgánicos volátiles (TVOC), el formaldehido (CH_2O) y el dióxido de carbono (CO_2).

Derivado de esto, este trabajo propone la creación de un modelo computacional para la evaluación de la calidad del aire en interiores, tomando como base el índice IMECA y el uso de técnicas de procesamiento digital de señales.

Como resultado se obtiene un indicador capaz de evaluar los 5 parámetros y establecer un nivel correspondiente a la buena o mala calidad del aire en espacios interiores.

2. Materiales y métodos

2.1 Parámetros de la calidad del aire

De acuerdo con la literatura, diversos organismos han dado seguimiento a la evaluación de la calidad del aire en interiores, estableciendo los materiales tóxicos en casas y edificios como son: los humos de tabaco, las partículas de materia de 10 y 2.5 micrómetros, compuestos orgánicos volátiles, monóxido de carbono, dióxido de carbono, metano, formaldehido, radón, fibras de asbesto, fungicidas, pesticidas, etc. Sin embargo, para estandarizar los contaminantes que existen en los lugares cerrados.

Se utilizaron los límites máximos permisibles propuestos por la USEPA, la ASHRAE y la LEED [13] para los principales contaminantes, como son: PM_{10}, $PM_{2.5}$, CO_2, CH_2O, TVOC's. Además, debemos señalar los principales daños a la salud que provocan cada uno de los contaminantes, como a continuación se detallan en la Tabla 1.

Tabla 1. Importancia de los parámetros de calidad del aire en el organismo humano.

Parámetro	Importancia
PM_{10}	Estas partículas se sedimentan en las vías respiratorias superiores, dando origen y penetran a los alvéolos pulmonares, causando el agravamiento de las enfermedades respiratorias y cardiovasculares, ya que pueden desencadenar efectos carcinógenos.
$PM_{2.5}$	Estas partículas pueden penetrar las vías respiratorias inferiores, causando diversas enfermedades de tipo respiratorio, como la bronquitis, y más recientemente se ha demostrado sus efectos sobre las enfermedades de tipo cardiovascular.
TVOC's	Estos compuestos orgánicos al ser inhalados se adhieren a los tejidos grasos y no pueden disolverse en nuestro organismo. Por tanto, estos pasan al torrente sanguíneo y se acumulan en diversos órganos del cuerpo y ponen en riesgo la vida, sobre todo en las mujeres embarazadas ya que afecta directamente al feto en el desarrollo embrionario.
CH_2O	El formaldehído en niveles bajos puede producir irritación a la piel, los ojos, la nariz y la garganta. Sin embargo, si se llega a beber grandes cantidades de formaldehído puede causar un profundo dolor, vómitos, coma, y posiblemente la muerte.
CO_2	Puede afectar la función respiratoria y provocar excitación seguida por depresión del sistema nervioso central.

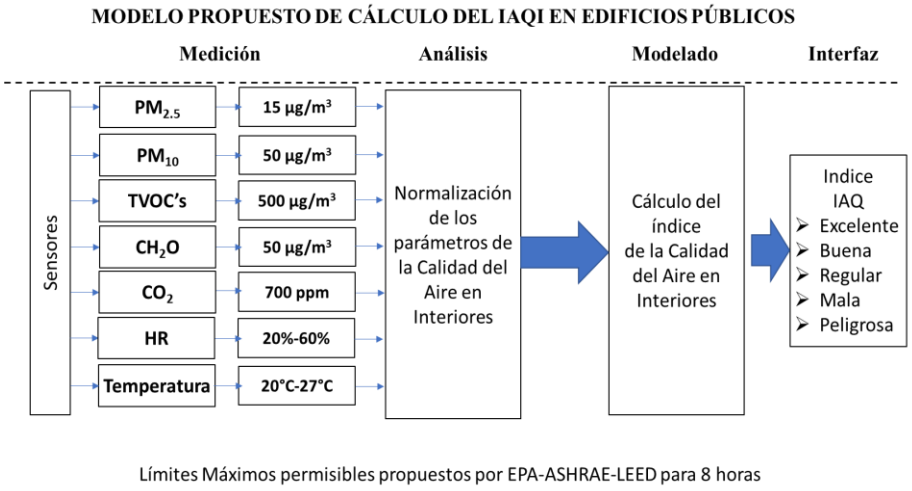


Fig.1. Arquitectura del sistema de monitoreo de la calidad del aire en interiores.

2.2 Modelo de la evaluación de la calidad del aire en interiores

El Modelo de Evaluación de la Calidad del Aire consta de 4 fases: medición, análisis, modelado y finalmente la interfaz de usuario que muestra el cálculo del Índice de Calidad del Aire en Interiores (IAQ) [14]. La fase de medición consta de un módulo de 7 sensores correspondientes a los parámetros necesarios para analizar la calidad del aire. En la fase de análisis, los niveles obtenidos son clasificados de acuerdo con los límites máximos permisibles.

En la fase de modelado los valores obtenidos son evaluados para obtener el índice de calidad del aire. En la última fase, la Interfaz gráfica muestra la clasificación y

Tabla 2. Intervalos de concentración para asignación de niveles de calidad del aire.

IQA	$PM_{2.5}$ $\left[\mu g/m^3\right]$	PM_{10} $\left[\mu g/m^3\right]$	TVOC´s $\left[\mu g/m^3\right]$	CH_2O $\left[\mu g/m^3\right]$	CO_2 [ppm]
0-50	0-7.5	0-25	0-250	0-25	0-350
51-100	7.6-15.2	26-50	251-500	26-50	351-700
101-150	15.1-22.5	51-75	501-750	51-75	701-1050
151-200	22.6-30.0	76-100.0	751-1000	76-100.0	1051-1400
>200	>30.0	>100.0	>1000	>100.0	>1400

resultados de las mediciones y calidad del aire obtenido. A continuación, en la Fig.1 se detalla el modelo de cálculo para la evaluación de la calidad del aire en interiores.

2.3 Sistema de monitoreo de la calidad del aire en interiores

a) Aspectos generales

Para la implementación del modelo de la calidad del aire en lugares cerrados, es necesario conocer los límites máximos permisibles y tiempos de exposición a los contaminantes. Sin embargo, debemos mencionar que en México no se han establecido las Normas Oficiales para evaluación de la calidad del aire en interiores. Por tanto, para proponer un modelo que permita evaluar la calidad del aire en interiores, se utilizaron los límites máximos propuestos por la USEPA, la ASHRAE y la LEED. Estas instituciones definen los límites máximos permisibles para los principales contaminantes del aire interior. Los límites máximos propuestos son: 15µg/m^3 para las $PM_{2.5}$, 50 µg/m^3 para las PM_{10}, 500µg/m^3 para el TVOC's, 50 µg/m^3.para el CH_2O, 700 ppm para el CO_2. Finalmente hay que comentar que la temperatura se estableció entre 20.3°C - 23.3°C en invierno y 23.9°C - 26.9°C en verano. La humedad relativa se estableció entre el 30% y 50%. Con estos límites máximos permisible se establece la fase de medición de los parámetros.

Además, para implementar la fase de la Interfaz de usuario, se establecieron 5 niveles de calidad del aire en interiores como son: excelente (0-50), buena (51-100), regular (101-150), mala (151-200) y peligrosa (>200) de acuerdo con la normalización utilizada en el índice IMECA como se muestra en la Tabla 2.

Cada parámetro es medido con una frecuencia de muestreo de 15 minutos. Sin embargo, en ocasiones las mediciones pueden perderse por diferentes cuestiones. Para obtener una base de datos consistente es necesario tener todo el conjunto de mediciones acorde a la frecuencia de muestreo, por lo que una interpolación lineal permitirá restaurar aquellos datos faltantes acorde con la siguiente expresión:

$$s = s_1 + (t - t_1)\frac{(s_2 - s_1)}{(t_2 - t_1)}, \tag{1}$$

donde s refiere a la concentración por calcular. Las mediciones se muestran por cada hora, por lo que antes de evaluar cada parámetro, es necesario realizar un promedio móvil de la siguiente forma:

$$\hat{s}(t) = \frac{1}{n}\sum_{t=1}^{n} s\,(t-1). \tag{2}$$

Para implementar la fase de análisis se hará la normalización de los valores se realiza mediante la aplicación de las fórmulas de forma similar a las utilizadas en la normalización del índice IMECA, como se detalla a continuación.

Indicador para partículas menores a 2.5 micras ($PM_{2.5}$):

$$I[PM_{2.5}] = C[PM_{2.5}] * \frac{100}{15}. \tag{3}$$

Indicador para partículas menores a 10 micras (PM_{10}):

$$I[PM_{10}] = C[PM_{10}] * \frac{100}{50}. \tag{4}$$

Indicador para compuestos orgánicos volátiles (TVOC):

$$I[\mathrm{TVOC}] = C[\mathrm{TVOC}] * \frac{100}{500}. \tag{5}$$

Indicador para formaldehído (CH_2O):

$$I[CH_2O] = C[CH_2O] * \frac{100}{50}. \tag{6}$$

Indicador para dióxido de carbono (CO_2):

$$I[CO_2] = C[CO_2] * \frac{100}{700}. \tag{7}$$

Finalmente, para implementar la fase de modelado hacemos el cálculo de la calidad del aire se puede definir como:

$$IAQ = max\{I[\mathrm{PM_{2.5}}], I[\mathrm{PM_{10}}], I[\mathrm{TVOC's}], I[\mathrm{CH_{20}}], I[\mathrm{CO_2}]\}. \tag{8}$$

3. Resultados

3.1 Adquisición de datos

Para la evaluación del modelo, se obtuvieron un total de 7**0**56 datos, correspondiente a una semana de medición. 6 datos por hora, y un dato por cada parámetro. Además, debemos señalar que, para obtener el dato de cada parámetro por hora, fue necesario

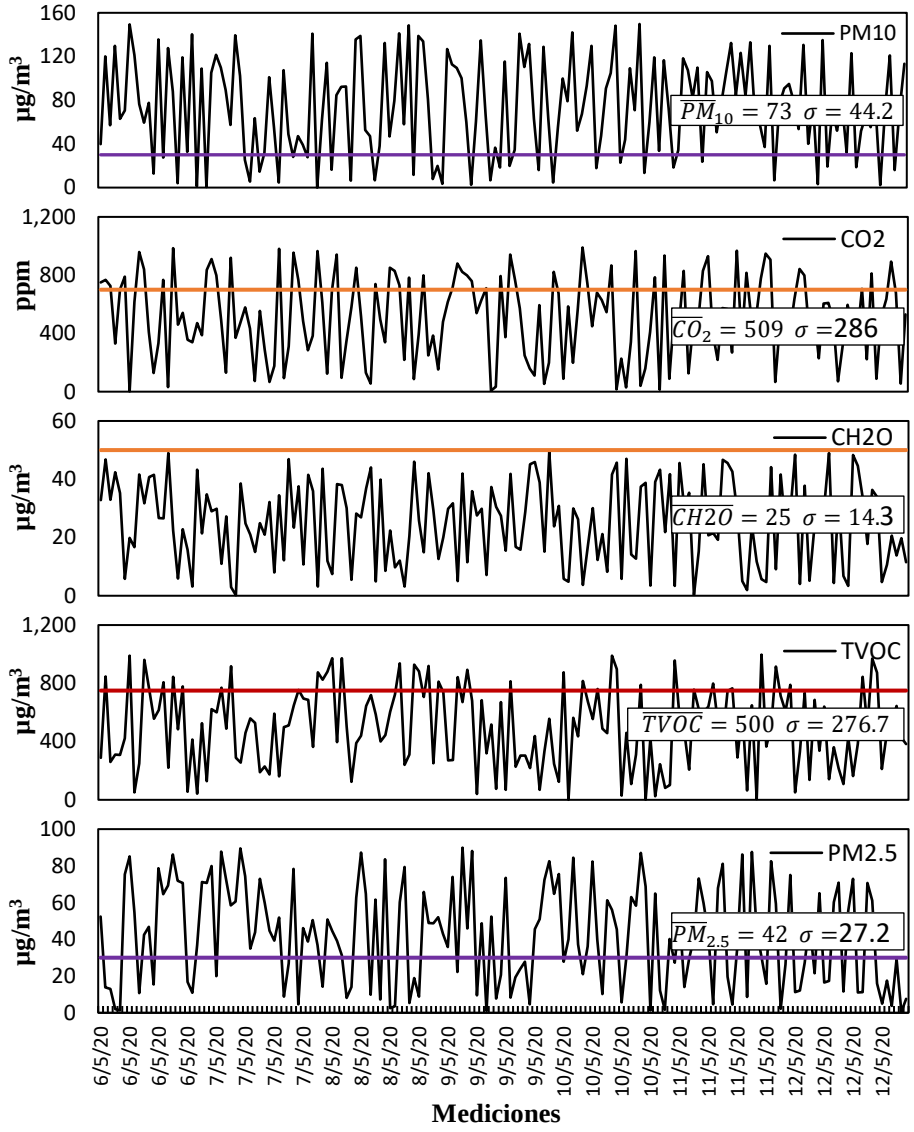


Fig. 2. Conjunto de mediciones de parámetros de la calidad del aire al interior durante una semana de mediciones.

tomar 6 muestras de cada parámetro a cada 10 minutos y obtener el promedio de las mediciones para evaluar el dato a cada hora durante los 7 días de la semana.

Obteniendo una base de datos con 168 datos por parámetro a cada hora y con un total de 1,176 datos. Para efectos de graficar los datos de manera representativa sólo se muestran 24 datos de un día en cada gráfico.

La arquitectura del sistema está basada en un módulo de sensores MoreSunsDIY, permite medir $PM_{2.5}$, PM_{10}, TVOC's, CH_2O, CO_2, Temperatura y Humedad Relativa.

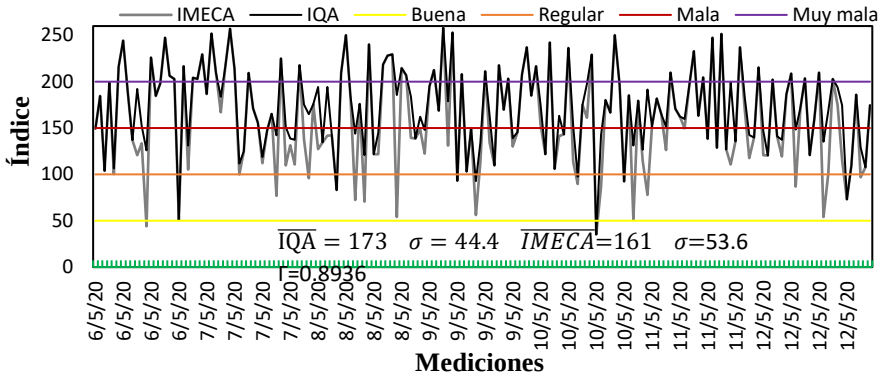


Fig. 3. Evaluación y comparación de la calidad del aire en interiores empleando el índice propuesto y el IMECA.

La medición de parámetros se obtiene con rangos de error de $PM_{2.5}$ (±10%), PM_{10} (±10%), TVOC's (±7%), CH_2O (±8%), CO_2 (±2%), Temperatura (±1%) y Humedad Relativa (±3%). Por los rangos de error, este es un sistema confiable para el monitoreo de la calidad del aire. Cuenta con una la interfaz FT232R USB que utiliza el protocolo UART con 5 volts de alimentación para hacer la conexión a la computadora, Mediante la Interfaz VISA de National Instruments es posible leer los datos mediante la configuración de un puerto serial a 9600 baudios, bit de paro, sin paridad y sin control de flujo de datos.

3.2 Análisis de parámetros

En la Fig. 2 se puede observar el conjunto de datos medidos durante un día. En términos generales podemos observar que el parámetro CH_2O es el que muestra valores más aceptables en un rango de *bueno*, CO_2 se presenta niveles de *regular*, TVOC *malo* y PM_{10} con $PM_{2.5}$ niveles *peligrosos*. Estos estados permiten deducir que el resultado de la calidad del aire presentara estados peligrosos para la salud humana.

3.3 Evaluación de la calidad del aire

En la Fig. 3 se puede observar los resultados de la evaluación del índice IAQ y la comparación contra el índice IMECA. En este caso se puede observar algunas diferencias con el IMECA, debido a que éste último solo considera tres de los 5 parámetros involucrados, mismo que a su vez muestra algunas diferencias en la evaluación, en donde valores menos penalizados se obtienen debido a la falta de las evaluaciones en los parámetros faltantes. De esa forma, se generan evaluaciones que podrían se catalogadas como de buena calidad del aire, cuando en las evaluaciones del IAQ se observa lo contrario.

En la Fig. 3 se puede observar el conjunto de datos medidos durante la semana de prueba y sus respectivas evaluaciones del IAQ, pudiendo observar claramente el comportamiento del IAQ esta por encima de los 150 puntos por tanto la calidad del aire fue mala durante el periodo de prueba. resultante.

4. Conclusiones

El presente trabajo pretende sentar las bases para mejorar la calidad del aire en edificios cerrados, en donde reciben grandes concentraciones de CO_2, $PM_{2.5}$ y PM_{10}. Actualmente, los esfuerzos por notificar acerca del estado de la calidad del aire se enfocan a exteriores, en donde se proponen programas de reducción de movilidad automotriz como método de contingencia. Sin embargo, para interiores aún hace falta mucho trabajo y difusión para generar normas y procedimientos que permitan mejorar las condiciones de salud en espacios cerrados. Las direcciones futuras en esta área de investigación estarán enfocadas a estudiar parámetros adicionales que intervienen en la calidad del aire al interior.

Referencias

1. Aserea, L., Mols, T., Blumberga, A.: Assessment of indoor air quality in renovated buildings of Liepłja municipality. Energy Procedia, 91, pp. 907–915 (2016)
2. Thach, T.Q., Mahirah, D., Dunleavy, G., Nazeha, N., Zhang, Y., Hui-Tan, C., Roberts, A.C., Christopoulos, G., Soh, C.K., Car, J.: Prevalence of sick building syndrome and its association with perceived indoor environmental quality in an Asian multi-ethnic working population. Building and Environment, 166, pp. 106420 (2019)
3. Meier, R., Schindler, C., Eeftens, M., Aguilera, I., Ducret-Stich, R.E.: Modeling indoor air pollution of outdoor origin in homes of Sapaldia subjects in Switzerland. Environment International, 82, pp. 85–91 (2015)
4. Hwang Sung Ho, Seo Sung Chul, Yoo Young, Ki Yeon Kim, Choung Ji Tae, Park Wha Me: Indoor air quality of daycare centers in Seoul, Korea. Building and Environment, 124, pp. 186–193 (2017)
5. Renaud-Salis, L.C., Abadie, M., Wargocki, P., Rode, C.: Towards the definition of indicators for assessment of indoor airquality and energy performance in low-energy residential buildings. Energy and Buildings, 152, pp. 492–502 (2017)
6. Asif, A., Zeeshan, M., Hashmi, I., Zahid, U., Faraz-Bhatti, M.: Microbial quality assessment of indoor air in a large hospital building during winter and spring seasons. Building and Environment, 135, pp. 68–73 (2018)
7. Hernández-Castillo, O., Mugica-Álvarez, V., Castañeda-Briones, M.T., Murcia, J.M., García-Franco, F., Briseño, Y.F.: Aerobiological study in the Mexico City subway system. Aerobiologia, 30, pp. 357–367 (2014)
8. Bin Xu, Jinliang Hao: Air quality inside subway metro indoor environment worldwide: A review. Environment International, 107, pp. 33–46 (2017)
9. Mugica-Álvarez, V., Figueroa-Lara, J., Romero-Romo, M., Sepúlveda-Sánchez, J., López-Moreno, T.: Concentrations and properties of airborne particles in the Mexico City subway system. Atmosphere Environment, 49, pp. 284–293 (2012)
10. Agencia de Protección Ambiental de Estados Unidos (USEPA): Indoor Air Quality (IAQ). https://epa.gov/indoor-air-quality-iaq (2020)
11. Sociedad Estadounidense de Ingenieros de Calefacción, Refrigeración y Aire Acondicionado (ASHRAE): Indoor air quality guide. https://ashrae.org/ technical-resources/bookstore/indoor-air-quality-guide (2020)
12. Liderazgo en Energía y Diseño Ambiental (LEED): What indoor air quality parameters do you need to know?. https://iotacommunications.com/blog/indoor-air-quality-parameters/ (2020)

13. Phillips, H., Handy, R., Sleeth, D., Thiese, M.S., Schaefer, C., Stubbs, J.: Taking the "LEED" in indoor air quality: Does certification result in healthier buildings?. Journal of Green Building, 15 (3), pp. 55–66 (2020)
14. Li, H., You, S., Zhang, H., Zheng, W., Zheng, X., Jia, J., Ye T., Zou, L.: Modelling of AQI related to building space heating energy demand based on big data analytics. Applied Energy, 203, pp. 57–71 (2017)

Data Migration in Graph-oriented Databases

Soumaya Boukettaya[1,3], Ahlem Nabli[1,2], Faiez Gargouri[1]

Sfax University, MIRACL Laboratory,
Institute of Computer Science and Multimedia,
Tunisia

Al-Baha University, Faculty of computer sciences and information technology,
Kingdom of Saudi Arabia

Faculty of Economic Sciences and Management, Sfax University,
Tunisia

faiez.gargouri@isims.usf.tn,soumayaboukettaya@gmail.com,
ahlem.nabli@fss.usf.tn

Abstract. Data is expanding at a rapid pace these days, and dealing with it has become incredibly challenging. Since they allow for the storage of various data structures, NoSQL graph databases are becoming more popular. Nonetheless, due to their schema-less nature, improper data migration and manipulation during the query phase might result in significant data loss. This paper deals with data migration within NoSQL graph databases in which we propose a graph matching algorithm based on similarity measures. We also adopt a lazy data migration approach to ensure a low cost of data migration and avoid critical data loss.

Keywords: Graph databases, data migration, similarity measures, graph matching, nodes similarity, relationships similarity.

1 Introduction

NoSQL data models are different from the relational model in terms of structure and capacity. NoSQL data models offer great flexibility due to their schema-free nature. As one of the NoSQL data models, the use of graph databases has increased significantly. Graph databases are database models that are based on the graph structure, essentially nodes and edges. NoSQL graph databases come with the benefit of handling large volumes of heterogeneous and semi-structured data. Yet, data management in such databases is a challenging task.

Data under databases with a pre-defined scheme, such as relational databases, can be migrated in a version-controlled sequence by saving each schema transformation alongside its data migration. While data under schema-free databases still require careful migration techniques.

Graph databases have a schema-free nature. Thereby, schemes are implicit and modified directly by managing data instances without any specific constraints on any such manipulations.

Primarily, a data management task of a graph database does not require a dedicated team of database administrators. It is often the responsibility of Application teams or business units.

A common practice to handle data and schema management is to write custom migration scripts to migrate data eagerly (all data migrated on one go when the database structure changes) or lazily (migrating only data being acceded to).

When migrating data eagerly, data is accessed all at once. That reduces the data latency. Nevertheless, migrating all data at once can take a long time and requires a shut down the access to the database. With lazy data migration, legacy data that no longer needs any changes is not accessed. Yet, this strategy has a high data latency.

This paper deals with data migration and evolution in graph databases. It propose an approach that helps to manage data migration taking into account not only the flexibility provided by such databases but also the nature of the graph databases.

The remainder of the paper is structured as follows: Section 2 overviews of the state-of-the-art that treated data migration in the field of NoSQL databases and specifically graph databases. Section 3 contains some required preliminary definitions. Section 4 explains our approach for data migration and details our process of lazy data migration.The proposed approach is based on similarity measures and graph matching. We performed experiments and addressed the evaluation results in section 5. Section 6 concludes the paper and presents some future works.

2 Related Works on Data Migration

Data migration has been an area of active research with a long history [10, 11, 13, 1]. There are essentially two main data migration strategies to guaranty accurate data migration and evolution and recently, three new data migration strategies are developed.

In the following, we start by presenting the data migration strategies. Then, we overview related works on data migration within the context of NoSQL databases.

2.1 Data Migration Strategies

There are essentially two data migration strategies to guaranty accurate data migration and evolution such as *Eager data migration* and *Lazy data migration*. The authors in [12] present three more migration strategies such as *incremental migration*, *predictive migration*, and *adaptive migration*.

Eager data migration: with this strategy, all entities within the database are migrated at once. Though this strategy has a low access cost (latency) , it procures a high migration cost as some of the data will be updated even though they will not be accessed in the future. This strategy is best when migrating

data from different DBMS (E.g., from a relational model to graph model). The key issue of this strategy is that the data, even data that may not be usable again in the future, must be kept up to date.

Lazy data migration: with this strategy, all data remain unchanged until being accessed. This strategy has no immediate migration costs and ensures flexibility to agile requirement changes. Lazy data migration strategy aims to minimize migration costs. Nevertheless, when compared to the eager data migration strategy, data access latency can be relatively high.

Incremental data migration: Incremental data migration strategy works similarly to lazy data migration strategy, i.e., only data that needs to be changed is accessed. Nonetheless, lazy migration periods are regularly interrupted to clean the database which reduce run-time overhead caused by updating legacy data on-the-fly. For these interruptions, eager data migration over legacy entities is performed regularly.

Predictive data migration: Predictive data migration strategy highlights the frequently accessed data. It keeps track of past data accesses while ordering the accessed entities accordingly via exponential smoothing. This established technique in time series data weighs the entities by their actuality and access frequency. [12] present a detailed study of the different data migration strategies.

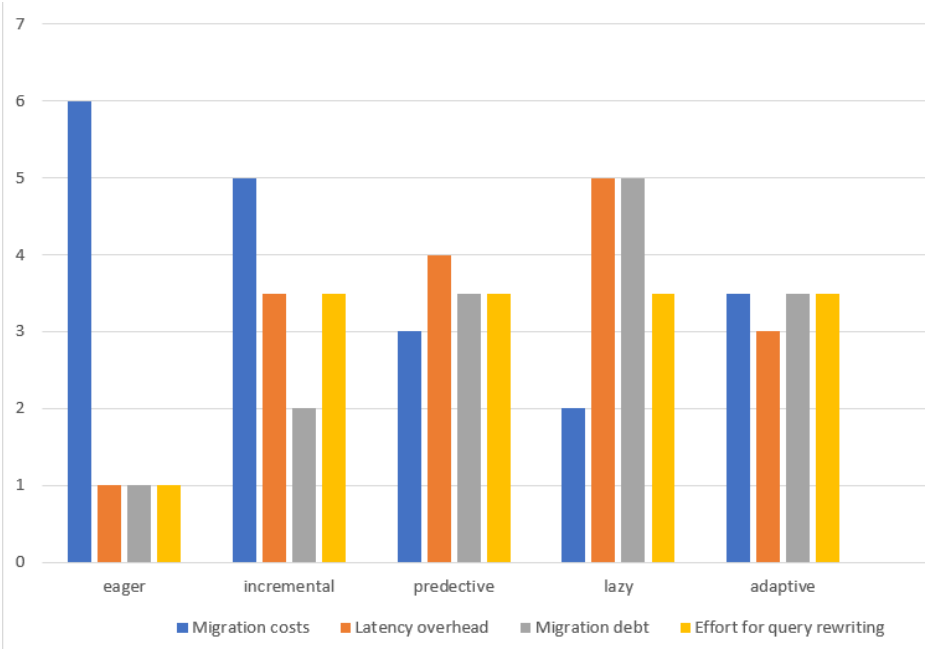


Fig. 1. Characteristics of the different data migration strategies.

Figure 1 compare the different data migration strategies with regards to the *migration costs*, *latency*, *migration debts* and *effort for query rewriting* [12].

The term *latency overhead* refers to the time needed to retrieve the data. In [12], *migration cost* is a term that refers to the charges occasioned by migrating the data, and *migration debt* refers to the changes needed to be invested to migrate data to a homogeneous data structure.

Both *eager data migration* and *incremental data migration* have a high migration cost due to migrating all entities at once. *Lazy data migration* has the lowest data migration costs. However, it has the highest latency overhead and migration cost.

2.2 Data Migration in NoSQL Databases

Data migration in NoSQL databases is a relatively new research field. Regardless, researchers conducted many works in this area, from managing JSON-based files to schema extraction and data migration.

The authors in [4, 5] present a framework τJSchema that for the definition and validation of temporal JSON documents that conform to a temporal JSON schema. The authors proposed a versioning technique that provides a complete set of low-level and high-level change operations. Both at the instance and schema levels, τJSchema fully supports temporal versioning of JSON-based Big Data.

The work presented in [14] details a study of the evolution of the domain model of applications built against a NoSQL data store. This methodology is applied to ten real-world database applications. They start by extracting the schema, then analyze the entire project history. Lastly, they analyze the evolution of the NoSQL database schema.

In [15], the authors propose a framework called *Datulation* that support *lazy* and *eager* data migration strategies using Datalog rules. Datalog is a programming language for deductive databases. The work highlights the merit of the lazy data migration strategy. The framework supports adding, renaming, or removing attributes for entities. When carrying out data migration eagerly, the framework evaluate Datalog rules bottom up. As for lazy migration, *Datulation* evaluates Datalog rules top-down, thus migrating only the legacy entities that are deployed by the application. One strong point of Datalution is that it can lazily roll forward chains of schema changes. Yet, this tool only supports data with prefixed schema and written in JSON format.

In [9], the authors demonstrate *ControVol Flex*, an Eclipse plugin for controlled schema evolution in Java applications backed by NoSQL document databases. Based on a lazy migration strategy. This tool helps to migrate the NoSQL by adding Morphia annotations to the code. It also supports an automatic version-numbering mechanism for the different stages of the schema evolution process. This tool keeps track of the various schema versions that occur in the production data store. It also support *lazy* and *eager* data migration strategies. The advantage of *ControVol Flex* is that it allows to carry out eager and lazy

data migration concurrently, which is vital for the continuous deployment of zero-downtime applications.

In [12], the authors present a methodology of self-adapting data migration which focuses on data migration itself. The framework presented is based on schema management middleware *Darwin* and a tool-based advisor *MigCast*. *Darwin* supports schema extraction, schema evolution, and data migration of data stored in a NoSQL database. *Darwin* support all four data migration strategies such as eager, lazy, incremental, and predictive strategy. *Darwin* also supports popular NoSQL database management systems, such as MongoDB, Couchbase, Cassandra, and the multi-model database ArangoDB. The choice of the best data migration to be applied is made by *MigCast*: a tool for self-adapting data migration strategy. *MigCast* helps to explore the different data migration strategies in order to examine the effects of the data migration strategies with regard to the metrics of migration costs, latency, and migration debt.

The work [6] emphasizes the importance of tracking the history of data changes within the graph database. The authors present a plug-in that delivers a novel representation of historical graph data using graph versioning techniques.This work features the specific structure of the graph database. It proposes an approach to represent history related to both (i) nodes (track down changes that occurred to the entity itself) and (ii) relationships (track down the changes to the relation including start or end node updates). The authors in this work present the different versions of data separately in another graph called *VersionGraph*. *VersionGraph* stores the history of different versions of a data graph.

The authors of [7] introduce a method based on in-memory architectures to retrieve the structural schema of a graph database. The authors focused on schema extraction in the context of semi-structured graph data. They extend the existing methods to manage large NoSQL graph databases. They introduce several types of summaries and provide the methods to extract them. The authors present four types of summaries, such as structural summaries, structural data summaries, structural data key property summaries, structural data key-value property summaries.

The work [16] presents a solution for missing data in the NoSQL graph databases by e introducing a novel approach for mining gradual patterns in the presence of missing values. The focus of this paper is the extraction of gradual patterns from property graphs on *IntraNode-Label* gradual patterns, that is the pattern extraction process is to be performed among the properties/attributes of the "same label" nodes. The work presented in [2] proposes a data definition language (DDL) schema for property graphs inspired by Cypher query language to handle schema validation and schema evolution for graph databases. The work presented a mathematical framework that allows enforcement schema and expresses propagation from schema to instance and vice versa. Table 1 summarizes data migration works in NoSQL databases.

Table 1 compares the presented works with regard to the data storage type, the different data migration strategies they support, and whether or not they

Table 1. Data migration and schema extraction in NoSQL databases.

Works	Data type	Migration Strategy	schema extraction	schema versioning
[4, 5]	JSON documents	-	✓	✓
[14]	NoSQL databases	Eager, Lazy, Incremental, Predictive	✓	-
[15]	NoSQL databases	Lazy	✓	-
[9]	NoSQL databases	Lazy	✓	-
[12]	NoSQL databases	Eager, Lazy, Incremental, Predictive, Self-adapting	✓	-
[6]	Graph databases	-	✓	✓
[7]	Graph databases	-	✓	-
[16]	Graph databases	-	✓	-
[2]	Graph databases	-	✓	✓

support schema extraction and schema versioning. being an important step of data management, all works presented offer schema extraction techniques. even though most of them support different types of NoSQL databases, yet only a few of them emphasize the graph database structure in terms of nodes and relationships.

2.3 Discussions

Most works use data stored in document-store databases or data of JSON format as input. Adopting this strategy when dealing with graph data excludes the benefits of using graph theory and graph algorithms that may help reduce the run time overhead. Moreover, they treat data entities similarly and don't consider the graph structure (nodes, and relationships). That is explained due to the structure of links (relationships) between different entities as they are different with each NoSQL data model. Thus, most works support data only and exclude the relationships between entities.

Additionally,they emphasize schema extraction or schema validation. Though schema proved to be relevant when dealing with data history or data migration, depending on schema extraction or schema validation to deal with data migration goes against the fundamental idea of having a schema-free database nature.

Graph matching methods are widely used for subgraph matching or extraction. They also proved to be very useful for pattern recognition and biological and biomedical database where graph representation of data is used. NoSQL graph databases are already based on graph theory. To make sure to highlight the structure of the graph database, proposing a solution based on the graph matching technique seems very promising. By definition, graph matching is the problem of finding a similarity between graphs. Thus, we propose a solution based on string similarity measures and graph matching to help migrate data correctly within the graph database.

3 Preliminary Definitions

A graph database (GDB) is formed by a set of nodes, relationships, properties, and labels. Both nodes and their relationships are named and can store properties. These properties are represented by key/value pairs. Nodes and relationships can be labeled. The edges between two graphs representing the relationships have two qualities: they always have a start node and an end node and are directed making the graph a directed graph. Relationships can also have properties.

Definition1 (graph): A graph G is defined by a pair (V, E), where V is a set of vertices and E is a set of edges with $E \subseteq (VxV)$

Definition2: (directed graph is defined by a pair (V, E), where V is a set of vertices and E is a set of edges with $E \subseteq (VxV)$ and where all the edges are directed from one vertex to another.

Definition 3 (Graph database): A graph database is a couple (N, R), where N is the total set of nodes and which form the entities of GDB and R as the set of relations that join the different nodes.

Definition 4 (Nodes): Each node n is composed of its identifier id_n, a set of properties P_n, and a set of labels L_n. It should be noted that the identifier does not contain any semantic information. Semantic is usually expressed through one or more labels and a single property is a couple of as (key/ value) pair. A node can be written as follows: $\forall n \in N; n = (id_n, P_n, L_n)$.

Definition 5 (Relationships): A relation r is defined as $(id_n, Strn, Endn, T, P)$ which contains the identifier id, the start node Strn/ end node Endn, the type of relation T, and its set of properties P.

4 LD-MIG: Graph Data Migration Approach

In order to control data migration under graph databases, we propose a lazy data migration approach based on graph matching. The data migration approach under graph database is a process composed of four phases (i): Analysis phase, (ii): Graph matching, (iii): MOs identification, and (iv): Graph merging as shown in figure 2. In the following, We assume that we have two graphs GDB (the database that is currently deployed) and the *operation-graph*(the new graph to add) where N (respectively M) are the sets of nodes of GDB (respectively the *operation-graph*) and R (respectively W) are the sets of relationships of GDB (respectively the *operation-graph*).

4.1 Analysis Phase

The first step in our approach is to examine the input queries responsible for any database changes. The aim of this phase is to generate the *Operation-graph* from the query to be applied on the database or the source code. This step intend to analyze the composition of different CRUD(create, read, update and delete) queries, extract the input data, and build a graph model based on the operations

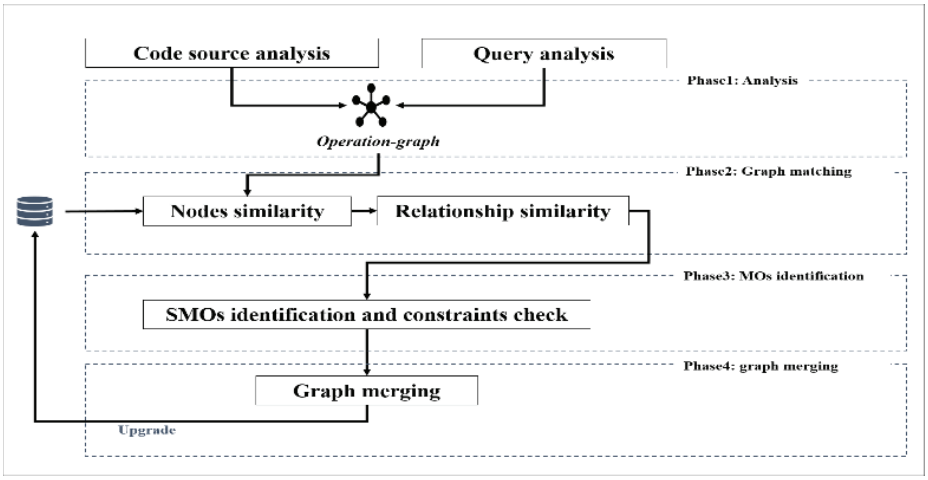


Fig. 2. Overall approach.

to be performed on the graph database (GDB). The generated graph is called an *operation-graph (OPG)*.

We choose to work with Neo4j as one of the most popular graph databases management systems that use Cypher a SQL- inspired query language that describes visual patterns in graphs using ASCII-Art syntax. Figure 3 presents an example of an *operation-graph* created based on a "Merge" query expressed by *Cypher* in *Neo4j*.

Fig. 3. Merge query with its corresponding *operation-graph*.

4.2 Graph Matching

In this step, we aim to extract a sub-graph from GDB that is most similar to the operation graph. The extracted sub-graph will serve as the entry of the third phase. A graph database presents the data as it is conceptually viewed in the form of nodes linked by relationships. Therefore, we propose a process composed of node-based similarity measures, relation-based similarity measures, and graph matching step based on Levenshtein edit distance to detect the most

similar sub-graph. As mentionned above, the Graph matching phase is composed by three main steps ***(i): nodes similarity, (ii): relationships similarity, (iii): sub-graph matching***.

Given the nature of the graph database, we can safely consider that node's labels, properties, and relationship types are presented as strings. Therefore, it is suitable to propose a string-based similarity, embedded in the process of structure graph matching, to identify the similarity between both graphs.

Nodes Similarity We assume that we have two graphs GDB (the database that is currently deployed) and the *operation-graph*. N (respectively M) are the sets of nodes of GDB (respectively the *operation-graph*) and R (respectively W) are the sets of relationships of GDB (respectively the *operation-graph*). The similarity between two nodes (m,n) depends on the similarity between their labels and their properties.

Labels similarity. To compute the similarity between two labels we simply apply the Levenshtein distance between them. Taking into account that a node may have more than one label, the similarity of the labels between two nodes is calculated by comparing each label of m to each label of n and then taking the maximum value obtained. Formula 1 determines the similarity between a label $l_m \in L_m$ of a node $n \in N$ with all the labels of a given node n and returns the maximum:

$$sim(l_m, L_n) = \max_{j=1}^{k} \big(1 - \frac{lev(l_m, l_{nj})}{max(length(l_m), length(l_{nj})}\big) \tag{1}$$

To determine the similarity between all labels L_m of a node m with a given node $n \in N$ we apply algorithm 1.
As input for algorithm 1we have two node's labels L_m and L_n and the overall

Algorithm 1: Simlabels.

Input: L_n, L_m
Output: simls
; ▷ `With` L_n `as the labels of the node n,` L_m `as the labels of the node m and simls as the similarity value between the labels of m and the labels of n`

1 $k \leftarrow \min\big(\mid L_m \mid, \mid L_n \mid\big)$
2 $x \leftarrow$ *the node with the minimum set of labels*
3 $y \leftarrow$ *the remaining node*
4 $simls \leftarrow simls\big(L_m, L_n\big) = \frac{\sum_{i=1}^{k} SimL(l_{xi}, L_y)}{k}$

similarity *simls* as an output, (line4) calculate the average similarity using the *simls* formula.

Properties similarity. Node property is composed of a key/ value pair. To measure the similarity between two properties we compute the similarity of

their keys and then for the maximum value obtained, we measure the similarity between their values. To calculate the similarity between two pairs of keys k_{mi} as a key of a property of the node m (respectively k_{nj} as a key of a property of the node n) we apply the following formula:

$$SimK(k_{mi}, k_n) = \max_{j=1}^{h} \big(1 - \frac{lev(k_{mi}, k_{nj})}{\max(length(k_{mi}), length(k_{nj}))}\big) \tag{2}$$

$$SimValue(v_{mi}, v_{nj}) = \big(1 - \frac{lev(v_{mi}, v_{nj})}{\max(length(v_{mi}), length(v_{nj}))}\big) \tag{3}$$

Formula 3 computes the similarity of property key k_{mi} of the node m to all properties of the node n. Formula 4 computes the similarity between the value of k_{mi} and the most similar key from the node n.

The global similarity between both of the properties is computed by the average of the keys and values similarity. Algorithm 2 describes the process of computing the properties similarity.

Algorithm 2: Simproperties.

Input: P_m, P_n
; ▷ P_m and P_n are the properties of the nodes m and n
Output: simps
; ▷ simps the similarity value between the properties of m and the properties of n

1 $z \leftarrow \min\big(\mid P_m \mid, \mid P_n \mid\big)$
2 $x \leftarrow$ *the node with the minimum set of labels*
3 $y \leftarrow$ *the remaining node*
4 **foreach** $i \in P_x$ **do**
5 **foreach** $j \in P_y$ **do**
6 $dictkey \leftarrow \{xkey : i, ykey : j, simkey : SimK(k_{xi}, k_y)\}$; ▷ dictkey is a dictionary that contains i (current key of x), j (current key of y) and their similarity value using SimK(kxi,ky)
7 $maxdictkey \leftarrow \{xkey : i, ykey : j, simkey : MaxSimK(k_{xi}, k_y)\}$; ▷ maxdictkey is a dictionary that contains the maximum similarity of i (current key of x) to all the keys of y
8 $maxdictkey \leftarrow \{xkey : i, ykey : j, simkey : MaxSimK(k_{xi}, k_y), simvalue : SimValue(v_{xi}, v_{yj}), prop_sim : maxpro\}$; ▷ add the value similarity (SimValue) and the overall similarity (max pro) to maxdict
9 $simps(P_m, P_n) \leftarrow \frac{\sum_{i=1}^{z} maxdict['maxpro']}{z}$

As an input for algorithm 2, we have the sets of properties of the nodes m and n. The output will be simps of all properties (similarity value between the properties of m and the properties of n). After determining the node with the minimum set of properties (lines:1,2,3) we calculate the maximum keys similarity and store

them in a dictionary where *(dictionaries are unordered, changeable collections that have key/value pairs used in the python programming language.)* (line 4 to 7) Then, for each key in Maxdict, the corresponded value similarity is computed (line 8). Finally (lines 9) computes the overall properties similarity.

Nodes similarity. The nodes similarity is calculated by invoking the two algorithms *Algorithm1* simlabels (L_m, L_n) and *Algorithm2* simproperties (P_m, P_n) as follows:

$$nodessim(m,n) = \frac{simlabels(L_m, L_n) + simproperties(P_m, P_n)}{2}. \quad (4)$$

4.3 Relationship Similarity

As stated previously, a graph database is characterized by its nodes and relationships. Relationships are the key entities used in the database to express semantic. A relationship is declared by a directed edge and defined by its type, start/end nodes, and eventually properties. We measure the similarity between the two relationships by measuring the similarity between their types (names) and properties.

Relationship type similarity. We assume that we have two relationships r = (id_r, $Strn_r$,$Endn_r$, T_r, P_r)and w = (id_w, $Strn_w$,$Endn_w$, T_w, P_w) respectively belongs to GDB and the *operation-graph*. The first step is to verify the existence of the relationship by measuring the similarity between both their types. We simply apply the Levenshtein distance function to compute the similarity between their types (names). It is to note that a relationship can have one type presented as a string. Formula 6 calculates the relationship-types similarity.

$$Simtyperelation(T_w, T_r) = \big(1 - \frac{lev(T_w, T_r)}{\max(length(T_w), length(T_r))}\big). \quad (5)$$

Properties similarity. To compute the similarity of the relationship's properties, we reuse *algorithm2* of the node's similarity.

In this section, we mainly focus on proposing a subgraph matching algorithm for specifying the similarity between the graph database (GDB) and *Operation-graph*. This algorithm 3 employs the similarity measures we previously suggested. Algorithm 3 have as input both GDB (the current database in use) and OPG (the *Operation-graph*. The algorithm go through every node in OPG and compare it with all nodes of GDB. Everytime a high similarity between two nodes m and n is detected (lines 3 and 4), we compute the similarity between their relationships (line 5 to line 12). The output is a dictioanary *SIMDICT* containing the entities of GDB that are most similar to OPG.

nodes similarity and graph matching phases are published recently in [3]. Further details concerning the nodes and relationships similarity are discussed in [3].

Algorithm 3: Graph matching.

Input: GDB, OPG
; ▷ with GDB as the database in use and OPG as the opertation-graph
Output: $SIMDICT$
; ▷ $SIMDICT$ is a dictionary contaning the diffrent nodes and relationships of OPG and the most similar nodes and relationships of GDB a long with their similarity values

foreach $m \in OPG$ **do**
  foreach $n \in GDB$ **do**
    $NSIM \leftarrow nodessim(m, n)$;
    while *(*$NSIM > threashold$*)* **do**
      $R \leftarrow$ extract all relationships of n and assign them to a list R;
      $W \leftarrow$ extract all relationships of m and assign them to a list W;
      foreach $w \in W$ **do**
        foreach $r \in R$ **do**
          $TRSIM \leftarrow Simtyperelation(T_r, T_w)$;
          $PSIM \leftarrow simproperties(P_r, P_w)$;
          ********** $ENSIM \leftarrow nodessim(Endn_r, Endn_w)$;
          $RelSim \leftarrow \frac{TRSIM+PSIM+ENSIM+NSIM}{4}$;
          while *(*$RelSim > threashold$*)* **do**
            $SIMDICT \leftarrow$
            $SIMDICTid_{mi}, w, id_n, r, NSIM(n, m), RelSim(w, r)$

4.4 MO's Identification Phase and Data Migration Process

Having a schema-free nature, while managing data in a graph database, we implicitly manage its schema. The most fundamental data manipulation that a graph database DBMS offers are:

- **Create/Add**: refers to create a new entity in a database.
- **Update**: refers to updating an existing entity in the database.
- **Delete**: refers to deleting an existing entity from the database.

GDBMS are error-prone due to the lack of schema restrictions. In the following section, we propose a set of operations for correctly migrating data in GDBMS and a global data migration algorithm. This step aims to identify the set of operations needed to convert an OPG entity to a GDB entity and ensure a correct data migration. Let OPG and GDB be two graphs where (M, W) are the sets of nodes and relationships belongs to OPG and (N, R) are the sets of nodes and relationships belongs to GDB.

A homomorphism $h : OPG \longrightarrow GDB$ is a function $h_\eta : M \longrightarrow N$ and a function $h_\varepsilon : W \longrightarrow R$, mapping nodes and relationships (M,W) of OPG to nodes and relationships (N,R) of GDB. It is to be noted that h_η and h_ε are the sets of operations required to safely migrate M and W to N and R. In the following, we present in detail a set of modification operations. In the scope of this paper, we cover the basic CRUDE operations such as Add and Update.

Add Operation. This Operation requires the non-existence of the entity in GDB. Let e be the entity from OPG to migrate e can be a node m or a relationship w.

Algorithm 4: Procedure: addEntity.

```
   Input: (e, GDB)
1  ∀n, r ∈ GDB,  e ∈ OPG  NodesSim(m, n) = 0 or RelSim(w, r) = 0 while
   (NodesSim(e,n)=0) do
2  |  Add m(L_m, P_m) to GDB.
3  while (RelSim(e,r)=0) do
4  |  if (Strn_w and Endn_w∃ GDB) then
5  |  |  Add only T_w and P_w to OPG
6  |  else
7  |  |  Add both nodes and the relationship to GDB
```

The addEntity aims to add a new entity from OPG that does not exist in GDB. If the entity to add is a relationship then, we verify the existence of its start/ end nodes first (lines 3 to 7).

Update operation : Updating an entity (a node or a relationship) can be done by adding, updating, retyping, or even deleting its elements (node labels, node properties,relationship properties or relationship types). Algorithm 5 illustrate an example of updating a node property.

Algorithm 5: Procedure: updateEntity.

```
   Input: (m, GDB)
1  ∀n, r ∈ GDB,  e ∈ OPG  NodesSim(m, n) >= threashold or
   RelSim(w, r) >= threashold while (NodesSim(m, n) >= threashold) do
2  if (p_m P_n) then
3  |  add the new property p_m to the existing node n
4  else if (p_n P_m) then
5  |  delete the property p_m from the existing node n
6  else
7  |  replace the existing property p_m with p_n
```

Data Migration Process : the data migration process takes as input the most similar entities from GDB to OPG, then it apply the diffrent modification operations over them. As it only deals with entities being modified, our data migration process is concidered to a lazy data migration process. algorithm 6 detilas the process to migrate data correctly.

Algorithm 6: Procedure: Lazy data migration.

Input: $(SIMDICT, GDB, OPG$
; ▷ `with GDB as the database in use, OPG as the opertation-graph and SIMDICTis a dictionary contaning the diffrent nodes andrelationships of OPG and the most similar nodes andrelationships of GDB a long with their similarity values`

1 **while** *(*$NodesSim(m,n) < threashold$*)* **do**
2 addEntity(m,GDB)
3 **while** *(*$RelSim(w,r) < threashold$*)* **do**
4 addEntity(w,GDB)
5 **while** *(*$NodesSim(m,n) >= threashold$*)* **do**
6 **if** *(*$\{P_m\} \cap \{P_n\}=\emptyset$*)* **then**
7 updateEntity(p_m, GDB)
8 **else if** *(*$\{L_m\} \cap \{L_n\}=\emptyset$*)* **then**
9 updateEntity(l_m, GDB)
10 **else**
11 break;
12 **while** $RelSim(w,r) >= threashold$ **do**
13 **if** *(*$NodesSim(Strn_w, Strn_r) < threashold)OR$ *(*$NodesSim(Endn_w, Endn_r) < threashold$*)* **then**
14 addEntity($Endn_m$,GDB)
15 **else**
16 updateEntity($Strn_w$,GDB);
17 updateEntity($Endn_w$,GDB)
18 **if** *(*$\{P_w\} \cap \{P_r\}=\emptyset$*)* **then**
19 updateEntity(p_w, GDB)
20 **else**
21 updateEntity(P_w,GDB)

Algorithm6 takes as input the simiarity dictioanry SIMSICT that contains the most similar entities of GDB and its similarity measures. Then, it compares each measure to a prefixed threshold. In case that the similarity between two entities is low, we then add the entity from OPG to GDB (lines 1 to 4). In case two entities have a high similarity values we then update the existed entity in GDB depending on the dissimilarity that both entities may have.

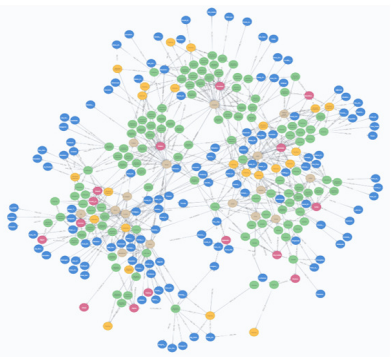
5 Evaluations: LDBC-SNB Benchmark

To evaluate the proposed approach, we carried out a set of experiments on the LDBC-SNB benchmark. First, we present an overview about the LDBC-SNB, we then present our evaluation results.

5.1 LDBC-SNB: Overview

LDBC's Social Network Benchmark (LDBC-SNB) is an effort intended to test various functionalities of systems used for graph-like data management. For this, LDBC SNB uses the recognizable scenario of operating a social network, characterized by its graph-shaped data.

A detailed description of the benchmark can be found in the initial publications [8, 17]. The model features individuals and their actions in a social network over time. It describes the structure of the data in terms of nodes and their relationship. The initial databases contain 29,192 nodes and 39,800 relationships.



Nodes	29192
Relationships	39800
Labels	18
Relationship Types	15
Property Keys	20

Fig. 4. Snapshot of the LDBC-SNB database.

5.2 Evaluations

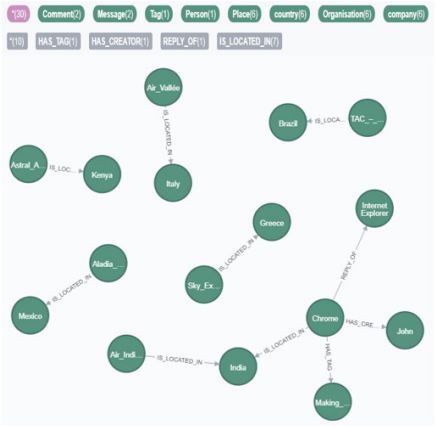
We carried out two evaluations: ***(i) an expremental evalutaion*** in which we extracted randomly ten relationships with their nodes (OPG1). We first applied some changes to the extracted entities (OPG1), then we carried off our program, and finally, we verified whether the changes made are correct in the GDB. And ***(ii) a second evaluation*** in which we variate the number of entities of the *Operation-Graph*. The evaluation is done based on the run time. the datasets used as *Operation-graphs* are:

- The first dataset (OPG1): contains 30 entities randomly extracted from the database GDB.
- The second, THird and fouth datasets (OPG2, OPG3,OPG4): contains respectively 60,150 and 200 entities.
- For the fifth, sixth and seventh datasets (OPG5, OPG6,OPG7), we increased the number of entities and alter more modifications to them. The datasets contains respectively 500, 700,1000 entities.

– The eighth, ninth and tenth datasets (OPG8, OPG9,OPG10) contains respectively 50, 100 and 200 entities.

It is to note that the entities of datasets (OPG1 to OPG7) are randomly extracted from the GDB. to be able to test our approach, we altered modifications to each of these datasets. Datasets (OPG8, OPG9 and OPG 10) contains entities that are very dissimilar to the original database GDB. We used Neo4j as a graph database management system and python 3.7 as a programming language. In this paper, we have carried out two types of experimentation. The *(i): first evaluation* in which we used the first dataset as *operation-graph* to evaluate the correctness of our data migration approach. And the *(ii): second evaluation* in which we used all the datasets and for each set, we compute the run time of graph matching and the data migration process.

Experimental Evaluation We extracted ten random relations with theirs corresponding nodes from the initial database GDB (OPG1-1). Figure 5 displays the nodes and relationships of OPG1-1.



Nodes	16
Relationships	10
Labels	8
Relationships types	4

Fig. 5. Snapshot of OPG1-1.

Then we altered some modifications over some of the nodes and relationships (OPG1-2). We changed about 30% of the dataset. An example of modifications are presented as follows:

– changing the label "Person" of some nodes to "User".
– changing the value of the property "creationDate" of some nodes.
– changing the relationship type "HAS-TAG" to "H-TAG".

We then performed our data migration process to check the data correctness and migration run time after migrating the data. Figure 6 is a showcase of the database after the data migration.

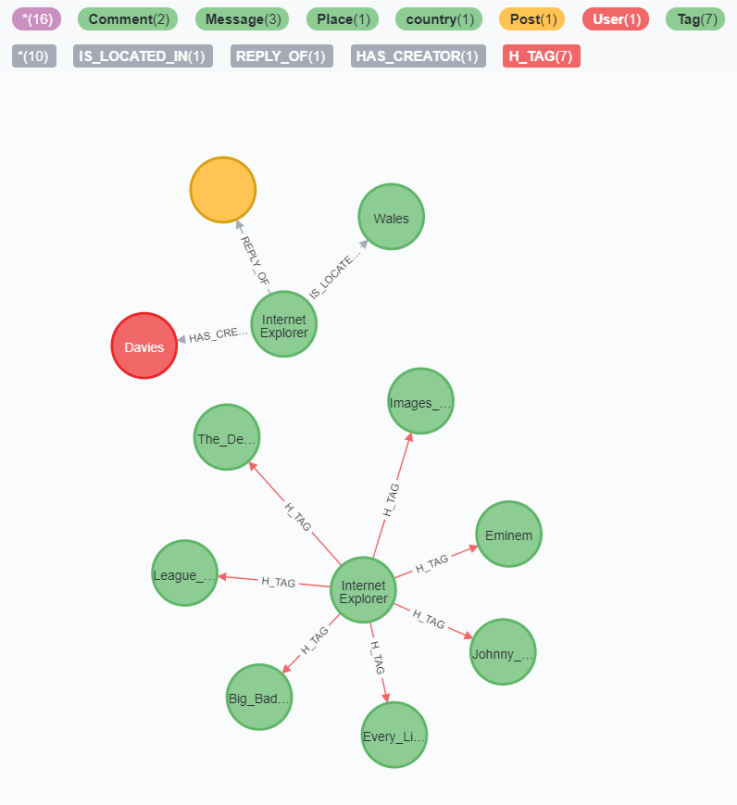


Fig. 6. Snapshot of OPG1-1 after data migration.

Table 3 summarizes the obtained results. The performance of our data mi-

Table 2. Results of the first evaluation.

OPG	Changes	latency	Data correctness	Run time	similarity value
OPG1-1	0%	0.017 s	100%	07.639049 s	1
OPG1-2	30%.	0.016 s	100%	0.783 s	0.93566120387549

gration approach shows that, though the runtime is relatively higher when the average similarity is ¡1 , legacy data is migrated correctly and no data loss is generated. it is not that data latency is almost the same for each dataset.

Second Evaluation For the second type of evaluation, we used all the datasets and for each set, we computed the run time of graph matching and the data migration process. We also computed data latency (number of data access for

each data migration). Table 2 represents a quantitative study for all datasets. Table 3 summarizes the results we obtained. We noticed that even though our

Table 3. Results of the second evaluation.

OPG	OPG entities	nodes	relations	properties
OPG1	30	20	10	128
OPG2	60	40	20	198
OPG3	150	100	50	630
OPG4	300	200	100	1180
OPG5	1500	1000	500	3750
OPG6	2100	1400	700	5009
OPG7	3000	2000	1000	630
OPG8	150	100	50	274
OPG9	300	200	100	563
OPG10	600	400	200	1142

Table 4. Results of the second evaluation.

OPG	Similarity value	Run time
OPG1	0.93566120387549	00:07.639049 m
OPG2	0.9549847478673998	00:17.390528 m
OPG3	0.9356612038754876	00:19.007823 m
OPG4	0.9549847478674043	00:18.727151 m
OPG5	0.8501378090918807	03:25.444671 m
OPG6	0.9281139995680916	04:59.606046 m
OPG7	0.8491987906273264	07:33.561546 m
OPG8	0.5036420203001845	00:12.009441 m
OPG9	0.5036420203001861	01:18.471618 m
OPG10	0.8491987906273264	07:33.561546 m

approach ensure a correct data migration aver small datasets, checking both graph similarity and running the data migration process take an important time. In fact, the runtime depends greatly on entities number and the modification rate, i.e., the more dissimilarity the higher the run time. In fact whenever the dissimilarity is higher, the more data to be migrated. Figure 7 illustrates the runtime of executing our program in regards to the number of entities and the average similarity.

6 Conclusion

Graph databases are widely used in recent years. For that, it is crucial to study its evolution. However, migrating data in such databases can be difficult given

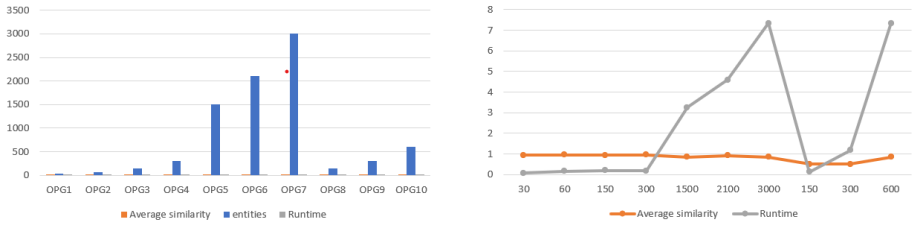


Fig. 7. Runtime by number of entities and average similarity.

their schema free nature. In the scope of this paper, we proposed an approach to safely migrate the graph database.

In this paper, We presented a Graph matching algorithm based on similarity measures in which we extracted a sub-graph from the initial database that is most similar to the *Operation-graph*.

We also presented a process based on the lazy data migration strategy in order to minimize data migration costs and reduce unnecessary data migration procedures (migrating data that will not be accessed in the future) and finally, we detailed the MO's identification phase in which we described specific scripts that help control the migration of legacy data within the graph database. As future work, we aim to propose more optimization by developing an approach based on deep learning algorithms in order minimize the runtime of our approach as it has an exponential complexity.

References

1. Abbes, H., Gargouri, F.: Modular ontologies composition: Levenshtein-distance-based concepts structure comparison. International Journal of Information Technology and Web Engineering (IJITWE) 13(4), 35–60 (2018)
2. Bonifati, A., Furniss, P., Green, A., Harmer, R., Oshurko, E., Voigt, H.: Schema validation and evolution for graph databases. In: International Conference on Conceptual Modeling. pp. 448–456. Springer (2019)
3. Boukettaya, S., Nabli, A., Gargouri, F.: Graph matching in graph-oriented databases. In: International Conference on Intelligent Systems Design and Applications. Springer (2020)
4. Brahmia, S., Brahmia, Z., Grandi, F., Bouaziz, R.: Temporal json schema versioning in the jschema framework. Journal of Digital Information Management 15(4) (2017)
5. Brahmia, S., Brahmia, Z., Grandi, F., Bouaziz, R.: Managing temporal and versioning aspects of json-based big data via the τjschema framework. In: International Conference on Big Data and Smart Digital Environment. pp. 27–39. Springer (2018)
6. Castelltort, A., Laurent, A.: Representing history in graph-oriented nosql databases: A versioning system. In: Eighth International Conference on Digital Information Management (ICDIM 2013). pp. 228–234. IEEE (2013)

7. Castelltort, A., Laurent, A.: Exploiting nosql graph databases and in memory architectures for extracting graph structural data summaries. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 25(01), 81–109 (2017)
8. Erling, O., Averbuch, A., Larriba-Pey, J., Chafi, H., Gubichev, A., Prat, A., Pham, M.D., Boncz, P.: The ldbc social network benchmark: Interactive workload. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. pp. 619–630 (2015)
9. Haubold, F., Schildgen, J., Scherzinger, S., Deßloch, S.: Controvol flex: Flexible schema evolution for nosql application development. Datenbanksysteme für Business, Technologie und Web (BTW 2017) (2017)
10. Herrmann, K., Voigt, H., Behrend, A., Rausch, J., Lehner, W.: Living in parallel realities: Co-existing schema versions with a bidirectional database evolution language. In: Proceedings of the 2017 ACM International Conference on Management of Data. pp. 1101–1116 (2017)
11. Herrmann, K., Voigt, H., Rausch, J., Behrend, A., Lehner, W.: Robust and simple database evolution. Information Systems Frontiers 20(1), 45–61 (2018)
12. Hillenbrand, A., Störl, U., Nabiyev, S., Klettke, M.: Self-adapting data migration in the context of schema evolution in nosql databases. Distributed and Parallel Databases pp. 1–21 (2021)
13. Mesiti, M., Celle, R., Sorrenti, M.A., Guerrini, G.: X-evolution: A system for xml schema evolution and document adaptation. In: International Conference on Extending Database Technology. pp. 1143–1146. Springer (2006)
14. Scherzinger, S., Sidortschuck, S.: An empirical study on the design and evolution of nosql database schemas. In: International Conference on Conceptual Modeling. pp. 441–455. Springer (2020)
15. Scherzinger, S., Sombach, S., Wiech, K., Klettke, M., Störl, U.: Datalution: a tool for continuous schema evolution in nosql-backed web applications. In: Proceedings of the 2nd International Workshop on Quality-Aware DevOps. pp. 38–39 (2016)
16. Shah, F., Castelltort, A., Laurent, A.: Handling missing values for mining gradual patterns from nosql graph databases. Future Generation Computer Systems 111, 523–538 (2020)
17. Szárnyas, G., Prat-Pérez, A., Averbuch, A., Marton, J., Paradies, M., Kaufmann, M., Erling, O., Boncz, P., Haprian, V., Antal, J.B.: An early look at the ldbc social network benchmark's business intelligence workload. In: Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA). pp. 1–11 (2018)

Normalized NoSQL Graph Data Warehouse

Amal Sellami[1], Ahlem Nabli[1,2], Faiez Gargouri[1]

[1] University of Sfax, MIRACL Laboratory,
Tunisia

[2] Al-Baha University,
Kingdom of Saudi Arabia

sellami.amal91@gmail.com, ahlem.nabli@fss.usf.tn,
faiez.gargouri@isims.usf.tn

Abstract. In the Big Data warehouse context, a graph-oriented NoSQL database system is considered as the storage model which is highly adapted to data warehouses and online analysis. Indeed, the use of NoSQL models allows data scalability and the graph store offers more flexibility when storing and managing massive data. We propose, in this paper, an approach to create a Graph-oriented Data warehouse by transforming Dimensional Fact Model into Graph Dimensional Model. Then, we implement the Normalized Graph Dimensional Model using java routines in Talend Data Integration tool (TOS). The resulting warehouse was evaluated in term of "Read Request Latency" using LDBC-SNB benchmark.

Keywords: Big data, NoSQL, graph databases, data warehouse, normalized transformation rules, extract transform and load.

1 Introduction

During the last decade the explosion of social media has led to the generation of massive volumes of user-generated data and consequently given birth to a novel area of research, namely social network Data Warehousing. This emphasizes how to extend classical data warehouse (DW) methodologies in order to deal with new features of social network data, such as volume, dynamicity and heterogeneity.

A data warehouse is a database for online analytical processing (OLAP) to support decision-making. It is often implemented in the relational database management system (RDBMS). Intuitively, a well-designed DW requires a well-planned logical design all updates and versions of a DW lead to a revision of the logical design. Generally, the mapping from the conceptual to the logical model is made according to three approaches: ROLAP (Relational-OLAP), MOLAP (Multidimensional-OLAP) and HOLAP (Hybrid-OLAP). However, all these models are inadequate when dealing with large amount of data which needs scalable and flexible systems. However, in a constantly connected world, data sources produce increasingly massive data, namely big data.

Traditional relational storage models have shown their limitations in terms of storing and managing big data. Indeed, major players of the web such as Yahoo, Google, Facebook, Twitter and LinkedIn were the first to point out the limitations of the relational model. They found that relational DBMSs are no longer adapted when dealing with an enormous amount of data in the context of distributed environments. Usually, classical DW and On Line Analytical Processing (OLAP) are comprised of a set of concepts like: facts, dimensions, measures and dimension hierarchies, those are used for structured schema representations. However, in case of web-scale applications, many of the dimensional information may not be available in regular structure. Consequently, decision makers are increasingly using NoSQL databases to implement their business solutions. Indeed, as NoSQL database offer great flexibility, they can improve the classic solution based on data warehouses (DW). In the recent years, many web applications are moving towards the use of data in the form of graphs.

For example, social media and the emergence of Facebook, LinkedIn and Twitter have accelerated the emergence of the NoSQL database and in particular graph-oriented databases that represent the basic format with which data in these media is stored. However, some new NoSQL (not-only-SQL) Database Management Systems (DBMSs) have been recently proved to be effective Business Intelligence solutions. They have proven some clear advantages with respect to relational database management systems. Nowadays, the research attention has moved towards the use of these systems for storing "big" data and analyzing it. Different families of NoSQL DBMSs exist: Key-value, Column, Document and Graph. A Key-value database is a collection of data without a schema and organized as a collection of key-value pairs. Data is accessed using the key and its value represents data. A Column database represents data with tables where each row can present different attributes (different columns). A Document database stores information as documents having a complex structure. A Graph database is suited for applications in which there are more interconnections between the data like social networks.

In this paper, we focus on one class of NoSQL stores, namely graph-oriented systems. Graph-oriented systems are used for managing highly connected data and perform complex queries over it. Not only data values but also graph structures are involved in queries. Specifying a pattern and a set of starting points, it is possible to reach an excellent performance for local reads by, first, traversing the graph, then collecting and aggregating information from nodes and edges. Graph-oriented databases are based upon graph theory (set of nodes, edges, and properties).

We recall that data warehousing relies mostly on multidimensional data modeling which is a conceptual model that uses facts to model an analysis subject and dimensions for analysis axes. This conceptual model must then be converted in a graph-oriented logical model. Mapping the multidimensional model to relational databases is quite straightforward, but until now there is no work that considers the direct mapping from the multidimensional conceptual model to NoSQL logical models. The objective of this paper is to model the

normalized logical model of graph data warehouse using java routines in TOS. Then we evaluated our resulting warehouse in term of "Read Request Latency".

This paper is organized as follows. Section 2 represents a literature review. Section 3 introduces an overview of our approach. Section 4 presents the input data source LDBC-SNB. Section 5 discloses our proposal for the data warehouse schem design. Section 6 presents the graph logical model and the transformation rules. Section 7 addresses the creation of Normalized GDM with TOS. Section 8 evaluates the created Normalized Grah Dimensional Model based on set of queries. Section 9 concludes this paper by giving some future research directions.

2 Literature Survey

In the most of existing studies, three variant of transformation approch are proposed (i) an approach that transforms a data warehousing concepts into relational logical model or (ii) an approach that transforms relational data model into NoSQL logical model; (iii) an approach that transfsorms a conceptual model into specific NoSQL DB.

Transformation of data warehousing concepts into relational logical model. Multidimensional databases are mostly implemented using RDBMS technologies. Mapping rules are used to convert structures of the conceptual level (facts, dimensions and hierarchies) into a logical model based on relations. Moreover, many researchers have focused on the implementation of optimization methods based on pre-computed aggregates (also called materialized views, or OLAP cuboids). However, R-OLAP implementations suffer from scaling up to very large data volumes (i.e. "Big Data"). According to Gartner, Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making and process automation. Big Data have led data warehouses towards to distributed environments to store and analyze the large amount of data. Research is currently under way for new solutions such as using NoSQL systems is detailled in [1].

Transformation of a relational data base into NoSQL logical model. Some studies have presented approaches that transform relational DB into NoSQL DB. In the literature, a number of researchers have recognized the deficiencies of the traditional ROLAP data storage and have proposed approaches for the migration from relational databases to NoSQL ones. For example, in [2] two proposals are defined which allow big data warehouses to be implemented under the column oriented NoSQL model. The first one (normalized approach) uses different tables for storing fact and dimension at physical level which requires to achieve the join between tables when aggregation is performed. The second one (denormalized approach) stores the fact and dimensions into one table, which allows to avoid performing join between tables. Furthermore, authors in [3] propose rules allowing the storage of a time dimension in HBase table. In [4], authors propose a set of transformation rules for translating a relational model to column-oriented model via HBase. In [5] an algorithm is introduced

for mapping a relational schema to a NoSQL schema in MongoDB [6], as document oriented NoSQL database. In [7], authors propose a method for transforming object-relational database to NoSQL databases, more especially to the document-oriented databases.

Transformation of a Conceptual Model into NoSQL DB. Few works have focused on the transformation of the multidimensional conceptual model to NoSQL logical one. In [8] authors tried to define logical models for NoSQL data stores (oriented columns and oriented documents). They proposed a set of rules to map star schema into two NoSQL models: column-oriented (using HBase) and document-oriented (using MongoDB). In [9] , authors have proposed a transformation rules that ensure the successful translation from conceptual DW schema to two logical NoSQL models (column-oriented and document-oriented). They also proposed two possible transformations namely: simple and hierarchical transformations. The first one stores the fact and dimensions into one column-family/collection. The second transformation uses different column-families/collections for storing fact and dimensions while explaining hierarchies. In [10] authors focused on simplifying the heterogeneous data querying in the graph-oriented NoSQL systems. In [11] authors give a solution to perform join between tables and performing aggregates from data warehouses implemented according the normalized approach. It consists in the integrating of software and tools such as Hive and Kylin in the ecosystem used for implementing the data warehouse, and use their cube building operators. A recent work, [12] proposed a data storage models for Graph cubes by introducing a document oriented model and a column oriented model for storing a graph cube data and implementing the roll up operation over the MongoDB document-oriented database and Cassandra Column-oriented database. Authors in [13] have proposed already two types of transformation from the multidimensional model to the graph-oriented model. An other approach proposed in [14] propose an approach to create a Graph-oriented Data warehouse by transforming Dimensional Fact Model into Graph Dimensional Model (Denormalized Transformation).

Table 1 summarizes our literature review, based on following four criteria.
C1: Describe the tranformation type of data base uses. (i) from the relational model to NoSQL; (ii) from the conceptual multidemnsional model to NoSQL.
C2: Describe the name of NoSQL database used.
C3: Describe the proposed set of rules for the transformation.
C4: Describe the experimentation used to evaluate his works.

To conclude on the literature review, the majority of approaches propose to transform and create the data warehouse under two NoSQL models (Column and Document oriented NoSQL model). There is no approach that directly transform a data warehouse multidimensional conceptual model into a Graph logical model in order to create a data warehouse using graph data base. The major interesting advantage of the graph oriented model is related to the ability for supporting complex queries without using joins. For that, we propose a new approach to create data warehouse under graph oriented NoSQL data base.

Table 1. Summary of the literature review.

Works	C1	C2	C3	C4
(Stonebraker 2012)	NoSQL	-	-	-
(Rocha L et al., 2015)	Relational to NoSQL	(Column-oriented)	(Simple, Hierarchical)	-
(Li et all., 2010)	Relational to NoSQL	(Column-oriented)	(Simple)	-
(Vajk T et all., 2013)	Relational to NoSQL	(Column-oriented)	(Simple)	-
(Dehdouh et all., 2014)	Relational to NoSQL	(Column-oriented)	(Simple)	✓
(Aicha A et all., 2020)	Relational to NoSQL	(Document-oriented)	-	-
(Chevalier et all., 2015) (a)	Conceptual to NoSQL	(Column-oriented, Document-oriented)	(Simple)	✓
(Chevalier et all., 2015) (b)	Conceptual to NoSQL	(Column-oriented)	(Simple,Hierarchical)	✓
(Dehdouh et all., 2015)	Conceptual to NoSQL	(Column-oriented)	(Simple)	✓
(Yangui et all., 2016)	Conceptual to NoSQL	(Column-oriented, Document-oriented)	(Simple, Hierarchical)	✓
(Elmalki et all., 2018)	Conceptual to NoSQL	(Graph-oriented)	-	-
(Challal et all., 2019)	Conceptual to NoSQL	(Column-oriented,Document-oriented)	(Simple,Hierarchical)	✓
(Sellami et all., 2018)	Conceptual to NoSQL	(Graph-oriented)	(Simple)	-
(Sellami et all., 2020)	Conceptual to NoSQL	(Graph-oriented)	(Simple)	✓

3 Graph NoSQL Warehousing: Approach Overview

In this section, we describe our new approach to design and create data warehouse building under graph-oriented system. This approach is composed of five phases as shown in Fig.1.

Conceptual phase. The conceptual model is designed based on a set of rules from Benchmark LDBC-SNB as data source.

Logical phase. The second phase ensure the transformation of the Conceptual model of DW into the graph-oriented model based on set of rules . We distingush two logical model(Normalized and Denormalized) based on the rules used to transform dimensions.

ETL phase. In the third phase, we are interested on the identification and the implementation of ETL operations. The ETL operations are implemented under TOS. The result of this phase is two logical models (Normalized and Denormalized) based on the graph paradigm.

Comparative Study. This phase has as input the two logical models and perform a comparative study in order to choose the best logical model based on two metrics: Write-Request-Latency (WRL) and Read-Request-Latency (RRL). WRL measures the loading time for a single write, and RRL measures the response time of a query.

Reporting Queries. In this phase, we propose to use Cypher graph query language to create and analysis a report of the graph oriented data warehouse. The visualization of the query is done using powerBI.

4 LDBC'S Social Network

As input data source we used the Linked Data Benchmark Council Social Network Benchmark. The LDBC SNB is generated using data generator (DATAGEN) evolved from the S3G2 generator.The LDBC SNB aims at being a comprehensive benchmark by setting the rules for the evaluation of graph-like data management technologies. LDBC SNB is designed to be a plausible look-alike of all the aspects of operating a social network site, as one of the most representative

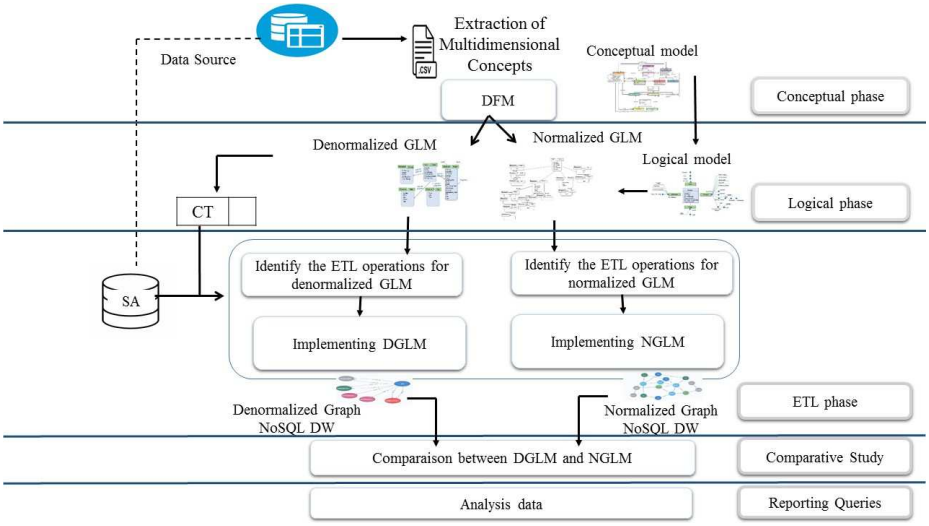


Fig. 1. Approach overview.

and relevant use cases of modern graph-like applications. Its schema has 11 entities connected by 20 relations, with attributes of different types and values, making for a rich benchmark dataset.

A detailed description of the schema is found at [15, 16]. Fig.2 shows the LDBC data schema in UML. The schema defines the structure of the data used in the benchmark in terms of entities and their relations. Data represents a snapshot of the activity of a social network during a period of time. Data includes entities such as Persons, Organisations, and Places. The schema also models the way persons interact, by means of the friendship relations established with other persons, and the sharing of content such as messages (both textual and images), replies to messages and likes to messages. People form groups to talk about specific topics, which are represented as tags.

5 Data Warehouse Schema Design

We propose, in this section, the design of the data warehouse schema based on a set of rules proposed in [17]. These rules applied on the LDBC-SNB Benchmark are used to identify the multidimensional concepts precisely of fact, measures,dimensions and hierarchies.

5.1 Determination of Fact and Measures

An analyzed subject represented by the concept of fact. Each fact characterized by one or more measure representing the indicators analyzed. To extract the fact and its measures, we apply the following rules:

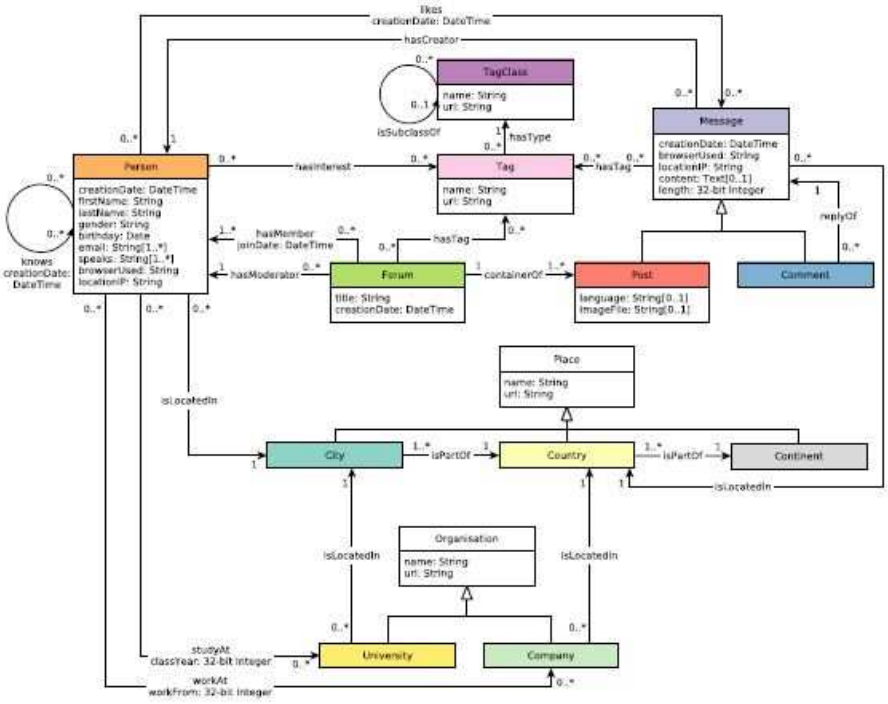


Fig. 2. The LDBC-SNB data schema.

– **Identification of a fact.** We are interested in our work in the analysis of the forum. So we obtain the fact forum.
– **Identification of measures.** The measures are generally numeric and correspond to the "how much" or "how many" aspects of a question. Each measure in a fact should have a default aggregation (or derivation) rule. The data retreived from LDBC-SNB allows us to obtain the following measures: Number of Post, Number of Member, Number of Moderator, Number of Tag, Number of comment and Number of like. Table 2 presents the name of the class / relationship / attributes in our benchmark, the name of the determined measure and its description.

Table 2. Forum Measures.

Class / RlationShip / Atributes Name	Measures Name	Description
Post / Container of	nbPost	The total number of a Post in forum
Person / Has member	nbMembre	The total number of member in forum
Person / Has moderator	nbModerator	The total number of forum moderator
Tag / Has tag	nbTag	The total number of Tag representing in the forum's
Message / Reply of Comment	nbComment	The total number of comment in forum

5.2 Determination of Dimensions

The extraction of dimensions is based on a type of object called basic object. A dimension represents a single set of objects or events in the real world. Dimensions are the qualifiers that make the measures of the fact table meaningful, because they answer the what, when, and where aspects of a question. Based on these questions, we obtain four dimensions for forum which are: Person, Date, Message and Tag (Table 3).

Table 3. Determination of dimensions from LDBC-SNB.

Questions	Basic Object
What contains the forum?	Message
Who moderats the forum?	Person
When the forum is published?	Date
Which words are used to describe the forum?	Tag

For each determined dimension, we detail in the following subsections its parameters. Each attribute or class not chosen as a measure can be an attribute for a dimension specially the categorical attributes.

Determination Date Dimension Attributes. The date is an information that is saved in each record of the data source. The date dimension is a mandatory dimension for analysis and interrogation. The definition of the granularity of the date dimension is based on the need of decision makers, according to which granularity leads its analysis.Table 4 shows the attributes composing the date dimension. The dimensional elements for the Date dimension are day, month and year. Day has a roll-up hierarchical relationship with month which has a roll-up hierarchical relationship with year. Day has an attribute of id date. Fig.3 (a) illustrates the date dimension according to the DFM formalism.

Table 4. Attributes of Date dimension.

Atributes Name	Description	Type
IdDate	The identifier of the date	Identifier
Day	Day of date	Level 2 parameter
Month	Month of date	Level 3 parameter
Year	Year of date	Level 4 parameter

Determination Person Dimension Attributes. The person is an abstract entity that represents a person. It contains various information about the person as well as network related information. The attributes of Person class are: id, firstName, lastName, gender, birthday, email, speaks, browserUsed, locationIP and creationDate. As relationships with person we retreived "islocatedin" and "studyat". The relationship "islocatedin" describe a person and their home

located. There is a class place with attributes name and url. City, country and continent are a sub-class of a place. The relationship "studyat" describe the organisation of the person studied. The class organization have the attributes name, url and have to sub-class university, company.

So, we can describe the Person dimension (D_{Person}) with parameters ID_{Person} along with the weak attributes ($First_{name}$, $Last_{name}$, Birthdate, E_{mail}, $Creation_{Date}$), organized using four hierarchies Hgendre, Hspeaks, Hplace and Horganisation. Table 5 illustrates the set of attributes making up the Person dimension. Fig.3 (b) illustrates the person dimension according to the DFM formalism.

Table 5. Attributes of Person dimension.

Atributes Name	Description	Type
IdPerson	The identifier of the person	Identifier
FirstName	The first name of the person	Weak attribute
LastName	The last name of the person	Weak attribute
Birthday	The birthday of the person	Weak attribute
Email	The email of the person	Weak attribute
CreationDate	The date the person joined the social network	Weak attribute
Gender	The gender of the person	Level 2 parameter
Speaks	The set of languages the person speaks	Level 2 parameter
OrganisationName	The name of the organisation	Level 2 parameter
Url	The url of the organisation	Weak attribute
Type	The type of the organisation	Level 3 parameter in the organisation hierarchie
Label	The label of the organisation	Weak attribute
PlaceName	The name of the place	Level 2 parameter
City	The city of the place	Level 3 parameter in the place hierarchie
Country	The country of the place	Level 4 parameter in the place hierarchie
Continent	The continent of the place	Level 5 parameter in the place hierarchie

Determination Message Dimension Attributes. The message is an abstract entity that represents a message created by a person. The attributes of Message entity: creationDate, browserUsed, locationIP, content and length. Post and comment are a sub-class of Message, it is defined as a type of the message.

Posts contain either content or imageFile, always one of them but never both. Table 6 shows all the set of attributes composing the message dimension. Fig.3 (C) illustrates the message dimension according to the DFM formalism.

Table 6. Attributes of Message dimension.

Atributes Name	Description	Type
IdMessage	The identifier of the message	Identifier
CreationDate	The date the message was created	Weak attribute
Content	The content of the message	Weak attribute
Length	The length of the content	Weak attribute
TypeMessage	The type of the message	Level 2 parameter
BrowserUsed	The browserused of sent the message	Level 2 parameter

Determination Tag Dimension Attributes. Tag is used to specify the topics of forums. The attributes of Tag entity: id, name and url. Table 7 shows all the set of attributes composing the Tag dimension. Fig.3 (d) illustrates the Tag dimension according to the DFM formalism.

Table 7. Attributes of Tag dimension.

Atributes Name	Description	Type
ID	The identifier of the tag	Identifier
Name	The name of the tag	Weak attribute
Url	The URL of the tag	Weak attribute
Type	The type of the tagclass	Level 2 parameter

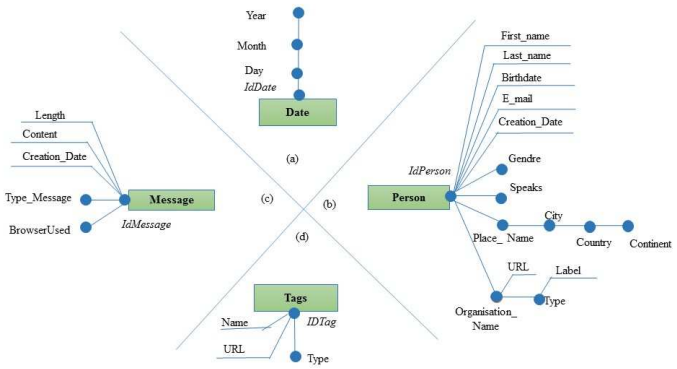


Fig. 3. Dimensions of our LDBC-SNB data warehouse.

From the previous steps, we obtain the conceptual model of data warehouse schema generated from the data source LDBC-SNB (Fig.4).
The mapping from the conceptual to the logical model is made according to three approaches: ROLAP (Relational-OLAP), MOLAP (Multidimensional-OLAP) and HOLAP (Hybrid-OLAP). All these models are inadequate when dealing with large amount of data which need scalable and flexible systems. As an alternative, NoSQL systems begin to grow. In our case we are oriented to use the NoSQL graph-oriented databases.

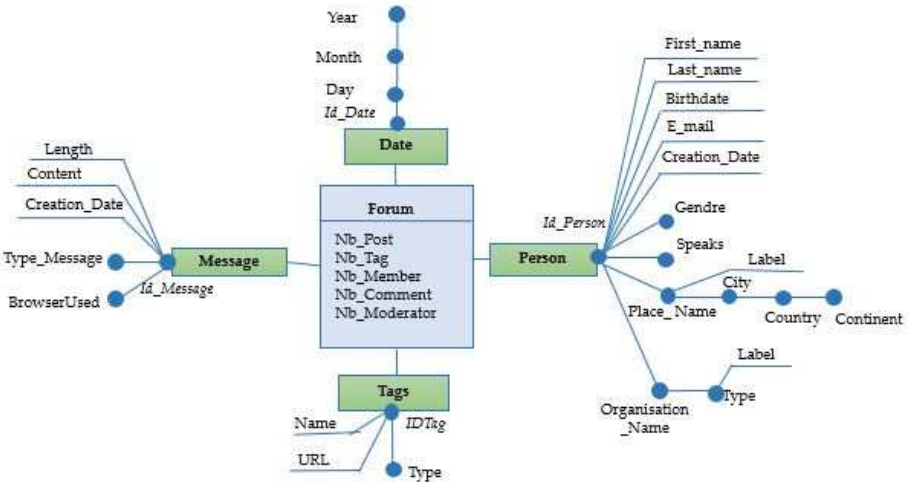


Fig. 4. Multidimensional Conceptual Model: DFM.

6 Graph Logical Model

Graph technology has been the fastest growing category of databases in recent years. In the world where connected data represents a new source for companies, graph technology appears as the obvious option. Therefore, NoSQL Graph-oriented databases are perfectly adapted to voluminous, heterogeneous and massively interconnected data due to its flexible structure capable to representing elegantly correlated and dynamic data.

This kind of database are based upon graph theory. Data is represented as nodes, edges and attributes, which allows the modeling of different interactions between data. Graphs modeling is ubiquitous in most social networks, semantic web and bio science (protein interactions ...) applications.

Graph-oriented systems belong within the "schema less" framework, that consists in writing data without any prior schema restrictions; i.e., each node and each edge have its own set of attributes, thus allowing a wide variety of representations. This flexibility generates heterogeneous data, and makes their interrogation more complex for users, who are compelled to know the different schema of the manipulated data.

It is useful for inter-connected relationship data. The relational database performed better on executing queries when the amount of data is relatively limited. However, as queries became complex, the graph database outperformed the relational one. For the conceptual modeling, an entity-relationship diagram is readily translated into a Property Graph Model, making a conceptual model for graph databases necessary. It helps to understand which entities can be logically connected to which other entities. Graph databases support only binary relation-

ships. On the other hand, graph modeling is much easier than for a relational data model because real world objects are explicit in terms of connections.

The data modeling in NoSQL graph-oriented systems consists in representing the database as a graph. The reason why graph databases are an interesting category of NoSQL is because, contrary to the other approaches, they actually go the way of increased relational modeling, rather than doing away with relations. that is, one to one, one to many, and many to many structures can easily be modeled in a graph-based way. In a way, a graph database is a hyper-relational database, where JOIN tables are replaced by more interesting and semantically meaningful relationships that can be navigated (graph traversal) and/or queried, based on graph pattern matching. The data modeling in NoSQL graph-oriented systems consists in representing the database as a graph.

Formally, we can represent a NoSQL **Graph-oriented database** as G (V, E, P) where:

- V is a set of node that represent the entities,
- $E^G = E_1, \ldots, E_y$ is a set of edges that represent the relation between the nodes,
- P is a set of properties attributed to each component of the graph-oriented database (node/arc). A property is formed by a couple of a key and value pair.

Node. Each node has property and label. Formally, a node, is defined by (id^V, P^V, L^V) where:

- id^V is the identifier of the nodes,
- P^V is a set of properties that describe a node,
- L^V is a set of labels or etiquette attached to the node. In order to express the semantic of the nodes, usually a node can have 0 or more labels written as $L^G = L_1, \ldots, L_q$.

Relation. The relations connecting the nodes can eventually have properties. Formally, a relation is defined by (id^R, Vi^R, Vo^R, T^R, P^R) where:

- id^R is the identifier of the relation,
- Vi^R is the identifier of the incoming node,
- Vo^R is the identifier of the outgoing node,
- T^R is the type of relation that bears the name of the relation,,
- P^R is a set of properties of a relation.

In order to implement the data warehouses within the graph-oriented NoSQL model, we propose two transformations namely DLM (Denormalized Logical Model), and NLA (Normalized Logical Model). Each one differs in terms of the structure and the attribute types used when mapping is performed. In the following we details the two transformations rules. Each transformation load to a graph logical model.

6.1 Transformation rules: Normalized logical model

Recall that a data warehouse schema consists of fact with measures, as well as a set of dimensions with attributes, we map the dimensions according to its attributes and the facts according to its measures. The normalized logical transformation ensures the mapping from the multidimensional model of DW to NOSQL Graph logical model, while explaining hierarchies. In this transformation each fact and dimension are transformed into nodes according the following rules:

Rule 1: Transformation of a fact and its measures to the graph-oriented model.

Fact/Measures transformation. Each fact is transformed into a node with the label of the node takes the type of the concept of the multidimensional model which is 'fact' then we add the name of the fact as a second label at the same node. Each measure is transformed by a property of Fact node.
RF.1. Each fact F $\in F^{MS}$ is transformed into a node, defined by V (id^V, P^V, L^V) where:

- Label l_1 is the type of the multidimensional concept: l_1='Fact' / L^V=$\{l_1\}$,
- Label l_2 is the name of the fact: l_2=N^F / L^V=$L^V \cup \{l_2\}$,
- Each measure $m_i \in M^F$ is represented as a property with p $\leftarrow m_i$/P^V=$P^V \cup$\{p\}.

Rule 2: Transformation of a dimension and its attributes(Strong and Weak) to the graph-oriented model.
Rule.2- Dimension/Parameters transformation. Each dimension is transformed into a node with the label of the node takes the name of the concept of the multidimensional model (in this case is the dimension). Then, we use the name of the dimension as a second label at the same node. After, the identifier is transformed into a property in the node. Finally, any weak attribute associated to the identifier is transformed into a property in the same node. After that, each weak attribute is represented in the form of property.
RD.1. Each name and identifier of a dimension is transformed into a node V (id^V, P^V, L^V) where:

- Label l_1 is the type of the multidimensional concept: l_1='Dimension' / L^V=$\{l_1\}$,
- Label l_2 is the name of the dimension: l_2=N^D / L^V=$L^V \cup \{l_2\}$,
- The identifier a_i modeling by a property p with p $\leftarrow a_i$/P^V=$P^V \cup$\{p\},
- Each weak attribute a_w associated to a_i is transformed into a property p with p $\leftarrow a_w$/P^V=$P^V \cup$\{p\}.

Rule.3- Hierarchies transformation.A hierarchy consists of a set of parameters and a link of precedence between parameters. Each parameter is transformed into a node with the label of the node takes the name of the concept of the multidimensional model (in this case is the parameter). Then, we allow the name of the parameter as a second label at the same node. After that, each weak attribute is represented in the node in the form of property. Finally, each link of precedence is transformed into a relation.

RH.1. Transformation of parameter / Modeling the link of precedence between parameter

RH.1.1 Transformation of parameter

Each parameter a_i is transformed into a node V (id^V, P^V, L^V) where:

- Label l_1 is the type of the multidimensional concept: l_1='Parameter' / L^V=$\{l_1\}$,
- Label l_2 is the name of the parameter: l_2=a_i / L^V=$L^V \cup \{l_2\}$,
- Each weak attribute a_w associated to a_i is transformed into a property p with p $\leftarrow a_w$/P^V=$P^V \cup$\{p\}.

RH.1.2. Transformation of link of precedence between parameter

Each $a_i \rightarrow a_{i-1} \subset H^D$ is transformed into a relation R, defined by (id^R, $V1^R$, $V\emptyset^R$, T^R, P^R) where:

- $V1^R$ is the node represented a_i,
- $V\emptyset^R$ is the node represented a_{i-1},
- The type t_1 is the name of the relation: t_1='Precede'/T^R=$\{t_1\}$.

Rule 4: Transformation of the link between the fact and dimension to the graph-oriented model.

Rule.4- Link fact-dimension transformation. Each link between fact and dimension is represented as a relation having as node source the node modeling the fact and as node destination the node modeling the dimension. The relation has as name 'link fact-dimension'.

RFD.1. Each link fact-dimension is transformed into a relation R, defined by (id^R, $V1^R$, $V\emptyset^R$, T^R, P^R) where:

- $V1^R$ is the node represented the fact,
- $V\emptyset^R$ is the node represented the dimension,
- The type t_1 is the name of the relation: t_1='link fact-dimension'/T^R=$\{t_1\}$.

As output of this transformation is normalized logical model for graph data base and a correponding table with full documentation of all transformation operations. This table will be used in ETL process for modeling and implementing the transformation rules within ETL. As NoSQL DBs are schema-less, this increases the need for extending the existing ETL tool in order to be able to create data warehouse while integrating data. ETL tool should be adapted with the constant changes, to produce and to modify executable code quickly. An example of the correpondence table for the normalized logical model is presented in Table 8.

Table 8. Example of correspondence table for NLM.

Object source	Type	Operation	Target	Type Data
Forum	Fact	Rule1: Fact Transf	Forum	Node
Nb_Tag	Measures	Rule1: Measures Transf	Nb_Tag	Property
Message	Dimension	Rule2: Dimension Transf	Message	Node
Place	Weak attribute	Rule3: Hierarchie Transf	Place	Node

The application of the normalized logical transformation rules on the dimensional fact model of Fig. 4 provides the logical model of a data warehouse using the graphic formalism illustrated by Fig.5.

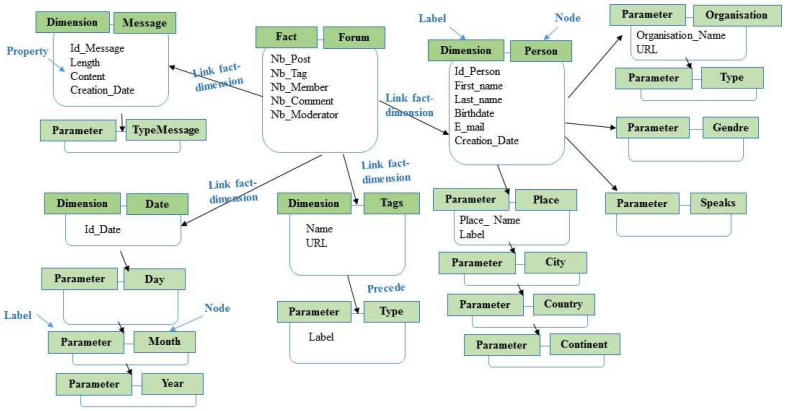


Fig. 5. Normalized Graph Dimensional Model.

6.2 Transformation Rules: Denormalized Logical Model

We recall that, the denormalized Logical transformation ensures the mapping to NoSQL model while highlighting the concepts of the Multidimensional Schema (MS) but without detailing the hierarchies. In this transformation we use 3 rules as follows:

Rule 1: Transformation of a fact and its measures to the graph-oriented model.
Rule.1- Fact/Measures Transformation. Each fact is transformed into a node with the label of the node takes the type of the concept of the multidimensional model which is 'fact' then we add the name of the fact as a second label at the same node. Each measure is transformed by a property of Fact node.

Rule 2: Transformation of a dimension and its attributes (Strong and Weak) to the graph-oriented model.
Rule.2: Dimension/Parameters Transformation.
Each dimension is transformed into a node with the label of the node takes the name of the concept of the multidimensional model (in this case is the dimension). Then, we allow the name of the dimension as a second label at

the same node. After, the identifier is transformed into a property in the node. Finally, any weak attribute associated to the identifier is transformed into a property in the same node. Each parameter is transformed into a property in the node (dimension). After that, each weak attribute is represented in the form of property.

Rule 3: Transformation of the link between the fact and dimension to the graph-oriented model.

Rule.3: Link fact-dimension Transformation.

Each link between fact and dimension is represented as a relation having as node source the node modeling the fact and as node destination the node modeling the dimension. The relation has as name 'link fact-dimension'.

The application of the proposed denormalized Logical transformation rules on the dimensional fact model is illustrated in Fig.6.

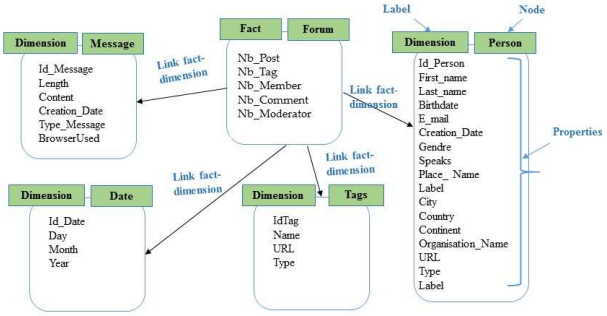


Fig. 6. Denormalized Graph Dimensional Model.

As we previously mentioned, in this level, the correspondence table is generated to keep trace of different transformations. Table 9 presents an excerpt of the generated correspondence table (CT). This table is useful for the ETL process.

Table 9. Example of correspondence table for DLM.

Object source	Type	Operation	Target	Type Data
Forum	Fact	Rule1: Fact Transf	Forum	Node
Nb_Post	Measures	Rule1: Measures Transf	Nb_Tag	Property
Message	Dimension	Rule2: Dimension Transf	Message	Node
IDPerson	Weak attribute	Rule2: Parameter Transf	Place	Property

7 Implementing the Normalized logical transformation Rules: ETL Process

Traditional ETL is a type of data integration from multiple sources (structured and semi-structured data), that follows three steps (extraction, transformation, and loading to a data warehouse or data mart). Its goals are the organization and the storage of data in unified format frequently as a data warehouse.
To load data in the Graph NoSQL DWs, we choose to use the data integration tool "Talend for Big Data". This tool allows extracting data from large and heterogeneous data sources and integrates them into NoSQL database. In the context of our work, data integration is done according to our transformation rules. These rules are implemented using ETL routines in the same tool. The key component of the ETL process is the Job. It is a graphical design, of one or more components connected together such as: tFileInputDelimited (PersonFile, DateFile, ect.), tMap, tNeo4jConnection, tNeo4jRow, tNeo4jOutputRelationship. This components are described as follows:
-tFileInputDelimited reads a given file row by row with simple separated fields. It purpose to open a file and read it row by row to split them up into fields then sends fields as defined in the Schema to the next Job component, via a Row link.
-tNeo4jOutputRelationship receives data from the preceding component, and writes relationships into Neo4j. It is used to output relationship into a Neo4j database.
-tNeo4jOutput receives data from the preceding component, and writes the data into Neo4j. It is used to write data into a Neo4j database, and/or update or delete entries in the database based on the index defined.
-tNeo4jRow is the specific component for this database query. It executes the stated Cypher query onto the specified database. The row suffix means the component implements a flow in the Job design although it doesn't provide output. It depending on the nature of the query, tNeo4jRow acts on the data (although without handling data).
-tMap is one of the core components of Talend Studio and is used very often in Jobs. The tMap component is primarily used for mapping input fields to output fields and transforming the input data in the Expression Builder of the corresponding output column.

For implementing the graph-oriented DW, we use Neo4j. The graph model in Neo4j consists of a Property, only edges can be associated with a type and Edges can be specified as directed or undirected. Neo4j uses the following index

mechanism: a super reference node is connected to all the nodes by a special edge type "REFERENCE". This actually allows to create multiple indexes to distinguish them by different edge types.

All these components are used to create the data warehouse as depicted in Fig.7.

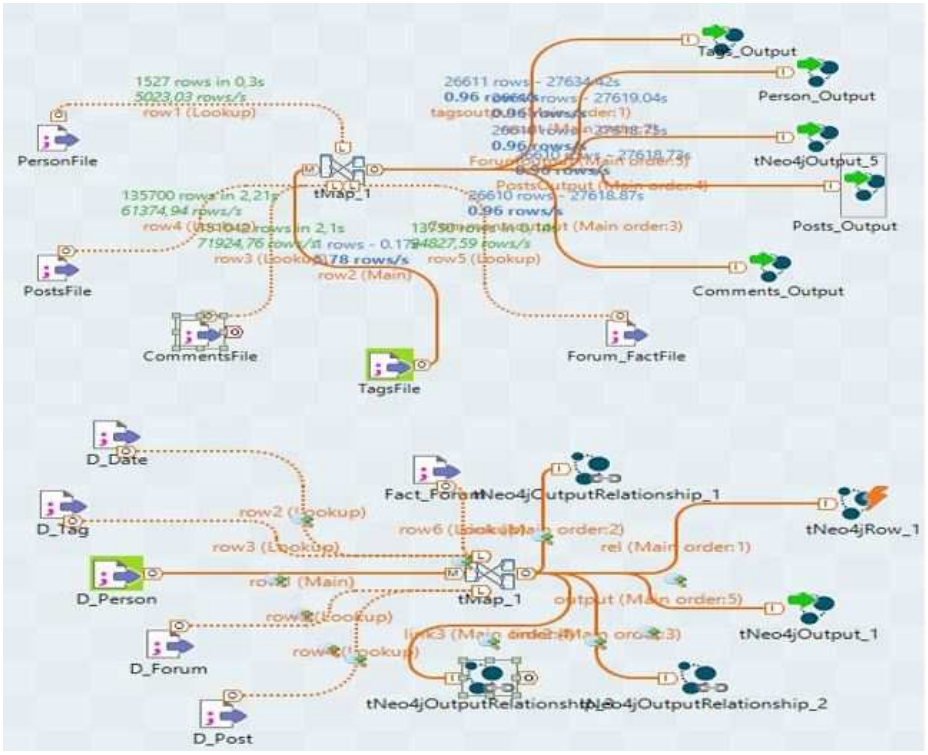


Fig. 7. Creation of Graph DW under TOS.

An example of the number of rows, reading time and loading time (in seconds) of some input files are detailled in Table 10.

The created data warehouse is visualized under Neo4j as presented in Fig.8. As shows in Fig.8, it is composed of 29192 noeuds and 39800 relationShips.

8 Evaluation

Cypher is Neo4j's graph query language that allows users to store and retrieve data from the graph database. The Cypher query language depicts patterns of nodes and relationships and filters those patterns based on labels and properties.

Like SQL, Neo4j CQL has provided some aggregation functions to use in RETURN clause. It is similar to GROUP BY clause in SQL. We can use this

Table 10. Obtained results.

Input File	Number Rows	Reading Time	Loading Time
PersonFile	1527 rows	0,3s	0,8s
TagsFile	16079 rows	0,72s	0,96s
PostFile	135700 rows	0,76s	0,96s
CommentsFile	151042 rows	0,9s	1,2s
ForumFile	13750 rows	0,5s	0,90s
PlaceFile	1495 rows	0,9s	1,2s
DateFile	1095 rows	0,2s	0,8s
OrganisationFile	2258 rows	0,6s	1,2s
Fact_ForumFile	965800 rows	4,98s	9,2s

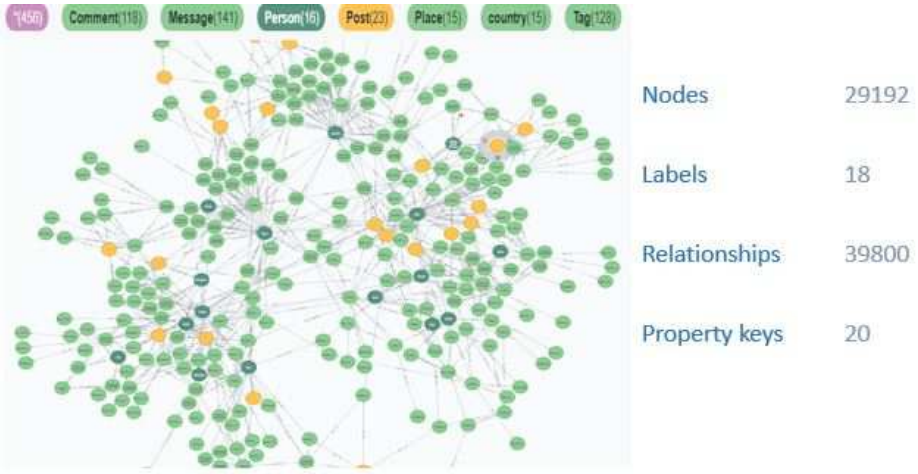


Fig. 8. Normalized Graph NoSQL DW.

RETURN + Aggregation Functions in MATCH command to work on a group of nodes and return some aggregated value.

The aggregate function can take multiple values and can calculate the aggregated values for them. Four levels of pre-aggregates are computed on top of the benchmark generated data. Precisely, at each level we aggregate data respectively on: the combination of 4 dimensions all combinations of 3 dimensions, all combinations of 2 dimensions, all combinations of 1 dimension, 0 dimensions (all data). At each aggregation level, we apply aggregation functions: max, min, sum and count on all dimensions.

In this paper we evaluate the created GDW based on 4 queries described in Table 11.

We measure the efficiency of the implemented NoSQL Graph DW with the metric Read-Request-Latency (RRL). RRL measures the response time of a query. Table 12, summarize the result of graph analysis oriented data warehouse using cypher query language (1 to 4).

Table 11. Query description.

Request	Neo4j Langage
Query 1	MATCH(p:Person) MATCH(Fact:FactForum) MATCH(Mes:Message) RETURN Fact.nbpost ORDER BY p.gender, Mes.browserUsed;
Query 2	MATCH(p:Person) MATCH(Fact:FactForum) MATCH(Mes:Message) RETURN Fact.nbpost ORDER BY p.langue, Mes.browserUsed;
Query 3	MATCH(p:Person) MATCH(Fact:FactForum) MATCH(Mes:Message) MATCH(D:Date) RETURN Fact.nbpost where p.gender='female', D.year='2012' ORDER BY p.gender, Mes.browserUsed, D.month;
Query 4	MATCH(T:Tag) MATCH(Fact:FactForum) MATCH(Mes:Message) RETURN Fact.nbtag ORDER BY T.nametag, Mes.browserUsed;

Table 12. Query analysis based in RRL.

Number Records	Request	RRL(s)
80000	Query 1	0,21
	Query2	0,25
	Query3	0,51
	Query 4	0,32
100000	Query 1	0,29
	Query2	0,33
	Query3	0,64
	Query 4	0,46

The visualization of the query analysis of graph oriented data warehouse using cypher query language is done using powerBI.

Query1: This query gives the number of post by gender and browserUsed. Fig.9 shows the result of Q1 inspired on the Graph DW. *Query2:* This query gives the number of post by langue, browserUsed and year. Fig.10 shows the result of Q2 inspired on the Graph DW. *Query3:* This query gives the number of post by month name and browserUsed, with a clause in the property gender and the year. Fig.11 visualize the result of Q3 inspired on the Graph DW. *Query4:* This query gives the number of moderator by name of the tag and browserUsed.

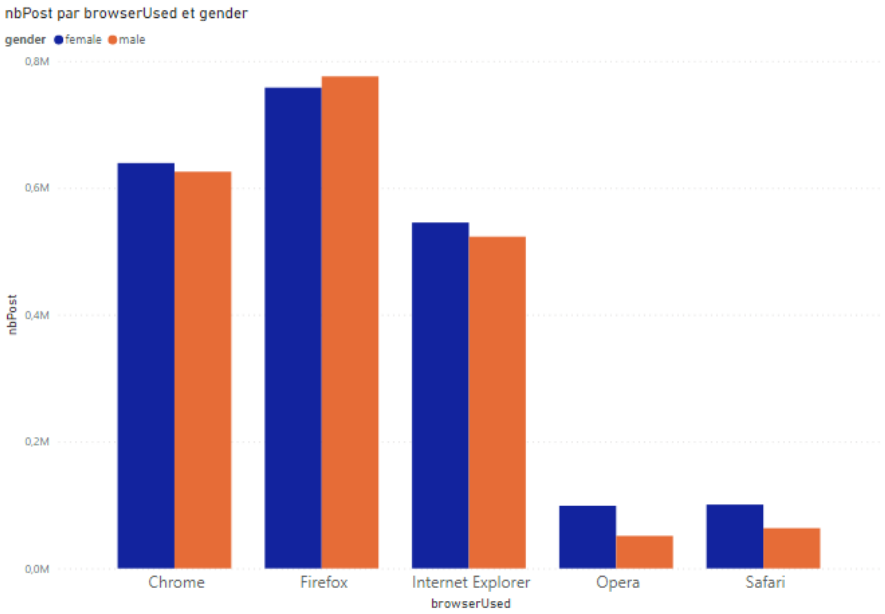


Fig. 9. Number of post by gender and browserUsed.

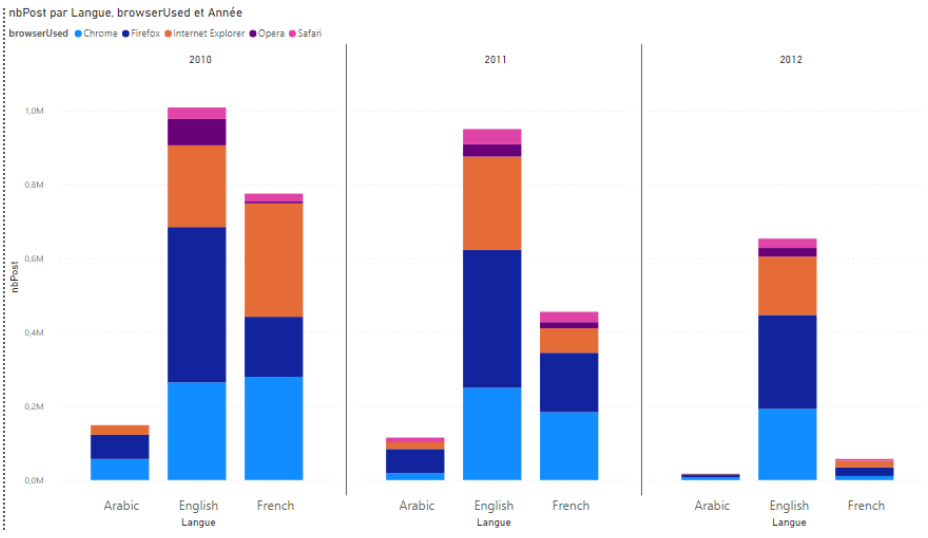


Fig. 10. Number of post by langue, browserUsed and year.

Fig.12 visualize the result of Q4 inspired on the Graph DW.

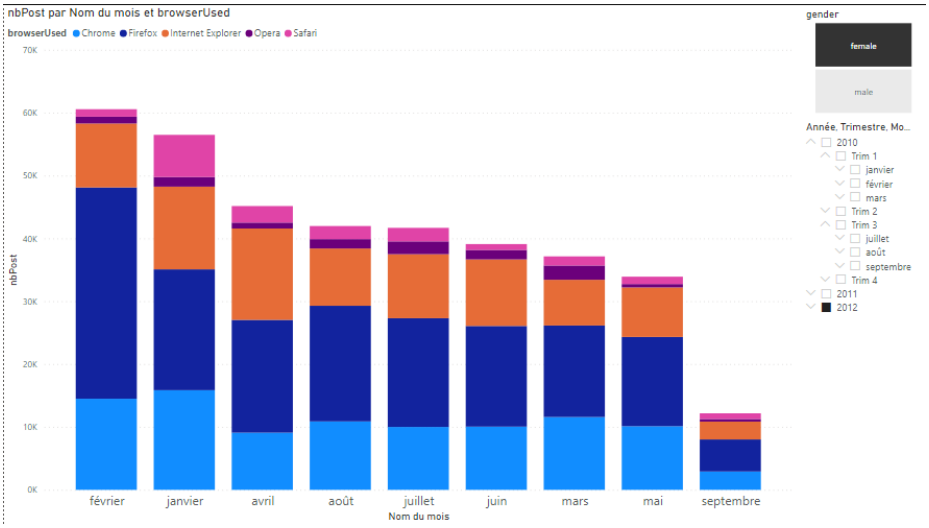


Fig. 11. Number of post by month name and browserUsed, with a clause in the property gender and the year.

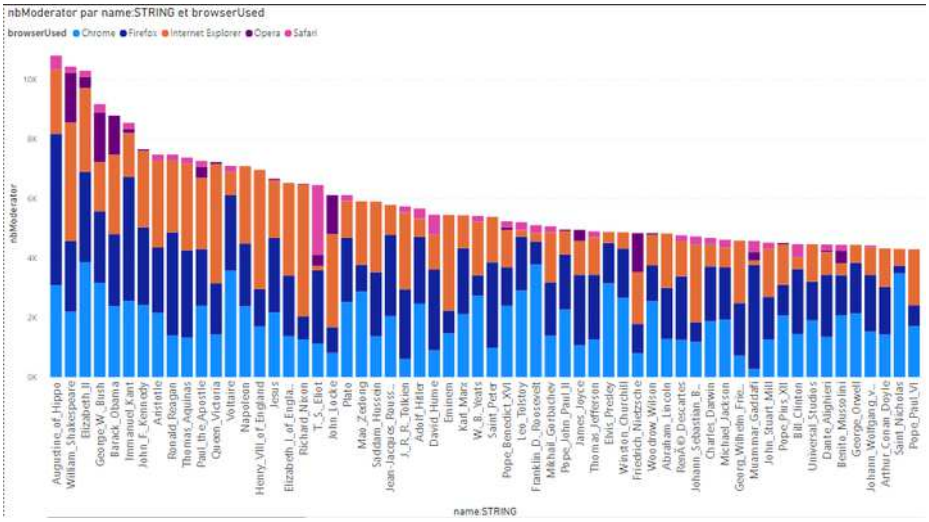


Fig. 12. Number of moderator by name of the tag and browserUsed.

9 Conclusion

As big data continues down its path of growth, a major challenge of the decisional information systems has become how to deal with the explosion of data and its analysis when the data warehouses are implemented.

Consequently, the implementations of data warehouses are oriented towards the new technologies in order to allow more scalability and flexibility for storing and handling data. Since the relational systems are lack of scaling and inefficient of handling big data it is vital to extract transform and loading the data into graph NoSQL data warehouse.

We propose, in this paper, an approach to create a Graph-oriented Data warehouse. We identified two transformations named normalized and denormalized. We have focused on the normalized transformation. Then, we have implemented the Normalized Graph Dimensional Model using java routines in Talend Data Integration tool (TOS).

After that, we evaluated our approach using a set of OLAP queris. As future work, we aim to carry a comparative study in order to choose the best transformation between normalized and denormalized one.

References

1. Stonebraker, M.: New opportunities for new SQL. Commun. ACM 55(11), 10–11 (2012). http://doi.acm.org/10.1145/2366316.2366319.
2. Rocha, L., Vale, F., Cirilo, E., Barbosa, D., Mourão, F.: A framework for migrating relational datasets to NoSQL. Procedia Computer Science 51, 2593–2602 (2015)
3. Li, C.: Transforming relational database into HBase: A case study. In: ICSESS'10. pp. 683–687. IEEE (2010)
4. Vajk, T., Fehér, P., Fekete, K., Charaf, H.: Denormalizing data into schema-free databases. In: CogInfoCom'13, pp. 747–752, IEEE (2013)
5. ONeil, P., ONeil, E., Chen, X., Revilak, S.: The star schema benchmark and augmented fact table indexing. In: Performance Evaluation and Benchmarking, vol. 5895, pp. 237–252. Springer Berlin Heidelberg (2009)
6. Dehdouh, K., Boussaid, O., Bentayeb, F.: Using the column oriented NoSQL model for implementing big data warehouses. In: Proceedings of the 21st International Conference on Parallel and Distributed Processing Techniques and Applications, pp. 469–475 (2015)
7. Aggoune, A., Namoune, M. S.: A Method for Transforming Object-relational to Document-oriented Databases. In: International Conference on Mathematics and Information Technology, Adrar, Algeria (2020)
8. Chevalier, M., El Malki, M., Kopliku, A., Teste, O., Tournier, T.: Implementing Multidimensional Data Warehouses into NoSQL. In: International Conference on Enterprise Information Systems David Harel. First-Order Dynamic Logic. Lecture Notes in Computer Science, Vol. 68. Springer-Verlag, New York, NY (1979). https://doi.org/10.1007/3-540-09237-4.
9. Yangui, R., Nabli, A., Gargouri, F.: Automatic Transformation of Data Warehouse Schema to NoSQL Data Base: Comparative Study. Procedia Computer Science, vol. 96, p. 255-264 (2016)
10. El Malki, M., Ben Hamadou, H., Chevalier, M., Péninou, A., Teste, O.: Querying Heterogeneous Data in Graph Oriented NoSQL Systems. Big Data Analytics and Knowledge Discovery - 20th International Conference, DaWaK (2018)
11. Chavan, V., Phursule, R.: Survey paper on big data. International Journal of Computer Science and Information Technologies, 5(6), 7932–7939 (2014)

12. Challal, Z., Bala, W., Mokeddem, H., Boukhalfa, K., Boussaidy, O., Benkhelifa, E.: Document-oriented versus Column-oriented Data Storage for Social Graph Data Warehouse. (2019)
13. Sellami, A., Nabli, A., Gargouri, F.: Transformation of Data Warehouse Schema to NoSQL Graph Data Base. In: 18th International Conference on Intelligent Systems Design and Applications (2018)
14. Sellami, A., Nabli, A., Gargouri, F.: Graph NoSQL Data Warehouse Creation. In: 22nd International Conference on Information Integration and Web-based Applications and Services (iiWAS) (2020)
15. Prat, A., Averbuch, A.: Benchmark design for navigational pattern matching benchmarking. http://ldbcouncil.org/sites/default/files/LDBC _D 3.3.34.pdf (2020)
16. Erling, O., Averbuch, A., Larriba-Pey, J., Chafi, H., Gubichev, A., Prat, A., Pham, Minh-Duc, Boncz, P.: The LDBC social network benchmark: Interactive workload. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 619–630 (2015)
17. Moalla, I., Nabli, A., Bouzguenda, L., Hammami, M.: Data warehouse design from social media for opinion analysis: the case of Facebook and Twitter. In: 13th ACS/IEEE International Conference on Computer Systems and Applications (2016)

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación