https://lajc.epn.edu.ec

ESCUELA POLITÉCNICA NACIONAL

VOLUME 8, ISSUE 1
JANUARY 2021
ISSN: 1390-9266
e-ISSN:1390-9134

LATIN-AMERICAN JOURNAL OF COMPUTING

IAIC

Vol VIII, Issue 1, January 2021

ESCUELA POLITÉCNICA NACIONAL

EPN

https://www.epn.edu.ec

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

MISIÓN

La Escuela Politécnica Nacional es una Universidad pública, laica y democrática que garantiza la libertad de pensamiento de todos sus integrantes, quienes están comprometidos con aportar de manera significativa al progreso del Ecuador. Formamos investigadores y profesionales en ingeniería, ciencias, ciencias administrativas y tecnología, capaces de contribuir al bienestar de la sociedad a través de la difusión del conocimiento científico que generamos en nuestros programas de grado, posgrado y proyectos de investigación. Contamos con una planta docente calificada, estudiantes capaces y personal de apoyo necesario para responder a las demandas de la sociedad ecuatoriana.

VISIÓN

En el 2024, la Escuela Politécnica Nacional es una de las mejores universidades de Latinoamérica con proyección internacional, reconocida como un actor activo y estratégico en el progreso del Ecuador. Forma profesionales emprendedores en carreras y programas académicos de calidad, capaces de aportar al desarrollo del país, así como promover y adaptarse al cambio y al desarrollo tecnológico global. Posiciona en la comunidad científica internacional a sus grupos de investigación y provee soluciones tecnológicas oportunas e innovadoras a los problemas de la sociedad.

La comunidad politécnica se destaca por su cultura de excelencia y dinamismo al servicio del país dentro de un ambiente de trabajo seguro, creativo y productivo, con infraestructura de primer orden.

ACCIÓN AFIRMATIVA

La Escuela Politécnica Nacional es una institución laica y democrática, que garantizar la libertad de pensamiento, expresión y culto de todos sus integrantes, sin discriminación alguna. Garantiza y promueve el reconocimiento y respeto de la autonomía universitaria, a través de la vigencia efectiva de la libertad de cátedra y de investigación y del régimen de cogobierno.

ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

https://fis.epn.edu.ec

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

MISIÓN

La Facultad de Ingeniería de Sistemas es el referente de la Escuela Politécnica Nacional en el campo de conocimiento y aplicación de las Tecnologías de Información y Comunicaciones; actualiza en forma continua y pertinente la oferta académica en los niveles de pregrado y postgrado para lograr una formación de calidad, ética y solidaria; desarrolla proyectos de investigación, vinculación y proyección social en su área científica y tecnológica para solucionar problemas de transcendencia para la sociedad.

VISIÓN

La Facultad de Ingeniería de Sistemas está presente en posiciones relevantes de acreditación a nivel nacional e internacional y es referente de la Escuela Politécnica Nacional en el campo de las Tecnologías de la Información y Comunicaciones por su aporte de excelencia en las carreras de pregrado y postgrado que auspicia, la calidad y cantidad de proyectos de investigación, vinculación y proyección social que desarrolla y su aporte en la solución de problemas nacionales a través del uso intensivo y extensivo de la ciencia y la tecnología.

CEDIA

CORPORACIÓN ECUATORIANA
PARA EL DESARROLLO DE LA
INVESTIGACIÓN Y LA ACADEMIA

Gonzalo Cordero 2-122 y J. Fajardo Esq.
Teléfono (+593) 7 407 9300
info@cedia.org.ec • Cuenca - Ecuador

www.cedia.edu.ec

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Promovemos la investigación y desarrollo de proyectos innovadores que vinculan a las instituciones mediante concursos e iniciativas. Impulsamos y facilitamos recursos a estudiantes, docentes, profesionales e investigadores con la impresión de esta recopilación científica.

Designed and Printed by
CEDIA

LATIN-AMERICAN JOURNAL OF COMPUTING

Vol VIII, Issue 1, January 2021

ISSN: 1390-9266 e-ISSN: 1390-9134

Published by:
Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas

Quito – Ecuador

Mailing Address

Escuela Politécnica Nacional,
Facultad de Ingeniería de Sistemas
Ladrón de Guevara E11-253, La Floresta
Quito-Ecuador, Apartado Postal: 17-01-2759

Web Address

https://lajc.epn.edu.ec/

E-mail

lajc@epn.edu.ec

Frequency

2 issues per year

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Published by

Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Ecuador

Editor in Chief

Editorial Committee

Associate Editors

Assistant Editors

EDITORIAL

Marco Santórum G. PhD.

Editor in Chief
Escuela Politécnica Nacional, Ecuador

¡Hemos alcanzado 100 artículos científicos publicados!

Una revista científica tiene por vocación la difusión del conocimiento asegurando el buen desarrollo del proceso de evaluación y garantizando la validez de su contenido.

El corazón de nuestra revista es motivar la producción científica de artículos originales y su divulgación. Así lo hemos plasmado en los 16 números publicados durante estos seis años, con una periodicidad semestral, en los cuales hemos alcanzado un registro total de 101 contribuciones.

La revista LAJC, auspiciada por la Facultad de Ingeniería de Sistemas de la Escuela Politécnica Nacional, desde su lanzamiento a finales de 2014, publica un promedio entre seis y siete artículos por edición. Más allá de las implicaciones causadas por la emergencia sanitaria COVID-19, presentamos el Volumen VIII, Número 1.

En septiembre 2019, se me confió la responsabilidad de ser el Editor de la revista, trasladando la responsabilidad a un equipo que vino a aportar una imagen renovada y vanguardista a la revista, acorde a las nuevas tecnologías que van surgiendo en nuestro tiempo.

Hemos conseguido exitosamente los indexamientos en la base de datos Latindex catálogo con metodología 2.0 y la inclusión, una vez más, en el Directorio de revistas de acceso abierto -DOAJ. Conforme a las exigencias internacionales de las bases de datos de indexamiento, hemos regulado la periodicidad. En resumen, le hemos dado un nuevo comienzo a la revista y, en este número, se publica el artículo número 100.

Estos resultados, además de alegrarnos e impulsarnos a ir por más, nos remite a la gratitud hacia quienes sentaron las bases de la revista, así como a quienes contribuyen día a día con LAJC.

Les invito cordialmente a aprovechar la lectura de este número que recoge ocho artículos de investigación relacionados con temas relevantes sobre: minería de procesos, análisis de datos, orquestación de servicios, tecnologías de asitencia AAL, lenguaje de procesamiento natural, migración de datos y ciberseguridad.

Finalmente, en nombre del equipo LAJC, nuestros mejores augurios para que durante este 2021 podamos continuar compartiendo la producción científica con ánimo renovado.

Marco SANTORUM G.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

We reached 100 scientific papers published!

A scientific journal aims to disseminate knowledge, ensuring the proper development of the evaluation process as an essential component to guarantee the validity of its content.

The heart of our journal is to motivate the scientific production of original articles and their dissemination. This is what we have reflected in the 16 issues published over these six years, on a semi-annual basis, in which we have reached a total record of 101 contributions.

The Journal LAJC, sponsored by the Faculty of Systems Engineering of the “Escuela Politécnica Nacional del Ecuador”, since its launch in late 2014, publishes an average of six to seven articles per issue. Beyond the implications caused due to the global health emergency caused by COVID-19, we present Volume VIII, Number 1.

In September 2019, I was entrusted with the responsibility of being the Editor of the journal, transferring the responsibility to a work team that contributed to the creation of a renewed and avant-garde image in accordance with the latest technological developments that are emerging in our time.

We have successfully achieved indexing in the Latindex catalog database with methodology 2.0, and inclusion, once again, in the Directory of Open Access Journals -DOAJ. In line with the international requirements for indexing databases, we have regulated the periodicity. In short, we are bringing a new beginning to the journal and, in this issue, the 100th article is published.

These results, in addition to rejoicing and encouraging us to go for more, remind us of gratitude to those who laid the foundations of the journal, as well as to those who contribute every day to LAJC.

I cordially invite you to take advantage of the reading of this issue that includes eight research articles related to relevant topics on: process mining, data analysis, service orchestration, AAL assistive technologies, natural processing language, data migration and cybersecurity.

Finally, on behalf of the LAJC team, our best wishes that during this 2021 we can continue to share scientific production with a renewed spirit.

Marco SANTORUM G.

LATIN-AMERICAN JOURNAL OF COMPUTING

We are most grateful to the following individuals for their time and commitment to review manuscripts for Latin American Journal of Computing – LAJC

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Reviewers

Patricia Acosta, PhD.
Universidad de Las Américas
Josafa Aguiar, PhD.
Escuela Politécnica Nacional
Roberto Andrade, MSc.
Escuela Politécnica Nacional
Boris Astudillo, MSc.
Escuela Politécnica Nacional
Fabián Astudillo, PhD.
Universidad de Cuenca
Lorena Barona, PhD.
Escuela Politécnica Nacional
Jhonattan Barriga, MSc.
Escuela Politécnica Nacional
Eduardo Benavides, MSc.
Escuela Politécnica Nacional
Nancy Betancourt, MSc.
Escuela Politécnica Nacional
Susana Cadena, PhD.
Universidad Central del Ecuador
Iván Carrera, MSc.
Escuela Politécnica Nacional
Juan Pablo Carvallo, PhD.
Universidad del Azuay
Heitor Costa, PhD.
Federal University of Lavras
Pablo del Hierro, PhD.
Escuela Politécnica Nacional
Robert Enríquez, PhD.
Universidad Central del Ecuador
Vera Ferreira, PhD.
Federal University of the Pampa
Denys Flores, PhD.
Escuela Politécnica Nacional, Ecuador
Miguel Flores, PhD.
Escuela Politécnica Nacional
Walter Fuertes, PhD.
Universidad de las Fuerzas Armadas
Julián Galindo, PhD.
Escuela Politécnica Nacional
José García Rodríguez, PhD.
Universidad de Alicante
Fredy Gavilanes, MSc.
Escuela Politécnica Nacional
Maria Hallo, PhD.
Escuela Politécnica Nacional
Myriam Hernández, PhD.
Escuela Politécnica Nacional
Cindy López, MSc.
Escuela Politécnica Nacional
Gabriel López, MSc.
Escuela Politécnica Nacional
Edison Loza, PhD.
Escuela Politécnica Nacional
Sergio Luján Mora, PhD.
Universidad de Alicante
Marco Molina, PhD.
Escuela Politécnica Nacional
Marcela Mosquera, MSc.
Escuela Politécnica Nacional
Rosa Navarrete, PhD.
Escuela Politécnica Nacional
Freddy Tapia, MSc.
Escuela Politécnica Nacional
Luis Tello Oquendo, PhD.
Universidad Nacional del Chimborazo
Edgar Torres, MSc.
Escuela Politécnica Nacional
Jenny Torres, PhD.
Escuela Politécnica Nacional
Henry Paz, MSc.
Escuela Politécnica Nacional
Diego Pinto, MSc.
Universidad de las Fuerzas Armadas ESPE
Lorena Recalde, PhD.
Escuela Politécnica Nacional
Henry Roa, PhD.
Pontificia Universidad Católica del Ecuador
Daniel Sanaguano, MSc.
Escuela Politécnica Nacional
Franklin Sánchez, MSc.
Escuela Politécnica Nacional
Manuel Sánchez Rubio, PhD.
Universidad Internacional de la Rioja
Sandra Sánchez, PhD.
Escuela Politécnica Nacional
Luis Urquiza, PhD.
Escuela Politécnica Nacional
Ángel Valdivieso, PhD.
Escuela Politécnica Nacional
Víctor Velepucha, MSc.
Escuela Politécnica Nacional
Diana Yacchirema, PhD.
Escuela Politécnica Nacional
Yasmina Vizuete, MSc.
Escuela Politécnica Nacional

TABLE OF CONTENTS

Arquitectura de Analítica de Big Data para Aplicaciones de Ciberseguridad 23

Big Data Analytics Architecture for Cybersecurity Applications

Roberto Andrade
Luis Tello-Oquendo
Susana Cadena-Vela
Patricia Jimbo-Santana
Juan Zaldumbide
Diana Yacchirema

Ataques Zero-day: Despliegue y evolución 39

Zero-day attack: Deployment and evolution

Xavier Riofrío
Fabián Astudillo-Salinas
Luis Tello-Oquendo
Jorge Merchan-Lima

Ciclos Autónomos de Análisis de Datos basados en la Minería de Procesos para el Estudio del Comportamiento Curricular de los Estudiantes 56

Autonomous Cycles of Data Analysis based on Process Mining for the Study of the Curricular Behavior of Students

Sonia Duarte
Jose Aguilar

Modelos de grafos para la detección de datos de texto no estructurados como el sarcasmo 71

Graph Model for Detection of text unstructured data such as Sarcasm

Axel Rodríguez-García
Armando Jipsion

Orquestación de Servicios RESTful y SOAP: Caso de estudio, asignación de visas Americanas. 93

RESTful and SOAP Services Orchestration: Case Study, American Visa Assignment.

Jhon Calle
Pablo Lója
Marcos Orellana
Priscila Cedillo

Proceso de migración de datos en la implantación de Aplicaciones Informáticas Empresariales 105

Data migration process in the implantation of Enterprise IT Applications

Elvis Moreta
Irving Reascos

Síntesis de Sistemas de Conmutación Mediante Permutación de Tablas de Código Gray (Método PGC) 119

Switching Systems Synthesis Method Using Permuted Gray Code Tables (PGC Method)

César Troya-Sherdek
Valentin Salgado Fuentes
Jaime Molina
Gustavo Moreno

Vehículo Eléctrico con Algoritmo de Control de Velocidad y Freno Regenerativo y Diseño de una Aplicación Web Móvil Basada en IoT 131

Electric Vehicle with Speed Control Algorithm and Regenerative Braking and the Design of a Mobile Web App based on IoT

Alex Pulamarin

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol VIII, Issue 1, January 2021

Arquitectura de Analítica de Big Data para Aplicaciones de Ciberseguridad

Big Data Analytics Architecture for Cybersecurity Applications

ARTICLE HISTORY

Received 01 October 2020
Accepted 02 November 2020

Roberto Andrade
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
roberto.andrade@epn.edu.ec

Luis Tello-Oquendo
College of Engineering
Universidad Nacional de Chimborazo
Riobamba, Ecuador
luis.tello@unach.edu.ec

Susana Cadena-Vela
College of Administrative Sciences
Universidad Central del Ecuador
Quito, Ecuador
scadena@uce.edu.ec

Patricia Jimbo-Santana
College of Administrative Sciences
Universidad Central del Ecuador
Quito, Ecuador
prjimbo@uce.edu.ec

Juan Zaldumbide
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
juan.zaldumbide@epn.edu.ec

Diana Yacchirema
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
diana.yacchirema@epn.edu.ec

Arquitectura de Analítica de Big Data para Aplicaciones de Ciberseguridad

Big Data Analytics Architecture for Cybersecurity Applications

Roberto Andrade
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
roberto.andrade@epn.edu.ec

Luis Tello-Oquendo
College of Engineering
Universidad Nacional de Chimborazo
Riobamba, Ecuador
luis.tello@unach.edu.ec

Susana Cadena-Vela
College of Administrative Sciences Universidad Central del Ecuador
Quito, Ecuador
scadena@uce.edu.ec

Patricia Jimbo-Santana
College of Administrative Sciences Universidad Central del Ecuador
Quito, Ecuador
prjimbo@uce.edu.ec

Juan Zaldumbide
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
juan.zaldumbide@epn.edu.ec

Diana Yacchirema
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
diana.yacchirema@epn.edu.ec

Resumen— Los cambios tecnológicos y sociales en la era de la información actual plantean nuevos desafíos para los analistas de seguridad. Se buscan nuevas estrategias y soluciones de seguridad para mejorar las operaciones de seguridad relacionadas con la detección y análisis de amenazas y ataques a la seguridad. Los analistas de seguridad abordan los desafíos de seguridad al analizar grandes cantidades de datos de registros de servidores, equipos de comunicación, soluciones de seguridad y blogs relacionados con la seguridad de la información en diferentes formatos estructurados y no estructurados. En este artículo, se examina la aplicación de big data para respaldar algunas actividades de seguridad y modelos conceptuales para generar conocimiento que se pueda utilizar para la toma de decisiones o la automatización de la acción de respuesta de seguridad. En concreto, se presenta una metodología de procesamiento masivo de datos y se introduce una arquitectura de big data ideada para aplicaciones de ciberseguridad. Esta arquitectura identifica patrones de comportamiento anómalos y tendencias para anticipar ataques de ciberseguridad caracterizados como relativamente aleatorios, espontáneos y fuera de lo común.

Palabras clave *— Big data, ciberoperaciones, ciberseguridad*

Abstract— The technological and social changes in the current information age pose new challenges for security analysts. Novel strategies and security solutions are sought to improve security operations concerning the detection and analysis of security threats and attacks. Security analysts address security challenges by analyzing large amounts of data from server logs, communication equipment, security solutions, and blogs related to information security in different structured and unstructured formats. In this paper, we examine the application of big data to support some security activities and conceptual models to generate knowledge that can be used for the decision making or automation of security response action. Concretely, we present a massive data processing methodology and introduce a big data architecture devised for cybersecurity applications. This architecture identifies anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

Keywords *— Big data, cyber operations, cybersecurity*

I. INTRODUCTION

The increase of digitalization processes, social networks, and interactions generated in digital environments has caused security management requires more complex processes. Besides, another reality has been added related to the diversity of data sources and diverse formats. In this context, security analysts need to implement more controls and methods to know the different attacks that may occur [1].

A challenge for security processes is to establish mechanisms that require processing a large amount of data to determine patterns or anomalies that activate alerts of possible attacks. Security data analysis is not a new field; this process has been growing and developing from data mining solutions, big data, automatic learning, high-performance computing, cloud,

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

and many available information resources to implement data science solutions. The implemented data analysis strategies offer to create a significant change for the treatment of the multiple security problems in both the training of the personnel in charge and how to analyze the companies' problems, thus the analysis of the data to generate contributions in the cybersecurity field.

The amount of data generated within the company operation is significant; therefore, the verification, analysis, and corresponding evaluation by the teams responsible for security become a challenge. Additionally, there may be the need to know different data analysis methods that respond to the security problems encountered. In an attack, the person in charge needs to review the relevant information in a short period and must analyze structured data such as the logs generated from the different infrastructure equipment (server logs, network hardware, personal user devices) and the applications of the implemented information systems; unstructured data such as those coming from websites, news, security feeds, and manufacturers' bulletins.

With this background, a proposal that allows security managers to work with these data types becomes relevant. This study is motivated by these premises and presents an analysis of big data's proposals in cybersecurity matters. A massive data processing methodology is presented jointly with an architecture proposal based on big data comprising five layers: extraction, load, transformation, analytics, and execution.

The rest of this study is organized as follows. Section II presents background information on the challenges of cybersecurity. Section III presents the related work and analyzes the different contributions that use big data in cybersecurity. Section IV discusses the massive data processing methodology that goes from the business problem to the analytical solution's value. Section V presents the results of the different contributions using a big data cybersecurity model and proposes an architecture based on big data to identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary. Finally, Section V concludes this study.

II. BACKGROUND

According to the report by [2], 45 percent of organizations are underprepared for dedicated cyber attacks, and 30 percent have still not fully implemented antimalware software. The adoption of emerging technologies such as bring your own device (BYOD), cloud, Internet of Things (IoT), among others, increases the amount of data and complexity of networks that exceed the human capabilities of the security analyst to make sense of interrelationships among data, systems, and users. According to [3], by 2020 is predicted over 40 trillion gigabytes of digital data or 5,200 gigabytes for every person on earth. In [2], the authors mention that IoT devices are attracted to cybercriminals to be used in their illegal activity. In the year 2016, home routers of a European telecom provider were successfully attacked by a version of the Mirai worm, that convert all compromised devices into an army of bots for massive DDoS attacks [6]. FBI Cyber Division mentions that prioritization of knowledge and emerging threats is significant since cyberactors adapt and alter their tactics and techniques rapidly [5].

Big data analytics focuses on knowledge discovery in structured and unstructured data using data science, advanced statistical functions, machine learning algorithms, and visualization tools. Big data presents new alternatives for the detection and prevention of cyber-attacks using the correlation of internal and external security data [12]. Through Big data, we can take data by twitter feeds and correlate with detected events with security news published on websites or specialized blogs [4]. NIST Information Access Division (NIST-IAD) promotes the development of data analytic methods for greater and more accurate access and understanding of the information contained in multimodal heterogeneous data [8]. On the other hand, [18] mentions some cybersecurity challenges that Big data can help to resolve:

- Data volume: Security analysts need to process large volume of data that demands efficient storage processes, high computer processing and fast access.

- Data inconsistency: Collected data from heterogeneous sources present different structure and format that require pre-processing to prepare the data.

- Data visualization: Visualize large data-sets in real-time with different types of data require an efficient technique of visualization to present all the information in customized dashboards.

Some working groups focused on the use of Big data for cybersecurity are:

- NIST Big Data Public Working Group [10];

•IEEE Special Interest Group (SIG) on Big Data for Cyber Security and Privacy [27];

•ITU Study Group 17 (SG17) [28];

•Cognitive Cybersecurity Intelligence (CCSI) Group [26];

•Microsoft Security and Privacy Group [34].

III. RELATED WORK

Some solutions that use big data applied to cybersecurity have been proposed in recent years. Table I presents these solutions in which the scope of the solution, the technology used, and additional techniques (e.g., statistical or machine learning) that complement the solution are highlighted.

Table I: Big Data Proposal.

ID	Scope	Technology	Complement	Author
S1	Anomaly detection	Hadoop	None	[20]
S2	Network analysis	Hadoop	None	[40]
S3	Alert correlation	Hadoop	None	[35]
S4	Intrusion detection	Hadoop	None	[45]
S5	Network analysis	Apache Spark	None	[37]
S6	Network monitoring	Hadoop	None	[32]
S7	Phising detection	Apache Spark	None	[7]
S8	DDoS detection	Hadoop	Neuronal network	[46]
S9	Intrusion detection	Hadoop	GPGNU	[13]
S10	Security events	Apache Spark	None	[19]
S11	Cyber Threat Intelligence	Hadoop	None	[42]
S12	DDoS detection	Apache Spark	Neuronal network	[22]
S13	Network monitoring	Hadoop	None	[16]
S14	DDoS detection	Hadoop	None	[47]
S15	Intrusion detection	Apache Spark	None	[24]
S16	Anomaly detection	Apache Spark	PCA	[39]
S17	DDoS detection	Apache Spark	None	[43]
S18	Anomaly detection	Apache Spark	Machine learning	[29]
S19	Anomaly detection	Apache Spark	Machine learning	[30]
S20	Anomaly detection	Apache Spark	Social Media	[31]

Fig. 1 summarizes the number of solutions found using Hadoop and Apache spark and those that have considered complementing the use of big data with other solutions such as: statistical processes or machine learning. From the literature review, it is observed that Hadoop and Apache Spark are Big data solutions mostly used for different scientific proposals; there is no substantial difference in the number of proposals using Hadoop or Apache Spark.

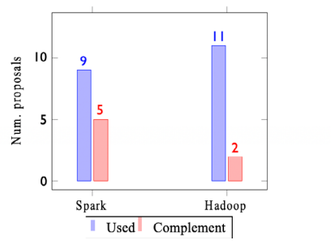


Figure 1: Comparative of Hadoop and Spark proposals.

Some solutions that use big data applied to cybersecurity have been proposed in recent years. Table I presents these solutions in which the scope of the solution, the technology used, and additional techniques (e.g., statistical or machine learning) that complement the solution are highlighted.
Additionally, Fig. 2 presents the cybersecurity operations such as anomaly detection (AD), network analysis (NA), alert correlation (AC), intrusion detection (ID), cyber threat intelligence (CTI), and attack detection (ATD) that are executed using Big data solutions. Security events and CTI have the same scope in the reviewed proposals, similar to network monitoring and network analysis. Proposals about DDoS and phishing detection are grouped into ATD. As observed, most cybersecurity operation applications mainly focus on anomaly and attack detection while AC and CTI are less developed.

A. Big data commercial solutions for cybersecurity

In the following, the commercial big data solutions focused on cybersecurity operations are reviewed.
Watson Cognitive Security [25] integrated two of its products: (i) Watson: a self-learning system that uses natural lan- guage processing to analyze unstructured data such as website information, and (ii) Qradar advisor: a security information and event management. Qradar correlates the events from different information sources such as firewall, server logs and machines. Using Watson allows correlating local security data in QRadar with unstructured data from sites such as blogs, websites or research articles.

In [23], a real-time cybersecurity platform is presented; it is composed by three macro components: telemetry data sources, telemetry data collectors, and a real-time processing engine. The latter is Apache Metron; it is composed of four modules: data collection, message queue, stream process and enrichment, and data access. Table II presents the solutions used in each module of Apache Metron.

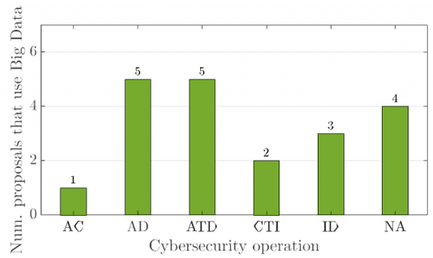


Figure 2: Application of big data in cybersecurity operations.

Table II: Apache Metron Modules.

Module	Solution
Data Collection	Pcap
Message Queue	kafka
Stream Process and Enrichment	Spark Storm
Data Access	Hive Solr Hbase

In [17], a real-time platform CDH based on Apache Hadoop is presented. Apache Hadoop is a software framework that supports distributed applications across clusters of computers to process large data sets using simple programming models [21]. The CDH configuration consists of three macro steps: configuring Apache Spot ODM in HDFS, installing Stream-Sets, and configuring StreamSets Data Collector Pipelines. CDH for data management is based on the Apache Spot Open Data Model (ODM), and considers the following data sources: Qualys KnowledgeBase, Qualys Vulnerability Scans, Windows Security Logs, Centrify Identity Platform Logs. CDH architecture defines six core database tables:

•event;

•vulnerability_context;

•user_context;

•endpoint_context;

•threat_intelligence_context;

•network_context.

SELKS [36] is an open distribution of linux based on the suricata ecosystem for the detection of intrusions, uses the ELK stack to correlate and display security events. The components of SELKS are:

•Suricata is a high-performance network IDS, capable of processing more than 10 Gbps.

• Logstash processes the different sources of information.

•Elasticsearch performs indexing from data events.

•Kibana is a visualization platform that allows customized dashboards, read information from elasticsearch component.

•Scirius is a web interface for Suricata that allows maps signatures from Scirus with Kibana.

•EveBox is a web-based event viewer to generate reports and alerts.

Table III: Relevant attributes of big data cybersecurity sol.

Attribute	Watson	Hortonworks	Cloudera	Selks
RTP	yes	yes	yes	yes
NLP	yes	yes	no	no
IDS	yes	no	yes	yes
ML	no	no	no	no
VA	yes	no	yes	no
CD	no	no	no	yes
ES	yes	yes	no	no
Open	no	yes	yes	yes
Core	Watson	Spark	Hadoop	ELK

Table III presents a consolidated of the attributes that we consider most relevant in each solution: Real Time Processing (RTP), Natural Language Processing (NLP), Intrusion Detection System (IDS), Machine Learning (ML), vulnerabiliLinksy analysis (VA), customize dashboard (CD), information from external sources (ES) (e.g., blogs, web pages), and security news.

IV. PROCESSING METHODOLOGY USING BIG DATA

Regarding massive data processing, a complete methodology should be considered that goes from the business problem to the analytical solution's value.

In general, a data processing model includes several phases such as the acquisition and registration (data understanding), extraction, cleaning and metadata, integration, aggregation, and representation (treatment), analysis and modeling, visualization and interpretation, communication (presentation of results), application and decision-making (enhancement).

The methodology for processing large volumes of information (big data), which allows the transformation of data into knowledge, has several components, which are:

- ***Business component:*** The business is the one that allows to address the problem and put in the value;
- ***Technology:*** it is one of the most important components, since here is the technology used and the way in which the information will be displayed;
- ***Scientific method:*** The models are built by applying the scientific method to the data. Its phases are data processing and data modeling;
- ***Communication:*** It is considered a key factor to transmit all the data in the clearest and most summarized way, it is important to consider that if the results are not communicated, value is lost.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Fig. 3 indicates the processing methodology used in big data environments, considering the four main stages: deal, technology, scientific method, communication [49], [50].
Within the Knowledge Discovery in Databases (KDD) process, data mining is considered the most important phase, since it brings together the techniques capable of modeling the available information. From the use or understanding of the generated model it is possible to extract knowledge. To be able to use data mining techniques (models) it is important to have a "minable view" of the information, (within the proposed architecture in Section V, it can be viewed as a transformation), which involves several stages within which we find the analysis of the distribution function of each attribute, in order to detect values.

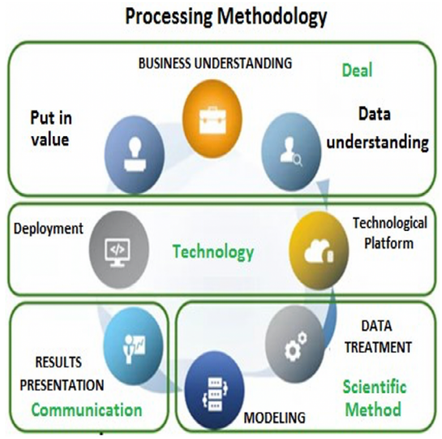


Figure 3: Processing methodology using big data.

Within the data processing stage, it is important to consider the following processes:

- Cleaning;
- Standardization;
- Transformation;
- Integration;
- Determination of missing values;
- Noise identification;
- Detection of anomalous values.

Variable transformation must be carried out, it must be indicated that it will depend on the type of problem to be solved and the data mining technique to apply if the values are ignored, discarded, or transformed. Fig. 4 indicates the steps to be followed to obtain the vitality [11].

V. CYBERSECURITY ARCHITECTURE BASED ON BIG DATA

In this section, the topics in which Big data analytics can contribute to the field of cybersecurity are presented. Then, an architecture is introduced; it comprises five layers: the extraction layer, the load layer, the transformation layer, the analysis layer, and the execution layer. This architecture pretends to identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

A.Big data in Cybersecurity topics

According to our study, Big data mainly focuses on detecting anomalies and attacks; however, these activities are passive cyber-defense strategies in which the objective is to generate alerts for the security analyst. Big data could establish proactive security strategies such as cyber-deception and threat

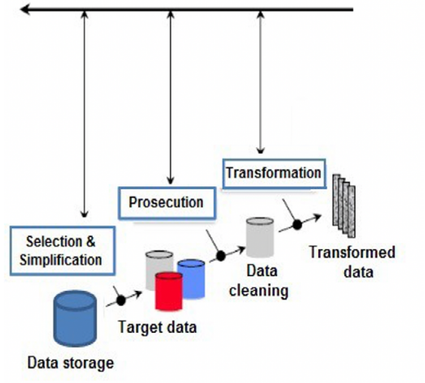


Figure 4: Mineable View.

Figure 5: Cybersecurity applications for big data analytics.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

hunting that allows predicting possible attacks in the future based on extensive information processing. By doing so, attack patterns and profiles of attackers can be determined to establish counterattack strategies. Big data allows analyzing structured and unstructured data like documents, images, and videos used as digital evidence in computer forensic processes. Fig. 5 illustrates an overview of the topics in which Big data analytics can contribute to the field of cybersecurity.

Forensic analysis focuses on the preservation, analysis, and interpretation of computer data. According to the Regional Computer Forensics Laboratory (RCFL) by FBI report in 2016, 17 088 evidence items were received. This generated 5 667 terabytes for digital forensic examinations. In [45], the authors define Big data forensics as a particular branch of digital forensics where the identification, collection, organization, and presentation processes deal with a large data-set. Also, they propose a conceptual model for Big data forensics based on Hadoop; the model considers a reduplication layer to remove redundant data. This is a crucial issue in Big data proposals for assuring data integrity and quality and avoiding incorrect results due to duplicate data. In [38], the authors mention that it is possible to reduce the time and improve the effectiveness to find suspicious files by applying visualization techniques. In the current information age, an analyst is faced with looking at large volumes of data in different heterogeneous sources. Big data solutions provide two fundamental approaches: (i) integrating information from different sources with structured and unstructured formats and different file types such as images, text, or videos; (ii) customized visualization tools that include geographic attributes that provide more significant aspects for visibility to the analyst.

Malware detection. In the first half of 2018, IoT devices were attacked with more than 120 000 modifications of malware [2]; so, considering the growth of data and the need to reduce processing times, it is necessary to analyze new technological alternatives. This context motivated the interest of several researchers in analyzing the use of Big data for malware detection. In [14], the authors propose a scalable clustering approach to identify and group malware with similar behavior for which they use more than 75 thousand samples and require three hours for the processing. In [48], the authors present a method for classifying malware by combining Big data analysis with machine learning, binary instrumentation, and dynamic instruction flow analysis. In [44], the authors present issues and challenges for malware detection, such as incremental learning, active learning, malware prediction, prevalence, adversarial learning.

Security offense. It consists of main techniques, namely cyber deception and threat hunting.

Cyber deception aims to detect attacks for establishing adaptive cyber defense techniques to confuse the attacker. Traditional cyber deception techniques use honeypots and honeynets, but some exciting motivations in this research field are to incorporate artificial intelligence, game theory, and Big data to enhance cybersecurity strategies against attackers [41]. Threat hunting is an iterative activity of active defense searching through the networks and security data to detect advanced threats, instead of waiting for attack alerts [33]. In [9], the deployment of threat hunting processes using GRR rapid response is discussed through two experiments that include tests for remote code execution and the clientside exploits. In [33], the authors present the differences between threat hunting and other cybersecurity activities such as cyber defense, penetration testing, forensics, IDS, and cyber intelligence.

From these two works, the most relevant contributions can be correlated and it is concluded that threat hunting is focused on detecting intruders and unknown threats. The identification of vulnerabilities and mechanisms that can be used by an attacker before an

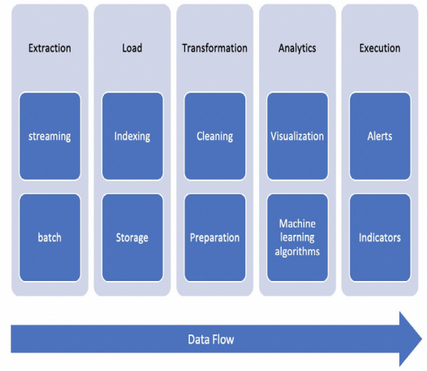


Figure 6: Cybersecurity big data architecture.

attack is made, using basic searching, statistical analysis, visualization techniques, aggregation, machine learning, and Bayesian probability. The process of threat hunting requires processing large amounts of information generated by the logs that exceed human capabilities. By using Big data solutions, it is possible to compensate for this limitation.

Attack detection. Security analysts need to detect attacks in the shortest time possible to reduce the time between detection and attack response. The effective attack detection requires a low false-positive rate. In [15], the authors propose two detection mechanisms: Multivariate Dimensionality Reduction Analysis (MDRA) and Principal Component Analysis (PCA). In [39], the authors propose unsupervised anomaly detection on Apache Spark using PCA for dimension reduction. Also, they mention that Big data implementations face the following challenges: selecting relevant features, scalability, and validation of learned knowledge.

B.Big Data Analytics Architecture for Cybersecurity Appli- cations

The proposed architecture, represented in Fig. 6, contains five functional layers: the extraction layer, the load layer, the transformation layer, the analysis layer, and the execution layer. The different layers are integrated to identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

- Extraction layer. The extraction layer is the foundation of the architecture since it allows us to connect to the source databases and extract the data from those sources, data streams that are received continuously (streaming), and dataset that has a beginning and an end (batch). The main objective of this layer is to extract the data from different sources. Among the data sources, the following can be considered:

  - Cyber simulations platforms;
  - Sensors;
  - Intrusion detection systems;
  - Vulnerability analysis;
  - Security portals, blogs, or feeds;
  - Netflow;
  - Servers and networking appliances logs.

It is worth noting that the data format can be in many forms, such as XML, JSON, CSV, and logs. The data can be received on a scheduled or real-time basis. To perform the extraction, many methods can be used for full or incremental loads. The extraction layer is made up of two sub-modules: streaming and batch, which are described below.

Streaming module: The data stream collected is generated in many formats, volume, and almost impossible to regulate or enforce a single data structure or control the data generated volume and frequency. The streaming submodule is in charge of obtaining the data from different data streams, using one data packet at a time, in sequential order. Each data packet includes the source of the data and a time reference to be used for loading.

Batch module: The batch sub-module is designed to obtain data from legacy batches collected from a group of events over a while (usually long). Batch data extraction is an efficient way to extract large volumes of data.

- Load layer. This layer is responsible for loading the data into the data lake for further transformation and analysis. This layer is made up of two sub-modules: storage and loading to carry out the data loading.

Storage module: This module facilitates data storage either on a local or remote platform. To allow large volumes of data to be loaded in a relatively short time, this process has been optimized so that, in the event of a load or storage failure, recovery mechanisms are triggered to restart from the point of failure without loss of data. NoSQL databases will be used to increase the responsiveness and flexibility of formats.

Indexing module: The objective of the data indexing module is to reduce the time it takes to see the results when generating a query for data with an unknown structure, especially in data that forms large tables and complex queries that involve data combinations in many cases. To carry out the indexing, this module uses some variables such as data type (file or in real-time), data size, and way of accessing the data (ad hoc or through structured application interfaces).

- Transformation layer. This layer is in charge of taking the stored data, cleaning it, and preparing it. Many of the indexed and stored data will come with empty or inconsistent fields. These incomplete tuples can affect the next layers of the architecture. The data must also be prepared in the necessary formats that will be inputs for the analysis layer.

Cleaning module: The data cleaning module will be in charge of taking the "raw" data and will be in charge of standardizing the content of the data, taking into account duplicate values, inconsistent heats, additional fields not taken into account, incomplete values, or meaningless fields.

Preparation module: This module is responsible for the preparation of clean data in aspects such as grouping, extrapolation, reduction, and increase of variables, dataset unification. Note that, although there are structured and unstructured data, it must have a logical and standardized structure.

- Analytics layer. This layer is responsible for the analysis of clean and organized data. Different machine learning and data exploration techniques will be applied here. One of the purposes of this layer is to find anomalous patterns and behaviors in them.

Visualization module: The visualization module takes care of a type of exploration based on different types of graphs so that the user can better assimilate the findings in the data.

Machine learning module: This module is designed to apply different machine learning algorithms. The purpose of the module is to find patterns and predict possible behaviors in the data. These anomalous behaviors will allow the next layer to automate early alerts.

- Execution layer. The execution layer is designed to offer different services and applications to generate alerts and perform indicators.

Alerts module: The alerts sub-module is responsible for identifying unusual or anomalous events detected in the data analysis and sending alert messages accordingly to timely inform those responsible for the data of the events detected.

Indicators module: The indicators module will allow the visualization of key performance indicators to obtain the near-real-time status of the data obtained.

Fig. 7 depicts the architecture devised based on ELK stack for covering the extraction, load, and execution layers. The number of collector servers depends on the number of data sources. The massive amount of data in cybersecurity could be a limit for the batch process. The streaming process is an adequate manner to extract data. Data is processed in real-time for collector servers. Cybersecurity data sources, such as logs, NetFlow, or beats, are relevant information sources to detect anomalous behavior patterns.

The visualization layer depends on two main factors: data and indicators of compromise (IoC). The first one is associated with the data type that could establish information relevant to anomalous behaviors. There is a lot of data sources such as firewall, routers, servers, and end devices. Try to process all this information to increase the numbers of collector servers and all big data architecture's total capacity.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

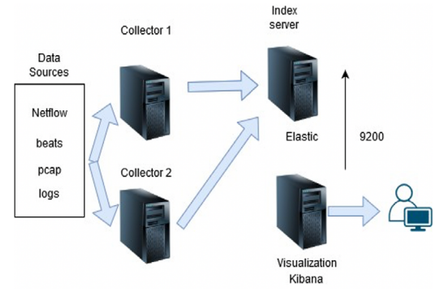


Figure 7: Architecture implementation.

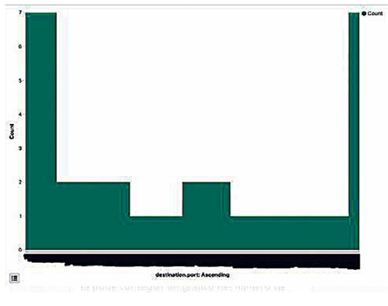


Figure 8: Number of ports during a period.

The second one depends on the first one; if the information generated based on data sources is not relevant is not possible to generate useful alerts about anomalous behaviors. The IoC allows the cybersecurity analyst to know if one event is malicious or not. For instance, Fig. 8 exhibits the number of ports that were used in a specif time; the security analyst could not identify to simple view if this is part of a cybersecurity attack. This figure was generated based on NetFlow traffic.

Another example is DNS traffic, as illustrated in Fig. 9. The ELK architecture can process this kind of data. However, without adequate IoCs, it is not possible to define by the cybersecurity analyst if a high number of connections are part of an event or not.

The analyst in this scenario needs to evaluate past events for identifying if this DNS high-rate count is normal or not. This could be a relevant factor in the streaming process, where the speed of processing is more critical than the storage of the data. To cover this lack, machine learning techniques could be considered as an alternative.

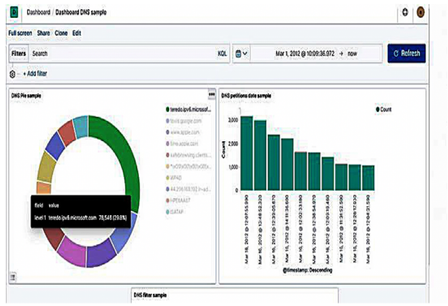


Figure 9: Number of DNS connections in a period.

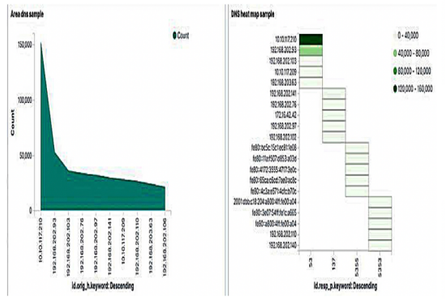


Figure 10: Number of DNS connections associated with an IP address.

Data could be combined in the ELK architecture. For instance, DNS data and NetFlow data could be associated; this allows the cybersecurity analyst to identify DNS requests with the IP address (see Fig. 10). This aspect could be relevant to establish the geo-reference of the attacker.

VI. DISCUSSION

The technological and social changes generate dynamic and complex environments that produce large amounts of data, posing new challenges to security analysts who must process this data to determine patterns or anomalies that allow identifying threats or security attacks. Big data analytics is proposed as a new alternative to improve security operations' effectiveness by processing large volumes of data of different formats in a short time.

In cybersecurity, big data is applied most to monitoring operations and detecting anomalies, focusing on reactive security strategies. However, big data analytics could enhance other security activities for proactive strategies such as threat hunting or cyber deception.

Big data can work with other solutions to complement its ability to process large amounts of data from heterogeneous sources to detect attack patterns. For instance, machine learning allows automating anomaly identification processes through training by the analyst, while natural language processing allows associate publications made in blogs or security news site blogs with detect patterns.

Note that the proposed big data architecture with ELK stack could process different types of data sources. However, the data needs a cleaning process. Another relevant aspect to consider is encrypted traffic because the ELK architecture, in its basic configuration, does not have a way to process this kind of data.

It is crucial to define the problem to be solved or countered with the architecture (e.g., DDoS, phishing, or botnets) because, depending on this, specific data sources and parameters will be necessary. It is recommended to follow the methodology outlined in Fig. 3; in particular, it is essential to understand the business component because it allows addressing the problem to be solved, and it is suggested to work in this phase with the business actors.

It is necessary to consider load balancing and fast-read disks in the architecture that facilitates processing large amounts of data from data sources such as communications equipment or server logs.

VII. CONCLUSIONS

The proposed model based on Big data covers the different components that must be considered for the generation of knowledge regarding the cybersecurity status (Cybersecurity Situation Awareness).

Implementing big data architecture is not enough to solve the problem of dealing with large amounts of data. We should identify reliable information sources, establish data quality control processes, generate safety commitment indicators, and define the update data times. The proposed cybersecurity architecture based on big data comprises five layers that identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

From the conducted literature review, it is evident that few contributions exist regarding threat hunting and cyber-deception, and through the use of Big data, these operations can be enhanced. Therefore, adjust the proposed architecture to these operations is

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

relevant, and it is proposed as future work for the project members. In the case of threat hunting, the architecture will allow identifying, in a predictive way, possible attacks by processing large amounts of data to implement security controls before an attack. In the case of cyber-deception, when identifying patterns of threats or attacks, we can change the functionality of security controls to prevent attack vectors.

ACKNOWLEDGMENT

The authors would like to thank the financial support of the Ecuadorian Corporation for the Development of Research and Academy (Red CEDIA) for the funding support of this work, under Research Group GT-II-2018 (Cybersecurity).

REFERENCES

[1] IBM. AI for cybersecurity. [Online]. Available: https://www.ibm. com/security/artificial-intelligence [Accessed: Nov.25, 2020].

[2] Kaspersky. New IoT-malware grew three-fold in H1 2018. [Online]. Availablet: https://www.kaspersky.com/ [Accessed: Nov.25, 2020].

[3] Microsoft. Enhancing Cybersecurity with Big Data: Challenges and Opportunities. [Online]. Availablet: https://query.prod.cms.rt. microsoft.com [Accessed: Nov.25, 2020].

[4] SK., Kamaruddin and V. Ravi, "Credit Card Fraud Detection using Big Data Analytics: Use of PSOAANN based One-Class Classification," In Proceedings of the International Conference on Informatics and Analytics (ICIA-16). ACM, New York, NY, USA, Article 33 , 8 pages, 2016.

[5] FBI. Audit of the Federal Bureau of Investigation's Cyber Threat Prioritization . [Online]. Available: https://oig.justice.gov/reports/2016/ [Accessed: Nov.25, 2020].

[6] Kaspersky. DDoS attacks in Q4 2016. [Online]. Available: https://securelist.com/ddos-attacks-in-q4-2016/77412/ [Accessed: Nov.25, 2020].

[7] P. Las Casas, V. Santos Dias, W. Meira Jr, and D. Guedes, "A Big Data Architecture for Security Data and Its Application to Phishing Characterization," pp.36-41, 2016

[8] NIST. Data Science. [Online]. Available: https://www.nist.gov/ programs-projects/data-science [Accessed: Nov.25, 2020].

[9] H. Rasheed, A. Hadi and M. Khader, "Threat Hunting Using GRR Rapid Response," International Conference on New Trends in Computing Sciences (ICTCS), Amman, 2017, pp. 155–160.

[10] NIST. Big Data Public Working Group. [Online]. Availablet: https://www.nist.gov/el/cyber-physical-systems/big-data-pwg [Accessed: Nov.25, 2020].

[11] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, 17(3), pp. 37–37, 1996.

[12] R. Alguliyev and Y. Imamverdiyev, "Big data: Big Promises for Information Security," IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, 2014, pp. 1–4.

[13] S.R. Bandre, and J.N Nandimath, "Design consideration of Network Intrusion detection system using Hadoop and GPGPU," 2015 International Conference on Pervasive Computing (ICPC), Pune, pp. 1– 6.

[14] Bayer, Ulrich, P. Comparetti, C. Hlauschek, Ch. Kruegel, and E. Kirda, "Scalable, behavior-based malware clustering," In NDSS, vol. 9, pp. 8–11. 2009.

[15] J. Bin, M, Yan, H. Xiaohong, L, Zhaowen and S. Yi, "A Novel Real-Time DDoS Attack Detection Mechanism Based on MDRA Algorithm in Big Data," Mathematical Problems in Engineering. 2016. pp. 1–10.

[16] Z. Chen, H. Zhang, W.G. Hatcher, J. Nguyen and W. Yu, "A streaming-based network monitoring and threat detection system," IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, 2016, pp. 31–37.

[17] Cloudera. Cloudera cybersecurity. [Online]. Available: https:// www.cloudera.com/ [Accessed: Nov.10, 2020].

[18] A. Dauda, S. Mclean, A. Almehmadi and K. El-Khatib, "Big Data Analytics Architecture for Security Intelligence," Proceedings of the 11th International Conference on Security of Information and Networks, 2018.

[19] L. Fetjah, K. Benzidane, H.E. Alloussi, O.E Warrak, S. Jai-Andaloussi and A. Sekkaki, "Toward a Big Data Architecture for Security Events Analytic," IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), Beijing, 2016, pp. 190–197.

[20] R. Fontugne, J. Mazel and K. Fukuda, “Hashdoop: A MapReduce framework for network anomaly detection,” IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, 2014, pp. 494–499.
[21] Hadoop. Apache Hadoop. [Online]. Available: https://hadoop. apache.org/ [Accessed: Nov.10, 2020].
[22] C. Hsieh and T. Chan, “Detection DDoS attacks based on neural- network using Apache Spark,” International Conference on Applied System Innovation (ICASI), Okinawa, 2016, pp. 1–4.
[23] Hortonworks. Ciberseguridad de los macrodatos. [Online]. Available: https://es.hortonworks.com/ [Accessed: Nov.10, 2020].
[24] G.P.Gupta and M. Kulariya, “A Framework for Fast and Efficient Cyber Security Network Intrusion Detection Using Apache Spark,” Procedia Computer Science.
[25] IBM. Watson and Cybersecurity: The Big Data challenge. [Online]. Available: https://www.ibm.com/blogs/think [Accessed: Nov.10, 2020].
[26] BM. Cognitive Cybersecurity Intelligence (CCSI) Group. [Online]. Available at: https://researcher.watson.ibm.com/researcher [Accessed: Nov.10, 2020].
[27] IEEE. IEEE Special Interest Group (SIG). [Online]. Available: http://computing.northumbria.ac.uk/staff/FGPD3/sigbdcsp/ [Accessed: Nov.10, 2020].
[28] ITU. Study Group 17. [Online]. Available: https://www.itu.int/ en/ITU-T/about/groups/Pages/sg17.aspx [Accessed: Nov.10, 2020].
[29] Z. Jia, C. Shen, X. Yi, Y. Chen, T. Yu and X.Guan, “Bigdata analysis of multi-source logs for anomaly detection on network based system,” 13th IEEE Conference on Automation Science and Engineering (CASE), 2017.
[30] Lighari, S. N., and Hussain, D. M. A. (2017). Testing of algorithms for anomaly detection in Big Data using apache spark. 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN).
[31] H.C. Manjunatha and R.Mohanasundaram, “BRNADS: Big data real-time node anomaly detection in social networks,” 2nd International Conference on Inventive Systems and Control (ICISC), 2018.
[32] S. Marchal, X. Jiang, R. State, R and T. Engel, “A Big Data

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

AUTHORS

Roberto Andrade

Roberto O. Andrade received the electronics and telecommunications engineering degree from the Escuela Politécnica Nacional (EPN) in 2007, and the master's degree in Network and Telecommunications Management from the Army Polytechnic School (ESPE), in 2013. He is currently PhD. Candidate in security systems with the School of Systems Engineering, EPN. He worked in the security areas of the Ministry of Education of Ecuador (MINEDUC) SENPLADES. He has been a certified technical instructor of CCNA, CCNP, and CCNA Security at EPN, since 2010.

Luis Tello-Oquendo

Luis Tello-Oquendo received the electronic and computer engineering degree (Hons.) from Escuela Superior Politécnica de Chimborazo, Ecuador (2010), the M.Sc. degree in telecommunication technologies, systems, and networks (2013), and the Ph.D. degree (Cum Laude) in telecommunications from Universitat Politécnica de Valencia (UPV), Spain (2018). He was Graduate Research Assistant with the Broadband Internetworking Research Group, UPV (2013 - 2018) and Research Scholar with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA, USA (2016-2017). He is currently an Associate Professor with the Universidad Nacional de Chimborazo. His research interest includes 5G and beyond cellular systems, IoT, machine learning.

Susana Cadena-Vela

Susana Cadena-Vela is Professor at the Central University of Ecuador (UCE), PhD in Computer Science, in the line of Data Quality and Open Data. Member of the research groups: Indicators for the Management of the Ecuadorian University, State of the IT of the Ecuadorian Universities sponsored by the Ecuadorian Consortium for the Development of Research and the Academy (CEDIA) and Group of Analytics and Big Data for the Cybersecurity, in addition to the Red Ecuatoriana de Datos y Metadatos (REDAM) groups and the Open Science Research Group.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Patricia Jimbo-Santana

Patricia Jimbo-Santana is Full Professor at the Central University of Ecuador, is an Engineer in Computer and Computer Systems, Computer Expert of the Criminology Institute of the Central University of Ecuador. She is PhD in Computer Science at the National University of La Plata, Argentina, and PhD in Computer Science and Mathematics of Security at the University of Virgina Rovaire of Spain, among her research lines are data mining, machine learning, big data, risk, information and communication technologies.

Juan Zaldumbide

Juan Pablo Zaldumbide is a professor at the Technologist Training School of the National Polytechnic School, in addition to the Master of Information Systems and Business Intelligence of the Armed Forces University (ESPE) and of the Big Data subject of the SEK International University. He obtained his degree in Computer and Computing Systems Engineer from the National Polytechnic School. Later, he obtained his master's degree in Systems Management at the ESPE. He then obtained his Master of Science (Computer Science) degree from the University of Melbourne - Australia, graduating with honors. He has been part of several research projects focusing on the area of Big Data, Data Analytics, Cloud Computing and Machine Learning. In addition, he has published several articles and has been part of conferences and talks in these areas. He has also served on several scientific committees and a peer reviewer for indexed journals.

Diana Yacchirema

Diana Yacchirema received the M.Sc. degree and Ph.D. degree (Cum Laude) in Telecommunications from Universitat Politècnica de València, Spain, in 2011 y 2019, respectively, and the M.Sc. degree (Hons.) in communications and information technology from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2009. She is currently researcher and professor of the Department of Informatics and Computer Sciences of EPN. Her research activities and interests include a wide range of subjects related to Internet of Things, sensor networks, big data, cloud computing, edge computing, and network security. She received the Best Academic Record Award from EPN in 2009.

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol VIII, Issue 1, January 2021

Ataques Zero-day: Despliegue y evolución

Zero-day attack: Deployment and evolution

ARTICLE HISTORY

Received 01 October 2020
Accepted 02 November 2020

Xavier Riofrío
Departamento de Eléctrica Electrónica y Telecomunicaciones
Universidad de Cuenca
Cuenca, Ecuador
xavier.riofrio@ucuenca.edu.ec

Fabián Astudillo-Salinas
Departamento de Eléctrica Electrónica y Telecomunicaciones
Universidad de Cuenca
Cuenca, Ecuador
fabian.astudillos@ucuenca.edu.ec

Luis Tello-Oquendo
College of Engineering
Universidad Nacional de Chimborazo
Riobamba, Ecuador
luis.tello@unach.edu.ec

Jorge Merchan-Lima
Departamento de Eléctrica Electrónica y Telecomunicaciones
Universidad de Cuenca
Cuenca, Ecuador
jorge.merchanl@ucuenca.edu.ec

Zero-day attack: Deployment and evolution

Ataques Zero-day: Despliegue y evolución

Xavier Riofrío
Departamento de
Eléctrica Electrónica y
Telecomunicaciones
Universidad de Cuenca
Cuenca, Ecuador
xavier.riofrio@ucuenca.edu.ec

Fabián Astudillo-Salinas
Departamento de
Eléctrica Electrónica
y Telecomunicaciones
Universidad de Cuenca
Cuenca, Ecuador fabian.
astudillos@ucuenca.edu.ec

Luis Tello-Oquendo
College of Engineering
Universidad Nacional de
Chimborazo
Riobamba, Ecuador
luis.tello@unach.edu.ec

Jorge Merchan-Lima
Departamento de Eléctrica
Electrónica y Telecomunicaciones
Universidad de Cuenca
Cuenca, Ecuador
jorge.merchanl@ucuenca.edu.ec

Resumen— En la ciberseguridad y la informática, el término "Zero-day" se relaciona comúnmente con problemas, amenazas y peligros, esto debido a la falta de conocimiento, experiencia o malentendidos relacionados. Un ataque de Zero-day se considera generalmente una nueva vulnerabilidad sin defensa; por lo tanto, el ataque consecuente tendrá una alta probabilidad de riesgo, y un impacto crítico. Lamentablemente, sólo unos pocos estudios están disponibles para comprender estas amenazas, y no bastan para construir nuevas soluciones para detectar, prevenir y mitigar estas dificultades. En este artículo, se presenta una revisión del ataque Zero-day, enfocándose en comprender su impacto real y algunas soluciones accesibles hoy en día. Este estudio presenta una referencia útil que proporciona a los investigadores conocimientos para comprender el problema actual relacionado con los ataques Zero-day. Este puede ser un punto de partida para desarrollar soluciones para combatir este problema.

Palabras clave— Zero-day, vulnerabilidad, ataque, impacto, implementación.

Abstract— In cybersecurity and computer science, the term “zero-day” is commonly related to troubles, threats, and hazards due to the lack of knowledge, experience, or misunderstanding. A zero-day attack is generally considered a new vulnerability with no defense; thus, the possible attack will have a high risk probability, and a critical impact. Unfortunately, only a few surveys on the topic are available that would help understand these threats, which are not enough to construct new solutions to detect, prevent, and mitigate them. In this paper, it is conducted a review of the zero-day attack, how to understand its real impact, and a few different accessible solutions nowadays. This study introduces a useful reference that provides researchers with knowledge to understand the current problem concerning zero-days attacks; hence they could develop solutions for facing them.

Keywords— Zero-day, vulnerability, attack, impact, deploy- ment.

I. INTRODUCTION

No operating system or software is entirely secure, humans develop them, and humans often make mistakes. In this sense, security is essential and constant updates are needed to cover emerging susceptibilities. These software holes are known as vulnerabilities; they can also result from misconfigurations or errors in the code, which create problems that could be exploited by several actors, such as cybercriminals, competitors, ethical hackers, or malicious people.

Zero-days are undiscovered vulnerabilities; this term was used initially for developers with “zero days” to fix a re- spective vulnerability. It demands their attention as urgent as possible, trying to avoid exposure as much as probable; although, usually at this stage, threat actors (hackers) have already taken advantage of it. They are dangerous because they are unknown, there are no preliminary data available, and these vulnerabilities are only known by threat actors. There are no updates available and no anti-virus scanners can detect them. Therefore, criminals are free to gain access through computer assets, getting benefits without

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

obstruction. Software with these bugs could be trendy, such as Microsoft systems, adobe software, or even security products as firewalls. It becomes even more critical and complex to control within web systems because they are built using several components or libraries from different vendors. It is a challenge to manage the different versions of these components or libraries and patch them; for example, a web application can be developed in Angular JS, which is a set of libraries in JavaScript with their modules or external add-ons (unknown sources) for complementary functions. In this case, the attack surface becomes enormous and it will be unmanageable for the development team to reduce the damage [1].

The zero-day term could be referred to as diverse ideas in the same context. Firstly, zero-day vulnerability refers to the software being exposed and further indicates that neither software ownership nor security products such as antivirus scanners knew its existence. Second, zero-day exploit refers to threat actors who have developed code or performs an action for this zero-day vulnerability to directly affects assets; generally is developed by the person who finds the fault. However, it could be exploited with negative or positive intentions; section II-C1 gives more details about the stakeholders. Finally, a zero-day attack consists of the direct abuse of a particular computer, application, system, or data, taking advantage of the zero-day vulnerability through a zero-day exploit. The latter represents the final objective of the two previous definitions [2], [3].

In general, the zero-day vulnerabilities are a problem with an underestimated impact. This problem is not considered extremely important for ordinary users because companies receive bug reports (or find their bugs) and just patch them. They minimize their errors, do not disclose related data, and avoid disclosing details as feasible. The reason to do this is to hinder cyber-criminals attention so that they do not take advantage of the exposure. Nevertheless, this will not prevent it from being exploited; usually, the exploit appears on the same day as announced, demonstrating that obscurantism is not a significant obstacle to threat actors. It is worth noting that an exploit is a malicious code that abuses flaws in software to infect, interrupt, or control a computer without the user’s consent and usually without their knowledge [4]. Furthermore, little analysis has been made of the real-life phases of the difficulties related to zero-day, which contributes to the fact that those in charge of computers are not seriously focused on addressing these issues [5].

In the following, two examples to illustrate zero-day attacks are described. The first one is Stuxnet; it was a type of zero-day attack and used as a digital weapon (a pioneer in this domain). This malware is classified as a virus/worm and was addressed at the uranium enrichment plant’s computers in Iran. It exploited five zero-day vulnerabilities to spread and gain privileged access to the systems. Microsoft patched one of the vulnerabilities on time; nevertheless, the Microsoft patch was not enough; criminals attacked, took control of the computers altering the plants’ settings, and achieved to shut down the nuclear plant. Although this happened in 2010, these vulnerabilities are still a threat today, especially CVE- 2010-25681, which is Windows Server 2003 vulnerability [7], [8].

The second is F5 BIG-IP, a modern one, which tries to demonstrate the problem in present days. This zero-day attack was disclosed in July 2020 and is a Remote Code Execution (RCE) vulnerability, which affects each product related to the BIG-IP for the company F5 Networks. This allows executing code in the vulnerable server by sending a specifically single HTTP request to the server hosting the traffic management user interface. The relevant role of the attack is the extensive vulnerable surface; the software is widely used around the world, and according to SHODAN2 there are more than 31000 recognizable devices of this type, as illustrated in Figure 1. This indicates that all of those are potentially vulnerable and need to be patched. Several authors developed exploits immediately; in less than one day, they were spread beyond the Internet (Twitter, Reddit, blogs, among others), demonstrating that this zero-day vulnerability could be exploited without too much knowledge. This example shows that a security hole could be exploited in hours, representing a significant threat to the valuables with critical impact [9], [10].

The concern about zero-day dangers is authentic and real. Researchers have focused on countering the problems and creating solutions considering the victim’s rapid reaction to minimize or disappear the risks presented by those vulnerabilities. Despite this, the most significant challenge when developing solutions is a lack of practical and concrete information; this is needed to test and find errors. Another limitation is the extremely low probability of finding a new bug; it takes millions of files to find a unique vulnerability; besides, false positives must be controlled. These reasons demonstrate the concern of a laboratory for investigating this issue because it allows us to have a better understanding of

how the attack is carried out and how it should be prevented and detected [2].
The main contributions of this study are the following:

(i) Creating a concrete and straightforward source of information to begin the understanding of what are the zero-day attacks.
(ii) Revealing the impact and defining the life cycle that could have an attack of this type.
(iii) Analyzing and comparing solutions existing nowadays to face these attacks.

The rest of this paper is organized as follows. Section II describes the real impact of these attacks in real environments; this is based on collected data from computers in use over the Internet. On the other hand, in Section III, some approaches that exist as countermeasures to prevent, detect, and mitigate this predicament are described. Section IV provides a general discussion of the zero-day, their current countermeasures solutions, and a brief analysis of the open issues and challenges that could be addressed in future works. Finally, conclusions and future work are explained in Section V.

II.ANALYSIS OF A ZERO-DAY

The effect and impact of a zero-day depend on the mode of detection, the affected product, who finds it, and other factors. These will mutate the difficulty depending on each unique scenario. This section explains some of the critical circumstances to analyze and consider a zero-day vulnerability's real impact. First, the deployment cycle of a zero-day attack is introduced; then, its lifetime cycle is explained; finally, the real-life impact of this attacks on several factors is discussed.

A.Zero-day deployment cycle

The deployment cycle for a zero-day attack could vary from each case. However, it is considered a common scenario with two significant phases that threat actors follow to proceed with the attack, as shown in Figure 2. The next steps are performed for white and black hat hackers to abuse weaknesses. These steps could be in a different order and may go through multiple repetitions [7], [12]. The term white and black hat hackers is a categorization where the principles of a hacker are focused. Both groups usually have extensive knowledge of how to break into applications, computer networks, and bypass security protocols. Black hat hackers can be involved in cyber espionage, terrorism, or just for challenging cybercrimes. Their primary motivation is financial, and they are responsible for writing malware and exploits. On the opposite, white hat hackers use their skills for the right team, called "ethical hackers" sometimes earn money for reporting bugs to the official sources [13].

1)Discovery phase: The goal of this phase is to find a zero-day vulnerability. The threat actor attempts to recognize, observe, detect, and even guess possible vulnerabilities of a respective surface target. Thus, with a clear idea of how the target is built and structured, the threat actor could audit and inspect it to determine a specific flaw and then move on to a triage stage to test their ideas and findings to generate an inherent exploit. In the following, it is presented each of the stages of this phase.

(i) Recognition: The initial action of exploitation is to discover what can be vulnerable, finding components to start searching defects issues. While more elements found, more chances of finding a security flaw. Therefore, security researchers typically use tools that help them search for these elements in an automated and agile method such a fuzzers or subdomain listers. However, they do not discard manual analysis that provides a more advanced strategy and adds the ability to go deeper into hidden vulnerabilities [14].

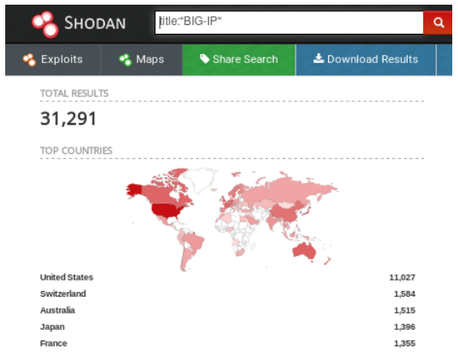


Figure 1: Number of BIG-IP devices connected and found in the internet [11].

(ii) Audit: The next action is to start looking for vulnerabilities in the components beforehand found. In this position, diverse operations may be performed; a threat actor can start analyzing the code directly, performing a binary analysis or reviewing the business logic. Although companies try to hide them, they will be accessible through disassembling tools. As previously affirmed, humans make mistakes, and the code is written for humans; consequently, it is a method to find their errors. Besides, other techniques are applied, such as binary analysis, fuzzing methods that consist of sending a set of payloads until finding one that harms the objective and performing a logical review of the application or system operation. The payload is part of an exploit, which is the portion of the code not related to propagation nor concealment, i.e., it is in charge of performing the malicious action. From the threat actor's point of view, this is where he takes advantage of the system [15].

(iii) Triage: This process involves identifying, tracking, and determining the root cause of the fault in the code; specifically, the part of the application code is being vulnerable and will be exploited. The reason is to exploit the error in the most optimal approach and have a significant and more harmful impact. In this step, there is usually a difference between white hat hackers, who will stop at one point not to damage critical assets, while black hat hackers will continue as much damage as possible because their goal is precisely that.

2)Exploitation phase: In this second phase, "Exploitation", researchers take their vulnerability found apriori and create a sufficiently functional exploit. For that, with the verdicts previously found, it starts to develop a potential exploit. It would help if they debugged with different techniques and types of attacks to take advantage of this flaw. Once the exploit is identified, the developed focus on their effectiveness and efficiency. Finally, concluding with the deployment of the PoC (Proof of Concept)3 in a real environment and with the risk of being compromised.

(i) Debug: At this point, the individual who found the vulner- ability should evaluate the techniques and approaches to exploit it and make the exploit effective. Once evaluated, the threat actor determines precisely what can do exploiting it, the potential impact, and other requirements needed to reduce the uncertainty and create a fully functional exploit. It is possible to return to previous phases, principally if it does not impact or find more complementary vulnerabilities. Generally, exploitation will consist of multiple vulnerabilities, such as Stuxnet's example, and each will contribute to the optimal exploitation.

(ii) Exploit: Once the accurate method is identified jointly with how and what will be exploited, it is necessary to begin applying and testing them, determining their effectiveness, and reviewing several scenarios to confirm the PoC. Initially, it will be a simple exploit, but as it is developed, it could increase the impact. For instance, to escalate privileges or automate actions. There is also the option to cover tracks and cleanup footprint to avoid an effortless discovery of this zero-day.
(iii) Deploy: Finally, the cycle continues to push this zero-day exploit in a real environment; notwithstanding secure laboratory variations, the real world may imply extra obstacles. For black hackers, they can do the damage to the compromised systems or sell it on the black market. However, for white hat hackers, it is up to them to create a tangible and real PoC or report the flaw directly to the corresponding entity. This topic is discussed in Section II-C2.

It is worth emphasizing that every zero-day vulnerability does not convert to a zero-day attack. Sometimes security issues do not lead to exploitable vulnerabilities, do not present a real impact, or they are identified for the company before someone can exploit them.

B.Lifetime cycle

Generally, a widespread belief is that a zero-day is working in the background for a short amount of time because vendors mitigate and release patches as soon as they appear, but this is not always the case. Furthermore, it is believed by the IT community that after being disclosed, this vulnerability will become obsolete or at least with a lower frequency of use [5]. This segment will answer these and other related myths, aiming to detail the authentic lifetime cycle of a zero-day, from its initial discovery until it is supposed dead.

The first point to answer is when a vulnerability appears in a production environment; due to improperly tested or ignored issues. Then, once in production, this error will be exposed for an indeterminate period, as shown in Figure 3. Besides, it will end when the vendor officially releases the patch, consequently going from being exposed indeterminately to being exposed for out-of-date software [16].

Nevertheless, the exposure time is not equal such as a zero-day attack sequence. When the vulnerability is found, criminals develop it, and the vulnerability will be exploited until official sources publish the respective CVE. Nevertheless, while criminals are abusing in

1Common Vulnerabilities and Exposures (CVE) is a document in a database with extensive information detailing vulnerabilities, technical issues, and the disclosure dates; this is a standard used and accepted for academia, govern- mental organizations, private developers, and the cyber-security industry [6]

2A search engine specialized in Internet-connected devices, used for secu- rity people to look for assets connected to the web

3A PoC in the cybersecurity field is used for demonstrating that an exploit is possible on an established system, which is an initial proof because sometimes it is not necessary to attack the objective; testers solely require to prove their idea [14].

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

this time interval, the vulnerability should be discovered by official sources or security researchers (company consulting team, bug hunters, or other community members). Finally, when the zero-day is disclosed publicly, security vendors such as antivirus scanners should develop their solution to find these new threats [17].

It should be highlighted that these actions frequently do not always occur in the corresponding order as Figure 3, although there is always an exposure time before the vulnerability disclosed publicly, and the patch released is always later or equal to this date.

Nonetheless, what happens after its disclosure? To have an answer, it is necessary to know when a vulnerability indeed dies. It is thought that it is dead when the patch is released, either because there is not enough information about its longevity or due to the fact that providers do not release this data for security reasons. However, the study of [7] shows that exploits have an average lifetime of 6.9 years, some of which remain active for more than ten years, new versions continually appear; therefore, they are considered 'immortal'. It also demonstrates that these zero-days will consider as a short lifetime cycle whether they have 1.51 years or less, but this occurs in hardly a quarter of the data analyzed. On the contrast, the longevity group will live more than 9.53 years, representing a 25 percent more longlasting exploits.

C.Real life impact

It has been already talked about attacks and how they are carried out, but is the impact visible and serious in real life?. Several factors influence in this topic, the next arguments are the most significant and are referenced to Figure 4.

1)Stakeholders in the zero-day surface: [13]. Security researchers concentrate on finding zero-day vulnerabilities and report them to the provider, sometimes for money and sometimes for fame. Usually, these vulnerabilities are released into the public domain through published vulnerability advisories, blogs, and news. Software vendors have their security team, but in most cases, this is not enough, then they launch an external consultancy for researchers.

These businesses are speedily expanding; programs such a HackerOne, BugCrown, or Vulnscope (for Latin America) pay independent researchers (called bug hunters) to find vulnerabilities in private programs. They do not exploit a full zero-day attack, instead of they develop a very basic PoC exploit for it and get a payment. In this way, vendors will have an external extensive security team that brings excellent results and they could offer more reliable products [14].

Lastly, enterprises specialized in zero-day such as Exodus Intelligence, ZDI and iDefense find these attacks and provide data for their subscribers to use for defensive testing and product protection measures. These groups belong to the white group.
In contrast, for the dark side, the stakeholders are nations, cybercriminals, competitors, or hobbyists with another motivation. They will sell their findings privately in different markets [13].

2)Zero day markets: In recent years, zero-day vulnerability markets have been growing exponentially and are divided by the buyer, the public vs. private, the vulnerability's nature, and the threat's objective. The subsequent categorization will focus on these points.
The first is a white market, used for bug hunters to report the found vulnerabilities over to the affected vendor. They use them for defensive purposes such as new patches or improvement of new versions. It depends on the vendors, whether should be disclosed or remains private.

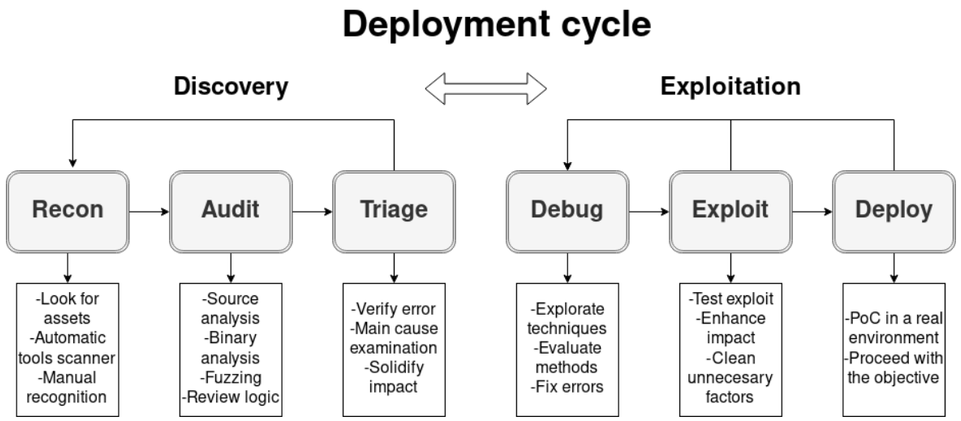


Figure 2: The zero-day attack deployment cycle.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

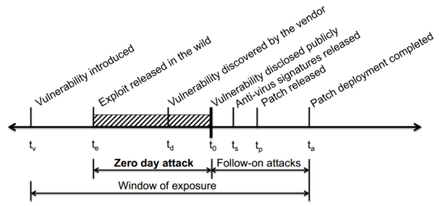


Figure 3: The time line for a zero-day attack [5].

Next is a grey market, vulnerabilities continue to be confidential among initiators and collaborators, which are used primarily for offensive attacks rather than defensive. They will remains in the background, used against the affected provider, although they are usually sold towards governments or national institutions which could use them in diverse purposes, for this reasons they shall not be divulged.

Finally, black markets are sold for criminals where the vulnerabilities are not disclosed because they will use them for illicit purposes. The buyers could be competitors vendors, cyber-terrorists or even nations. This market is the most profitable market because it is illegal and may pay large amounts of money for exploits capable of damaging an organization [13].

3)Evolution of the zero-day: The paper has been discussed the development of a zero-day attack and its lifetime cycle, without specifying what indeed occurs and how changing in these stages in a wild scenery. Here arises a point of inflection where the exploitation rises exaggeratedly, this point is after its disclosure: Zero-day vulnerabilities before disclosure regularly rises and remains running in a context such as the black market members, security researchers, or small groups of hackers.

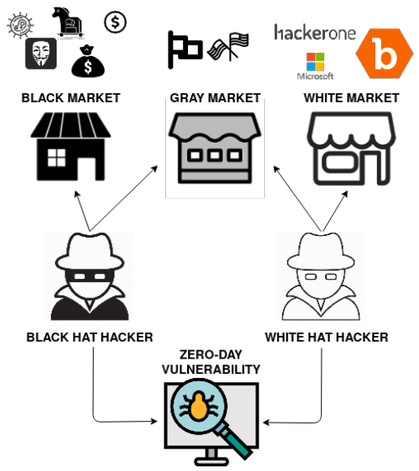


Figure 4: Markets and stakeholders related with zero-days.

Each group wants to avoid discovered their goal is to take advantage of the vulnerability as much as possible (for white and black hat hackers). Therefore, the number of attacks at this stage is low for two reasons: the number of threat actors is low, and the number of targets is limited. Most zero-day attacks do not exceed 1000 attempts, as shown in Figure 5. This point is directly proportional to how is the exploit evolution and variations (Figure 6); the malware remains hidden and continues without threat; therefore, it has no problem attacking, do not need to change to be effective.

On the other hand, once the flaw is publicly exposed, ***Zero-day vulnerabilities after disclosure*** increased logarithmically. This fact is produced because whether a system vulnerability is revealed or widespread, each actor in this environment will have the possibility of exploiting that (they indeed would attack). Although no extensive information is revealed in the CVEs, exploits and attacking methods are immediately developed. Figure 6 demonstrates that malware variations also present a logarithmic growth. Consequently, victims of this vulnerability are more exposed at this point and have a significant probability of being attacked, representing an extreme boosting number of attacks after t0. This behavior is exhibited in Figure 5.

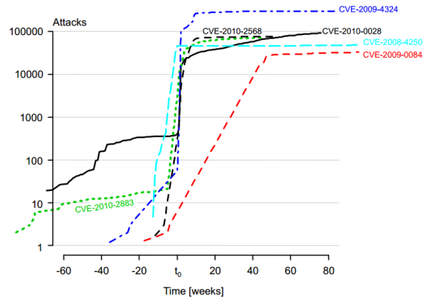


Figure 5: Number of attacks before and after the CVE disclosure. t0 is the disclosure date [16].

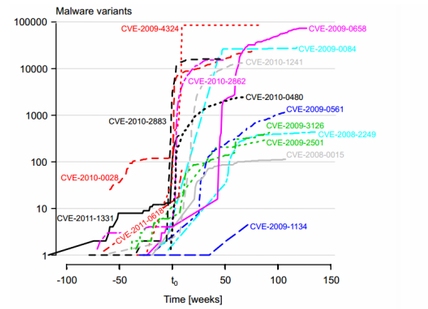


Figure 6: Number of exploit variations before and after the CVE disclosure. t0 is the disclosure date [16].

In this segment, the dangerousness of a zero-day was exposed, in both cases, before and after its disclosure. To emphasize, after the CVE (or another method) is published, it is just a matter of time for someone develops an exploit and spread over the web, facilitating the exploitation of a resource and increasing the attack rate. This amount of attacks will increment and last over time until patches and security solutions will be implemented. Furthermore, this would remain for a considerable amount of time due to the exploit evolution and variations with the same CVE.

III. COUNTERMEASURES AND TECHNIQUES

Security mitigation and countermeasure perform an essential role in the exploitability of a vulnerability. Exploitable vulnerabilities and affected services can be retained or at least deferred over time. There are different ways to counteract the direct influence on the threat implications of zero-day attacks. Standards such as ISO 27000 or NIST4 have various approaches to dealing with computer issues. In the case of ISO 27001, section A-12 refers to "Detection, prevention and mitigation controls to protect against malware shall be implemented, combined with appropriate user awareness" which is also applicable for the case of zero-day [18]. The explanation of them is below, followed by contemporary examples in Section IV-A.

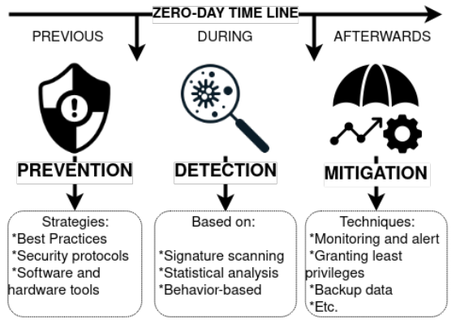


Figure 7: Countermeasures details for zero-days expressed in a timeline.

A.Detection

Any feature with Internet access is exposed to zero-day attack, but antivirus scanners, Internet Detection System (IDS), and other mechanisms are commonly employed to detect possible threats in a system or organization. Nevertheless, these do not operate for new vulnerabilities; some strategies have been developed to dispense or to limit the damage caused by this problem. Figure 7 illustrates that they act during the attack and use the methods explained as follows:

(i) Signaturebased: It is a method regularly employed by antivirus scanners and attempts to discover nonpolymorphic malware (worms and viruses). A signature is a unique part of an exploit; it is usually a string with an impossible date or a hash value. Anti-viruses scanners and security software have signatures databases that are used to find malware [4], [15].

A database contains the signatures previously found in old files, and these are compared with new data to find evolved malware and harmful files. For zero-day, a payload comparison of old malware is carried out. As previously mentioned, a zero-day is exploited with several vulnerabilities, but these are not necessarily zero-day. These libraries are continually updated. The comparison procedure depends on the algorithm used. For instance, in [19], an approach based on the decoding and encoding of the requests received (used as a signature) is presented; these requests are analyzed by a neural network to detect anomalies in the requested parameters.

(ii) Statistical-based: Statistic techniques are one of the most popularly used methods for detecting difficulties in a network, software or unknown intruder. This technique saves computer records of occurred events (conditions, performances, network or memory statistics, among others) that left a trace in the past and compares them with a new running record. If a record does not match, an alert is triggered when different stats are discovered, and they are blocked whether it is necessary. Consequently, if more data would gather, the detection algorithm will be more reliable and more precise.
In the case of zero-day, it can be contradictory because there is no preliminary data. Nevertheless, the method can be applied with procedures based on statistics profiles from system's data to detect new actions in an identical scenario. For example, in [12] a probabilistic approach is presented to detect abnormal patterns in a network, uncovering these dilemmas.

(iii) Behavior-based: They are based directly on the conduct of a system or application, trying to forecast an expected behavior to detect anomalies, whether it would differ or not. Behavior-based strategies use unusual marks that can leave an attack, but the problem with this method is that a large number of false positives (or false negatives) may occur. For this field, it could be beneficial to rely only on the regular conduct of the defended system because these vulnerabilities would not exist in a database. The methodology proposed in [20] aims to analyze flows in real-time, finding anomalies that may be zero-day attacks, but as mentioned above, they have a large number of false positives/negatives anomalies that can be counterproductive.

4The National Institute of Standards and Technology from the United States

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

B.Prevention

In business, zero day vulnerabilities can be chaotic, causing severe consequences for companies. Although this has already been mentioned and sounding repetitive,decisions and security efforts must be applied to prevent zero-day attacks (among others).

If an entity has a well established and conscious security structure and policies, a zero day attack is likely to be considerably less damaging. On the other hand, not having defenses will open up the possibility of unfortunate accidents with economic, structural and reputation effects [21]. Consequently, it is indispensable to take into account the probability of occurrence and common safety recommendations. This countermeasure is performed previous the assault, and below is a brief explanation of strategies to prevent them:

(i) Security best practices manuals and resources: The standards are studied, created, and imposed for the prevention of failures that cause problems; for this explanation, it is necessary to apply these standards for preventing.

(ii) Security patching and updates: The new versions of the software are bringing for a reason; vendors should know their vulnerabilities and updates can appropriately protect the system from zero-day exploitation.

(iii) Advance security hardware and software: Secure programming, Quality Assurances (QA), and other measures to regulate the responsibility of programmers are not enough. Therefore, it is necessary to use tools that complement this and reduce zero-day attacks, for example, secure code scanners could be used.

(iv) Security protocols: Inventing and creating something that already exists is entirely insecure; the protocols are tested and approved for several years, entities should not implement algorithms or methods that are easy to implement. For example, the cryptographic algorithms. Generally, they are ordinary security operations of an organization; hence there are not many approaches that perform prevention in this field. However, in [21], they perform a risk assessment for these threats focusing on the method of attack graph-based security metric, which analyzes the risks quantitatively, examining access vector, access complexity, among others factors. This solution contributes to having a risk control for unknown vulnerabilities, preventing and reducing the impact on an organization.

On the other hand, as future work, we will propose creating a tool for the prevention of zero-days attacks that will focus on day zero (disclosure day) to control and reduce the risk on the vulnerable assets of an organization.

C.Mitigation

Having a zero-day issue will not last just for a day, an organization can suffer from these attacks and not find a solution (an official patch) for a considerable amount of time, as mentioned in II-B. In this time gap, damage mitigation is needed; thus, security flaws may be less damaging or nonexistent until a solution would be available. Therefore, the zero-day mitigation approach will focus on the point where it begins until the point where an official patch is provided, in Figure 3 these points are represented from t_a to t_e.

To apply mitigation into the real world, different methodologies and standards can pursue that apply measures for different environments. In this aspect, the leading best practices should be implemented: monitoring the behavior of a resource, granting least privileges, only relying on verified sources, using white lists, and finally having backup measures in case of data loss.

An example that applies mitigation to zero-day is found in [22]. They propose an approach using a critical data sharing protocol in the scenarios with a potential zero-day threat, evaluating the risk that can categorize them to establish a level of confidence. In this process, the case of a zero-day attack is to guarantee that the least important data will be exposed, expecting an early detection that would not compromise more critical assets.

IV. DISCUSSION

The full zero-day cycle explains the whole process involved, the attacker's mind, and the interests behind these types of attacks. The complete cycle is not short, as is generally believed; it takes time to find a vulnerable point and, consequently, develop an exploit for these attacks. Stakeholders are involved

in this branch not only because of damaging or accessing susceptible systems. Behind it, the main interest is the money, as analyzed in section II-C2. Vulnerability markets move a massive amount of capital, regardless of the legal or illegal team.

It is not simple to join this world; researchers require a great set of skills and experiences to develop a zero-day attack. It is also essential an adequate computing power and resources for performing actions at this level and analyzing the different factors that a system may have for a potential flaw. Although this becomes more feasible today through virtualization and other solutions, it is a fact that the resulting attack change in a real environment, then it is necessary to perform tests and attacks on existing real assets. Thus, the threat actors must have security measures to hide his identity and conceal their attack (even more critical) because if the zero-day vulnerability is detected, it will become public and patched.

The life extension of a zero-day exploit is much longer than commonly thought. For this reason, in IT management, it could be necessary to take certain precautions respecting to vulnerabilities. Having out of date systems might remain dangerous for up to 10 years after the zero-day exploit has been launched (and patched). That is why it would be convenient to have a tool which could alert as soon as possible when a zero-day appears, and whether possible assets vulnerable are present to a new potential attack. In the next section are presented the existing solutions.

A.Existing solutions

Currently, exceptional studies exist that try to counteract the predicament of zero-day attacks; each individual is attempting to resolve this problem in a particular direction by concentrating on its objective with diverse methodologies. In Table I, some of the contemporary studies are presented, identifying their main aspects and approaches.

Firstly, Table I shows that most studies are focused on the network environment, and the main objective protected will be through it. Here, distinct techniques and mechanisms are used for distinct solutions; however, only [20] tries to give a real-time solution, which should be the most effective.

Furthermore, it reflects that the most of approaches propose to perform a zero-day detection([12], [19], [23], [24], [25], 26]); this is logical because if it could detect 100% of them, zero-days would cease to exist. However, it is not possible due to this vulnerability's nature; consequently, it is essential to think about other containment methods. Finally, this table shows that different detection strategies are employed as a solution; most of them were mentioned in III-A.

A single prevention method is used, which focuses on analyzing the risks that certain assets may offer to find preventive methods that follow specific standards. On the other hand, there are a few more solutions for mitigation ([22], [27], [28]), and it is essential to highlight the solution presented in [28], which includes two countermeasures in one, detection and mitigation. They decide to detect zero-day but are aware that their tools could fail, then they propose to have a mitigation mechanism through reliable protocols and different treatment to avoid having more vulnerable data exposed.

To conclude, there are limited approaches related to web attacks; this is "Tang2020" focuses on WAFs, which tries to detect web-only vulnerabilities such as SQLI (SQL injection), XSS (Cross-Site-Scripting), RCE (Remote Code Execution), among others. In the future, zero-day web-based solutions will be necessary to develop, as all assets are currently being moved to the cloud systems and related operations.

B.Open issues and research challenges

The main open issues are web applications, cloud computing, virtualization, and others omitted in common zero-day studies. The reason is that defending against cyberattack's surface represents a fundamental challenge, and the main issue is recognizing the point of attacks and the system vulnerabilities that cybercriminals could exploit [29]. These are the current trends and they are growing exponentially; that is why it is imperative to start studying these areas. However, massive companies related to this field have their research programs and their bug bounty programs, but this may not be enough in the real world because threat actors are constantly innovating and finding new ways of exploiting them, resulting in critical future problems that have not been analyzed yet.

Since 2017, an exponential increase in ransomware5 has risen, where different types of zero-day vulnerabilities have been exploited to create this malware. The losses are millionaires affecting companies as large as small, and almost 50 percent of these attacks end up with organizations losing

5A ransomware is considered a type of malware that implements cryptog- raphy to harm data from a device. It encrypts the victim's data with a secret key to block access from a genuine user [15].

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Table I: Various measures to counteract a zero-day attack.

Comparison of multiple countermeasures projects related to zero-day

Model	Countermeasure	Oriented to	Mechanism or Technique	Year
Towards probabilistic identification of zero-day attack paths [12]	Detection	Networks	* Networks based data * Statistical-based * Compute probabilities with a Bayesian network	2017
ZeroWall: Detecting Zero-Day Web Attacks through Encoder-Decoder Recurrent Neural Networks [22]	Detection	Web Application Firewalls (WAFs)	* Signature-based * Semantic patters in requested parameters	2020
The Performance of Machine and Deep Learning Classifiers in Detecting Zero-Day Vulnerabilities [23]	Detection	Networks Operating system	* 34 Machine/deep learning classifiers	2019
An Adaptive Real-Time Architecture for Zero-Day Threat Detection [20]	Detection	Honeypots in the networks	* Behavior-based * Real-time processing and classification	2018
Repids: A multi tier real-time payload-based intrusion detection system [24]	Detection	Networks	* Behavior-based * Reliable data * Low false alarms	2018
An Attack Graph Based Procedure for Risk Estimation of Zero-Day Attacks [21]	Prevention	Networks	* Attack Graph Based Security Metric * Using standards and good practices * Risk analysis	2016
LISABETH: automated content-based signature generator for zero-day polymorphic worms [25]	Detection	Malware Worms	* Devising algorithms * Signature and content based * Low false alarms	2008
A case study of unknown attack detection against zero-day worm in the honeynet environment [26]	Detection	Honeypot in networks Worms	* Traffic monitoring * Polymorphic recognition * Signature generation and based	2009
A framework for mitigating zero-day attacks in IoT [22]	Mitigation	IoT Networks	* Identified critical data * Reliable protocols	2018
Cyber resilience recovery model to combat zero-day malware attacks [27]	Mitigation	Networks	* Intrusion detection methods * Incident rate control * Using NIST SP-800-61 standard	2016
A Consensus Framework for Reliability and Mitigation of Zero-Day Attacks in IoT [28]	Detection Mitigation	IoT Networks	* Signature-based * Reliable data * Context behavior	2017

their data or paying criminals to recover them [30]. This problem results in a research challenge for controlling the future. Therefore, it is fundamental to consider this situation to mitigate and detect these growing and critical threats. future. Therefore, it is fundamental to consider this situation to mitigate and detect these growing and critical threats.
Deep learning and related methods usually require massive data, which means a powerful capacity of processing and other high computing characteristics. Consequently, trying to detect an anomaly when an attack is running could be useless or unlikely. New strategic approaches should be considered to defend and counterattack because traditional tactical strategies would not work in the future, such as an IDS with an approach in both detection and prevention capabilities [31].

V. CONCLUSIONS AND FUTURE WORK

Zero-day vulnerabilities and attacks can be highly critical for indefinite fields of computing. They operate a malicious behavior before the disclosure day, but can continue after their patch or until other solutions are released. Although the human factor may be the fault factor, prevention and mitigation could minimize the risk and avoid these issues.

This paper presented a detailed study of how a zero-day behaves and operates, from discovering the vulnerability to the attack performed in a real environment. This learning process includes a background of essential knowledge, and the zero-day cycle is developed in a general idea. Besides, understanding the threat actor's role is addressed to explain how it works behind the scene and assimilates all the back- ground motivational factors.

A state-of-the-art comparison was also exhibited regarding countermeasures, which explains the current solutions' approach and mechanisms. It shows that most of these solutions focus on detection, while prevention is underestimated, and limited solutions are available. It is advisable to continue digging on this approach. Finally, the open problems were exposed jointly with research challenges for future work to counteract the impact.

Conclusively, we plan to develop a zero-day laboratory that prevents and detects attacks launched on a day zero based on the reasons outlined in this study. Building on this can test several scenarios to avoid attacks by searching for assets, finding a vulnerability, and reducing the damage exposed there.

Furthermore, this archetype will be divided into two main modules. The first module has the task of preventing attacks published and dispersed throughout the web, analyzing their relevance and the assets' attack surface to be protected in an entity. However, this will be complemented by a detection module. The use of big data will be integrated for detection to analyze massive data to find behavioral anomalies that may present specific patterns of Zero-day attacks.

This proposal will differentiate this approach by having a hybrid model of prevention and detention simultaneously, besides applying big data to find spontaneous and random information not analyzed in standard strategies mentioned before. This method will aim to provide a solution that can detect anomalies with a low rate of false positives or false negatives.

Finally, the comparison in Table I showed that there is no systematic review of literature exclusive related to Zero-day studies. Therefore, conducting a SLR in this field is relevant and it is proposed as future work for the project members.

ACKNOWLEDGMENT

The authors would like to thank the financial support of the Ecuadorian Corporation for the Development of Research and Academy (RED CEDIA) for the development of this work, under Research Team GT-II-2018 (Cybersecurity). The research team was co-financed by the Research Department of the University of Cuenca (DIUC), Cuenca-Ecuador.

REFERENCES

[1] E. Chien, and L. O'Murchu, "Zero-day vulnerability: What it is, and how it works" [Online]. Available: https://us.norton.com/internetsecurity-emerging-threats-how-do-zero-day-vulnerabilities-work-30sectech.html [Accessed: Nov.25, 2020].

[2] S. Akshaya and G. Padmavathi. "A Study on Zero-Day Attacks," In Proceedings of International Conference on Sustainable Computing in Science (SUSCOM), pp. 2170–2177, 2019.

[3] A. Ye, Z. Guo, and Y. Ju, "Zero-Day Vulnerability Risk Assessment and Attack Path Analysis Using Security Metric," International Conference on Artificial Intelligence and Security, 11635(2016), pp. 266–278, 2019.

[4] P. Szor. "The art of computer virus research and defense". Pearson Education, 2005.

[5] L. Bilge, and T. Dumitras, "Investigating zero-day attacks," the magazine of USENIX & SAGE, 2013.

[6] MITRE. "Common Vulnerabilities and Exposures - CVE: The Standard for Information Security Vulnerability Names", 2019.

[7] L. Ablon, and A. Bogart, "Zero Days, Thousands of Nights: The Life and Times of Zero-Day Vulnerabilities and Their Exploits," Rand corporation, 2017.

[8] National Institute of Standards and Technology. "NVD – CVE-2010-2568" [Online]. Available: https://nvd.nist.gov/vuln/detail/CVE-2010-2568 , [Accessed: Nov.25, 2020].

[9] National Institute of Standards and Technology. "NVD - CVE-2020-5902"[Online]. Available: https://nvd.nist.gov/vuln/detail/CVE-2010-5902 [Accessed: Nov.25, 2020].

[10] F5 Networks. "Article: K52145254: TMUI RCE vulnerability CVE-2020-5902"[Online]. Available: https://support.f5.com/csp/article/K52145254 [Accessed: Nov.25, 2020].

[11] SHODAN Search engine. "BIG-IP affected Software", 2020.

[12] X. Sun, J. Dai, P. Liu, A. Singhal and J. Yen, "Towards probabilistic identification of zero-day attack paths," IEEE Conference on Communications and Network Security, CNS 2016, pp. 64–72, 2017.

[13] L. Ablon, M. Libicki, and A. Abler "Markets for Cyber-crime Tools and Stolen Data: Hackers' Bazaar," Rand Corporation, 2014.

[14] T. Walshe and A. Simpson, "An Empirical Study of Bug Bounty Programs," In IBF 2020 - Proceedings of the 2020 IEEE 2nd International Workshop on Intelligent Bug Fixing, 2020.

[15] X. Riofrío, F. Salinas Herrera and D. Galindo, "A Design for a Secure Malware Laboratory," In Advances in Intelligent Systems and Computing, volume 1099, pp. 273–286, 2019.

[16] L. Bilge and T. Dumitras, "Before we knew it: An empirical study of zero-day attacks in the real world," In Proceedings of the ACM Conference on Computer and Communications Security, 2012.

[17] L. Glanz, S. Schmidt, S. Wollny and B. Hermann, "A vulnerability's lifetime: Enhancing version information in CVE databases," In ACM International Conference Proceeding Series, volume 21-22-Octo, 2015.

[18] International Organization for Standardization. "ISO/IEC 27001:2013". Information technology — Security techniques — Information security management systems — Requirements, 2013.

[19] R. Tang, Z. Yang, Z. Li, W. Meng, H. Wang, Q. Li, Y. Sun, D. Pei, T. Wei, Y. Xu and Y. Liu, "ZeroWall: Detecting Zero-Day Web Attacks through Encoder-Decoder Recurrent Neural Networks," IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, pp. 2479–2488, 2020.

[20] A. Lobato, M. Lopez, I. Sanz, A. Cardenas, O. Duarte, and G. Pujolle, "An Adaptive Real-Time Architecture for Zero-Day Threat Detection," IEEE International Conference on Communications, 2018-May:1–6, 2018.

[21] M. Keramati, "An attack graph based procedure for risk estimation of zero-day attacks," In 2016 8th International Symposium on Telecom- munications (IST), pp. 723–728. IEEE, sep 2016.

[22] V. Sharma, J. Kim, S. Kwon, I. You, K. Lee and K. Yim, "A framework for mitigating zero-day attacks in IoT," eprint arXiv:1804.05549, pp. 1–4, 2018.

[23] F. Abri, S. Siami-Namini, M. Adl Khanghah, F. Mirza-Soltani and A. Siami-Namin, "The Performance of Machine and Deep LearningClassifiers in Detecting Zero-Day Vulnerabilitie," In Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, 2019.

[24] A. Jamdagni, Z. Tan, X. He, P. Nanda and R.Ping Liu, "RePIDS: A multi tier Real-time Payload-based Intrusion Detection System," Computer Networks, 2013.

[25] L. Cavallaro, A. Lanzi, L. Mayer and M. Monga, "LISABETH: Automated content-based signature generator for zero-day polymorphic worms," In Proceedings - International Conference on Software Engineering, 2008.

[26] I. Kim, D. Kim, B. Kim, Y. Choi, S. Yoon, J. Oh and J. Jongsoo "A case study of unknown attack detection against zero-day worm in the honeynet environment," In International Conference on Advanced Communication Technology, ICACT, 2009.

[27] H. Tran, E. Campos-Nanez, P. Fomin and J, Wasek, "Cyber resilience recovery model to combat zero-day malware attacks," Computers and Security, 2016.

[28] V. Sharma, K. Lee, S. Kwon, J. Kim, H. Park, K. Yim and S. Young Lee, "A Consensus Framework for Reliability and Mitigation of Zero-Day Attacks in IoT," Security and Communication Networks, 2017.

[29] M. Conti, T. Dargahi, and A. Dehghantanha. "Cyber threat intelligence: Challenges and opportunities". In Advances in Information Security. Springer, 2018.

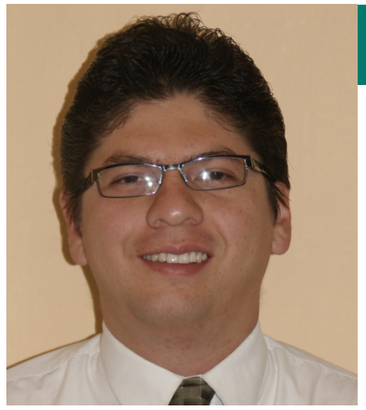
[30] A. Fagioli, "Zero-day recovery: the key to mitigating the ransomware threat," Computer Fraud and Security, 2019.

[31] K. Kim, M. Erza-Aminanto and H. Chandra, "Summary and further challenges," In Network Intrusion Detection using Deep Learning, Springer, pp. 69–70, 2018.

AUTHORS

Xavier Riofrío

Xavier Riofrío, a graduate computer science engineer from Universidad de Cuenca and a master of Cybersecurity with honours at the University of Birmingham certified by the UK government. The specialisation is in computer security and highly qualified for Penetration Testing, Ethical Hacking, amongst others. This leads me to be interested and enthusiastic about everything that entails the research field and Cybersecurity. That is why I have been part of diverse projects at the University of Cuenca.



Fabián Astudillo-Salinas

Darwin F. Astudillo-Salinas received the B.S.E (C.S) degree from Universidad de Cuenca, Cuenca, Ecuador, in 2007, and the M.S. and Ph.D. degrees from the "Institut National Polytechnique de Toulouse", Toulouse, France, in 2009 and 2013, respectively. Since 2013, he has been a Full-Time Researcher with the Department of Electrical, Electronic, and Telecommunications Engineering, Universidad de Cuenca, Cuenca, Ecuador. His research interests include network coding, wireless sensor networks, vehicular networks, networked control systems, simulation of networks, performance of networks and cybersecurity.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Luis Tello-Oquendo

Luis Tello-Oquendo received the electronic and computer engineering degree (Hons.) from Escuela Superior Politécnica de Chimborazo, Ecuador (2010), the M.Sc. degree in telecommunication technologies, systems, and networks (2013), and the Ph.D. degree (Cum Laude) in telecommunications from Universitat Polit\'ecnica de Val\'encia (UPV), Spain (2018). He was Graduate Research Assistant with the Broadband Internetworking Research Group, UPV (2013 - 2018) and Research Scholar with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA, USA (2016-2017). He is currently an Associate Professor with the Universidad Nacional de Chimborazo. His research interest includes 5G and beyond cellular systems, IoT, machine learning.

Jorge Merchan-Lima

Jorge Merchan-Lima,Since 2017 it has actively participated in research projects focused on digital signal processing, energy efficiency, data analysis, cybersecurity, and information security. I have written and published technical, scientific articles in national and international conferences.
Collaborator in the analysis of computer and electronic components in "FUNCIONA," a member of the cybersecurity working group in CEDIA, collaborator in a private company in compliance with ISO 27001, 27002, 27701, PCI, threat/vulnerability analysis, offensive security, and data protection.

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol VIII, Issue 1, January 2021

Ciclos Autónomos de Análisis de Datos basados en la Minería de Procesos para el Estudio del Comportamiento Curricular de los Estudiantes

Autonomous Cycles of Data Analysis based on Process Mining for the Study of the Curricular Behavior of Students

ARTICLE HISTORY

Received 01 September 2020
Accepted 02 November 2020

Sonia Duarte
Universidad Pedagógica Experimental Libertador
Instituto Pedagógico Rural Gervasio Rubio
Rubio, Venezuela.
sduarte@iprgr.upel.edu.ve

Jose Aguilar
CEMISID, Universidad de los Andes,
Mérida, Venezuela.
GIDITIC, Universidad EAFIT,
Medellín, Colombia
aguilar@ula.ve

Ciclos Autónomos de Análisis de Datos basados en la Minería de Procesos para el Estudio del Comportamiento Curricular de los Estudiantes

Autonomous Cycles of Data Analysis based on Process Mining for the Study of the Curricular Behavior of Students

Sonia Duarte
Universidad Pedagógica Experimental Libertador.
Instituto Pedagógico Rural Gervasio Rubio.
Rubio, Venezuela.
sduarte@iprgr.upel.edu.ve

Jose Aguilar
CEMISID, Universidad de los Andes,
Merida, Venezuela.
GIDITIC, Universidad EAFIT,
Medellin, Colombia
aguilar@ula.ve

Resumen— En este trabajo se evalúa el comportamiento curricular de los estudiantes de una carrera de maestría, a través de la Minería de Procesos. Específicamente, se analiza lo relacionado a la determinación de los factores internos y externos que inciden en la prosecución de sus estudios. Para comprender el comportamiento del estudiante, se usa la metodología MIDANO, la cual ha sido usada para el desarrollo de aplicaciones de analítica de datos. En particular, se especifican los Ciclos Autonómicos de tareas de análisis de datos que permiten estudiar el abandono de la maestría durante la escolaridad o durante el desarrollo del trabajo de grado, con el fin de determinar las causas o problemas que se presentan durante la prosecución de los estudios. Se obtuvieron resultados muy alentadores sobre las causas del abandono de la maestría que descubren los ciclos autónomos.

Palabras claves— *Minería de Procesos, Analítica de Datos, comportamiento curricular, Analítica de Aprendizaje*

Abstract— In this work, the curricular behavior of the students of a master's degree program is evaluated through Process Mining. Specifically, what is related to the determination of the internal and external factors that affect the pursuit of their studies is analyzed. To understand student behavior, the MIDANO methodology is used, which has been used for the development of data analytics applications. In particular, it is specified the Autonomous Cycles of data analysis tasks that allow studying the dropout of the master's degree program during schooling or during the development of graduate thesis, in order to determine the causes or problems that arise during the pursuit of the studies. Very encouraging results were obtained on the causes of the dropout of the master's degree program, which discover the autonomous cycles..

Keywords— *Process Mining, Data Analytics, curricular behavior, Learning Analytics.*

I. INTRODUCCIÓN

El currículo es diseñado por las Instituciones Educativas, para lograr objetivos que se han planteado. En tal sentido, para su logro se establecen restricciones, como, por ejemplo, los estudiantes deben considerar unos cursos específicos antes de ver otros, determinadas unidades de crédito por lapso académico, la disposición de unidades curriculares electivas, entre otros aspectos. Aunado a lo expuesto, se establecen las ofertas de las unidades curriculares por lapso académico, lo que puede ocasionar que estudiantes tomen diferentes caminos una vez se inicie la carrera. De esta manera, para algunos estudiantes, el proceso de prosecución de sus estudios puede resultar más exitoso que para otros. En consecuencia, resulta conveniente recomendar que los estudiantes tomen caminos más apropiados. Para ello, se puede recurrir al análisis de datos de los historiales académicos, con la finalidad de extraer conocimiento.

Por otra parte, a partir de esos datos, su análisis es una necesidad, a través de la aplicación de las diferentes técnicas de minerías, bien sea de datos, semántica, de proceso, entre otras, ya que apertura posibilidades para responder a interrogantes como: ¿Son necesarios los requisitos exigidos para cursar una materia u otra?, ¿Qué tan probable es que un discente finiquite su carrera o la abandone?, ¿Existen otras posibilidades para terminar la carrera?
Desde estas perspectivas, este artículo presenta una herramienta de analítica de datos, para la evaluación del comportamiento curricular de una carrera de maestría, basada entre otras cosas, en la minería de procesos, con el fin de determinar los factores que pudieran incidir y coartar el finiquito de su carrera. Para ello, en este trabajo se lleva a cabo la evaluación comparando el modelo curricular formal del programa mencionado con el modelo real que siguen los estudiantes. Con ello, se determinan los factores a los cuales están expuestos los estudiantes, que en definitiva marcan su prosecución.

En general, es difícil encontrar trabajos cercanos a nuestra propuesta. Algunos trabajos similares al nuestro son los siguientes: el trabajo de Bogarín [1] propone patrones seguidos por los estudiantes durante el proceso de aprendizaje, aplicando técnicas de la minería de procesos sobre los datos generados de la interacción de los estudiantes con el entorno virtual de aprendizaje (EVA) Moodle. Por su parte, Wang et al [2] examinan los caminos exitosos que los estudiantes deben tomar para lograr la consecución de sus estudios, descubriéndolos por medio de simulaciones, y considerando los cursos tomados por los estudiantes.

Por su parte, en Pechenizkiy et al. [3] descubren diferentes estilos individuales de navegación de los estudiantes, aplicando dos pruebas en dos EVAs diferentes, Moodle y Sakai. Para ello, aplican la minería de procesos, y proponen estrategias para reducir la carga cognitiva, y mejorar la facilidad de uso y la eficiencia de aprendizaje de los sistemas de e-Learning. Por su parte, en [4] aplican la minería de procesos a la capacitación profesional, y definen métodos genéricos para ser aplicados en materia de formación profesional en e-learning. Por otro lado, también se ha estudiado el rendimiento del aprendizaje, como el trabajo de Sedrakyan et al. [5], quienes identificaron patrones del proceso de aprendizaje, identificando los peores o mejores resultados del rendimiento de aprendizaje. En [6] presentan el proceso de desarrollo del plan de estudios de una carrera de Ingeniería del Tecnológico de Monterrey, en México, mediante el uso de mapas conceptuales para ayudar a caracterizar los cursos y sus interconexiones. Usan una herramienta llamada STAUNCH, que permite evaluar la cobertura de los cursos y su contribución individual y colectiva al plan de estudios desde una perspectiva sistémica.

En cuanto al uso de la analítica de datos para entornos educativos, actualmente existen una variedad de estudios, entre los que se puede mencionar el que analiza el rendimiento de la plataforma Moodle, de Moreno, et al. [7], quienes usaron MIDANO para especificar Ciclos Autonómicos (CAs) de tareas de análisis de datos para estudiar los subprocesos de Moodle relacionados con la carga de datos y descarga de archivos, con el fin de mejorar el rendimiento de los medios de almacenamiento. Por otro lado, Aguilar et. al [8] incorporaron la Analítica Social de Aprendizaje (SLA) en aulas inteligentes (SaCI), para mejorar sus procesos de enseñanza-aprendizaje, usando el concepto de CAs de análisis de datos y MIDANO, lo que permite construir un mejor perfil de aprendizaje de un estudiante, y de esta manera, detectar estilos de aprendizaje para grupos de cursos, e incluso carreras. Por otro lado, en [9] se examina el aprendizaje y el análisis académico y su relevancia para la educación a distancia en programas de pregrado y posgrado. El objetivo del trabajo es explorar los datos como predictores del éxito de los estudiantes e impulsores del plan de estudios del programa. [10] explora la aplicabilidad de la analítica de aprendizaje para la predicción del desarrollo de dos competencias transversales: trabajo en equipo y compromiso, basado en el análisis de los registros de datos de interacción de Moodle en un programa de Maestría en la Universidad a Distancia de Madrid (UDIMA).

Particularmente, en este trabajo se especifica la arquitectura computacional del sistema de análisis del comportamiento curricular de los estudiantes, y se realiza el desarrollo de un prototipo del mismo. Seguidamente, se toma como caso de estudio, la evaluación del comportamiento curricular de un Programa de maestría, denominado: Innovaciones Educativas, para descubrir los cuellos de botella existentes en el proceso educativo, y los nodos problemáticos que presenta el estudiante en la carrera. En dicho caso, se realizan experimentos que permitan evaluar el comportamiento curricular del programa educativo con nuestra herramienta. Para el desarrollo de este trabajo se utilizó la metodología de MIDANO [11], la cual ha sido usada para el desarrollo de aplicaciones de analítica de datos, la cual permite el desarrollo de CAs de tarea de análisis de datos (AdD), introducidos en [8, 12], con el fin de integrar y automatizar las actividades de analítica de datos que permiten descubrir y utilizar el conocimiento del proceso analizado, para incidir en el mismo.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

II. CICLO AUTONÓMICO DE TAREAS DE ANALISIS DE DATOS

La analítica de datos es usada para examinar los datos, con la finalidad de buscar conocimiento. Al respecto, en [13] se señala que "es la ciencia de la recogida, almacenamiento, extracción, limpieza, transformación, agregación y análisis de datos con el fin de descubrir información y conocimiento". Por otro lado, toda actividad de analítica de datos debe integrarse y automatizarse. Es por ello que se ha desarrollado el concepto de CAs de tareas de AdD para cada objetivo estratégico del proceso a estudiar. Al respecto, en Aguilar et al. [14, 15] señalan que un "Ciclo Autonómico de Tareas de Analítica de Datos puede definirse como un conjunto de tareas de Análisis de datos cuyo propósito es mejorar el proceso que estudia. Este conjunto de tareas interactúa entre ellas, y tienen diferentes roles: observar el proceso, analizar e interpretar lo que pasa en él, y tomar decisiones para mejorarlo" (ver Fig. 1).

Fig. 1. Ciclo autonómico de tareas de análisis de datos.

La integración de las tareas en un ciclo, visto como un lazo cerrado, permite la resolución autónoma de problemas. En general, los posibles roles de cada tarea de AdD del ciclo son [12, 15]:

- Tareas de monitoreo: son responsables de obtener las variables del proceso a estudiar. Estas tareas monitorean el proceso, ven su comportamiento, y extraen los datos que los describen.

- Tareas de Análisis del sistema: son las responsables de la interpretación de lo que sucede en el proceso. Con ellas, se puede diagnosticar, entender, analizar, entre otras cosas, lo que sucede. Para ello, se construyen modelos de conocimiento (por ejemplo, de predicción, descripción, etc.).

- Tareas de toma de decisiones: definen las acciones a ejecutar a efectos de mejorar el proceso, considerando para ello el objetivo planteado para el CA.

III. CASO DE ESTUDIO: COMPORTAMIENTO CURRICULAR DE LOS ESTUDIANTES

Para el desarrollo de la arquitectura computacional del sistema, para analizar los problemas de abandono, se usa la metodología MIDANO. Como un primer paso, se analizan los procesos del programa educativo, con la finalidad de determinar los procesos viables para la aplicación de analítica de datos, vinculados al problema a analizar. Los procesos del postgrado de innovaciones educativas se muestran en la Tabla I.

De la Tabla I, claramente se identifica el proceso de prosecución del estudiante (modelo curricular), como el proceso objeto de estudio. Se procede a realizar los Ciclos autónomos para dicho proceso.

Tabla I. procesos de innovaciones educativas.

PROCESOS	
Pre-inscripción	PI
Inscripción curso introductorio	ICI
Inscripción	INSC
Prosecución del estudiante (Modelo curricular)	PE
Asesorías	AS
Emisión constancias	EC
Retiro materias	RM
Retiro semestre o Universidad	RSU
Reingreso	REI
Intención investigativa	IIN
Transcripción de calificaciones	TC
Culminación de estudios Grado	CEG

A. Ciclo Autonómico "abandono de maestría durante la escolaridad"

Este ciclo tiene como objetivo descubrir el patrón de comportamiento curricular que sigue un estudiante (caminos exitosos y no exitosos), destacando las causas o problemas que se le presentan en su prosecución. En tal sentido, sus tareas de análisis de datos son especificadas en la Tabla II y Fig. 2.

Tabla II. Grupo de tareas del ciclo autonómico: abandono de la maestría durante la escolaridad.

	Nombres	Fuentes generales de datos requeridas	Indicadores generados	Efectos esperados sobre objetivo estratégico
Tareas de AdD de observación	Caracterizar los estudiantes de maestría.	Base de datos (histórico académico) con registro de ID del estudiante, materias cursadas, fecha de inscripción de materia, fecha de finiquito de materia, reingresos, retiros de materias o semestre. Base de datos (datos externos): movilidad o transporte, acceso a internet, electricidad, influencia de otros compañeros, apoyo y motivación por parte de terceros. Base de datos personales: Si trabaja o no, si tiene recursos económicos o no, si es cabeza de hogar, si recibe ayuda de los padres, si tiene recursos computacionales.	Modelo de Minería de procesos-descubrimiento y conformidad	Organización y definición de los datos para la observación del comportamiento real de los estudiantes
Tareas de AdD de análisis	Identificar los estudiantes exitosos y no exitosos.			Agrupación de los estudiantes e identificar patrones
	Estudiar el flujo académico.			Descubrimiento de cuellos de botella, nodos críticos.
	Comparar los flujos para determinar el patrón de no exitosos.			Confrontación de las situaciones generadas en los caminos no exitosos que considera el estudiante en la prosecución de sus estudios.
Tareas de AdD de toma de decisiones	Determinar las causas o problemas que aparecen en el flujo de los no exitosos.			Generar la necesidad de documentar al estudiante para inducirlo por el camino más corto y correcto para el feliz término de la maestría.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Fig. 2. CA abandono de maestría durante la escolaridad.

Siguiendo la metodología de MIDANO, se describen las tareas de análisis de datos (Ver Tablas III, IV, V, VI, VII) para los diferentes CAs.

Tabla III. Tarea de observación 1 del ca abandono temporal o definitivo de la maestría durante la escolaridad.

Nombre de la tarea	Caracterizar los estudiantes de maestría
Descripción	Se seleccionarán los estudiantes, determinando mecanismos de extracción, muestra de estudiantes a usar, organizando los datos en formatos óptimos para ser utilizados en Minería de Procesos.
Tipo de tarea de analítica de datos	Clasificación
Técnicas de analítica de datos	Minería de datos
Tipo de modelo de conocimiento	Modelo de clasificación

Tabla IV. Tarea de análisis 2 del ca: abandono temporal o definitivo de la maestría durante la escolaridad.

Nombre de la tarea	Identificar los estudiantes exitosos y no exitosos
Descripción	Se obtiene información de lo que sucede realmente con los caminos exitosos y no exitosos tomados por los estudiantes durante la maestría (comportamiento real del estudiante).
Tipo de tarea de analítica de datos	Clasificación
Técnicas de analítica de datos	Minería de datos y Minería de Procesos
Tipo de modelo de conocimiento	Modelo de clasificación

Tabla V. Tarea de análisis 3 del ca: abandono temporal o definitivo de la maestría durante la escolaridad.

Nombre de la tarea	Estudiar el flujo académico
Descripción	Se obtiene información de lo que sucede realmente durante la prosecución del estudiante (modelo curricular real), en que momentos se quiebra la continuidad de los estudios de maestría.
Tipo de tarea de analítica de datos	Asociación
Técnicas de analítica de datos	Minería de procesos
Tipo de modelo de conocimiento	Descubrimiento

Tabla VI. Tarea de análisis 4 del ca: abandono temporal o definitivo de la maestría durante la escolaridad.

Nombre de la tarea	Comparar los flujos para determinar el patrón de no exitosos.
Descripción	Verificar el modelo formal y el patrón que sigue el estudiante a fin de diagnosticar el comportamiento del estudiante y las realidades que les aqueja.
Tipo de tarea de analítica de datos	Comparación
Técnicas de analítica de datos	Minería de procesos
Tipo de modelo de conocimiento	Conformidad

Tabla VII. Tarea de toma de decisiones 1 del ca: abandono temporal o definitivo de la maestría durante la escolaridad.

Nombre de la tarea	Determinar las causas o problemas que aparecen en el flujo de los no exitosos.
Descripción	Se obtiene información de los elementos que inciden en los estudiantes que abandonan la maestría durante la escolaridad.
Tipo de tarea de analítica de datos	Asociación
Técnicas de analítica de datos	Analítica de datos social-estadística
Tipo de modelo de conocimiento	Estadística

El modelo de datos multidimensional del ciclo autónomo de abandono temporal o definitivo de la maestría durante la escolaridad se muestra en la Fig. 3. La tabla de hechos es "abandono_ escolaridad" y a su vez, tiene cinco tablas de dimensiones que definen características específicas.

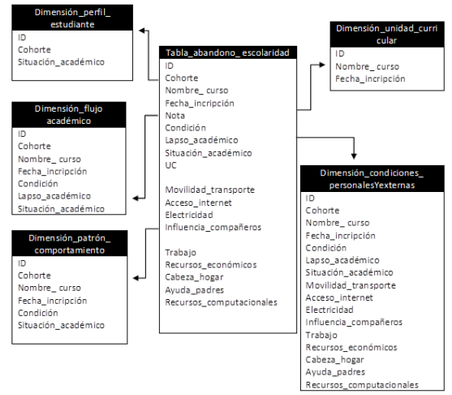


Fig. 3. Modelo de datos CA: abandono temporal o definitivo de la maestría durante la escolaridad.

El modelo de la Fig. 3 incluye datos del postgrado, y datos extraídos de Facebook, WhatsApp (redes sociales). Cada una de esta información está incluida en una dimensión diferente en el modelo de datos. Así, tenemos:

- Dimensión perfil estudiante: Son almacenados datos que definen al estudiante, como la cohorte en que inicia sus estudios y en qué situación académica se encuentra.

- Dimensión flujo académico: Permite conocer el camino que el estudiante recorre durante la prosecución de sus estudios.

- Dimensión patrón de comportamiento: Muestra los grupos de estudiantes en sus caminos exitosos y no exitosos

- Dimensión unidades curriculares: Muestran las unidades curriculares que ha tomado el estudiante en su recorrido durante la maestría.

- Dimensión experimentos: almacena las condiciones de los estudiantes: su situación académica, tiempos de permanencia en la maestría, entre otros.

- Dimensión condiciones externas y personales: almacena los factores del entorno que pueden afectar la prosecución del estudiante.

B. Ciclo Autonómico "Abandono temporal o definitiva de la maestría durante la intención investigativa y trabajo de grado

Este ciclo tiene como objetivo descubrir el patrón de comportamiento que sigue el estudiante durante el trabajo de grado, a fin de llevar a cabo el análisis del modelo curricular, destacando las causas o problemas que se presentan en el finiquito de la maestría. La Tabla VIII y Fig. 4 muestran las tareas de análisis de datos.

Tabla VIII. Grupo de tareas del ciclo autonómico de intención investigativa y trabajo de grado.

	Nombres	Fuentes generales de datos requeridas	Indicadores generados	Efectos esperados sobre objetivo estratégico
Tareas de AdD de observación	Caracterizar los estudiantes de maestría que están en la última etapa curricular	Base de datos con registro de fechas de inscripción de intención investigativa o trabajo de grado, reingresos.	Modelo de minería de procesos-descubrimiento	Organización y definición de los datos para la observación del comportamiento real de los estudiantes.
Tareas de AdD de análisis	Identificar los estudiantes exitosos y no exitosos de la última parte			Agrupación de los estudiantes e identificar patrones
	Estudiar la última parte curricular		Estadísticas de minería de procesos	Descubrimiento de cuellos de botella, nodos críticos.
	Analizar las variables del entorno externo			Confrontación de las situaciones generadas en los caminos no exitosos que considera el estudiante en la prosecución de sus estudios.
Tareas de AdD de toma de decisiones	Determinar las causas que afectan el finiquito de la intención investigativa o trabajo de grado.		Implicaciones	Generar la necesidad de documentar al estudiante para inducirlo por el camino más corto y correcto para el feliz término de la maestría.

Tarea 3. Descubrir patrones en los flujos académicos de los estudiantes: A partir del flujo general de la Fig. 6, donde se consideran todos los flujos de los estudiantes exitosos, se trata de construir el patrón de flujos exitosos en la Fig. 7. Se considera como patrón de los exitosos al flujo de los casos que tomaron más el mismo camino (es lo que muestra la Fig. 9). El promedio de duración es equivalente a 1,99 años.

Fig. 4. Representación gráfica CA abandono de maestría durante trabajo de grado o intención investigativa.

La especificación detallada de las tareas, y del modelo de datos, se hace de una manera similar a la indicada en la sección CA "abandono de maestría durante la escolaridad" para el primer ciclo autónomo. Para las siguientes secciones, se especifica cómo implementar el primer ciclo autónomo.

IV. EXPERIMENTACIÓN

Para los experimentos, se trabaja con la base de datos SIPOST del Instituto Pedagógico Rural Gervasio Rubio, alimentada por los eventos que se registran durante la prosecución de los estudiantes desde el año 2015 al 2019, considerándose en su totalidad 166 casos o estudiantes y 1955 eventos. Por otra parte, se consideran datos personales y del entorno del estudiante, obtenidos del sistema académico o redes sociales.

Se consideran, simulaciones para tres escenarios. Un primer experimento analiza si los factores externos son causas considerables en los tiempos de permanencia del estudiante durante la escolaridad de la maestría. En un segundo experimento, se busca determinar la incidencia de las características personales en el éxito o no del estudiante; para finalizar con un tercer experimento, donde se determina la incidencia de las unidades curriculares y unidades crédito cursadas por lapso académico.

A. Escenario experimental 1

En este experimento, se determina si el flujo curricular de cada estudiante está condicionado por su entorno: transporte, situación social: luz, internet, gas, agua, gasolina, y situación política del país: marchas, concentraciones, paros. La hipótesis es: ¿Son los factores externos causas considerables en los tiempos de permanencia del estudiante durante la escolaridad de la maestría? A continuación, se presenta el comportamiento de cada tarea del ciclo en este escenario experimental.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Tarea 1: Caracterizar los datos de los estudiantes:

Para la selección de estudiantes, se considera su situación académica: egresado, graduando, regular, participante especial, sin inscripción, retiro permanente. De esta manera, determinar su estatus.

Tarea 2. Identificar caminos exitosos y no exitosos: se identifican los estudiantes exitosos (que culminaron sus estudios). Primero que nada, se determinan los caminos tomados por los estudiantes; sin embargo, se recurre a varios filtros. Se considera el tiempo promedio de 2 años efectivos para el finiquito de unidades curriculares, para ser considerado un camino como exitoso (finiquito de la escolaridad de manera exitosa). De esta manera, se logra determinar los casos que cumplen con el requerimiento de haber cursado todas las unidades curriculares hasta trabajo de grado inscrito, generando el modelo de procesos (descubrimiento) que se muestra en la Fig. 5 para los flujos exitosos.

En la Fig. 5, se puede visualizar en los círculos

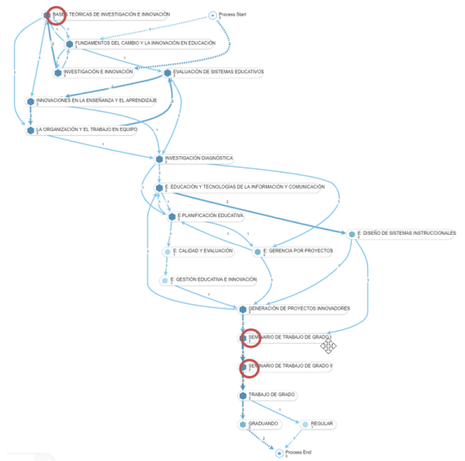


Fig. 5. Modelo de Procesos general de los caminos exitosos de los estudiantes.

rojos los nodos (cursos) más comunes seguidos por los estudiantes. El resto de flujos son los flujos no exitosos. Se consideran caminos no exitosos, los casos con más de 2 años efectivos en la escolaridad. En la Fig. 6, se muestra el modelo de procesos (descubrimiento) de los casos no exitosos definidos por esta tarea. Igualmente, los círculos rojos representan los nodos (cursos) más seguidos en los caminos no exitosos.

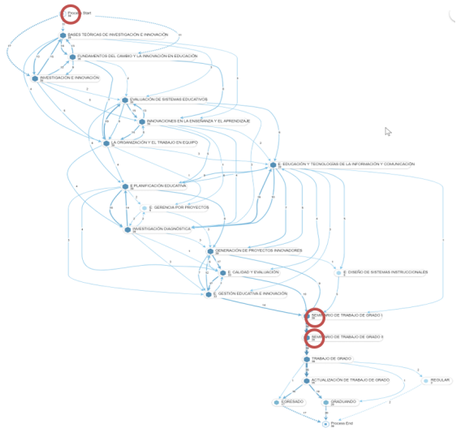


Fig. 6. Modelo de Procesos general de los caminos no exitosos durante la escolaridad.

Fig. 7. Patrón casos estudiantes exitosos.

El mismo procedimiento se sigue para determinar el patrón de los caminos no exitosos.

Tarea 4. Comparar con el modelo curricular: En esta tarea se realiza la comparación del patrón del flujo de los estudiantes no exitosos con el del pensum de estudio. Para ello, se contrasta el grafo del pensum con el grafo que siguen los estudiantes no exitosos, como se muestra en la Fig. 8.

En la Fig. 8, se aprecia la condición de los estudiantes, reflejándose en círculos azules cuantos egresados y cuantos en condición de preparación del trabajo de grado (lado izquierdo). Además, se compara ambos grafos o modelos, determinándose discrepancia en la cantidad de unidades curriculares que se deberían cursar por lapso académico (círculos azules del grafo del lado derecho) y las que realmente son tomadas por los estudiantes (círculos rojos del grafo del lado izquierdo, resaltándose el lapso académico adicional). En este sentido, se determina que existe un desplazamiento de un lapso académico adicional, que genera incidencia en la prosecución del estudiante durante los estudios de la escolaridad.

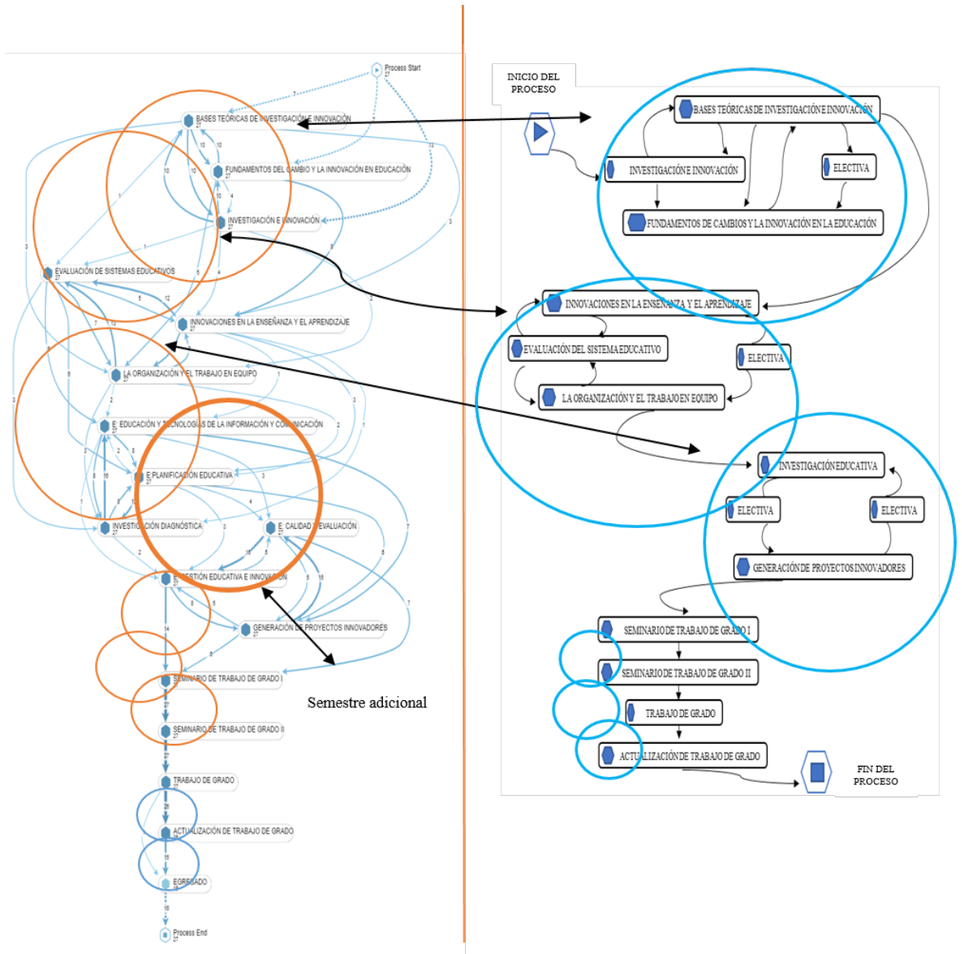


Fig. 8. Patrón de los no exitosos Vs Modelo del pensum de estudio.

Tarea 5: Determinar posibles causas de no éxito: Con el patrón de flujo de los estudiantes no exitosos determinados en la tarea 4, se pasa a obtener información de tipo social a través de datos personales y externos de estos estudiantes, los cuales se obtienen de las planillas de inscripción que reposan en control de estudios y en las diferentes redes sociales (Facebook, WhatsApp, Twitter). Para ello, se consideran aspectos del entorno del estudiante, como: situación social: transporte público, luz, internet, gas, agua, gasolina y situación política del país:marchas, concentraciones, paros. Las estadísticas de los estudiantes de los casos no exitosos se muestran en la Tabla 9, ordenada por su grado de afectación.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

De la Tabla IX, se puede destacar que los factores electricidad, transporte público, acceso a internet, agua, y paros tienen una incidencia alta (entre 85 y 100%). Por otra parte, el factor gasolina incide en un 74%. Además, la falta del gas tiene una incidencia del 44%. Por último, la incidencia de marchas y concentraciones es baja, con un 37%. En conclusión, en este escenario se determina que la mayoría de los factores del entorno del estudiante considerados como posible causa han incidido en la prosecución del estudiante de manera alta durante su carrera. Ahora bien, el uso del gas, marchas y concentraciones tienen menor incidencia. Por otra parte, los factores de agua, y electricidad tienen la incidencia más alta. En cuanto a la electricidad, significa que los estudiantes se ven más afectados y no logran desarrollar sus trabajos a tiempo, y en cuanto a los paros, atrasa sus tiempos de prosecución, extendiéndose sus períodos académicos.

Tabla IX. Incidencia del entorno del estudiante.

Factor	% Incidencia
Electricidad	100
Paros	96
Agua	93
Acceso a internet	89
Transporte público	85
Gasolina	74
Gas	44
Marchas	37

B. Escenario experimental 2

El objetivo de este escenario es determinar si el flujo de cada estudiante está condicionado por los hijos, trabajo, alimentación, ubicación, cabeza de hogar, dependencia económica, computador. La hipótesis es ¿Son las situaciones personales causas de los tiempos de permanencia del estudiante durante la escolaridad de la maestría?

En general, se recaba la información de la misma manera que para el experimento 1 (las primeras 4 tareas son iguales al caso anterior), evaluándose en esta oportunidad los aspectos personales y su incidencia, obteniéndose la Tabla X, que corresponde al estudio estadístico que representan el patrón de los estudiantes no exitosos.

En la Tabla X, se puede apreciar el porcentaje que representa cada uno de los factores personales del estudiante, destacándose que los factores de poseer computador y si trabaja tienen una incidencia del 96%, las cuales son más altas, infiriéndose que casi en su totalidad los estudiantes deben trabajar para mantenerse, presentando dificultad para desarrollar sus actividades académicas. También, el hecho de poseer un computador, inciden mucho en la prosecución del estudiante. En cuanto al hecho de tener hijos, cabeza de hogar, dependencia de su alimentación, dependencia de la alimentación de la familia, representan del 70 al 78%. Por último, el aspecto de dificultad en alguna unidad curricular sólo representa el 11 % de incidencia, que significa que hubo un pequeño porcentaje de estudiantes que presentaron problemas al momento de cursar materias.

Tabla X. Incidencia de los aspectos personales del estudiante.

Factor	% Incidencia
Trabajo	96
Computador	96
Cabeza de hogar	78
Dependencia de la alimentación de la familia	78
Dependencia de su alimentación	76
Hijos	70
Dificultad en unidad curricular	11

C. Escenario experimental 3

El objetivo es determinar si el flujo que sigue cada estudiante no exitoso está condicionado por unidades curriculares y/o unidades de créditos cursadas por lapso académico. La hipótesis es ¿Es el flujo que sigue el estudiante el que repercute en su éxito?

Las cuatro primeras tareas son las mismas de los experimentos 1 y 2. Así en esta sección solo se presenta el desarrollo de la tarea 5. Para ello, se analiza el flujo académico de los estudiantes que resultaron no exitosos. Particularmente, se comparan los patrones de los estudiantes no exitosos de la tarea 4 con el modelo formal del pensum, detectándose que existe discrepancia entre lo que se establece en el pensum y la prosecución real del estudiante, como se muestra en la Fig. 9.

En la Fig. 9, se pueden apreciar los lapsos académicos cursados por los estudiantes no exitosos durante la escolaridad (grafo a la derecha, círculos rojos intensos). Se determina que cursaron 6 lapsos académicos, lo que incide notoriamente en los tiempos de permanencia durante la escolaridad, ya el estudiante debe cursar un semestre adicional, un tiempo no previsto según el pensum de estudios, el cual establece 5 lapsos académicos (grafo a la izquierda, círculos rojo pálido). Ahora bien, viendo más de cerca el pensum, se logra detectar que en los primeros lapsos académicos debería verse una unidad curricular electiva (círculos en verde), y ésta no se ve en los dos primeros lapsos académicos. Igualmente, en el tercer lapso deberían verse 2 electivas y se ve 1, lo que desenlaza en un definitivo mayor tiempo durante la escolaridad.

Así, efectivamente el flujo que sigue el estudiante repercute en su éxito, ya que se evidencia que los estudiantes cursan menos unidades curriculares de las que se debería en los primeros lapsos, de acuerdo al Reglamento de Estudios de Postgrado [16].

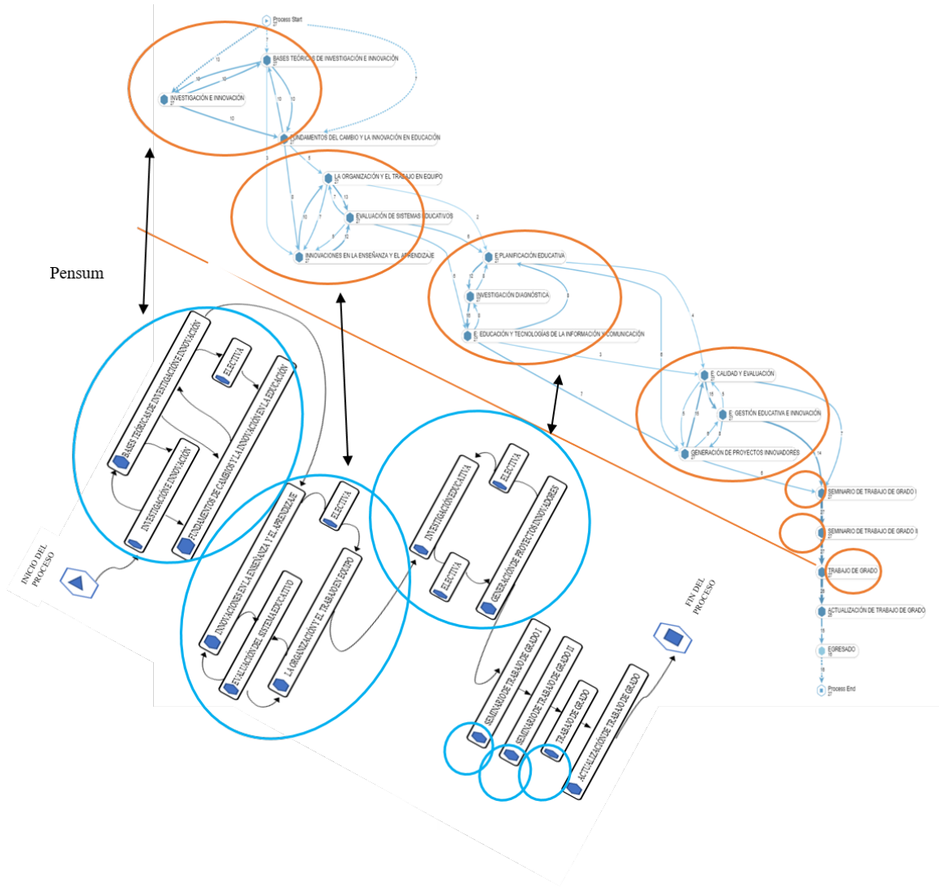


Fig. 9. Comparación de patrón no exitoso Vs Pensum formal.

D.Comparación con otros Trabajos similares

En general, los modelos curriculares no han sido estudiados ampliamente, a través de la minería de procesos. En cuanto al estudio de modelos curriculares por medio de minería de procesos, sólo un estudio previo consideró examinar caminos exitosos en la consecución de los estudios a través de la minería de procesos [4], sin embargo, no considera ninguna otra minería. El uso de minería de procesos ha sido empleada como única minería en varias investigaciones previas. En [1] se limitan a sólo mostrar diferentes modelos de descubrimiento aplicando varios algoritmos, y posteriormente los comparan. En [1] aplican minería de procesos para descubrir modelos y flujos de trabajo, para investigar los comportamientos de los usuarios. Por otro lado, en [2] se aplica minería de procesos a fin de determinar caminos exitosos. Asimismo, [5] sólo la aplican para analizar los registros de eventos de los trabajos grupales de los estudiantes.

Por otro lado, los ciclos autónomos de tareas de análisis de datos han sido empleados en [7], para desarrollar las implementaciones específicas de optimización de un EVA, para acelerar los procesos de transmisión de archivos y evitar la pérdida de datos. A su vez, [8] lo emplean para mejorar el proceso de enseñanza-aprendizaje que se da en un salón inteligente, al obtener el estilo de aprendizaje para un grupo de cursos. Finalmente, [12, 13] diseñan ciclos autónomos de tareas de analítica de aprendizaje.

Existen propuestas de un enfoque de minería de cualquier cosa (minería de datos, MP, minería semántica), en el proceso de aprendizaje. En [6] ha sido propuesta la minería de datos, tales como árboles de decisión, reglas de asociación, redes neuronales, entre otros, porque se requiere la optimización del proceso de evaluación. También, [8] emplea

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

minería de cualquier cosa para obtener el estilo de aprendizaje de los estudiantes en un salón de clases. Se aplica la analítica social de aprendizaje para determinar el estilo de aprendizaje adecuado para los estudiantes del curso seleccionado, usando datos de redes sociales [8].

Por otro lado, investigaciones previas usan fuentes externas de datos, como por ejemplo [7], quienes al analizar un EVA pueden considerar para la optimización, variables personales y externas que inciden en el proceso. Por su parte, [1], al presentar la rutas de aprendizaje, pueden escalar la investigación. Al comparar el presente estudio con investigaciones previas, se pueden destacar las diferencias de nuestra propuesta:

- Hace un análisis de modelos curriculares, lo cual ha sido poco estudiado, encontrándose sólo una propuesta.

- No sólo se aplica técnicas de minería de procesos, sino que va más allá, incorporando minería de datos, analítica de datos social-estadística, entre otras tareas.

- Al incorporar datos personales y del entorno para determinar las causas, se permite no sólo analizar los grafos de la prosecución personalizados, como lo hacen la mayoría de las investigaciones, pero además, se consideran otros elementos que afectan a los estudiantes y que influyen en el éxito de sus estudios.

Por último, la investigación es escalable porque se puede extender con otras tareas de análisis de datos para estudiar otros aspectos más específicos, como por ejemplo, factores locales de donde se vive.

V. CONCLUSIONES

En el trabajo se propuso un enfoque que usa la minería de procesos, entre otros enfoques de minería, para evaluar el comportamiento curricular que siguen los estudiantes de la maestría de innovaciones educativas. El estudio se desarrolla incorporando el concepto de CA de tareas de análisis de datos, el cual permite alcanzar objetivos específicos de análisis, para ayudar a tomar decisiones estratégicas.

La incorporación de la Minería de cualquier cosa, incluyendo la MP, en los CAs de tareas para analizar el abandono de la maestría durante la escolaridad, permite descubrir los modelos de los flujos que sigue el estudiante en la prosecución de sus estudios. En particular, para la experimentación se seleccionó el Ciclo “Abandono de la maestría durante la escolaridad”. Se desarrollaron 3 escenarios, un primer escenario tenía como objetivo determinar si el flujo de cada estudiante está condicionado por su contexto social, un segundo objetivo tenía como objetivo determinar si el flujo de cada estudiante está condicionado por entorno personal, y un tercer escenario tenía como objetivo determinar si el flujo que sigue cada estudiante está condicionado por el modelo académico de la maestría.

De acuerdo con la revisión de la literatura, no hay precedentes del uso de la metodología MIDANO para estudios en modelos curriculares. En particular, el ciclo de abandono de la maestría durante la escolaridad permite estudiar plenamente las causas que aquejan al estudiante. Con esta investigación, se crea el precedente de que con el uso de ciclos basado en tareas de minería de cualquier cosa (incluyendo la MP), es posible generar modelos de comportamiento curricular que emprenden los estudiantes en instituciones educativas. Todo ello, basado en los registros de datos académicos, eventualmente enriquecidos con datos del entorno (redes sociales, etc.). Así, es posible determinar los recorridos curriculares más exitosos o no, aspectos que puedan influir en esos recorridos exitosos o no, con el fin de determinar eventuales acciones correctivas. Un ciclo autónomo es una gran ayuda en la toma de decisiones correctivas, debido a que es capaz de generar conocimiento, y con ello, ayuda a determinar las decisiones que favorezcan el desempeño estudiantil.

Trabajos futuros estarán dedicado a estudiar diferentes estrategias de Minería de Procesos en Mallas Curriculares de carreras de pregrado basadas en el concepto de CAs, para automatizar tareas de detección de recorridos académicos de estudiantes anormales, proponer nuevos modelos curriculares según comportamientos de los estudiantes o necesidades en el mercado, entre otras cosas. Para ello, esos CAs serán enriquecidos con analítica de aprendizaje, análisis de redes sociales, etc., para enriquecer la información que usarán.

REFERENCIAS

[1] A. Bogarín, C. Romero, R. Cerezo and M. Sánchez-Santillán. “Clustering for Improving Educational Process Mining”. Proc. 4th Intl Conf. Learning Analytics and Knowledge, pp. 11–15. 2014

[2] R. Wang and O. Zaïane. “Discovering Process in Curriculum Data to Provide Recommendation.” EDM. Pp. 580-581, 2015

[3] M. Pechenizkiy, N. Trcka, E. Vasilyeva, W. Van der Aalst and P. De Bra. “Process Mining Online Assessment Data”. International Working Group on Educational Data Mining. 2009

[4] A. Cairns, B. Gueni, M. Fhima, S. David and N. Khelifa. “Process mining in the education domain”. International Journal on Advances in Intelligent Systems, vol. 8, pp. 219-232. 2015.

[5] G. Sedrakyan, M. Snoeck and J. De Weerdt. “Process mining analysis of conceptual modeling behavior of novices–empirical study using JMermaid modeling and experimental logging environment”. Computers in Human Behavior, vol. 41, pp. 486-503. 2014

[6] F. Lozano and R. Lozano, “Developing the curriculum for a new Bachelor's degree in Engineering for Sustainable Development”, Journal of Cleaner Production, vol. 64, pp. 136-146, 2014.

[7] Y. Moreno, C. Aguilar and F. Hidrobo. “Análisis de los Problemas de Rendimiento en un EVA a través de la Extracción de Conocimiento”. Revista Ingeniería al Día, vol. 4, pp. 3-24. 2018

[8] J. Aguilar, O. Buendía, A. Pinto and J. Gutiérrez. Social learning analytics for determining learning styles in a smart classroom, Interactive Learning Environments. 2019.

[9] K. Mattingly, M. Rice and Z. Berge, “Learning analytics as a tool for closing the assessment loop in higher education”, Knowledge Management & E-Learning: An International Journal, vol..4, pp. 236-247, 2012.

[10] S. Iglesias-Pradas, C. Ruiz-de-Azcárate, and Á. Agudo-Peregrina, “Assessing the suitability of student interactions from Moodle data logs as predictors of cross-curricular competencies”, Computers in Human Behavior, vol. 47, pp. 81-89, 2015.

[11] F. Pacheco, C. Rangel, J. Aguilar, M. Cerrada and J. Altamiranda, “Methodological framework for data processing based on the Data Science paradigm”, XL Latin American Computing Conference (CLEI). 2014

[12] J. Aguilar, O. Buendia, K. Moreno and D. Mosquera, “Autonomous cycle of data analysis tasks for learning processes, In International Conference on Technologies and Innovation, Communications in Computer and Information Science Series”, vol. 658, pp. 187-202. 2016

[13] J. Aguilar, M. Sánchez, M, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán and L. Chamba-Eras. “Learning analytics tasks as services in smart classrooms”. Universal Access in the Information Society, vol. 17, pp. 693–709, 2018

[14] J. Aguilar, J. Cordero and O. Buendía. “Specification of the Autonomic Cycles of Learning Analytic Tasks for a Smart Classroom”, Journal of Educational Computing Research, vol. 56, pp. 866-891. 2018

[15] M. Sánchez., J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, and L. Chamba-Eras “Cloud Computing in Smart Educational Environments: Application in Learning Analytics as Service”. In: Rocha Á., Correia A., Adeli H., Reis L., Mendonça Teixeira M. (eds) New Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing, vol. 444,. pp. 993-1002 2016.

[16] Reglamento de Estudios de Postgrado. Universidad Pedagógica Experimental Libertador. Gaceta 1-2018. Resolución Nº 2018.488.056. 2018

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

AUTHORS

Jose Aguilar

Ingeniero de Sistemas, Universidad de los Andes-Mérida-Venezuela, Maestría en Informática, Universidad Paul Sabatier-Toulouse-France, y Doctorado en Ciencias Computacionales, Universidad Rene Descartes-Paris-France. Postdoctorado en el Departamento de Ciencias de la
Computación de la Universidad de Houston, en el Laboratorie d'Automatique et Analyses de Systemes-CNRS-Toulouse-France y en el Departamento de Automática, Universidad de Alcalá-España (actualmente). Profesor Titular del Departamento de Computación de la Universidad de los Andes-Venezuela, Profesor Contratado de la Universidad EAFIT-Colombia. Miembro de la Academia de Mérida.
.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Sonia Duarte

Ingeniero de Sistemas, Universidad de los Andes-Mérida-Venezuela, Especialista en Planificación Educacional, Universidad Valle del Momboy, Magister en Innovaciones Educativas, Universidad Pedagógica Experimental Libertador. Profesora en la Especialidad de informática, Universidad Pedagógica Experimental Libertador

Modelos de grafos para la detección de datos de texto no estructurados como el sarcasmo

Graph Model for Detection of text unstructured data such as Sarcasm

ARTICLE HISTORY

Received 11 October 2020
Accepted 02 November 2020

Axel Rodríguez-García
Facultad de Ingeniería de Sistemas Computacionales
Universidad Tecnológica de Panamá
Ciudad de Panamá, Panamá
axel.rodriguez2@utp.ac.pa

Armando Jipsion
Facultad de Ingeniería de Sistemas Computacionales
Universidad Tecnológica de Panamá
Ciudad de Panamá, Panamá
armando.jipsion@utp.ac.pa

Modelos de grafos para la detección de datos de texto no estructurados como el sarcasmo

Graph Model for Detection of text unstructured data such as Sarcasm

Axel Rodríguez-García
Facultad de Ingeniería de Sistemas Computacionales
Universidad Tecnológica de Panamá
Ciudad de Panamá, Panamá
axel.rodriguez2@utp.ac.pa

Armando Jipsion
Facultad de Ingeniería de Sistemas Computacionales
Universidad Tecnológica de Panamá
Ciudad de Panamá, Panamá
armando.jipsion@utp.ac.pa

***Abstract*—** Sarcasm is frequently characterized as verbal incongruity to communicate scorn. It is a nuanced type of language with which people express something contrary to what is suggested. Perhaps the greatest test in building frameworks to consequently recognize unstructured information, for example, mockery, is the absence of huge, commented on informational indexes. We propose a diagram-based procedure in building conservative language models for sarcasm recognition. This strategy is likewise intended to utilize little information, it could help in different regions like disdain discourse, counterfeit news, and so forth. This charting strategy permits specialists to explore different parts of NLP without obtaining a huge dataset. These days, it still remains a challenge to unmistakably distinguish human slants and feelings by utilizing AI. Associations can use a superior philosophy to settle on proactive choices in basic circumstances. A definite investigation of our examination would hoist the current content mining applications and may help understand better the effect of mockery from the customers and partners communicated in a web-based media climate. We exhibit that straightforward classifiers worked from the model can recognize mockery very well, which they sum up 5 % better than those of the cutting edge.

***Keywords*—** *Unstructured Data, NLP, sarcasm, modelo de grafo.*

***Resumen*—** El sarcasmo se define a menudo como una ironía verbal para expresar desprecio, un lenguaje matizado con el que los individuos expresan lo contrario de lo que está implícito. Uno de los mayores retos en la construcción de sistemas para detectar los datos no estructurados como el sarcasmo, es la falta de grandes conjuntos de datos anotados. Proponemos un método basado en grafos para la construcción de modelos de lenguaje compacto para la detección del sarcasmo. Este método está diseñado para usar pocos datos, y podría ayudar a detectar fake news, hate speech, etc. Permite además a los investigadores analizar otros aspectos del NLP sin tener que obtener un conjunto de datos gigante. Hoy en día, sigue siendo un desafío identificar claramente los sentimientos y emociones humanos mediante el uso de Inteligencia Artificial. Una exploración detallada de nuestra investigación elevaría las aplicaciones actuales de minería de textos y podría ayudar a comprender mejor el impacto del sarcasmo de los clientes y las partes interesadas, expresado en un entorno de redes sociales. Demostramos que los clasificadores simples construidos a partir del modelo pueden detectar bastante bien el sarcasmo, que generalizan un 5% mejor que los del estado del arte.

***Palabras clave*—** *Datos no estructurados, lenguaje de procesamiento natural (NLP), sarcasmo, graph model.*

I. INTRODUCCIÓN

¿Qué es el sarcasmo? Es una ironía verbal que expresa el desprecio hacia una persona o la ridiculiza. Su naturaleza figurativa dificulta el análisis de sentimientos. Es también una forma de lenguaje matizada en la que los individuos afirman lo contrario de lo que se implica. En el lenguaje hablado, el sarcasmo se puede identificar por el tono del interlocutor, pero solo en el texto, detectar el sarcasmo se vuelve muy difícil. En términos generales, el sarcasmo implica un sentimiento negativo, pero a menudo muestra un sentimiento positivo en en cuanto a

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

mejorar la atención del cliente [1]. Por ejemplo: cuando un cliente descontento, que no pudo obtener un servicio satisfactorio de Amazon, hizo un comentario en Twitter publicando: "Buen trabajo@ AmazonHelp". O cuando una persona le dice a su amigo que: si este es un hotel de 5 estrellas, entonces "Sí, y yo soy la Reina de Inglaterra ". Tanto "Grantrabajo@..." como "Sí, y soy..." son sarcasmos en el contexto.

La detección automática del sarcasmo es una tarea para predecir el sarcasmo en el texto. Esto se ha vuelto cada vez más importante para mejorar el rendimiento de los sistemas de análisis de los sentimientos. Aparte del desafío de disuadir; o minar el verdadero sentimiento en una frase sarcástica [1], otro desafío importante para la detección automática del sarcasmo es la falta de grandes y fiables conjuntos de datos anotados [2]. La mayoría de los conjuntos de datos que existen fueron creados a partir de textos en inglés, lo que deja a otros idiomas con pocos recursos para abordar esta difícil tarea.

En este artículo, proponemos una metodología basada en grafos para construir modelos compactos de lenguaje para la detección del sarcasmo. Utilizando la supervisión a distancia mediante el uso del hashtag #sarcasmo, se recolectó un pequeño conjunto de datos de unos pocos tweets de mil arenas. Los textos fueron luego convertidos a forma de grafo, y con el uso de técnicas de análisis de grafos, los patrones fueron descubiertos automáticamente. Luego se extrajeron las representaciones vectoriales de los patrones; utilizando un enfoque para aprender las representaciones latentes de los vértices en una red.

El modelo de lenguaje compacto resultante se utilizó luego para construir un clasificador detectando los patrones en los textos de un conjunto de entrenamiento, obteniendo sus incrustaciones (embeddings) y representando cada texto de muestra como el promedio de todas las incrustaciones. Comparamos nuestro método con las líneas de base, que se basaban en las últimas técnicas de construcción de modelos de lenguaje y en los enfoques más avanzados para la detección del sarcasmo. Demostramos experimentalmente que con pocos datos, nuestro enfoque de construcción de modelos, puede generar clasificadores aceptables que generalizan mejor a los métodos más avanzados. Este enfoque podría extenderse a otros idiomas utilizando la traducción apropiada para "sarcasmo".
Las principales contribuciones de nuestro trabajo son:

1) Un enfoque independiente del idioma para construir un modelo para la detección del sarcasmo.

2) El enfoque propuesto puede construir modelos efectivos con pocos datos utilizando la supervisión a distancia.

3) Debido al tamaño compacto del modelo, puede construirse sin necesidad de una gran potencia de cálculo.

4) El modelo de lenguaje puede ser usado para obtener características para clasificadores simples basados en la Regresión Logística o Máquinas Vectoriales de Apoyo.

5) El modelo lingüístico es lo suficientemente expresivo como para ayudar a los clasificadores construidos sobre él, y para generalizar mejor al de los enfoques más complejos.

El resto de esta invetigación está estructurada de la siguiente manera: La sección 2 puntualiza el marcco teórico. En la sección 3 se especifica la metodología. Los experimentos serán presentados previamente en la sección 4. Por último, las conclusiones y el trabajo futuro figuran en la sección 5.

II. MARCO TEÓRICO

La tarea de clasificación es la metodología más común de la detección automática del sarcasmo. Varios estudios han abordado la cuestión de la detección del sarcasmo en textos cortos (por ejemplo, tweets). Al tratar esos textos, algunos investigadores anotaron manualmente los tweets como obras sarcásticas o no sarcásticas [1]. Se basaron en la supervisión a distancia para crear conjuntos de datos. El trabajo en [3] utiliza un conjunto de datos creados mediante la obtención de tweets que contienen hashtags, como #sarcasmo, #sarcasmo, #no. Otros estudios que utilizan la supervisión a distancia incluyen [4],[5],[6],[7],[8],[9],[10],[11],[12].

Otros estudios que utilizan este enfoque incluyen el trabajo en [3], con la diferencia de que sólo mantienen los tweets que contienen el hashtag en cuestión al final. No todos los textos que contienen sarcasmo son tweets o mensajes cortos de medios sociales. Los foros de debate y las reseñas de productos, que suelen ser más extensos, también pueden contener muchos casos de sarcasmo que vale la pena detectar. En la obra presentada en [11]. En [11] se utiliza un conjunto de datos de mensajes de foros de debate con múltiples etiquetas que incluyen el sarcasmo. De acuerdo con la investigación en [13] se utiliza un conjunto de datos de reseñas

de películas, libros y artículos de noticias marcados con sarcasmo y otras etiquetas de sentimiento. Otros trabajos de investigación que utilizan reseñas etiquetadas con sarcasmo incluyen [14], [15], [16], [17],[18].

Por último, dado que el sarcasmo aparece a menudo en conversaciones reales, también se han realizado investigaciones sobre transcripciones y diálogos. Estos pueden ser divididos en transcripciones de centros de llamadas [17] y frases extraídas de las conversiones de ciertos programas de televisión [18], [19]. La mayoría de los trabajos sobre detección automática de sarcasmo se han basado únicamente en el texto mismo. Más recientemente, los estudios han experimentado con información adicional para proporcionar el texto de la convención. Esto incluye el contexto específico del autor. Considerando la historia del autor, [20], [21] incorpora en el conjunto de datos tweets pasados como contexto. Otra forma de contexto cuando se trata de conversaciones incluye frases que ocurren antes de la frase a clasificar de una conversación [21], [22],[23],[24],[25] o comentarios relacionados con un post en foros de discusión [21]. Otro enfoque interesante para obtener el contexto es la idea de que ciertos temas evocan más sarcasmos que otros. Con eso en mente, los estudios presentados en [25], [26], intentan crear modelos de temas para ayudar a predecir el sarcasmo. Estudios más nuevos y avanzados han llevado el contexto a un nivel más profundo al incorporar metadatos como seguidores, ubicación, edad de la cuenta, número de mensajes, gustos, etc., de redes sociales como [15]. Se han utilizado diferentes tipos de características y algoritmos de aprendizaje para abordar la detección automática del sarcasmo. La mayoría de los estudios se basan en modelos de lenguaje de bolsa de palabras, aunque otros han hecho uso de patrones con coincidencia parcial o total.

La mayor parte de los trabajos efectuados en este campo se basan en diferentes formas de algoritmos tradicionales de aprendizaje de máquinas como Support Vector Machines (SVM) o Logistic Regression (Regresión Logística) en [1], [3], [15], [17], [23], [27], [28], [29]. Estudios más recientes han comenzado a utilizar cada vez más los enfoques de aprendizaje profundo para abordar esta difícil tarea [21],[30],[31],[32].

III. METODOLOGÍA

A. Vision General

Proponemos un enfoque novedoso para construir modelos de lenguaje para la detección del sarcasmo convirtiendo el conjunto de datos en un grafo y extrayendo de él representaciones de nodos (embeddings). Las representaciones resultantes pueden utilizarse para entrenar a un clasificador. Aunque este documento se centra en la detección de datos no estructurados como el sarcasmo en Twitter, el mismo enfoque podría ajustarse fácilmente para otras tareas.

Para generar el modelo, es necesario seguir los siguientes pasos. En primer lugar, el conjunto de datos se transforma en un grafo considerando las palabras (y en algún caso especial, secuencias cortas de palabras llamadas patrones) como nodos, y se crean conexiones entre los nodos (arcos o aristas) para las palabras que aparecen una al lado de la otra (o dentro de una distancia determinada) en los textos originales. La fig.1 y fig. 2 muestran ejemplos de este proceso de sustracción de grafos, en donde el resultado final es un grafo subjetivo después de eliminar los bordes.

A continuación, se analizan los grafos iniciando recorridos aleatorios desde cada nodo y generando secuencias aleatorias de los nodos que pueden verse como frases sintéticas. Se utiliza una técnica de generación de modelos lingüísticos basados en el texto para extraer incrustaciones (embedding) de nodos, tratando los patrones como una oración regular. El modelo de lenguaje resultante puede entonces ser usado con los clasificadores tradicionales de aprendizaje automático (Machine Learning).

B. Construcción del grafo: El primer paso para construir modelos es generar versiones de grafos de los textos. En este estudio, se investigan tres formas diferentes de construir dichos grafos.

1) Pathways Graphs (Grafos de Caminos): El primer enfoque para construir un grafo se basa en el enfoque descrito en [33]. Los pasos de preprocesamiento incluyen:

1. Normalizar el texto poniendo en minúsculas el conjunto de datos

2.Eliminar las palabras de parada (opcional)

3.Rama (opcional)

Tradicionalmente, la etapa de preprocesamiento, o de limpieza, incluye la normalización de las palabras mediante la corrección de errores ortográficos o la reducción de fichas con letras repetidas (por ejemplo wooooow!!!!!). No incluimos esto como parte del preprocesamiento; estas fichas especiales pueden ayudar a la detección del sarcasmo.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Una vez que el conjunto de datos ha sido preprocesado, el grafo se construye escaneando los textos, creando un nodo para cada palabra única, y bordes entre las palabras de co-ocurrencias dentro de un espacio de n palabras. El hueco es pre-especificado antes de la construcción y se pueden utilizar diferentes tamaños del hueco.

Un frase en el idioma inglés como: "today is very sunny and hot" generaría los siguientes aristas si la brecha se fijara en 1: today→ is, today→ very, is →ivery, is→ sunny, etc.

Aunque el ejemplo anterior muestra un→
para indicar el borde, esto es sólo para la representación, ya que el grafo real no está dirigido y no está ponderado.

Definición 3.1 Grafo de Palabra -Word Graph :Definimos el grafo de palabras de la siguiente manera:
donde V denota una agrupación de nodos de

$$G = (V, E, W) \qquad (1)$$

palabras, E representa la agrupación de las relaciones entre dos nodos de palabras, y W representa el conjunto de pesos de sus bordes (edge).

Definición 3.2. Peso de la arista - Edge Weight: Para cada arista $e_{vi,vj}$ ϵ E conección de palabras v_i y v_j, el peso del borde $w_{vi,vj}$ es calculado como:
donde freq ($e_{vi,vj}$}) denota como la frecuencia

$$W_{vi,vj} = \frac{freq(e_{vi,vj})}{\max(freq(E))} \qquad (2)$$

de palabras vi and vj co-ocurrencia y freq(E) denota el conjunto de frecuencias entre todos los pares de palabras.

En el caso de los Pathways Graphs (Gráfos de Caminos), no hay pesos para las aristas.

2) Minusnet Graphs: El segundo tipo de grafo se basa en partes de los métodos presentados en[34]. Su método se usó para crear clasificadores de emociones y estaba destinado a ser un enfoque independiente del idioma. Dado que el sarcasmo es un tipo de emoción, y dado que el paso particular descrito en ese documento podría funcionar en general para cualquier tarea, decidimos adaptarlo para construir nuestro segundo tipo de grafo.

La idea general detrás de los Grafos Minusnet es que el grafo resultante del conjunto de datos original contendrá muchos nodos y bordes inútiles. Podemos pensar en estos objetos sin importancia como palabras y co-ocurrencias que son concurrentes para cualquier lenguaje dado y no transmiten información útil para la tarea en cuestión. Por lo tanto, eliminar estos objetos del grafo podría hacer el modelo más compacto y aumentar su poder de predicción.
Para crear un Grafo Minusnet, se necesitan dos conjuntos de datos. El primero es el conjunto de datos relacionados con la tarea en cuestión, en el caso de este trabajo es un conjunto de datos con textos sarcásticos. El segundo conjunto de datos debe ser más general (neutral en el caso de la detección de emociones), y no contener instancias del primer conjunto de datos. Ya que el sarcasmo es un tipo de emoción, usamos un conjunto de datos neutrales (titulares de noticias) como en [34].

Para construir el grafo, ambos conjuntos de datos se convierten primero en un grafo ponderado utilizando el enfoque del Grafo de Caminos descrito en la definición del grafo de palabra. La principal diferencia esta vez es que los bordes tendrán un peso igual a su frecuencia dividido por la máxima frecuencia de borde en el grafo como en la Definición de Peso del arco. Por lo tanto, el borde más frecuente tendrá un peso de 1,0, y cada otro borde tendrá un peso entre 0,0 y 1,0. Después de completar este proceso obtenemos dos grafos Gn=(Vn,En,Wn) y Gs=(Vs,Es,Ws) donde el subíndice n denota neutralidad y s sarcasmo.

A continuación, un enfoque de sustracción de grafos introducido en [34] se aplica para generar el grafo final de Minusnet. La sustracción simplemente encuentra los bordes que coinciden en ambos grafos y resta el peso del borde en el grafo neutro del que está en el grafo de sarcasmo.

Definición 3.3 [Graph Subtraction- Substracción del Grafo] Dado un grafo de sarcasmo Gs=(Vs,Es,Ws) y un grafo neutral Gn=(Vn,En,Wn), nuevos pesos Ws se calculan para Gs de la siguiente manera:

$$Ws_{vi,vj} = \begin{cases} Ws_{vi,vj} - Ws_{vi,vj} & if\ vi,vj \in Vs \cap Vn \\ Ws_{vi,vj} & otherwise \end{cases} \qquad (3)$$

Después de actualizar un peso en el grafo de sarcasmo, si el valor resultante está por debajo de un cierto umbral; el arco correspondiente se elimina del grafo. Si, después de eliminar los arcos, algún nodo (palabra); ya no está conectado al grafo, el nodo se elimina. Esto da como resultado un grafo y las figuras 2 y 3 del Minusnet muestran un ejemplo de este proceso.

3) Patterns Graphs (Grafos de Patrones): El tercer y último tipo de grafo es el único en el que los nodos no son fichas de una sola palabra. En su lugar, los nodos son secuencias cortas de fichas de palabras y un comodín. Tabla ref{patterns-examples} muestra algunos ejemplos de patrones.

La idea es encontrar los patrones más comunes que se producen en los datos, y para aumentar el poder de generalización, sustituir algunas palabras por un comodín (.*) que coincida con cualquier palabra. Como por ejemplo en la fig. 3. Esto podría verse como un tipo de n-gram con un poder extra dado por el comodín. Aparte del comodín, lo que hace que los patrones sean diferentes de los n-gramas tradicionales es la forma en que se descubren.

El enfoque de descubrimiento de patrones se introdujo por primera vez en [34] y funciona de la siguiente manera: Un grafo Minusnet se construye a partir de una agrupación *de datos objetivos y una agrupación de datos neutral*. Dos tipos de palabras se extraen del grafo mediante técnicas de análisis de grafos. El primer tipo es *Connector Palabras*, que son frecuentes y también centrales en el grafo. El segundo tipo es *Palabras del tema*, que son menos frecuentes y tienden a agruparse con otras palabras relacionadas con los mismos temas. Utilizando las dos listas se construyen diferentes combinaciones de palabras de longitud de variables (instancias). Se mantienen todas las secuencias de instancias que se producen en el conjunto de datos y se hace un seguimiento de su frecuencia. Las instancias frecuentes tendrán su palabra temática reemplazada por el comodín para convertirse en un patrón. Después de este reemplazo, diferentes instancias se convertirán en el mismo patrón *"I love this"* y *"I hate this"* se convertirán en "I .* this" si "love" y "hate" son ambos *Topic Words (Palabras del tema))*.
Los patrones y su frecuencia se mantienen. Luego, se btiene una lista final de patrones seleccionando aquellos cuya frecuencia está por encima de un umbral determinado. A partir de la lista final de patrones se construye un Grafo de Patrones utilizando una versión modificada de la construcción del Grafo de Patrones. Más detalles para cada uno de los pasos anteriores son los siguientes:

Para encontrar el *Connector-Conector y Topic Words-Palabras* Tópicas, El análisis de los grafos es necesario. *Connector Words (conector de palabras)* son un tipo especial de palabras de parada, sin embargo, no están predefinidas y no se consideran una molestia. En cambio, *Connector Words* son descubiertos a partir del conjunto de datos; son frecuentes y son centrales, por lo que desempeñan un papel importante en el poder de generalización de los patrones resultantes. Para descubrir estas palabras, *betweenness centrality (Análisis de Centralidad)* el análisis se realiza en el grafo. En la teoría de los grafos, betweenness centrality es una medida de la centralidad en un grafo basado en los caminos más cortos. Una palabra con mayor betweenness centrality sería un buen candidato para un componente de patrón general, porque más n-gramas lo contendrán. La razón para usar *Connector Words*, en lugar de palabras de parada es que las palabras de parada son limitadas en número y expresión. *Betweenness centrality* se define como:

Definición 3.4 [Betweenness Centrality (Medida de Centralidad)] Por casa palabra en el nodo vu, la medidad de centralidad es calculada como:

$$BCv_u = \sum_{v_u \neq v_i \neq v_j} \frac{\sigma_{vi,vj}(v_u)}{\sigma_{vi,vj}} \qquad (4)$$

where $\sigma_{vi,vj}$ es la cantidad total de caminos más cortos desde el nodo vi al nodo vj y $\sigma_{vi,vj}(v_u)$ es el número de esos caminos que pasan por v_u.

a) Finding Topic Words (Encontrando palabras temáticas) : Las palabras temáticas se consideran como aquellas que tienen un alto grado de significado. En un gráfo de palabras, se espera que las palabras temáticas que tienen temas similares tiendan a agruparse. También se cree que si las palabras llevan menos significado, o en nuestro caso están menos relacionadas con la tarea de detectar el sarcasmo, estarán más aisladas. Para medir el grado de agrupación en un grafo, y detectar Topic Words, se calcula el coeficiente de agrupación de cada nodo. Siendo el coeficiente de agrupación una medidad del grado, en que los nodos de un grafo tienden a agruparse. Podemos definir coeficiente de agrupación de acuerdo a la formula # (5). De la siguiente manera:

Definición 3.5 [Clustering Coefficient (Coeficiente de la agrupación)] Por cada palabra del modo vu , el coeficiente es calculado como:

$$CCv_u = \frac{2T(Vu)}{nei(Vu)(nei(Vu)-1} \qquad (5)$$

Donde T (v_u) es el número de triángulos a través de los vértices v_uy nei(v_u) es el números de vecinos de v_u.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

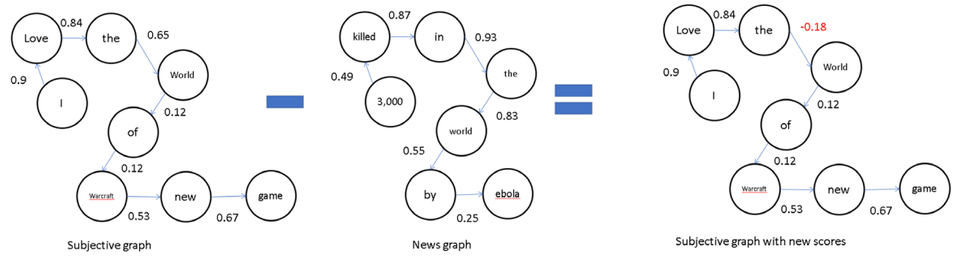


Fig. 1. Sustracción entre un grafo subjetivo y un grafo de titulares de noticias.

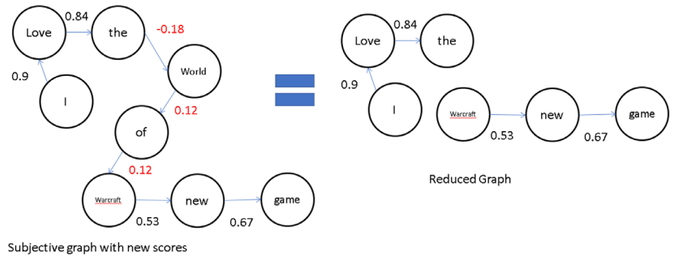


Fig. 2. Grafo subjetivo final después de eliminar los bordes (y los nodoscorrespondiente) con un peso inferior a un valor de umbral de 0,2.

Pattern	Example sentence match
was .* 🤣	That was professional 🤣
👍 .* .	Well done 👍 Trump.
loves .* and	He loves burgers and pizza

Fig. 3. Ejemplo de patrones y frases que los contienen. La coincidencia está subrayada en la frase.

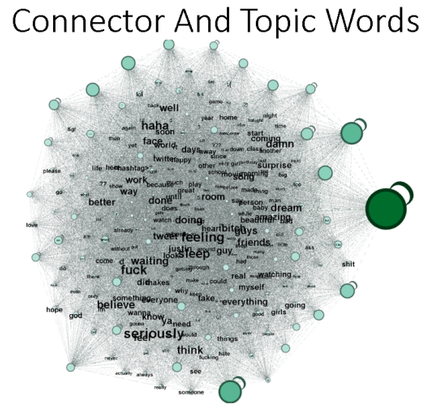


Fig. 4. Conector y palabras del tema dentro del grafo. El tamaño del nodo se determina por la medida de la centralidad mientras que el tamaño de la etiqueta se determina por el coeficiente de agrupación. Se puede observar como las palabras con etiquetas más grandes tienen más significado.

La fig. 4 muestra algunos ejemplos del conector y el tema Palabras encontradas en un gráfico Minusnet.

b) Extracting Instances (Extracción de instancias) : Los patrones en el conjunto de datos se descubren de abajo hacia arriba. La idea es encontrar primero el texto de los patrones potenciales, donde por instancias nos referimos a una secuencia de palabras que se corresponderían con tales patrones, y luego deducimos los patrones a partir de las frecuencias de instancias. El primer paso es definir un modelo de instancias, o del acuerdo a la invesigación original de acuerdo a [34] llamado *meta-patterns.* A *meta-pattern "CW-TW-CW"* indica que instancias debe construirse probando todas las combinaciones de Conector de Palabras Connecting Words, seguido de *Topic Words (Palabras Tópicas)*, y por último seguido finalmento por un *Conector de Palabras.*

Meta-patterns definir tanto el orden en que aparecen los dos tipos de palabras, como la longitud de la *instancias.* A *meta-pattern* es al menos 2 palabras y debe contener al menos una de cada tipo de palabras. Todas las *instancias* Todos meta-patterns (patrones meta). Para determinadas instancias (se puede pensar en ellos como n-gramas), calculamos su frecuencia en el conjunto de datos. Fig. 5 muestra ejemplos de ambos tipos de palabras y las *instancias* que puede resultar de ellas.

Topic Words
love
hate
gift
weather
...

Connector Words
this
got
my
pain
...

Instances
"hate this weather"
"lots of pain"
"got my gift"
"love this drawing"
"got my price"
"am in pain"
"kill this idiot"
"finish this task"

Fig. 5. Dos tipos de palabras y algunas instancias.

c) Generating Patterns (Generación de patrones): Desde las instancias con un proceso sencillo. En primer lugar, todas las instancias con la misma longitud y teniendo el mismo *Connector Word (Conector de palabra)* (la misma ficha de superficie en la misma posición) se agrupan. Sus frecuencias se agregan en la frecuencia del grupo. Entonces *pattern (el patrón)* se crea reemplazando los Topic Words *(Palabras Tópicas)* (que dentro de un grupo debe estar siempre en la misma posición para cada instancia) con una ficha especial que representa el comodín (e.g. * or.*).
Después de que cada grupo se ha convertido en un *pattern (patrón)*, no de frecuencias *patterns(patrones)* se filtran. Las Fig. 6 y 7 muestran los dos pasos que acabamos de describir.

Instances	Count
"hate this weather"	5
"lots of pain"	4
"got my gift"	7
"love this drawing"	2
"got my price"	1
"am in pain"	3
"kill this idiot"	1
"finish this task"	4

Connector Words
this
got
my
pain
...

Groups	Count
"Hate this weather", "love this drawing", "kill this idiot", "finish this task"	12
"lots of pain", "am in pain"	7
"got my gift", "got my price"	8
...	...

Fig. 6. Agrupando las instancias mediante el uso del Connector Word.

Pattern	Groups	Freq.
* this *	"Hate this weather", "love this drawing", "kill this idiot", "finish this task"	12
* * pain	"lots of pain", "am in pain"	7
got my *	"got my gift", "got my price"	8
...	...	...

Connector Words
this
got
my
pain
...

Fig. 7. Obtener patrones manteniendo las Palabras de Conexión y reemplazando las Palabras Temáticas con un comodín.

La idea principal es que construyendo un grafo reducido de palabras (Minusnet) que sea altamente representativo de la tarea en cuestión, y seleccionando de él palabras que sean a la vez centrales (comunes y a través de las cuales pase mucho significado) y que estén agrupadas (ellas mismas llevan mucho significado), podemos descubrir combinaciones que luego pueden llegar a ser muy expresivas *patterns (patrones)*. Reemplazando el original Topic Word (conector de palabra) con un comodín que coincida con cada palabra tiene el poder potencial de capturar otras palabras temáticas que no estaban presentes en el conjunto de datos original, pero puede ser importante para la tarea en cuestión. En la fig. 8 podemos apreciar el proceso parcial de construcción del grafo Esto es particularmente importante cuando se trata de pequeños conjuntos de datos. En la fig. 8 se puede apreciar el proceso parcial de construcción del grafo observando de dos a tres gramos consecutivos.

@PoptartLudwig Great selection of commentators on this sports channel!

k = 3

First 3-gram
@poptartludwig great selection
Pattern [Matched portion of the 3-gram]
.+ great ['@poptartludwig great']
great .+ ['great selection']
.+ great .+ ['@poptartludwig great selection']
Second 3-gram
great selection of
Patterns [Matched portion of the 3-gram]
.+ .+ of ['great selection of']
.+ of ['selection of']
great .+ of ['great selection of']
great .+ ['great selection']
great .+ .+ ['great selection of']
Edges between patterns with matches in first 3-gram and second 3-gram
.+ great ----> .+ .+ of
.+ great ----> .+ of
.+ great ----> great .+ of
.+ great ----> great .+
.+ great ----> great .+ .+
great .+ ----> .+ .+ of
great .+ ----> .+ of
.+ great .+ ----> .+ of
...

Fig. 8. Proceso parcial de construcción del grafo observando de dos o tres gramos consecutivos. Por cada 3 gramos, las líneas de abajo muestran el patrón y la porción del 3º gramo que fue emparejado. Después de mirar los dos 3 gramos consecutivos, se crea una conexión entre los patrones que coinciden con el primero y los patrones que coinciden con el segundo.La flecha es solo → para ilustración; ya que el grafo no está dirigido.

d) Patterns Graph Construction-Construcción de Grafos de Patrones: Una vez que el *patterns (patrón)* se extraen, pueden utilizarse para construir un segundo grafo final que constituirá el modelo de lenguaje (en forma de grafo; hay otro paso más adelante para conseguir el embedding o incrustación). El proceso de construcción de este grafo es una versión modificada del enfoque en la construcción de los Grafos de Caminos 3.2.3.

La principal diferencia viene del hecho de que *patterns o patrones* tienen una longitud

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

variable a diferencia de las palabras sueltas.
Para construir el grafo, todos los *k-grams*, conociendo la longitud de los patrones más largo, se extraen de cada texto en el conjunto de datos y el *k-grams* son inspeccionados en secuencia. Típicamente, patterns tienen una longitud de 2 o 3 pero podrían ser más, dependiendo de cómo meta-patterns se definen. La coincidencia de expresiones regulares puede utilizarse para encontrar los patrones que coinciden con un *k-gramo* dado.
La coincidencia puede ser parcial como algunos patterns pueden ser más corto que *k*. Cada dos patterns que coinciden con dos consecutivos *k-grams* se añadirán al grafo como nodos y se generará un borde correspondiente. Las fig. 7 ilustra este proceso y la fig. 9 muestra algunos patterns o caminos clasificado por su *Clustering Coefficient (Coeficiente de la Argupación) dentro de el Patterns Graph (Grafo de patrones).*

Pattern	Clustering Coefficient
best .+	1
good .+	1
away their .+	1
.+ 😁 .+	1
... .+	1
.+ 😂 .+	1
.+ would	1
.+ 🤓 .+	1
.+ !	1
great .+	1
happens ... .+	1
.+ 😳 .+	1
ever was .+	0.964285714
will .+ this	0.945454545
an .+ in	0.933333333
in these .+	0.928571429
already .+ the	0.927272727
.+ love the	0.923809524
at .+ best	0.916666667
are too .+	0.916666667
but a .+	0.916666667
he be .+	0.916666667
her so .+	0.916666667
so .+ much	0.909090909
a title .+	0.904761905
.+ imagine why	0.904411765
just .+ how	0.891666667
can't imagine .+	0.888888889
last .+ years	0.884615385

Fig. 9. Algunos patrones clasificados por su coeficiente de agrupación dentro del grafo. Los valores van de 0.0 a 1.0.

4) Latent Representations Extraction (Extracción de representaciones latentes)
En la sección anterior, describimos cómo construir diferentes tipos de grafos de texto para representar nuestros modelos de sarcasmo. En esta sección, describimos el proceso de extracción de incrustaciones (embeddings) para las diferentes características de words (las palabras) o patterns (patrones) de los grafos. El enfoque descrito en [35] se utiliza para obtener las diferentes incrustaciones de nodos del grafo. En Deepwalk: Aprendizaje en línea de las representaciones sociales [35], Las representaciones sociales de los vértices de un grafo se aprenden modelando recorridos cortos aleatorios. Las representaciones sociales son características latentes de los vértices que capturan la similitud del vecindario y la pertenencia a la comunidad.

Hemos seleccionado *DeepWalk* como nuestro enfoque en la obtención de incrustaciones para las fichas de nuestro grafo por las siguientes razones:

- ***Adaptabilidad:*** El lenguaje humano está en constante evolución y los temas de moda cambian todo el tiempo. Lo que es una forma popular de expresar el sarcasmo, o un tema popular de evocación de sarcasmo hoy en día, probablemente será más neutral en un futuro próximo. Debido a la forma en que DeepWalk en [34],crea las incrustaciones, las nuevas fichas y conexiones añadidas al grafo pueden ser fácilmente consideradas la próxima vez que se ejecute el proceso de extracción de incrustaciones.

- ***Low dimensional (baja dimensión):*** Como han afirmado los autores del documento en [34] , cuando los datos etiquetados son escasos, los modelos de baja dimensión se generalizan mejor y aceleran la convergencia y la inferencia. Es bien sabido que los datos etiquetados para la detección del sarcasmo son limitados, por lo que se prefieren los modelos de baja dimensión.

El proceso de generación de las incrustaciones con *DeepWalk* comienza con una impulso corto *random walks* utilizado para extraer información de la red. El ciclo de caminos aleatoria toma nuestro grafo de texto G y, en un bucle, toma cada nodo vi de V como la raíz de un camino aleatoria. Para realizar un recorrido, un vecino del último nodo visitado es muestreado uniformemente hasta que se alcanza la máxima longitud del recorrido. La idea principal detrás de *DeepWalk* es que se considerará la secuencia generada por el recorrido como una frase. Esta frase sintética es entonces alimentada al conocido SkipGram algoritmo [34] para generar los embedding (incrustaciones).
Definición 3.6 [Synthetic Sentences as a Random Walk-(Sentencias sintéticas como un paseo aleatorio)]
Sentencias sintéticas como un paseo aleatorio

Por cada raíz del nodo v_i, its k^{th} los caminos aleatorios de longitud n + 1 es definido como: donde v_{s1} es tomado aleatoriamennte desde el la vecindad v_i, v_{s2} , es tomada aleatoriamente desde la vecindad de v_{s1}y así que.

$$rw_{i,k} = v_{i,}.v_{s1,}\ v_{s2\ldots\ldots,}\ v_{sn} \qquad (6)$$

5) The Classifier (El Clasificador)

El objetivo del método propuesto es construir un modelo de lenguaje compacto para la detección del sarcasmo que pueda construirse a partir de pequeños conjuntos de datos y utilizarse con algoritmos sencillos de aprendizaje automático para obtener buenos resultados. Con un modelo de baja dimensión como el descrito en las secciones anteriores, una de las mejores opciones de algoritmo para construir un clasificador es la Regresión Logística (Logistic Regression). Por lo tanto, usamos el modelo para entrenar un clasificador de Regresión Logística y determinamos los mejores parámetros para dicho clasificador experimentalmente.

Para entrenar al clasificador, los tweets en un conjunto de entrenamiento son pre-procesados y marcados como en el paso de construcción del grafo. Los patrones se identifican de la misma manera. Luego, para cada palabra/patrón de un tweet, se extrae la incrustación correspondiente, si existe, y se añade a una lista de incrustaciones para ese tweet en particular. Si cada elemento del tweet no tiene incrustación, el tweet se descarta. La incrustación media del tweet se construye entonces calculando un vector medio obtenido por cálculo de la media a lo largo de las dimensiones de incrustación de la lista de incrustaciones.

El vector medio se convierte entonces en un vector unitario y se utiliza como la incrustación final del tweet. El mismo proceso se aplica cuando se valida.

Definición 3.7 [Final Embedding (Incrustración Final)] Para a tweet T, en el que es compuesto de items *(palabras or caminos)* $w_1, w_2, \ldots w_n$. Un embedding para T es entonces definido como: Where $u_{w1}, u_{w2},\ldots, u_{wn}$ son las incrustraciones (embedding) desde nuestro modelo de

$$u_T = \frac{1}{n}\sum_{i=1}^{n} u_{wi} \qquad (7)$$

construcción para cada item de T.

IV. EXPERIMENTOS Y RESULTADOS

En este estudio se realizaron diferentes experimentos. El objetivo de la primera serie de experimentos es determinar qué tipo de grafo, Pathways, Minusnet, o Patterns, tiene el mejor rendimiento. Para llevar a cabo el experimento, se construyen diferentes grafos de cada tipo cambiando los parámetros de pre procesamiento, aplicando la derivación, se eliminaron las palabras de parada y el tamaño del espacio entre las palabras, etc.

Después de determinar qué tipo de grafo funciona mejor, el segundo conjunto de experimentos tiene como objetivo comparar nuestro mejor modelo con el estado del arte (SOTA, *state of the art*) basadas en técnicas de última generación, así como métodos SOTA para la detección del sarcasmo. Se comparan tres líneas de base y un método SOTA con nuestro enfoque, con el objetivo de determinar qué enfoque generaliza mejor los datos no vistos.

A. Comparando los tipos de grafos

En este estudio adoptamos un modelo de lenguaje compacto, basado en grafos de texto, para la detección del sarcasmo. Proponemos tres formas diferentes de construir los gráficos, a saber: *Pathways Graphs-Grafos de caminos, Minusnet Graphs, y Patterns Graphs-Grafo de patrones.* En los dos primeros grafos los nodos son palabras, mientras que en el último los nodos son secuencias especiales de palabras llamadas *patterns o (patrones)*. En esta subsección describimos los experimentos realizados para determinar qué tipo de grafo funciona mejor cuando se utilizan incrustaciones (embeddings) extraídas de él para entrenar un clasificador.

1) Datasets (Conjunto de Datos)

Se utilizan diferentes conjuntos de datos durante estos experimentos. Los conjuntos de datos se describen a continuación:

• Conjunto de generación de grafos: este conjunto de datos consiste en 9,550 tweets recopilados con supervisión a distancia mediante el uso del hashtag #sarcasm. Para evitar recopilar por error tweets no sarcásticos utilizando #sarcasmo como parte del texto, nos aseguramos de que todos los tweets tuvieran el hashtag dentro de las últimas 3 palabras. Todos los tweets se clasificaron y recopilaron desde el 11/21/2019 hasta el 02/27/2020.

• Conjunto de datos neutrales: el conjunto de datos neutral consta de 34,047 titulares de noticias en forma de tweets recopilados de las cuentas de los principales medios de comunicación en inglés. Las cuentas provienen de: *The Wall Street Journal, The New York Times, The Guardian, etc.*

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

• Conjunto de datos de emociones: Este contiene 52,207 tweets recopilados con supervisión a distancia usando más de 200 hashtags relacionados con las emociones. Algunos de los hashtags incluyen *#enfado, #excitación, #gozo, #miedo, etc.*

• Conjunto de datos de sarcasmo: Este contiene 18,953 tweets recopilados con supervisión a distancia utilizando el *#sarcasm* hashtags. Estos tweets son diferentes de los c*onjuntos de generación de grafos and #sarcasm* puede aparecer en cualquier parte del tweet.

2) Configuración del experimento

Para realizar los experimentos, el conjunto de datos de generación de grafos se utiliza para generar los grafos y extraer las incrustaciones o embeddings. Los otros tres conjuntos de datos se combinan y luego se dividen en un conjunto de entrenamiento y validación para entrenar y validar un clasificador de Regresión Logística. Un 70 %-30 % de la división es utilizada.

Tabla I describe los hiperparametros que han cambiado para construir los grafos. Por ejemplo, cuando construimos *Pathways Graph (El grafo de camino)*, si el paso 2 es seleccionado con cada juego de parámetros verdaderos, entonces por cada *tweet @ AlDeTocqueville* por que este resulta entonces adecuado para nosotros en los pocos años pasados#sarcasm, AlDeTocqueville podría llegar ha *<usermention>, #sarcasm convertir <hashtag>*, las palabras *que, porque , nosotros, etc.* debería ser borrada, y la palabra resultante debería deribarse. Luego, al generar el gráfico, una arista entre los nodos *past y <hashtag>* existiría; ya que podemos tener un espacio de como máximo 2 elementos.

Tabla I. Los hiperpara´metros para el proceso de generación de grafos.

Parameter	Description	Graph Types
stepn	The gap between consecutive words/patterns that will be linked by an edge	All
nostopwords	If true, stop-words are removed	All
stem	If true, words are stemmed	All
noentities	If true, hashtags are changed to a keyword ¡*hashtag*¿ and user mentions are changed to ¡*usermention*¿	All
diff th	Minimum weight to keep an edge as part of the *Minusnet graph*	Minusnet, Patterns
cc th	Minimum *Clustering Coefficient* to consider a word a *Topic Word*	Patterns
centrality th	Minimum *Betweenness Centrality* to consider a word a *Connector Word*	Patterns
min freq	Minimum frequency to retain a *pattern*	Patterns

Tabla II. Parametros utilizados por deepwalk.

Parameter	Description	Default Value
randwalks	Random walks numberS per root node	10
walklen	random walk length	40
winsize	Window size used for *SkipGram*	5
embsize	Size of the resulting embedding	64

La incrustación de cada nodo se calculó usando *DeepWalk* y Tabla II describe los parámetros utilizados. Para todos los experimentos de esta sección, *embsize* se fijó en 64 y 128 mientras se mantenían los otros valores por defecto. Los clasificadores fueron entonces entrenados para clasificar los tweets en sarcásticos, emocionales, or neutrales utilizando el enfoque descrito en 3.4.

3) Resultados: Varios Grafos de camino se construyeron cambiando los parámetros descritos en la Tabla I y utilizando el primer conjunto de datos descrito en 4.1.2. La Fig. 10 muestra clasificadores basados en los Pathways Graphs clasificadas por su precisión en el conjunto de validación.

La fig. 10 resume el resultado de nuestros experimentos en la misma podemos observar que los 3 primeros clasificadores son los que tienen sufijos*_step3_stem_128, _step2_ stem_128, y _step3_noentities_128 con una precisión de 75 %.* Un rápido vistazo a la figura también muestra que, en general, la eliminación de entidades y un tamaño de incrustación de 128 da mejores resultados. También se puede ver que la eliminación de las palabras de parada perjudica la clasificación.

De manera similar, varios *Grafo Minusnet* se construyeron cambiando los parámetros de la misma de manera descrita anteriormente. *El diff_*th se fijó en 0.0 para todos los grafos, lo que significa que los bordes (y potencialmente las palabras) sólo se eliminaron si su peso después del paso de reducción se volvió negativo. Este valor tiene sentido para los conjuntos de datos pequeños, de lo contrario el grafo resultante sería demasiado pequeño.

Podemos observar en la fig. 11 que los 3 primeros clasificadores son los que tienen sufijos *_ step2_noentities_128, _step1_noentities_128, y _step3_stem_noentities_128 todas con precisión alrededor del 75%.* Precisamente como *Grafos de camino*, La eliminación de entidades y un tamaño de incrustación de 128 dan los mejores resultados. La eliminación parece tener un menor impacto y la eliminación de las palabras de parada no ayuda.

Finalmente, varios grafos se construyeron y probaron clasificadores basados en *Grafos de camino.* Este tipo de grafo se introduce unos pocos parámetros nuevos y requiere más tiempo para su construcción. Por lo tanto, no pudimos probar tantas combinaciones de parámetros como deseábamos. En este estudio probamos con *cc_th ...a* 1,0, 2,0 o 3,0, *centrality_th* siempre se ajusta a 0.0001 y min_

freq ...en 1 o 3. La Fig. 12 muestra clasificadores basados en los Grafos de camino clasificados por su precisión en el conjunto de validación.

El mejor resultado para Grafo de caminos se obtuvo con sólo una variación y una diferencia de 2 palabras. El Camino-Los parámetros específicos fueron de *0,3 para cc_th, 0.0001 para centralidad_th y 1 para min_freq. La mejor precisión fue ligeramente superior al 75%,* haciendo de este nuestro mejor clasificador basado en gráficos por un pequeño margen.

En general, los tres tipos de grafos se comportaron de forma similar, obteniendo alrededor de un 75% de precisión en una tarea de detección de sarcasmo muy desafiante y de múltiples clases. Las fig. 13, fig. 14. y fig. 15 muestran las matrices de confusión para los modelos superiores de cada tipo de gráfico.

A partir de las figuras se puede ver cómo todos los métodos funcionaron de manera similar, incluso a nivel de clase. Debido a que los Patrones o caminos y porque creemos que los patrones son más expresivos y se adaptan mejor a los nuevos conjuntos de datos no vistos, elegimos el mejor nuevo conjunto de datos de prueba compuesto principalmente por titulares de noticias, algunos sarcásticos y otros neutrales.

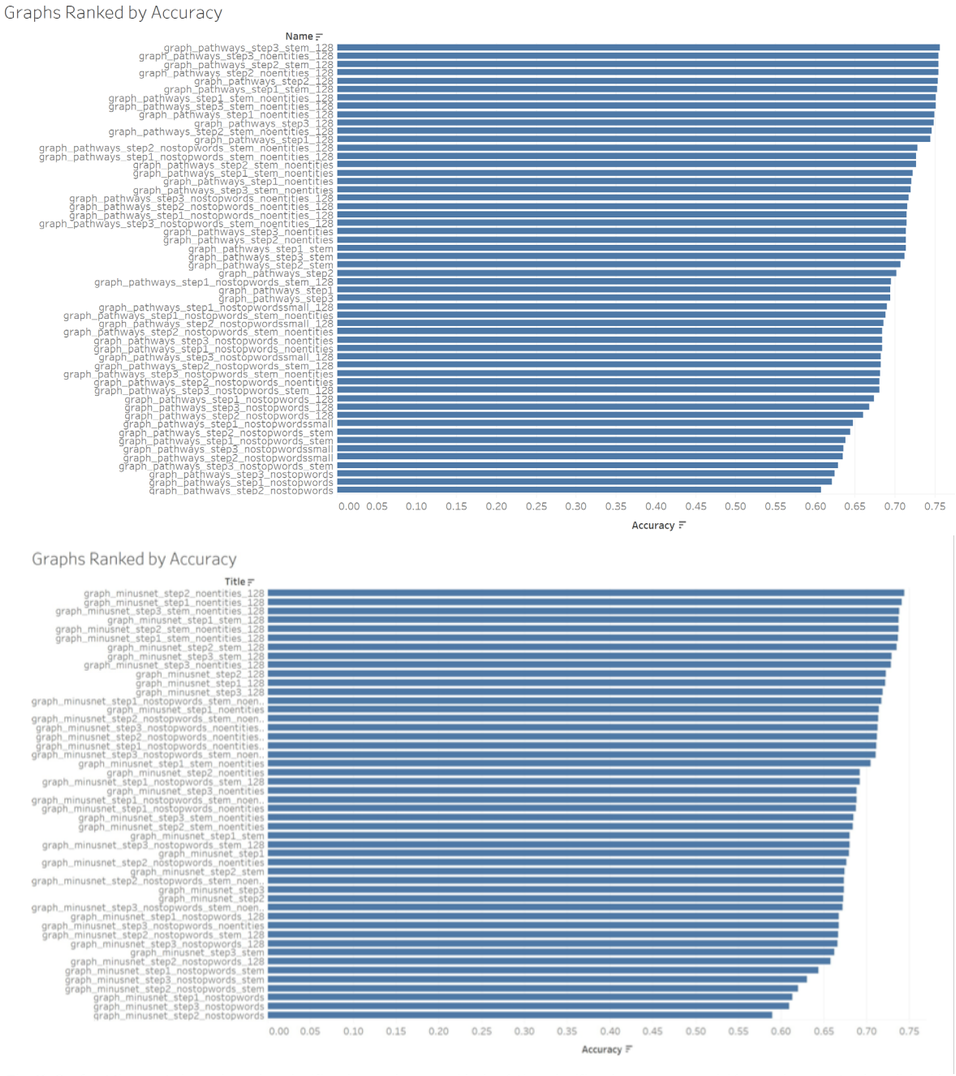


Fig. 11. Grafos Minusnet- basado en los clasificadores de regresión logística clasificados por su precisión en el conjunto de validación.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

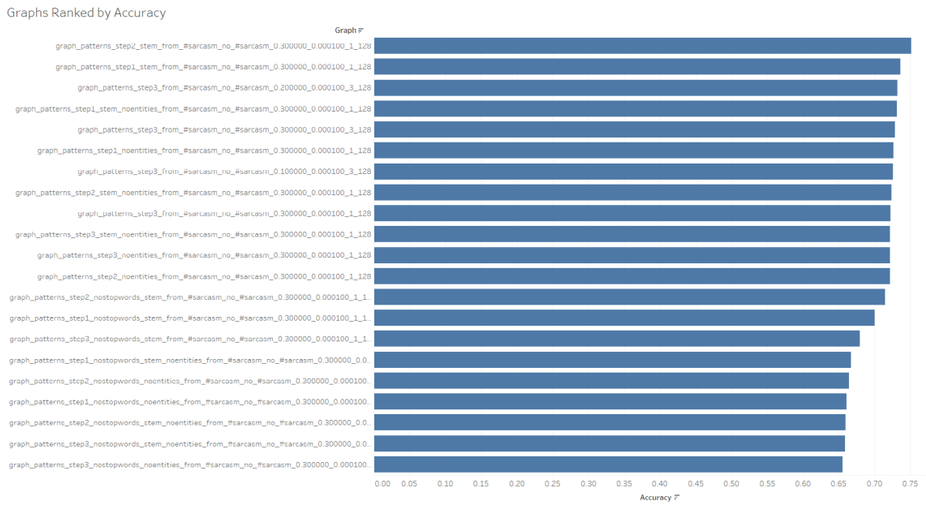


Fig. 12. Grafos de Patrones- basado en los clasificadores de regresión logística clasificados por su precisión en el conjunto de validación.

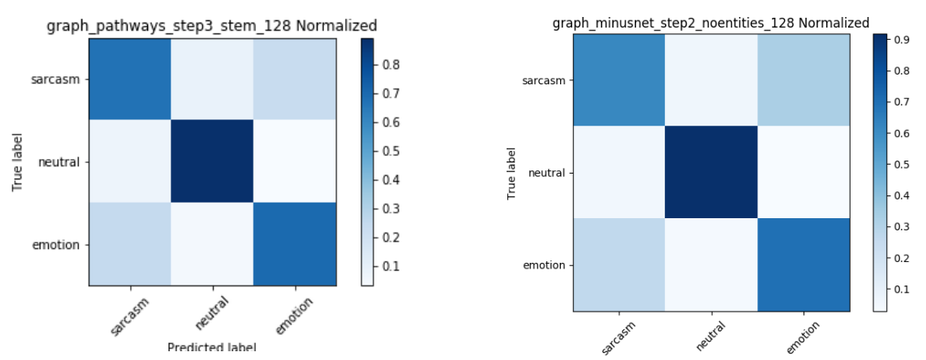


Fig. 13. Matriz de confusión para el mejor modelo de caminos basado en grafos.

Fig. 14. Matriz de confusión para el mejor modelo de caminos basado en grafos de Minusnet.

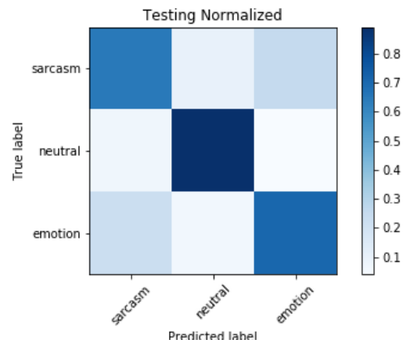


Fig. 15. Matriz de confusión para el mejor modelo basado en grafos de Patrones.

Patterns Graph clasificador que se comparará en la siguiente sección con otros métodos.

B. Comparación con otros métodos

En la sección anterior comparamos los tres tipos de grafos propuestos y determinamos que los *Patrones*- se basan en la tecnología son más precisos por un pequeño margen. También encontramos que, en general, el freno ha ayudado a que los tres enfoques funcionen mejor. En la segunda parte, queremos comparar nuestro mejor clasificador con algunas líneas de base y con uno de los más avanzados en detección de emociones y sarcasmo.

El principal aspecto que queremos medir es la capacidad de cada método para adaptarse a datos completamente diferentes de los conjuntos de entrenamiento y validación. Para lograrlo, todos los clasificadores se entrenan y validan con los mismos datos de Twitter, pero se prueban con un nuevo conjunto de datos de prueba compuesto principalmente por titulares de noticias, algunos sarcásticos y otros neutrales.

1) Conjunto de datos: Se utilizaron diferentes conjuntos de datos durante estos experimentos. Los conjuntos de datos se describen a continuación:

- ***Conjuntos de datos de entrenamiento y validación:*** este conjunto de datos está compuesto por el neutral, emociones, y el conjunto de datos de sarcasmo descritos en 4.1.2 y es usado para entrenar el otro método. La idea es que todos los métodos sean entrenados y validadas con la misma data de twitter.

- ***Titulares sarcásticos:*** este dataset consiste de 13,634 titulares de *The Onion. The Onion* es una satírica compañía y periódico de medio digital americano presentando mundanos, eventos cotidianos como noticia de una manera sarcástica. Por ejemplo, Un frustrado CEO admite que Pfizer descubrió la vacuna contra el virus Covid-19 hace meses, pero aún no se pone de acuerdo sobre la campaña publicitaria.

- ***Titulares neutrales:*** este conjunto de datos consiste en 14,985 titulares de *The Huffpost. The Huffpost* es un agregador de noticias y un blog americano.

- ***Tweets de emociones:*** este conjunto de datos contiene 18,018 tweets portadores de emociones, pero son diferentes de los descritos en el experimento anterior.

Como puede verse, el conjunto de datos de las pruebas es significativamente diferente de los conjuntos de entrenamiento y validación. En particular, los textos sarcásticos esta vez son titulares de noticias que son muy diferentes de los tweets sarcásticos. Esto hace que la tarea de clasificación sea mucho más difícil y constituye una muy buena manera de medir el poder de generalización de un clasificador.

2) Baseline Models (Modelos de línea Base): Queremos comparar nuestro mejor enfoque con algunas líneas bases (SOTA) construidas con técnicas avanzadas y de potencia. Las siguientes son las líneas bases utilizadas en este experimento:

- ***TV-C1D-D:*** este modelo está basado en el reciente TensonFlow T*extVectoritation (TV)* que crea capas que luego se alimentan a las redes neuronales para crear clasificadores. *TV-C1D-D* es compuesto de la capa de *TextVectorizer seguido* de la capa de *Convolution1D* y finalmente la *Densa Capa de Red Neuronal.*

- ***TV-BG-TFIDF-D:*** Al igual que el modelo anterior, este también se basa en el *TextVectorizer* pero luego usa una capa de *Bigram TF-IDF* Red Neuronal densamente-conectada como el clasificador.

Tabla III. Los resultados de la clasificación en el conjunto de pruebas para nuestro mejor enfoque y otros enfoques sota.

Model Name	Accuracy
Jammin Patterns Graph	**46 %**
DeepMoji-Finetune-Chain-Thaw	41 %
DeepMoji Finetune-Last	39 %
TV-C1D-D	38 %
DeepMoji-Embs-LogReg	35 %
DeepMoji Finetune-Full	35 %
Sarcasm-RoBERTa	33 %
TV-BG-TFIDF-D	31 %

Tabla IV. Informe de clasificación para el modelo basado en el modelo de grafos de patrones más alto.

	Precision	*Recall*	*F1 Score*
emociones	0.78	0.45	0.57
neutral	0.50	0.76	0.60
sarcasmo	0.16	0.15	0.16
Promedio Macro	0.48	0.45	0.44
Promedio del peso	0.51	0.46	0.46

3) Modelos de línea base o estado del arte (SOTA)

Para validar la afirmación de que nuestro enfoque puede adaptarse a los datos no vistos, mejor que otros enfoques; también queremos comparar nuestro mejor clasificador con algunos de los más recientes y aclamados estudios que han logrado resultados SOTA (*state of the art* - estado del arte).

- ***Sarcasm-RoBERTa:*** Desde el paper U*n enfoque de pre-entrenamiento del BERT robustamente optimizado [36]*, el modelo *RoBERTa* ha ha sido recientemente almacenado en SOTA para una amplia gama de tareas. *Sarcasm-RoBERTa* es una versión afinada y adaptada para la tarea de detección del sarcasmo.

- ***DeepMoji-Embs-LogReg:*** desde el paper *Usando millones de ocurrencias de Emoji para aprender cualquier representación de dominio para detectar el sentimiento, la emoción y el sarcasmo* [31] que nos dio el ampliamente popular modelo *DeepMoji*, implementamos una línea de base y tres de las variaciones que proponen en el documento original para la detección del sarcasmo. *DeepMoji-Embs-LogReg* es la línea de base que trata *DeepMoji* como un extractor de incrustaciones y añade un clasificador de regresión logística superior.

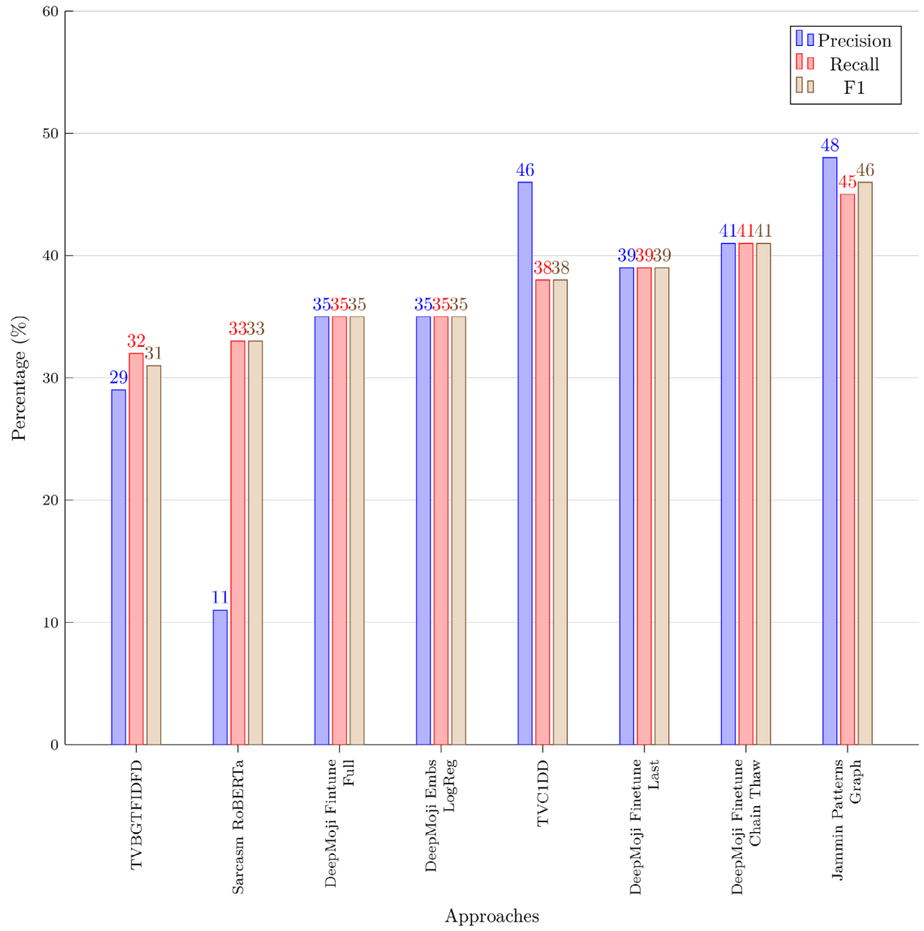


Fig. 16. Precisión, Recall, y puntajes de F1 para todos los métodos comparados.

- ***DeepMoji Finetune-Last:*** propuesto en [31], este método perfecciona *DeepMoji* congelando todas las capas excepto la última.

- ***DeepMoji Finetune-Full:*** propuesto en [31], este método afina todas las capas del original *DeepMoji*.

- ***DeepMoji-Finetune-Chain-Thaw:*** propuesto en [31] y también su enfoque de mejor rendimiento, este método afina cada capa por separado.

4) Configuración del experimento: El objetivo de esta segunda ronda de experimentos es demostrar que nuestro enfoque basado en

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

grafos, para la detección del sarcasmo, puede adaptarse mejor a nuevos conjuntos de datos que otros métodos. Cada una de las líneas de base y métodos propuestos que se compararán con nuestro clasificador de sarcasmo tienen una estructura similar al enfoque propuesto. Todos ellos convierten los textos en incrustaciones y construyen clasificadores encima de eso. Las principales diferencias son el enfoque para extraer las incrustaciones embedding) y el tipo de clasificador. Todos los métodos se entrenan y validan con los conjuntos de datos descritos en 4.1.2. y probado con los conjuntos de datos descritos en 4.2.1.

La tabla III, resume nuestros resultados, basados en nuestros experimentos, con nuestro clasificador *Jammin Patters Graph;* el cual tiene una mayor precisión comparado con los otros clasficadores en la desafiante tarea de detectar el sarcasmo de textos completamente diferentes a los métodos usados para entrenar a los clasificadores. De manera similar, la Fig.16 representa los resultados de nuestros experimentos en cuanto a las medidas de Precisión, Recall y F1 para los demás enfoques, mostrando nuevamente que nuestro enfoque basado en el clasificador de grafos supera a cualquier de los otros métodos. Es particularmente interesante apreciar que nuestro método superó significativamente el mejor metodo en un *5% DeepMoji,* Fine Tuning approach, a saber: *DeepMoji-Finetun e-Chain-Thaw*, que ha sido el método SOTA para la detección de emociones y sarcasmos durante unos años.

Por último, un análisis más profundo del rendimiento de nuestro método se sumariza en la tabla IV en donde se presentan los resultados por clase. En esta tabla se puede distinguir que producto de los resultados de nuestro experimento, la precisión para los textos de emociones presenta valores altos, posiblemente debido al hecho de que el clasificador fue entrenado y probado con tweets del conjunto de dtos de emociones. Del mismo modo, los titulares neutrales tuvieron la puntuación más alta en la F1, posiblemente debido a la similitud entre los titulares de las noticias tradicionales y los titulares de los tweets. Finalmente, y no es sorprendente, que el sarcasmo tuvo el peor desempeño; ya que este es el tipo de datos que fue el más diferente entre el entrenamiento y el conjunto de pruebas, haciendo una tarea; ya aún más difícil de clasificar.

V. CONCLUSIONES Y TRABAJO FUTURO

En la presente invesitgación, hemos planteado un método novedoso basado en grafos para construir modelos de lenguaje compacto y expresivo para la tarea de detecció n automática del sarcasmo sin tener que obtener un enorme conjunto de datos. En particular, nuestro método logra las siguientes caracterısticas diferenciables a otros métodos:

1. Independencia del lenguaje: si no se utiliza la eliminación de palabras de parada o de contención, las técnicas de generación de grafos y de extracción de incrustaciones (embedding) pueden aplicarse a los textos en cualquier idioma.

2. Pequeños requisitos de datos: utilizando #hashtags para la supervisión a distancia, unos pocos miles de tweets pueden ser agrupados rápidamente para crear los grafos. En comparación con los enfoques como DeepMoji, que necesitó miles de millones de tweets para alcanzar el estatus de SOTA, esto es muy conveniente

3. No requiere extensivos recursos: como no se necesitan grandes conjuntos de datos ni complejas redes de aprendizaje en profundidad, cualquier modelo puede construirse rápidamente con un hardware básico.

4. Agnóstico de tareas: al igual que ser agnóstico del lenguaje, el mismo enfoque también puede aplicarse a otras tareas utilizando un conjunto diferente de hashtags para la supervisión a distancia.

5.Expresivo: el modelo de lenguaje es lo suficientemente expresivo como para ayudar a los clasificadores construidos sobre él. Por lo tanto, puede generalizar mejor que los que tienen enfoques más complejos. En particular, creemos que el uso de patrones puede generalizarse mejor mediante el uso de comodines.

Hemos demostrado experimentalmente que, al encontrar un conjunto de datos experimentales diferentes del conjunto de datos utilizados durante la fase de entrenamiento, nuestro enfoque (Jammin Pattern Graph) supera a otras técnicas avanzadas en un promedio de 5 % en F1, Recall and Precision. Estos resultados se sumarizan en la Fig. 16.

En cuanto al trabajo futuro, nos gustaria probar más combinaciones de parámetros para los graafos basados en patrones.Tambie´n queremos demostrar que nuestro enfoque puede ser utilizado en diferentes idiomas y para diferentes tareas realizando experimentos con otros idiomas como el italiano y el francés. Estos incluyen tareas tan diversas como la clasificación de temas, la detección de discursos de odio, etc.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

REFERENCIAS

[1] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, “Sarcasm as contrast between a positive sentiment and negative situation,” in Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 704–714, 2013.

[2] D. Davidov, O. Tsur, and A. Rappoport, “Semi-supervised recognition of sarcasm in twitter and amazon,” in Proceedings of the fourteenth conference on computational natural language learning, pp. 107–116, 2010.

[3] A. Reyes, P. Rosso, and D. Buscaldi, “From humor recognition to irony detection: The figurative language of social media,” Data & Knowledge Engineering, vol. 74, pp. 1–12, 2012.

[4] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes, “Semeval-2015 task 11: Sentiment analysis of figurative language in twitter,” in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp. 470–478, 2015.

[5] C. Liebrecht, F. Kunneman, and A. van Den Bosch, “The perfect solution for detecting sarcasm in tweets# not,” 2013.

[6] F. Barbieri, H. Saggion, and F. Ronzano, “Modelling sarcasm in twitter, a novel approach,” in Proceedings of the 5th Workshop on Computatio- nal Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 50–58, 2014.

[7] G. Abercrombie and D. Hovy, “Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations,” in Proceedings of the ACL 2016 student research workshop, pp. 107–113, 2016.

[8] F. Barbieri, F. Ronzano, and H. Saggion, “Italian irony detection in twit- ter: a first approach,” in The First Italian Conference on Computational Linguistics CLiC-it, vol. 28, 2014.

[9] S. K. Bharti, K. S. Babu, and S. K. Jena, “Parsing-based sarcasm sentiment recognition in twitter data,” in 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1373–1380, IEEE, 2015.

[10] S. Lukin and M. Walker, “Really? well. apparently bootstrapping im- proves the performance of sarcasm and nastiness classifiers for online dialogue,” arXiv preprint arXiv:1708.08572, 2017.

[11] A. Reyes and P. Rosso, “On the difficulty of automatically detecting irony: beyond a simple case of negation,” Knowledge and Information Systems, vol. 40, no. 3, pp. 595–614, 2014.

[12] E. Filatova, “Irony and sarcasm: Corpus generation and analysis using crowdsourcing.,” in Lrec, pp. 392–398, Citeseer, 2012.

[13] K. Buschmeier, P. Cimiano, and R. Klinger, “An impact analysis of features in a classification approach to irony detection in product reviews,” in Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 42–49, 2014.

[14] O. Tsur, D. Davidov, and A. Rappoport, “Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews,” in fourth international AAAI conference on weblogs and social media, 2010.

[15] P. Liu, W. Chen, G. Ou, T. Wang, D. Yang, and K. Lei, “Sarcasm detection in social media based on imbalanced classification,” in Interna- tional Conference on Web-Age Information Management, pp. 459–471, Springer, 2014.

[16] J. Tepperman, D. Traum, and S. Narayanan, "yeah right": Sarcasm recognition for spoken dialogue systems,” in Ninth international confe- rence on spoken language processing, 2006.

[17] R. Rakov and A. Rosenberg, “sure, i did the right thing”: a system for sarcasm detection in speech.,” in Interspeech, pp. 842–846, 2013.

[18] A. Joshi, V. Tripathi, P. Bhattacharyya, and M. Carman, “Harnessing sequence labeling for sarcasm detection in dialogue from tv series ‘friends’,” in Proceedings of The 20th SIGNLL Conference on Compu- tational Natural Language Learning, pp. 146–155, 2016.

[19] A. Rajadesingan, R. Zafarani, and H. Liu, “Sarcasm detection on twitter: A behavioral modeling approach,” in Proceedings of the eighth ACM international conference on web search and data mining, pp. 97–106, 2015.

[20] D. Hazarika, S. Poria, S. Gorantla, E. Cambria, R. Zimmermann, and R. Mihalcea, “Cascade: Contextual sarcasm detection in online discussion forums,” arXiv preprint arXiv:1805.06413, 2018.

[21] D. Bamman and N. A. Smith, “Contextualized sarcasm detection on twitter,” in Ninth International AAAI Conference on Web and Social Media, 2015.

[22] A. Joshi, V. Sharma, and P. Bhattacharyya, “Harnessing context in-congruity for sarcasm detection,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 757–762, 2015.
[23] B. C. Wallace, E. Charniak, et al., “Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1035– 1044, 2015.
[24] Z. Wang, Z. Wu, R. Wang, and Y. Ren, “Twitter sarcasm detection exploiting a context-based model,” in international conference on web information systems engineering, pp. 77–91, Springer, 2015.
[25] A. Joshi, P. Jain, P. Bhattacharyya, and M. Carman, “Who would have thought of that!’: A hierarchical topic model for extraction of sarcasm-prevalent topics and sarcasm detection,” arXiv preprint ar- Xiv:1611.04326, 2016.
[26] R. Gonza´lez-Ibánez, S. Muresan, and N. Wacholder, “Identifying sarcasm in twitter: a closer look,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 581–586, 2011.
[27] A. Reyes and P. Rosso, “Making objective decisions from subjective data: Detecting irony in customer reviews,” Decision support systems, vol. 53, no. 4, pp. 754–760, 2012.
[28] R. Kreuz and G. Caucci, “Lexical influences on the perception of sarcasm,” in Proceedings of the Workshop on computational approaches to Figurative Language, pp. 1–4, 2007.
[29] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, “A unified deep learning architecture for abuse detection,” in Proceedings of the 10th ACM Conference on Web Science, pp. 105– 114, 2019.
[30] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm,” arXiv preprint arXiv:1708.00524, 2017.
[31] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman, “Are word embedding-based features useful for sarcasm detection?,” arXiv preprint arXiv:1610.00883, 2016.
[32] D. Paranyushkin, “Identifying the pathways for meaning circulation using text network analysis,” Nodus Labs, vol. 26, 2011.
[33] C. Argueta, E. Saravia, and Y.-S. Chen, “Unsupervised graph-based pat- terns extraction for emotion classification,” in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 336–341, 2015.
[34] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 701–710, 2014.
[35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.
[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

AUTHORS

Axel Rodríguez-García

Axel Rodríguez nació en Panamá. Recibió el título de Licenciado en Ingeniería de Sistemas Computacionales en la Universidad Tecnológica de Panamá, Panamá, en 1999; y una Maestría en Ciencias del Departamento de Ingeniería Industrial en la Universidad de Lousiville, en Los Estados Unidos, en el año 2000. Actualmente es candidato a doctor en el programa de doctorado en Ingeniería de Proyectos en la Universidad Tecnológica de Panamá, Panamá.

Armando Jipsion

Armando Jipsion nació en Panamá. Recibió el título de Licenciado en Tecnología en Programación y Análisis de Sistemas en la Universidad Tecnológica de Panamá, 1988. Obtuvo el Técnico en Ingeniería con Especialización en Programación y Análisis de Sistemas Universidad Tecnológica de Panamá, 1988. Obtuvo el Doctorado en Ingeniería de Proyectos de la Universidad Tecnológica de Panamá, 2012. La Maestría en Administración de Sistemas de Información en la Universidad Santa María La Antigua, 1995.
Tiene 30 años de ser Profesor Regular Titular.

A. Rodríguez-García and A. Jipsion, "Modelos de grafos para la detección de datos de texto no estructurados como el sarcasmo", Latin-American Journal of Computing (LAJC), vol. 8, no. 1, 2021.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol VIII, Issue 1, January 2021

Orquestación de Servicios RESTful y SOAP: Caso de estudio, asignación de visas Americanas.

RESTful and SOAP Services Orchestration: Case Study, American Visa Assignment.

ARTICLE HISTORY

Received 15 September 2020
Accepted 02 November 2020

Jhon Calle
Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay
Cuenca, Ecuador
john6ero@es.uazuay.edu.ec

Pablo Lója
Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay
Cuenca, Ecuador
arevalop17@es.uazuay.edu.ec

Marcos Orellana
Laboratorio de Investigación y Desarrollo en Informática (LIDI)
Universidad del Azuay
Cuenca, Ecuador
marore@uazuay.edu.ec

Priscila Cedillo
Laboratorio de Investigación y Desarrollo en Informática (LIDI)
Universidad del Azuay
Cuenca, Ecuador
priscila.cedillo@ucuenca.edu.ec

Orquestación de Servicios RESTful y SOAP: Caso de estudio, asignación de visas Americanas.

RESTful and SOAP Services Orchestration: Case Study, American Visa Assignment.

Jhon Calle
Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay
Cuenca, Ecuador
john6ero@es.uazuay.edu.ec

Pablo Lója
Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay
Cuenca, Ecuador
arevalop17@es.uazuay.edu.ec

Marcos Orellana
Laboratorio de Investigación y Desarrollo en Informática (LIDI)
Universidad del Azuay
Cuenca, Ecuador
marore@uazuay.edu.ec

Priscila Cedillo
Laboratorio de Investigación y Desarrollo en Informática (LIDI)
Universidad del Azuay
Cuenca, Ecuador
priscila.cedillo@ucuenca.edu.ec

Resumen— Este artículo presenta el análisis de un caso de asignación de visas americanas mediante la extracción de información de diferentes entidades. Para este problema se utilizan las siguientes fuentes de datos: i) información personal del solicitante, ii) información de los bienes inmuebles de la persona, y iii) el historial crediticio proveniente de entidades bancarias. Para la extracción de la información se crearon interfaces que simularon el funcionamiento de dichas entidades, así como sus respectivas bases de datos y servicios web. Para el envío de datos, se implementó en la aplicación, el uso de una arquitectura orientada a servicios (SOA) mediante el desarrollo de un sistema de orquestación de servicios. En este contexto, se aplicaron conceptos del lenguaje BPEL, cuya finalidad es demostrar la importancia de una orquestación de servicios en empresas u organizaciones y la gran utilidad de disponer de servicios integrados. Los conceptos aplicados en el desarrollo del sistema final, corroboran su utilización, puesto que los módulos desarrollados, pueden integrarse a otros sin que exista la necesidad de volver a diseñarlos desde cero, lo que representa una ayuda sustancial en la búsqueda de soluciones óptimas para infraestructuras TI.

Palabras Clave— *BPEL, Orquestaciόm, Servicios Web, SOA, SOAP, Visas.*

Abstract— This article presents the analysis of a case of U.S. visa assignment by extracting information from different entities. The following data sources are used for this issue: (i) personal information of the applicant, (ii) information on the person's real estate, and (iii) credit history from banks. For the extraction of the information, interfaces were created that simulated the operation of these entities and their respective databases and web services. For data submission, the use of a service-oriented architecture (SOA) by developing a service orchestration system was implemented in the application. In this context, BPEL language concepts were applied, which aim to demonstrate the importance of service orchestration across enterprises or organizations and the great utility of having integrated services. The concepts applied in the development of the final system corroborate its use, since the modules developed can be integrated with others without the need to redesign them from scratch, which is a substantial aid in the search for optimal solutions for IT infrastructures.

Keywords— *BPEL, Orchestration, SOA, SOAP, web services, Visas.*

I. INTRODUCTION

En la actualidad, los servicios web son de gran utilidad para el desarrollo e integración de diversas plataformas y entornos heterogéneos distribuidos. A lo largo de estos años, los servicios web han permitido el acceso a la información requerida, para tratar con todo tipo de negocio en cualquier momento y en cualquier lugar; estos servicios se pueden describir como una aplicación o parte de la misma que la podemos encontrar en la web [1].

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Constituyen entonces, módulos de software que se auto describen y que poseen autonomía; permiten realizar tareas concretas, además de caracterizarse por un fácil despliegue ya que se basan en tecnologías estándar HTML y HTTP [2].

Si bien, la orquestación de servicios web puede ser descrita como una integración de servicios; ésta puede ser interpretada también como una implementación de procesos por combinación de servicios en operaciones comerciales [3], permitiendo la creación de los mismos en alto nivel. Con estas características, la orquestación se basa en las interacciones entre servicios internos y externos que se dispongan para la ejecución de diversas tareas, mismas que son necesarias para el correcto funcionamiento de un sistema de información [4].

Un punto a considerar son las limitaciones de tiempo para establecer una consistencia temporal [4]. Entonces, es de suma importancia el uso de métodos unificados con sintaxis y semántica definida, con el objetivo de verificar y validar las propiedades que estos cumplen; para ello, es necesaria una arquitectura orientada a servicios (SOA), la cual permite crear sistemas escalables, facilitando la interacción ya sea en sistemas propios o de terceros, proporcionando ventajas; entre ellas: reutilización, agilidad y acoplamiento; esto, con la ayuda de una colección de servicios [5].

La implementación de una arquitectura basada en procesos, se ha venido implementando desde hace algunos años en organizaciones TI. Los principales beneficios que brinda SOA son la agilidad comercial y la reutilización. Cuando los requerimientos del negocio cambian, el costo de mantener la solución sincronizada puede ser muchas veces mayor que el de construir la solución desde sus componentes modulares. Si una solución se entrega con un enfoque SOA, junto con la gestión de procesos empresariales y las tecnologías de servicios web, se puede gestionar ágilmente con relación a las ya construidas con un enfoque tradicional [6]. El uso de una orquestación en modelos empresariales ya definidos, conlleva un cambio radical en los mismos, debido a que estos tienen que ser creados de forma general [1].

Un uso práctico que se puede dar a la orquestación es la inclusión de flujos de trabajos, muy usados dentro de los procesos de negocios. Gracias a que la orquestación está basada en una arquitectura de servicios, puede aplicarse a una arquitectura empresarial (EA - Enterprise Architecture); ésta coopera con una visión más abstracta de los activos de una empresa de TI. El enfoque SOA permite crear redes de sistemas que interactúan entre sí, y a los mismos se les puede asignar un flujo de trabajo, es decir, dotarlos de una jerarquía dentro de la empresa. El orquestador contiene la definición de flujos de trabajo y los ejecuta. La primera categoría del sistema representa la capa de flujo de trabajo, mientras que la segunda, se encuentra representada por la capa SOA, que expone los servicios a la capa de flujo de trabajo. De acuerdo a este modelo de orquestación para la automatización de flujos de trabajo, todo el sistema puede considerarse como un gráfico dirigido en el que el conductor central organiza los flujos entre varios sistemas participantes [7].

Para la realización de una orquestación en el análisis de este caso particular, se realizó un enfoque centrado en los servicios, es decir, los mismos deben poseer interfaces WSDL, para que de este modo puedan ser incluidos dentro del motor BPEL, dado que los servicios RESTful no utilizan WSDL, se presentaron dificultades al momento de integrarlos al BPEL [8] [18]. Mediante WSDL se puede definir de mejor manera un mensaje SOAP, y en caso de interactuar con un servicio del mismo tipo ya creado, solo es necesario que se proporcione el WSDL que funge como contrato entre cliente y proveedor. El uso de servicios SOAP es muy útil en entornos de tipo empresarial distribuidos, en cambio REST, está más encaminado a una conexión directa punto a punto. Además como punto a favor de SOAP, mediante la seguridad origen-destino, que la misma proporciona, es de mayor utilidad en la creación y uso de servicios de tipo asincrónico.

En este artículo, se presenta la simulación de un sistema de obtención de visas, el cual, según el historial crediticio de las personas, calificará a una persona a fin de saber si esta puede obtener una visa Americana. Para la esta simulación, se dispone de un sistema de registro, el cual contiene la información completa de la persona, a su vez, un sistema de comprobación de deudas, donde se verifica su historial crediticio, esto ayudará a determinar si puede o no calificar para la visa Americana. Los indicadores que se tomarán en cuenta son: las propiedades hipotecadas que disponga el solicitante y las deudas pendientes que disponga con alguna entidad bancaria; en ambos casos, si el solicitante dispone de al menos una hipoteca o varias deudas pendientes, cabe la posibilidad que la persona trata de evadir estas responsabilidades saliendo del país, por lo que la solicitud será cancelada hasta que sus movimientos financieros mejoren. Cada sistema se desarrollará en diferentes ambientes de programación y contarán con el

uso de servicios RESTful para la obtención de la información requerida, servicios SOAP para la creación de los micro servicios necesarios, y un sistema gestor de base de datos en donde se almacenará toda la información necesaria y requerida.

Para cumplir con este objetivo, es necesario que los servicios trabajen de forma conjunta, ordenada, y eficaz; lo que en conjunto forma parte de un proceso global que cumple con el objetivo primordial. Es necesario resaltar que si existen fallas en la implementación de la orquestación, se podría generar errores en el sistema tales como: fallos en el envío y/o recepción de información, o el envío de información errónea.

Partiendo de estas premisas el este artículo se encuentra organizado de la siguiente manera. La sección 2 describe todo el procedimiento y las herramientas que fueron necesarias tanto para la construcción del entorno de simulación, así como el desarrollo de la orquestación de los servicios involucrados, detallando de manera concisa cada una de las etapas a seguir. La sección 3 describe los resultados y observaciones que se obtuvieron al realizar la simulación. Finalmente, la sección 4 presenta las conclusiones y el trabajo futuro.

II. MATERIALES Y MÉTODOS

A.Modelamiento de la solución.

Para el desarrollo del caso de estudio, se debe considerar las siguientes restricciones: una persona calificará para la obtención de una visa, dependiendo de su historial crediticio en todas las entidades bancarias en las que posea una cuenta, y de las propiedades que la persona disponga y si éstas a su vez, no se encuentran hipotecadas.

Para obtener estos datos, se requiere consultas de las siguiente entidades: Registro Civil, Registraduría de Propiedad, y Superintendencia de Bancos. Para ejecutar la simulación, se han creado bases de datos en distintos gestores, y aplicaciones que simulan su funcionamiento. Para obtener la información de cada fuente, se crearon servicios web, que se encargaron de realizar la funcionalidad de cada entidad.

La orquestación de estos servicios se realizó con el lenguaje denominado BPEL, lo que permite que los servicios estén disponibles y a la espera que sean utilizados por el orquestador. A su vez se devuelve el conjunto de información para su posterior evaluación, es decir, si la persona califica o no para la visa (Fig. 1.).

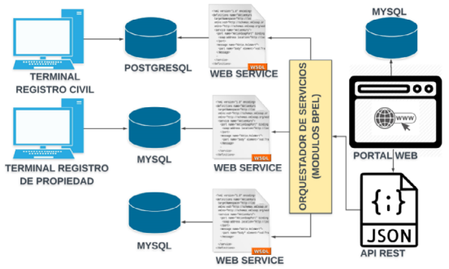


Fig. 1. Arquitectura Servicios.

La Fig. 1. muestra la topología utilizada para la orquestación, en donde se observa los diferentes elementos que la componen, entre ellos, los diferentes terminales correspondientes a las entidades: Registro Civil, y Registraduría de propiedad, y Portal web del sistema de asignación de visas, cada una con su base de datos y sus servicios web. Y en el centro el orquestador que es el encargado de dar funcionalidad al sistema.

B. Creación de las Bases de Datos

Con relación a los gestores de bases de datos, se ha utilizado dos diferentes, MySQL y Postgres, los cuales almacenan la información pertinente a cada entidad (ver Fig.1).
La siguiente sección, realiza una descripción detallada de la composición de cada una de las bases de datos creadas:

•Aplicativo Registro Civil: Utilizar Postgres como gestor de base de datos, misma que cuenta con un objeto de tipo tabla denominado persona, la cual almacena la información de la misma, los campos que componen dicha tabla son: id, número de identificación, nombres, apellidos, sexo, instrucción, nivel académico y estado civil.

•Aplicativo Registraduría de Propiedad: En este caso el gestor usado es MySQL. La base de datos cuenta con dos objetos de tipo tabla: persona y bienes, la primera registra la información básica de un usuario (cedula, nombres, apellidos, entre otros) mientras que, la segunda contiene la información de bienes; en ella se pueden apreciar los siguientes campos: código de propietario, dirección, tamaño, cotización, fecha de registro, tipo de inmueble y un campo adicional sobre el estado del inmueble, es decir, si posee alguna hipoteca.

•Base de datos de créditos: Se refiere al almacén de datos en donde se encuentra la información crediticia de los usuarios, esta al igual que la tabla anterior se encuentra en el gestor MySQL y cuenta con una sola tabla, misma que contendrá los siguientes campos: id, nro de identificación de usuario, banco asociado, tipo de cuenta o responsabilidad y finalmente el estado de la misma.

•Aplicativo web sistema de visas: Para este aplicativo se ha usado como gestor a MySQL, este almacena información del usuario, y la información referente a la decisión de otorgar o no la visa de acuerdo a los factores evaluados, es decir, su estado de aprobación.

Cabe recalcar que para la simulación e implementación se han creado tablas con la cantidad de columnas necesarias; en la realidad estos sistemas requieren más columnas por cada usuario ingresado y por ende de un esquema de mayor magnitud. Además, se debe también considerar que al ser entidades diferentes que gestionan su propia información, pueden llegar a ser similares muchas de las columnas, por el hecho mismo de que gestionan su propia información.

C.Desarrollo de Aplicativos

El desarrollo de los aplicativos se ha realizado usando diferentes lenguajes de programación:

•Java: Con este lenguaje se ha desarrolLado la aplicación de escritorio para el Registro Civil. Para este desarrollo se ha empleado el entorno de desarrollo Netbeans; además, esta aplicación contempla el paradigma de programación orientado a objetos (POO) para el manejo de información de usuarios. Además, para la conexión a la base de datos Postgres se ha usado el driver denominado PostgreSQL JDBC, permitiendo realizar un mantenimiento (CRUD - Create, Read, Update, Delete) de las tablas [9].

•Visual Basic: Se ha realizado el aplicativo de escritorio en donde se registra la propiedad del usuario, el entorno de desarrollo usado para la implementación es Visual Studio. De manera similar, se ha usado POO para gestionar de manera fácil y correcta la información, tanto de usuarios como de bienes [10].

Finalmente el desarrollo del aplicativo web está compuesto de la siguiente manera:
•Backend: Se ha utilizado PHP como lenguaje de programación para accionar los diferentes componentes HTML. En este caso, para consultar la información del usuario y almacenarla [11].

•FrontEnd: Adicionalmente, para el diseño se ha utilizado un Framework CSS denominado Bootstrap, que ofrece clases con estilos previamente realizados, y que pueden aplicarse fácilmente, esto con el objetivo de que el usuario disponga de una mejor interacción con el sistema [12].

D.Creación de Servicios Web

Una vez creados los diferentes aplicativos que simulan las entidades a ser evaluadas, se procede con el desarrollo de los servicios web que solicitarán y proporcionarán la información correspondiente. Para ello, se ha optado por crear servicios de tipo SOAP y REST.

Para consumir la información del Registro Civil, Registraduría de propiedad e historial de crédito, se ha realizado servicios del tipo SOAP y la codificación en lenguaje php, ya que este ofrece un kit de herramientas para desarrollo e implementación de este tipo. A este tipo de herramienta se le conoce como NuSOAP, la misma que está basada en SOAP 1.1, WSDL 1.1 y HTTP 1.0/1.1. Además, la herramienta está compuesta con una serie de clases que facilitan la implementación tanto de servidores (proveen servicios) como de clientes (consumen servicios). Algo particular y a considerar es que se ha optado el uso de esta herramienta, debido a que el componente BPEL proporcionado por OpenESB no acepta otras versiones de SOAP.

Los servicios mencionados, se han creado únicamente para consulta, estos tienen como parámetro de entrada el número de identificación del usuario, en caso de existir se devuelve la información completa del mismo. Por ejemplo. En el primer caso se devuelve toda información de usuario que se encuentra en el Registro Civil, por otra parte, el segundo devuelve un listado de bienes registrados correspondientes al usuario, y finalmente se devuelve todo su historial crediticio. Toda esta información coopera para tomar la decisión en cuanto a la aprobación de visas.

Por otra parte, el aplicativo web implementa servicios de tipo REST, con el objetivo de no centrarse en un solo tipo de servicio, ya que una de las principales ventajas que BPEL ofrece es la integración de servicios previamente creados. Entonces, para la implantación, es necesaria la instalación de un micro framework denominado Slim con el administrador de dependencias denominado Composer, luego es necesaria la configuración de un fichero denominado .htaccess el cual contiene las directivas de comportamiento del servidor. A continuación, se indica la definición del archivo:

```
RewriteEngine On
RewriteCond %{REQUEST_FILENAME} !-f
RewriteCond %{REQUEST_FILENAME} !-d
RewriteRule ^ index.php [QSA,L]
```


ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

La primera línea activa el motor de redirecciones, las siguientes dos indican que, si el archivo con el nombre especificado en el navegador no existe, o el directorio en el navegador no existe, entonces proceda a la regla de reescritura siguiente.

Luego de configurado el fichero se procede a crear el servidor, es importante definir los tipos de métodos aceptados por el servicio, en este caso se ha definido de la siguiente manera: ('Access-Control-Allow-Methods', 'GET, POST, PUT, DELETE, PATCH, OPTIONS').

Finalmente, se procede a la creación de los métodos que componen los servicios y registrarlos en el aplicativo; lo que comprende dos principales, el primero, enviará el número de identificación al WSDL proporcionado por BPEL y retornará la información del usuario en un texto con un formato de tipo JSON, mismo que será útil para visualizarlo en el aplicativo web y posteriormente generar el reporte final. Por otra parte, el segundo servicio es de ingreso de datos, es decir, una vez generado el reporte y asignado el estado de aprobación se procede a guardar los datos en la base local.

E.Implementación de BPEL

La etapa relevante es la implementación de la orquestación de servicios; como se mencionó con anterioridad, se ha utilizado un conjunto de herramientas que permiten su resolución y aplicación. Es importante recalcar que el módulo BPEL que se encuentra en OpenESB ofrece los tipos de variable estándar, de modo que, su uso es limitado, debido a normalmente se cuenta con una estructura de datos compleja, es decir, pueden ser objetos o un conjunto de ellos, sin embargo, la herramienta ofrece la posibilidad de crear esquemas XML propios con el fin de definir a criterio propio los tipos de datos a utilizar. Además, el módulo BPEL ofrece una vista de diseño de la aplicación, lo cual facilita la creación de estos procesos, algo a tomar en cuenta es que todos los servicios externos que se han añadido al módulo están descritos en WSDL (Servicios SOAP).

Como primer paso de la implementación se han agregado los diferentes servicios web tipo SOAP, mismos que se encargan de validar y enviar los datos desde las siguientes entidades: Registro Civil, Registraduría de Propiedad y la Superintendencia de Bancos. Cada uno de ellos tiene un parámetro de entrada con el cual retorna la información si encuentra coincidencia en el identificador. Posteriormente, se procede a crear un documento WSDL al cual se le asigna un nombre, el mismo que es necesario para su posterior invocación, como recomendación es preferible un nombre acorde al funcionamiento del servicio, en este estudio se definirán tanto el parámetro de entrada como el de salida y se elegirán los tipos de datos a usar.

En BPEL existen varias actividades que se pueden realizar. A continuación, se listan las utilizadas para la realización del proceso de Consulta de Datos:

- ***Receive:*** Es el que recibirá el parámetro de entrada principal, es decir una vez creado el módulo por completo, este necesitará de un parámetro para realizar todo el proceso, mismo que corresponde al número de identificación del usuario, que será enviado por el aplicativo web correspondiente al sistema de asignación de visas.
- ***Assign:*** Esta actividad se encarga de asignar valores a las variables, en ella, existe la posibilidad de aplicar operaciones lógicas, con cadenas, entre otras funcionalidades. Es decir, puede contener una o más operaciones elementales.
- ***Invoke:*** Encargado de invocar los diferentes servicios web.
- ***Reply:*** Retorna el mensaje o la información de respuesta al documento WSDL una vez que se ha cumplido toda la secuencia de operaciones.
- ***If:*** Estructura condicional de tipo Booleana.
- ***Sequence:*** Realiza actividades en orden secuencial, se presenta luego de una estructura condicional.

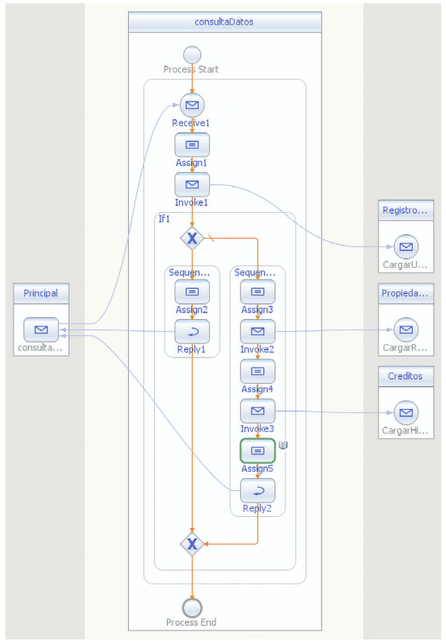


Fig. 2. Proceso Módulo BPEL.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

En la Fig. 2. se aprecia la estructura del proceso, en donde se usa la actividad de tipo Receive para obtener el código ingresado por el usuario y almacenarlo, luego existe una actividad Assign, la cual une el parámetro almacenado con el parámetro de entrada del elemento Invoke, este a su vez hace la petición al Registro Civil, la respuesta recibida por Registro Civil, será interpretada con la actividad condicional "if", esta actividad retorna un valor booleano, en caso de ser negativo se asigna la salida a un elemento Reply y este envía el resultado a la respuesta del proceso, terminando de esta manera el proceso de consulta de datos. Caso contrario, empieza la siguiente secuencia, es decir, una vez verificada la existencia del usuario en el Registro Civil, se procede a llamar a los servicios restantes para obtener los datos de bienes y del historial de crédito. Una vez realizadas todas las peticiones correspondientes, se procede a asignar la respuesta del proceso, en este momento se combinan los resultados de los tres servicios web invocados y se los envía a la salida del proceso.

Una vez ejecutado todo el proceso de asignación de variables y validaciones se ha concluido con la creación e implementación del módulo. Cabe mencionar que el BPEL está diseñado para este tipo de escenarios, lo cual hace que los tiempos de implementación sean más cortos [13].

Finalmente, para realizar pruebas de funcionamiento del módulo, no basta con solo crearlo, es necesaria la creación de una Composite Application (aplicación compuesta) en la cual se pueden incluir múltiples módulos BPEL u otros tipos de módulos de Java Business Integration (Integración Empresarial de Java). Una vez agregado el módulo BPEL el cual se ha creado con anterioridad, se procede a agregar un conector, existen varios tipos: SOAP, REST, FTP, HTTP, entre otros. Para este caso particular ha utilizado uno de tipo SOAP. Como paso final, se realiza el despliegue de la Aplicación, el WSDL creado por OpenESB correspondiente al servicio de orquestación se lo puede encontrar en la siguiente ruta: localhost:9080/CompositeVisasService1/ConsultaUsuario?WSDL. Esta ruta, contiene el servidor local, ya que fue en este sitio donde se publicaron, además, contiene un puerto dedicado en este caso 9080. De esta manera el servicio orquestado está listo y a la espera de ser consumido.

F.Integración BPEL con aplicativo web

Cuando se han realizado todos los procedimientos anteriores de implementación y sus respectivas pruebas, es necesaria la fase de puesta en producción de los servicios web, la orquestación, y en este caso del aplicativo principal, que es el desarrollado para la web. Se realiza entonces la invocación del servicio usando el lenguaje PHP, el mismo que contiene un formulario, en donde es necesario ingresar un campo de entrada: el número de identificación de un ciudadano. Posteriormente, este dato es enviado a través del servicio web creado y antes mencionado, mismo que invocará al recurso de orquestación, que a su vez hará la petición de la información a los servicios correspondientes. Obteniendo finalmente respuesta del mismo, los datos obtenidos son necesarios para la realización de los procesos internos, que verificarán y generarán el reporte final, posteriormente, se procede a guardar la información en la base de datos correspondientes.

III. RESULTADOS Y DISCUSIÓN

El modelamiento de la solución se realizó de manera efectiva utilizando el lenguaje BPEL, ya que permite crear un modelo de empresa multifuncional y adaptable a cualquier tipo de situación; es decir, permite su reutilización, con lo cual se evita crear una aplicación en específico para cumplir ciertas funciones puntuales requeridas por una empresa. La simulación de las entidades del Registro Civil, y Registraduría de Propiedad, plantea un entorno de simulación semejante a la realidad. Varias organizaciones sean estas públicas o privadas, realizan consultas a otras empresas para obtener información sea de personas, productos, etc. Esta información, se encuentra en diversos gestores de bases de datos, utilizando diversos entornos de desarrollo en múltiples lenguajes de programación. En el presente caso, la obtención de la información del solicitante se realiza mediante servicios web creados específicamente para cada entidad, esto facilita su uso dentro del modelo BPEL, ya que una de las ventajas que lo caracteriza como es la integración de servicios ya creados. Una orquestación provee de un ahorro de tiempo a la organización, mediante una sola interfaz, obtiene los datos relevantes acerca del solicitante y a su vez, según los mismos, lo califica apto o no para la obtención de una visa. Además, al estar disponible toda la información en una sola interfaz, se reduce el número de empleados en el proceso, lo cual aumenta la productividad dentro de la empresa, ya que se puede centrar esfuerzos en realizar otras actividades más complicadas, lo cual mejora considerablemente la calidad de servicio.

La utilización de la herramienta Open ESB, potencia el uso de recursos; los componentes son de fácil instalación y manejo. Este

software ofrece una interfaz amigable, su diseño es de gran similitud al IDE NetBeans, y, los aplicativos pueden ser creados, ya sea mediante componentes o código XML puro, evidentemente siguiendo una sintaxis y semántica adecuada. Adicionalmente, brinda una gran ayuda para empezar a desarrollar en BPEL, ya que dispone de una considerable cantidad de herramientas más que suficientes si el modelo a crear no requiere muchas especificaciones, si este es el caso, el uso de herramientas más completas, como puede ser Oracle JDeveloper resultaría más conveniente.
El sistema de otorgamiento de visas puede ser reutilizado en otros entornos que lo necesiten. La obtención de información y procesamiento, tanto del Registro Civil, y de la propiedad, pueden ser usados en numerosos procesos, por ejemplo: al solicitar un crédito en cualquier entidad bancaria, en donde es obligatorio conocer las deudas tanto del solicitante, o en caso de que el solicitante esté casado, se requerirá información de su cónyuge, y así determinar si califica o no al crédito [14].

En esta comparativa observamos que el sistema de créditos, es un sistema mucho más complejo, por lo que su desarrollo se podría realizar en la herramienta Oracle JDeveloper, debido a que este es un sistema más completo con respecto al usado en el presente caso de estudio. Se puede destacar las similitudes entre ambos modelos, ya que se sigue una estructura similar, y los pasos necesarios para la orquestación. Indistintamente de la herramienta utilizada, el caso de una aplicación de créditos, posee una construcción más extensa.

Un punto clave a considerar, son las capacidades que tenga el equipo en el cual se realice la orquestación de servicios. En el caso de Open ESB, basta con solo un equipo para el desarrollo de la aplicación en BPEL; en caso que se requiera integrar procesos más extensos, y se requiera el uso de Oracle JDeveloper, es recomendable el uso de varios computadores para equilibrar el uso de recursos. Adicional, como se requiere el uso del gestor de base de datos Oracle XE, es necesario analizar si no se encuentran versiones previas instaladas en el computador, y comprobar que los puertos de escucha se encuentren habilitados.
En estos dos casos, orquestar ayuda a procesar servicios que dispongan de un alto coste de ejecución; generalmente, en grandes empresas, la obtención de información, mediante servicios, tarda desde horas hasta días y en caso de no disponer de una arquitectura orientada a servicios, se dedicaría mucho tiempo a monitorizar los procesos. Para esto, conviene el uso de mensajes dentro del orquestador, mediante los cuales, se informe de manera rápida y concisa sobre los tiempos que se demora en obtener los datos, tal caso, que se sobrepase los mismos, alertará al usuario [14].
Una complicación que se puede encontrar al usar esta herramienta, es la no compatibilidad con servicios creados en VB.NET, ya que los servicios tipo SOAP que este ofrece están basados en SOAP 1.1 y 1.2.

IV. CONCLUSIONES

La orquestación de servicios al ser un proceso central que toma el control de todos los servicios que se encuentren conectados con un propósito común, brindan beneficios a empresas TI, ya que mejora la efectividad de las mismas.

Las opciones en orquestación que brinda Open ESB, permiten crear un ambiente de desarrollo entendible usando módulos BPEL, además de ofrecer otro tipo de módulos para el manejo de estructuras de información, mismos que pueden ser integrados para trabajar en conjunto, aunque existen ciertas restricciones en comparación con otras herramientas.

Además, como característica opcional, es de fácil instalación y su modo de diseño es entendible, este ofrece un servidor que se encarga de publicar nuestros servicios para su posterior prueba y uso.

El uso de BPEL, al ser un lenguaje ya estandarizado a nivel mundial, ayuda a reducir la complejidad que requiere realizar una orquestación. Las herramientas y componentes usados, están disponibles en cualquier software que permita su implementación. Al disponer ya de un estándar, facilita a las empresas su uso y a su vez reduce costos de desarrollo. El no disponer de un estándar obligaría a cada organización o empresa a crear su propio conjunto de reglas y esto produciría poca colaboración entre diversos servicios web.
Para el caso particular de estudio, la simulación de un entorno con diferentes entidades creadas en diferentes lenguajes de programación y usando diferentes gestores de datos, ayuda en gran manera a comprobar los beneficios que conlleva una orquestación, estos se ven reflejados en el tiempo que toma la verificación de datos del solicitante, la reducción de empleados al realizar este proceso, y al ser reutilizable, se pueden automatizar procesos repetitivos.

Para concluir, se puede apreciar, el potencial que brinda a una empresa cambiar a una arquitectura basada en servicios (SOA). Independientemente de la herramienta que se use, el realizar una orquestación brinda nuevos

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

horizontes, apegados a una mejor calidad de servicio para sus clientes y empleados, siendo esto de vital importancia en un mundo competitivo donde un error, sea este grande o pequeño, puede significar pérdidas económicas o mucho peor, el cierre definitivo de la misma. Un posible trabajo futuro para la arquitectura SOA con el uso de BPEL, sería interesante la implementación de un inventario de funcionalidades con categorizaciones para un desarrollo con asistencia de la inteligencia artificial. La idea central sería el desarrollo de un recomendador que guíen al desarrollador de un sistema en la utilización de componentes de otros sistemas que aporten en el objetivo del sistema final.

RECONOCIMIENTO

Los autores desean agradecer al Vicerrectorado de Investigaciones de la Universidad del Azuay por el apoyo financiero y académico, así como a todo el personal del Laboratorio de Investigación y Desarrollo en Informática (LIDI).

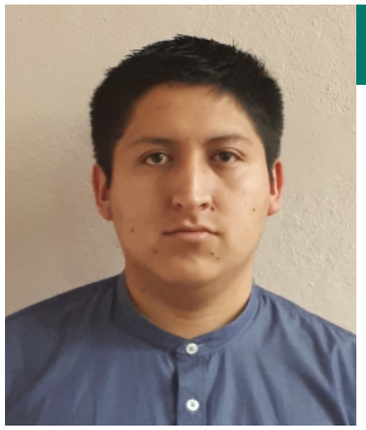
REFERENCIAS

[1] C. G. Bernardo, “Loan system in brazilian financial institution - A SOA application,” Proc. 9th Int. Conf. Inf. Technol. ITNG 2012, pp. 293–298, 2012, doi: 10.1109/ITNG.2012.50.
[2] Y. Chtouki, H. Harroud, P. O. Box, and A. H. Ii, “Service Orchestration Algorithm for Web Services : Evaluation and Analysis,” vol. 10, no. 5, pp. 208–218, 2013.
[3] V. W. Chu, R. K. Wong, S. Fong, and C.-H. Chi, “Emerging Service Orchestration Discovery and Monitoring,” IEEE Trans. Serv. Comput., vol. 10, no. 6, pp. 889–901, 2015, doi: 10.1109/tsc.2015.2511000.
[4] K. Benghazi, M. Noguera, C. Rodŕiguez-Domínguez, A. B. Pelegrina, and J. L. Garrido, “Real-time web services orchestration and choreography,” CEUR Workshop Proc., vol. 601, pp. 142–153, 2010.
[5] S. Kumari and S. K. Rath, “Performance comparison of SOAP and REST based Web Services for Enterprise Application Integration,” 2015 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2015, pp. 1656–1660, 2015, doi: 10.1109/ICACCI.2015.7275851.
[6] N. Zhou and L. J. Zhang, “Analytic architecture assessment in SOA solution design and its engineering application,” 2009 IEEE Int. Conf. Web Serv. ICWS 2009, pp. 807–814, 2009, doi: 10.1109/ICWS.2009.117.
[7] T. Ploom, A. Glaser, and S. Scheit, “Platform based approach for automation of workflows in a system of systems,” c2013 IEEE 7th Int. Symp. Maint. Evol. Serv. Cloud-Based Syst. MESOCA 2013, pp. 12–21, 2013, doi: 10.1109/MESOCA.2013.6632730.
[8] K. He, “Integration and orchestration of heterogeneous services,” 2009 Jt. Conf. Pervasive Comput. JCPC 2009, pp. 467–470, 2009, doi: 10.1109/JCPC.2009.5420139.
[9] Oracle Corporation, “¿Qué es Java y para qué es necesario?,” Oracle Corporation, 2018. https://www.java.com/es/download/faq/whatis_java.xml (accessed Jun. 12, 2020).
[10] Microsoft, “Herramientas de desarrollo e IDE gratuitos | Visual Studio Community.” https://visualstudio.microsoft.com/es/vs/community/ (accessed Jun. 11, 2020).
[11] R. Lerdorf, H. Magnusson, P. Olson, and L. Kahwe Smith, “PHP: ¿Qué es PHP? - Manual,” Http://Php.Net/, 2001. https://www.php.net/manual/es/intro-whatis.php (accessed Jun. 15, 2020).
[12] Bootstrap Team, “Download · Bootstrap v4.5.” https://getbootstrap.com/docs/4.5/getting-started/download/ (accessed Jun. 17, 2020).
[13] OpenESB, “OpenESB Home.” https://www.open-esb.net/ (accessed Jul. 18, 2020).
[14] M. Orellana and L. M. Arévalo, “Orquestación de servicios web aplicado a una solicitud de crédito comercial utilizando la herramienta Oracle BPEL Process Manager.,” pp. 1–137, 2013.

AUTHORS

Jhon Calle

Estudiante de la Carrera de Ingeniería de Sistemas y Telemática de la Universidad del Azuay. Certificado obtenido del VII Congreso Ecuatoriano de Tecnologías de la información y Comunicación – TICEC 2019.



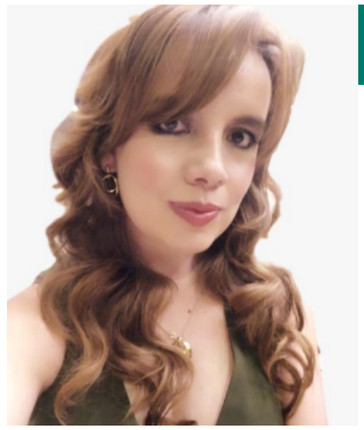
Pablo Lója

Estudiante de la Carrera de Ingeniería de Sistemas y Telemática de la Universidad del Azuay. Certificado obtenido del VII Congreso Ecuatoriano de Tecnologías de la información y Comunicación – TICEC 2019.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Marcos Orellana

Ingeniero en Sistemas de la Universidad del Azuay. Master en Gestión de Sistemas de Información e Inteligencia de Negocios de la Universidad de las Fuerzas Armadas (ESPE) y Master en Docencia Universitaria de la Universidad del Azuay. Candidato a doctor en Ciencias Informáticas en la Universidad Nacional de la Plata Argentina. Docente-Investigador en las líneas de Ciencia de los Datos e Inteligencia Artificia en la Universidad del Azuay. Responsable y Director del Laboratorio de Investigación y Desarrollo en Informática (LIDI).



Priscila Cedillo

Ingeniera de Sistemas por la Universidad de Cuenca, Máster en Ingeniería de Software, Métodos Formales y Sistemas de Información por la Universitat Politécnica de Valencia en España, Magister en Telemática por la Universidad de Cuenca, Ph. D. en Informática por la Universitat Politécnica de Valencia. Estancia de investigación en el National Institute of Informatics en Tokyo y un Posdoctorado en la Universidad Politécnica de Valencia. Autora de aproximadamente 80 artículos indexados. Sus líneas de investigación: Cloud Computing, Calidad de Software, Model Driven Engineering, Internet of Things y Ambient Assisted Living.

Proceso de migración de datos en la implantación de Aplicaciones Informáticas Empresariales

Data migration process in the implantation of Enterprise IT Applications

ARTICLE HISTORY

Received 12 October 2020
Accepted 02 November 2020

Elvis Moreta
Carrera de Ingeniería en Sistemas Computacionales
Universidad Técnica del Norte
Ibarra, Ecuador
edmoretam@utn.edu.ec

Irving Reascos
Carrera de Ingeniería en Sistemas Computacionales
Universidad Técnica del Norte
Ibarra, Ecuador
imreascos@utn.edu.ec

Proceso de migración de datos en la implantación de Aplicaciones Informáticas Empresariales

Data migration process in the implantation of Enterprise IT Applications

Elvis Moreta
Carrera de Ingeniería en Sistemas Computacionales
Universidad Técnica del Norte
Ibarra, Ecuador
edmoretam@utn.edu.ec

Irving Reascos
Carrera de Ingeniería en Sistemas Computacionales
Universidad Técnica del Norte
Ibarra, Ecuador
imreascos@utn.edu.ec

Resumen—La migración de datos se define como la trasferencia de datos entre sistemas gestores de bases de datos. Existen escasas metodologías, estrategias y técnicas sobre la migración de datos, sin embargo, estas no se aplican, son desconocidas o en su mayoría están diseñadas para grandes empresas. La presente investigación tiene como finalidad elaborar una descripción del proceso de migración de datos en la implantación de Aplicaciones Informáticas Empresariales (AIE). La metodología que usamos es un estudio de campo, el cual está basado en la metodología de estudio de caso propuesto por Yin. Este estudio de campo consistió en entrevistas a consultores expertos y a personal de las PyMEs que participaron en un proceso de migración de datos. El análisis cualitativo de los datos recolectados se realizó con la herramienta MAXQDA, siguiendo las recomendaciones de Kuckartz. El proceso resultante detalla las fases, actividades y actores involucrados en la migración de datos.

Palabras clave—*proceso de migración de datos, migración de datos, PyMEs, análisis cualitativo, ERP.*

Abstract— Data migration is defined as the transfer of data between database management systems. There are few methodologies, strategies and techniques on data migration, however, these are not applied, are unknown or mostly designed for large companies. The purpose of this research is to develop a description of the data migration process in the implantation of Enterprise IT Applications (EITA). The methodology we use is a field study, which is based on the case study methodology proposed by Yin. This field study consisted of interviews with expert consultants and staff of the SMEs that participated in a data migration process. The qualitative analysis of the data collected was carried out with the MAXQDA tool, following Kuckartz's recommendations. The resulting process details the phases, activities and actors involved in the data migration.

Keywords— *data migration process, data migration, SMEs, qualitative analysis, ERP.*

I. INTRODUCCIÓN

La migración de datos es la transferencia de datos entre diferentes tipos de formatos de archivo, sistemas gestores de bases de datos (SGBD), sistemas de almacenamiento o aplicaciones informáticas empresariales (AIE); en el ámbito empresarial representa un 60% de cualquier proyecto de tecnologías de la información (TI) [1]. Leguizamon [2] define a la migración de datos como el proceso para extraer información útil y comprensible en diferentes formatos, que se realiza por varios motivos, tales como: cambios, actualizaciones y problemas de rendimiento de las AIE, entre otras causas.

La pérdida de información, pérdidas económicas, el desprecio de datos históricos, el retraso en proyectos que dependen de la migración de datos y el fracaso en la transición de AIE constituyen los principales factores críticos de éxito y a su vez motivaciones al momento de aplicar un proceso de migración de datos en las pequeñas y medianas empresas (PyMEs) [3].

En la literatura existen escasas metodologías, estrategias y técnicas que facilitan la migración de datos, sin embargo, no se aplican, son desconocidas o diseñadas para grandes empresas. Márquez, et al. [4], afirman que existe una necesidad urgente de proporcionar metodologías, técnicas e instrumentos que faciliten el proceso de migración de datos a nuevas plataformas y arquitecturas.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

La presente investigación tiene como finalidad describir el proceso de migración de datos en las PyMEs a través de un estudio de campo, tomando como base la metodología de investigación empírica estudio de caso de Yin [5], para comprender como se está realizando esta actividad. La estructura del artículo es la siguiente: en la sección I se presenta la introducción a la unidad de análisis, en este caso el proceso de migración de datos. La sección II describe metodologías y estrategias existentes en la literatura. La sección III habla sobre la metodología empleada para el desarrollo del artículo. La sección IV describe los resultados obtenidos. En la sección V se establecen las conclusiones del trabajo de investigación y se presenta una discusión del tema.

II. ESTADO DEL ARTE

Para realizar la migración de datos en sistemas de legado Frey [6], describe las metodologías: Butterfly y Chicken Little.

Butterfly consta de seis fases: preparación, comprender la semántica del sistema de legado y desarrollar el esquema de datos de destino, construir una base de datos de prueba, migrar todos los componentes (excepto los datos) del sistema de legado, migrar gradualmente los datos del sistema de legado al sistema destino y empezar a usar el sistema objetivo. Esta metodología se usa especialmente en migraciones que requieren tener funcionando en paralelo tanto el sistema de legado como el destino, al menos durante el proceso de migración.

Chicken Little, compuesta de once pasos: analizar de forma incremental el sistema de legado, descomponer progresivamente el sistema de legado, diseñar de forma incremental las interfaces, aplicaciones y SGBD de destino, instalar de forma incremental el entorno, crear e instalar las puertas de enlace necesarias, migración incremental de los SGBD, aplicaciones e interfaces de legado y usar gradualmente la información de destino. Es considerada una técnica de migración incremental y tiene una característica particular que es utilizar puertas de enlace para la transferencia de datos entre el sistema de legado y el destino.

Russom [7] afirma que cada proyecto de migración de datos tiene una combinación única de AIE de origen, requisitos de la AIE destino y usuarios. Además, existen técnicas que prevalecen en la industria como: la extracción, transformación y carga de datos (ETL); codificación manual; replicación de bases de datos y la integración de AIE. Cabe mencionar que en este articulo los autores usan las siglas ETL haciendo referencia a la técnica de migración "Extract, Transform and Load", que son más conocidas que las siglas en español.

El proceso de migración de datos parece simple de realizar, sin embargo, deben considerarse los cambios que conlleva, como por ejemplo, SGBD, AIE, entre otros; además, es necesario un análisis dentro de la empresa para definir la estrategia apropiada. En la mayoría de los casos, los proyectos de migración de datos forman parte de la implantación de una AIE, por eso, la estrategia de migración de datos debe alinearse con la estrategia de implantación de la AIE.

En Fig. 1 se ilustran las estrategias de migración de AIE. Motiwalla y Thompson [8] clasifican a las estrategias en: por fases (Phased), donde la organización realiza la migración de manera gradual desde los sistemas de legado existentes hacia la aplicación destino; piloto (Pilot), se utiliza para garantizar que el sistema final es apropiado para la empresa; en paralelo (Parallel), tiene el mayor costo inicial porque ambos sistemas se usan a la vez, se emplea cuando el riesgo de fracaso del proyecto es inminente; y Big Bang, donde la empresa implanta la aplicación nueva de manera inmediata y directa dejando de usar el sistema de legado.

Las PyMEs amplían sus capacidades de gestión con la implantación de AIE y por efecto las capacidades de almacenamiento deben ajustarse a las necesidades de las PyMEs. A través de las AIE, las empresas hacen un uso completo de los datos para impulsar las decisiones empresariales y adaptarse al crecimiento de la información.

III. METODOLOGÍA

En esta investigación, utilizamos un enfoque cualitativo ya que este tipo de investigación se utiliza para comprender problemas o situaciones investigando las perspectivas y el comportamiento de las personas en estas situaciones y el contexto en el que actúan. Para lograr esto, la investigación cualitativa se lleva a cabo en entornos naturales y utiliza datos en forma de palabras en lugar de números. Los datos cualitativos se recopilan principalmente a partir de observaciones, entrevistas y documentos, y se analizan mediante una variedad de técnicas sistemáticas. Este enfoque es útil para comprender los procesos causales y para facilitar la acción basada en los resultados de la investigación [9].

El estudio de caso es una metodología de investigación empírica donde preguntas "cómo" y "por qué" son frecuentes a la hora de plantear la unidad de análisis [5]. La investigación descriptiva resulta útil para estudiar problemas prácticos, situaciones determinadas y comprender el proceso que se usa [10]. Por eso, la metodología empleada en esta investigación toma como base al estudio de caso para comprender como se realiza la migración de datos en las PyMEs. Es decir, usamos un estudio de campo (basado en las técnicas de estudio de caso), donde varios expertos en migración de datos y diferentes actores de TI, son entrevistados.

En Fig. 2 se detallan las fases de la metodología empleada para la investigación. Las fases son: (A) Planificación, (B) Recolección de datos, (C) Análisis de datos y (D) Presentación de resultados.

A.Planificación

En esta fase se define la unidad de análisis en base a las preguntas de investigación, se preparan los instrumentos para la recolección de datos, se buscan profesionales expertos en el área y se agendan las entrevistas a los participantes.

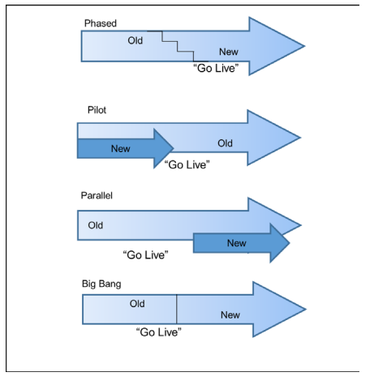


Fig. 1. Estrategias de migración de AIE [8].

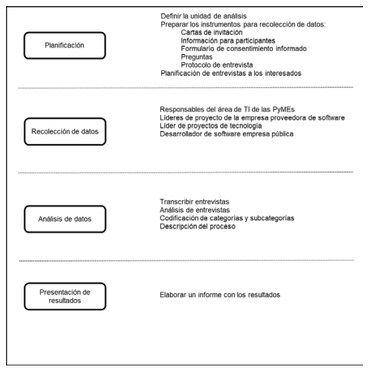


Fig. 2. Metodología para el estudio de campo adaptada de la metodología de estudio de caso propuesta por Yin [5].

La unidad de análisis es: "proceso de migración de datos", la cual se define en base a la pregunta de investigación detallada en la Tabla I.

La siguiente actividad es preparar los instrumentos para la recolección de datos. Los instrumentos elaborados fueron: cartas de invitación, información para participantes, formulario de consentimiento informado, preguntas y protocolo de entrevista.

Las preguntas de la entrevista se preparan en base a la unidad de análisis. Para afinar las preguntas, se hicieron iteraciones con entrevistas piloto para determinar si están bien formuladas. En la Tabla II se listan las preguntas realizadas a los entrevistados.

Una vez determinadas las preguntas se procede a contactar y agendar a los interesados en colaborar con la investigación empleando los instrumentos elaborados previamente.

B.Recolección de datos

La recolección de datos permite profundizar y comprender la unidad de análisis planteada durante la investigación. En esta fase se realizaron nueve entrevistas a diferentes actores involucrados en proyectos de migración de datos, estos actores fueron: responsables del área de TI de las PyMEs (2 entrevistados), líderes de proyecto de la empresa proveedora de software (2 entrevistados), líder de proyectos de tecnología (1 entrevistado), desarrollador de software empresa pública (3 entrevistados), desarrollador de software empresa privada (1 entrevistado). Las entrevistas, fueron grabadas con el consentimiento del entrevistado y tuvieron una duración promedio de 36 minutos, siendo la más corta de 12 minutos y la más larga de 89 minutos.

C.Análisis de datos

En esta fase se realiza la transcripción de las entrevistas, para lo cual se utilizan las herramientas oTranscribe (https://otranscribe.com/), para escuchar la entrevista y Google Docs para el dictado por voz.

Posteriormente siguiendo la guía de métodos y buenas prácticas para análisis de textos cualitativos de Kuckartz [11], se empieza con la codificación de categorías y subcategorías, usando MAXQDA 2020, software de análisis cualitativo.

Las entrevistas transcritas se analizan tomando en cuenta la pregunta de investigación. Se leen los textos en su totalidad y se resalta términos claves o conceptos. MAXQDA facilita la segmentación de varios textos en un solo proyecto y permite agregar notas para su posterior análisis. Se determina un código apropiado a los pasajes de los textos para la

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

codificación de los mismos. La construcción de categorías se basa en teorías (deductivo) y en los datos que se reflejan en las entrevistas (inductivo). La combinación de estos métodos permite definir el sistema de códigos.

En Fig. 3 se encuentra el sistema de códigos de la investigación. De la interpretación de categorías y subcategorías establecidas emergen el modelo del proceso de migración de datos en la implantación de AIE y también se responden las preguntas de investigación.

Tabla I. Pregunta de investigación.

Número	Pregunta de investigación	Motivación
PI1	¿Cómo se realiza el proceso de migración de datos en las PyMEs y cuáles son las principales motivaciones y dificultades?	Identificar las tareas y actividades que se llevan a cabo para realizar el proceso de migración de datos. Conocer las motivaciones y dificultades implicadas en el proceso.

Tabla II. Preguntas de la entrevista.

Número	Pregunta de la entrevista
PE1	¿Cuáles son las principales motivaciones por las cuales las pequeñas y medianas empresas realizan migración de datos?
PE2	¿Cuáles son las principales dificultades que han encontrado durante este proceso?
PE3	¿Cómo se realizó el proceso de migración de datos?
PE4	¿La empresa tenía definido el proceso de migración de datos o fue necesario empezar desde ahí?
PE5	¿Usaron alguna metodología para realizar la migración de datos o es un proceso empírico?
PE6	¿Existe alguna documentación que quedó para la empresa o que documentación se generó, en caso de existir, nos puede comentar acerca de contratos o formatos que se hicieron?
PE7	¿Tiempo estimado en el que realizaron la migración?
PE8	¿Qué recomendaciones haría para futuros procesos de migración de datos dentro de la empresa o con alguna otra organización?
PE9	¿Hubo asesoramiento de alguna entidad externa o ente regulatorio durante el proceso de migración de datos?
PE10	¿Cómo se hace la planificación de un proceso de migración?
PE12	¿Manejan alguna política para el tratamiento de la información?
PE13	¿Usaron aplicaciones informáticas para realizar la migración de datos?
PE14	¿Cómo se realiza la depuración de datos, se emplean estrategias?
PE15	¿Cómo manejan los históricos de los clientes?

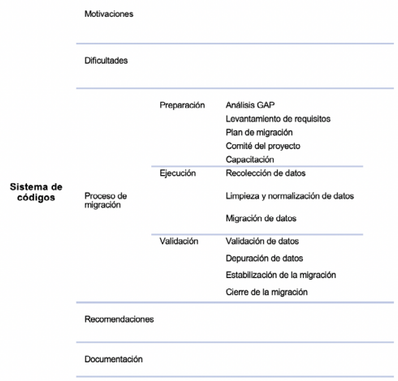


Fig. 3. Sistema de códigos de la investigación en MAXQDA.

D.Presentación de resultados

El proceso resultante del análisis cualitativo de las entrevistas consta de tres fases: preparación, ejecución y validación. Los resultados obtenidos y el proceso se detallan en la sección IV del artículo.

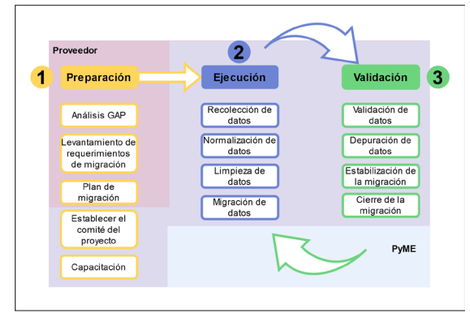


Fig. 4. Proceso de migración de datos

IV. RESULTADOS

A. Proceso de migracion de datos

La descripción del proceso de migración de datos se realiza en base al análisis cualitativo de los datos recolectados con la ayuda de la codificación de categorías y subcategorías. En la Fig. 4 se muestra el proceso de migración de datos, con sus respectivas fases, actividades y actores involucrados.

En la Tabla III se describen las actividades del proceso de migración de datos.

1) Fase de Preparación

Durante la primera fase tanto el cliente como el proveedor entran en un proceso de preparación de las empresas para tener claro el alcance, tecnologías a usarse, situación actual, responsables y estructura del proyecto.

a) Análisis GAP

El proceso conocido como análisis GAP busca reducir cualquier brecha existente entre dos entes, estos pueden ser, la empresa y su competencia, una AIE y otra, procesos o situaciones actuales de la empresa [12].

Analizar las deficiencias o brechas antes de comenzar el proyecto es más eficaz en cuanto a costos y la aplicación de un análisis GAP permite a la empresa tener un esfuerzo distribuido [13]. Addagada [14] afirma que las empresas usan el análisis GAP debido a cambios estratégicos, en la condición del mercado, de sus productos o buscando una mejora de procesos.

Es necesario analizar a la empresa a través de su misión, objetivos, estrategias y tácticas para comprender hacia donde se dirige y como ayudar. En este caso el proveedor debe

realizar una investigación a fondo acerca del cliente con la finalidad de entender como está estructurada la organización y como se manejan los procesos. Para esto, las entrevistas y reuniones entre los actores involucrados ayudan a sobrellevar esta fase inicial.

Con esta información, se construye un modelo del proceso de negocio que le sirve al proveedor para tener claro los objetivos que debe cumplir el proyecto, es recomendable que la empresa proveedora tenga experiencia en el área de negocios del cliente porque se facilita la consecución de fases con mayor agilidad.

b) Levantamiento de requerimientos de migración

Una vez comprendida la situación actual del cliente el proveedor procede con el levantamiento de requerimientos del proyecto de migración. Los requisitos funcionales y no funcionales son detallados en un acta firmada por los actores involucrados para evitar que el proyecto se extienda indefinidamente, malinterpretación por parte del personal y trabajar en base a lo requerido. Es importante recalcar que, si en fases posteriores el cliente tiene requerimientos adicionales, estos se detallen en un acta firmada en caso de que los actores involucrados estén de acuerdo.

c) Plan de migración

El plan de migración se compone de actividades, responsables, estrategias y el alcance del proyecto. En este caso, los proveedores de la AIE cuentan con un plan que se puede adaptar a diferentes situaciones y empresas.

El proveedor debe conocer a fondo a la empresa cliente, como se organiza y las personas a cargo del actual sistema, módulo o datos a migrar para establecer un plan de acuerdo con la situación actual del cliente. Aquí se especifican plazos a cumplir, involucrados y metas del proyecto en curso.

Por lo general, el plan usa herramientas como el cronograma y las reuniones para la consecución del proyecto. Los datos se almacenan en SGBD y se estructuran en módulos, así que la gran mayoría de planes de migración toman como punto de partida la cantidad de módulos del sistema o AIE a migrar para definir las etapas del proyecto y consideran la prioridad y complejidad de cada uno de ellos ya que puede haber módulos pequeños y otros con una gran cantidad de tablas.

d) Establecer el comité del proyecto de migración

Establecer un comité para el proyecto es de vital importancia. Si se presentan dificultades o ambigüedades son resueltas por este comité. Además, se encarga de la toma de decisiones y cumplimiento de la planificación realizada previamente. El seguimiento del proyecto se hace a través del cronograma y se valida los avances en las reuniones planificadas.

El comité debe ser multidisciplinario y contar con personas tanto de la empresa proveedora como del cliente. Existen aspectos técnicos que solo el personal del área tecnológica puede resolver y aclarar; y asimismo aspectos administrativos o de cualquier otra índole que en su momento pueden generar dudas o complicaciones.

e) Capacitación

Previo al inicio de la fase de ejecución se realizan capacitaciones de inducción al cliente para explicar el funcionamiento de la AIE, la estructura de los datos, los scripts de validación, los ambientes de pruebas, entre otras actividades que ayudan al cliente a implantar el software con éxito y migrar los datos de una manera efectiva y segura.

Tabla III. Actividades del proceso de migración de datos.

Fases	Actividades	Descripción
Preparación	- Análisis GAP	La empresa proveedora analiza al cliente para comprender su situación actual y hacia dónde quiere llegar. Responsable: Líder del proyecto de migración (Empresa proveedora)
	- Levantamiento de requerimientos de migración	Se detallan los requerimientos funcionales y no funcionales para delimitar el alcance el proyecto. Responsable: Líder del proyecto de migración (Empresa proveedora)
	- Plan de migración	El plan está compuesto de actividades, responsables, estrategias y el alcance del proyecto. Responsable: Líder del proyecto de migración (Empresa proveedora)
	- Establecer el comité del proyecto de migración	Se establece un comité multidisciplinario que cuenta con personas tanto de la empresa proveedora como del cliente y se encarga de solventar los problemas que se presenten. Responsable: Líderes del proyecto de migración, Gerente de la PyME
	- Capacitación	Se instruye al personal de TI de la empresa cliente en las funcionalidades de la AIE, estructura, validaciones y todo acerca del SGBD y AIE destino. Responsable: Usuarios finales, Personal de TI (Proveedor).
	- Recolección de datos	Iniciada la fase de ejecución, es necesario recolectar los datos que van a ser migrados y de ser necesario se involucran actores externos al cliente. Responsable: Personal de TI (PyME), Usuarios finales, Líderes del proyecto de migración.
Ejecución	- Normalización de datos	Se realiza una reestructuración de los datos. Responsable: Personal de TI (PyME), Líderes del proyecto de migración.
	- Limpieza de datos	Los campos incompletos, erróneos son corregidos. Responsable: Personal de TI (PyME), Líderes del proyecto de migración.
	- Migración de datos	Los datos son migrados hacia los ambientes de prueba de la AIE y/o SGBD destino. Responsable: Personal de TI (PyME), Líderes del proyecto de migración.
Validación	- Validación de datos	Automatización de las validaciones a través de scripts. También pueden validarse funcionalidades con grupos de usuarios. Responsable: Personal de TI (Proveedor, PyME).
	- Depuración de datos	Los errores o inconsistencias encontradas en la validación de datos se corrigen con la depuración. Responsable: Personal de TI (Proveedor, PyME), Usuarios finales.
	- Estabilización de la migración	Verificar que las iteraciones previas hayan conseguido solventar las dificultades tanto de la fase de ejecución como la de validación. Responsable: Personal de TI (Proveedor, PyME), Usuarios finales.
	- Cierre de la migración	Finalización del proyecto y firma de actas. Puede realizarse un cierre al final de cada iteración si el proyecto está conformado de varias fases o módulos. Responsable: Personal de TI (Proveedor, PyME), Usuarios finales.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

La experiencia del personal de la empresa cliente juega un papel importante durante las capacitaciones ya que se pueden incluir sesiones de análisis y compatibilidad de herramientas, soluciones a posibles errores o inconvenientes, entre otros. Para validar los conocimientos del personal se puede usar pruebas calificadas y pruebas piloto en ambientes controlados.

2) Fase de Ejecución

En la fase de ejecución inician las actividades que involucran a los datos. El proveedor toma el rol de guía para el cliente durante esta fase ya que los datos son el bien más preciado de una organización y contienen información sensible y confidencial, en la mayoría de los casos. Dependiendo de la planificación es posible realizar iteraciones entre la fase de ejecución y validación.

Además, es importante tomar en cuenta la complejidad y compatibilidad de las AIE para usar diferentes estrategias como ETL, soluciones de codificación manual, replicación de bases de datos, integración de aplicaciones empresariales, o técnicas como por fases, piloto, en paralelo o big bang. En este caso, las actividades de la fase de ejecución hacen alusión a ETL. Los datos se extraen en la recolección de datos (a), se transforman con la normalización (b) y limpieza de datos (c) y se cargan en el SGBD destino con la actividad denominada migración de datos (d).

a) Recolección de datos

En la empresa cliente empieza la fase de recolección de datos, para lo cual se involucran varios departamentos, y de ser necesario inclusive la alta gerencia. Siguiendo los parámetros establecidos durante los pasos previos, se decide que datos migrar con el apoyo del personal de TI (si existiera).
Dependiendo de la estructura de la empresa cliente y de las AIE, módulos o SGBD existentes es posible que la extracción o recolección de datos tenga que vincular a un actor externo al proyecto, ya sea para soporte o acceso a los datos, pero por lo general, el personal de TI conoce la situación actual de la organización y las tecnologías usadas.

b) Normalización de datos

Por lo general, la estructura de los datos antes del proyecto suele presentar redundancia e inconsistencia de datos. Si el SGBD destino no aplica las formas normales es recomendable aplicar la normalización tanto en el SGBD origen como en el SGBD destino. Es posible que las inconsistencias se presenten también como parte del modelo de negocio de la empresa o relacionado a alguna de las áreas de la organización.

c) Limpieza de datos

Previo a la migración de los datos hacia el sistema destino, hay veces en las que se realiza una limpieza de datos. Es decir, aquellos datos que están incompletos o erróneos tienen que corregirse para encajar perfectamente y no representar un problema aún mayor en futuras fases.

d) Migración de datos

Los datos son migrados hacia los ambientes de prueba de la AIE destino. Es importante tener en cuenta que no todos los escenarios de migración de datos son iguales, por lo tanto, pueden presentarse problemas de compatibilidad de herramientas, AIE, SGBD, entre otros.

Es importante tener en cuenta el tema de seguridad de datos. Durante la fase de ejecución se exponen los datos de la empresa cliente, así que tanto los ambientes de prueba como los medios en los que se transmite la información, deberán estar cifrados y contar con varias capas de seguridad.

3) Fase de Validación
La fase final de validación busca solucionar las dificultades previas y satisfacer las motivaciones que encausaron a la empresa a involucrarse en un proyecto de migración.

a) Validación de datos
La validación se realiza con la ayuda de scripts automatizados que verifican que son datos reales. Comprueban tipos de datos, reglas de validación y arrojan errores en caso de no cumplirse.

Si los datos migrados corresponden a información contable, existe aún mayor rigurosidad durante esta fase, en ocasiones se involucran actores externos como entes regulatorios públicos o privados dependiendo del área de negocio de la organización. Para evitar problemas, es recomendable revisar manualmente la información sensible en la que los procesos automatizados sean complejos de implementar.

Si existen nuevas funcionalidades o reportes específicos, las validaciones se realizan con grupos de usuarios para verificar que los datos sean íntegros y verdaderos.

b) Depuración de datos
Los errores o inconsistencias encontradas en la validación de datos se corrigen con la depuración. Es decir, si las validaciones, tanto automatizadas como manuales, fallan, comienza una iteración para resolver las inconsistencias.

La depuración de datos se encarga en gran medida de la estructura y errores relacionados con los SGBD destinos, mientras que la limpieza de datos busca solventar problemas del sistema o base de datos anterior.

c) Estabilización de la migración
En la estabilización de la migración se comprueba que las iteraciones previas hayan conseguido solventar las dificultades tanto de la fase de ejecución como la de validación. De ser necesario se replantea el cronograma y se proponen nuevas reuniones para la revisión.

Los tiempos de ejecución y características o procesos que se requieran mejorar son medidos y comparados en esta fase. También se analizan las soluciones propuestas durante la limpieza y depuración de datos para saber si es necesario o no una nueva iteración.

d) Cierre de la migración
La etapa de cierre se lleva a cabo una vez finalizado el proyecto o al final de cada iteración. Dependiendo del proyecto o empresa, es necesario agrupar los sistemas, módulos o datos a ser migrados en diferentes iteraciones o fases, para no generar problemas al cliente y tomar decisiones en base a los errores que se presenten en las fases anteriores.

También se firman actas que validan la consecución de un porcentaje o la totalidad del proyecto, los responsables y el comité del proyecto determinan si los objetivos trazados fueron alcanzados y si es necesaria una nueva iteración.

Durante esta etapa se debe establecer o al menos considerar una estrategia o qué hacer con el antiguo sistema, AIE o base de datos. Es decir, dejar de usar o eliminar los accesos a las herramientas usadas previo al inicio del proyecto, de ser necesario su uso; también establecer reglas o situaciones en las que se puede consultar los datos antiguos. Los respaldos finales e iniciales de ambas plataformas son necesarios para evitar futuros inconvenientes y tener un registro del proyecto.

B.Motivaciones y dificultades
Las motivaciones y dificultades presentes en el proceso de migración se agrupan en tres contextos: Tecnológico, Organizacional y Entorno; basados en el framework TOE que analiza la adopción e implantación de innovaciones tecnológicas en diferentes organizaciones [15].

1) Motivaciones
Del análisis y codificación de las entrevistas emergen factores tecnológicos, organizacionales y de entorno. Las motivaciones tecnológicas para que la empresa decida implantar una AIE y pasar por un proceso de migración de datos son: la integración con servicios de otras empresas, mejorar o cambiar las tecnologías que usa la empresa en ese momento por unas nuevas, con más prestaciones y que solucionen problemas anteriores.

En cuanto a lo organizacional, la transformación digital y la mejora continua de procesos figuran entre las principales motivaciones para iniciar el proyecto. En el contexto de entorno, las causas pueden ser regulaciones del estado, la reputación de la empresa y no relegarse de la competencia. Es decir, si existe una tendencia de las empresas de un determinado sector por adquirir un nuevo software o mejorar sus procesos, esto hace que el resto de las empresas del sector sigan estas tendencias.

2) Dificultades
Así mismo, se presentan dificultades, que emergen del análisis y codificación de las entrevistas, agrupadas en tecnológicas, organizacionales y de entorno. Entre las dificultades relacionadas con el contexto tecnológico esta la estructura de los datos, la compatibilidad de los SGBD que usan, la redundancia de datos, código fuente de la AIE no disponible, AIE con diferentes tecnologías, y también hay dificultades una vez iniciado el proceso de migración de datos como, por ejemplo, la calidad y depuración de los datos, herramientas de migración de datos poco eficientes y depuración de estas.

En cuanto a lo organizacional, el compromiso de la empresa, cambio de personal, procesos no definidos, gestión de procesos, apoyo de la alta gerencia, resistencia al cambio, retrasos en el proyecto y la estructura de la empresa. En el contexto de entorno, la empatía y las relacionales sociales con los proveedores suele determinar qué tan rápido y exitoso será el proyecto. Además, el tiempo y la interacción entre personas son factores a tomar en cuenta, ya que si se presentan dificultades es necesario solventarlas lo más rápido posible.

En la Tabla IV se destacan citas de las entrevistas realizadas durante el trabajo de investigación propuesto.

Tabla IV. Frases relevantas de las entrevistas.

Código	Cita	Entrevistado
C1	"En el GAP de migración se analizan fortalezas o debilidades que pueda tener cada institución."	Líder de proyecto de la empresa proveedora de software (1)
C2	"Se establece primero un plan de migración, dentro de este plan hay una etapa donde se verifican los posibles riesgos."	Desarrollador de software empresa pública (1)
C3	"Debe haber un Comité que defina el proceso de migración, el plan de migración, que valide la información capaz de que no existan repercusiones cuando haya el cambio de un sistema hacia el otro."	Desarrollador de software empresa pública (1)
C4	"Se realiza capacitaciones de inducción para que ellos como tal extraigan la información, coloquen en formatos establecidos y posterior la entrega a nuestro lado."	Líder de proyecto de la empresa proveedora de software (1)
C5	"El sistema anterior y el diseño de la base de datos tenía ciertas falencias. Había hasta cierto punto algunas tablas que estaban sin relaciones"	Responsable del área de TI de la PyME (1)
C6	"Definir qué información existe y si esa información va a ser útil o la podemos descartar."	Líder de proyectos de tecnología (1)
C7	"Recibimos la información y ejecutamos procedimientos de validación automáticos."	Líder de proyecto de la empresa proveedora de software (3)

V. CONCLUSIONES, DISCUSIÓN Y TRABAJOS A FUTURO

El estudio de campo planteado permite comprender como se realiza el proceso de migración de datos en las PyMEs. El proceso consta de tres fases: Preparación, Ejecución y Validación. Cada una de ellas incluye subfases o pasos, la primera análisis GAP, levantamiento de requerimientos, plan de migración, comité del proyecto, capacitación; la segunda recolección de datos, limpieza de datos y migración de datos; y la tercera validación de datos, depuración de datos, estabilización de la migración y cierre de la migración. Los actores involucrados en el proceso son el proveedor y el cliente (PyME).

Este proceso permite plantear objetivos alcanzables en los proyectos que involucran migración de datos y disminuir el índice de fracasos. Así mismo, ayuda a empresas de desarrollo de software a comprender el proceso de migración de datos y no sobrellevar el proyecto de manera empírica. De igual manera, las PyMEs pueden aplicar el proceso en caso de que el proveedor no cuente con un proceso o metodología a seguir.

El proceso de migración de datos propuesto aporta mayor detalle en cuanto a las fases y actividades a desarrollarse durante el proceso, siguiendo las buenas prácticas como ambientes de prueba antes de salir a producción que considera Butterfly en sus fases. A diferencia de Chicken Little, acompaña la fase de preparación con capacitaciones al personal de TI que se encarga de la migración de datos.
El objetivo de la investigación se cumple con el proceso descrito en la sección anterior. Además, se identifican las principales motivaciones y dificultades que se presentan durante el proceso.

La presente investigación no considera escenarios que involucren más actores, o microempresas que no cuenten con el personal necesario para ejecutar este proceso. No se cubrieron aspectos de migración a Cloud Computing, aunque el modelo resultante podría servir de referencia. La dificultad en esta investigación fue el acceso a expertos en el tema de migración de datos. La limitación fue un número limitado de actores entrevistados. Para trabajos a futuro se puede ampliar o desarrollar un nuevo proceso que resuelva este problema, por ejemplo, un proceso de migración de datos para entornos Cloud. El alcance de este trabajo es comprender como los diferentes actores están realizando la migración de datos. En un futuro, se tendrá que mejorar la propuesta y realizar la validación respectiva.

RECONOCIMIENTO
Universidad Técnica del Norte, Facultad de Ingeniería en Ciencias Aplicadas, Carrera de Ingeniería en Sistemas Computacionales.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

REFERENCIAS

[1] N. Tehreem, «Data Migration – The Why, The What, and The How,» 2019. [En línea]. Disponible: https://www.astera.com/type/blog/data-migration-software/.

[2] A. Leguizamon, "Pautas para una correcta migración de datos," 2017.

S. Vidacic, I. Pihir and R. Fabac, "Method of data migration from one ERP system to another in real time," Proceedings of the 21st Central European Conference on Information and Intelligent Systems, 2010.

[4] L. Márquez, D. G. Rosado, H. Mouratidis, D. Mellado and E. Fernández-Medina, "A Framework for Secure Migration Processes of Legacy Systems to the Cloud," Advanced Information Systems Engineering Workshops, 2015.

[5] R. K. Yin, Case Study Research Design and Methods, 5th ed., Thousand Oaks, CA: Sage, 2014, p. 282 páginas.

[6] S. Frey, "Migration of software systems to platform as a service based cloud environments (Doctoral dissertation, Kiel University)," 2011.

[7] P. Russom, «Best Practices in Data Migration,» The Data Warehousing Institute, 2006.

[8] L. Motiwalla y J. Thompson, Enterprise Systems for Management, Boston, MA: Pearson, 2012.

[9] B. Kaplan and J. A. Maxwell, "Qualitative research methods for evaluating computer information systems," Evaluating the organizational impact of healthcare information systems, pp. 30-55, 2005.

[10] B. Yazan, "Three Approaches to Case Study Methods in Education: Yin, Merriam, and Stake.," The Qualitative Report, vol. 20, pp. 134-152, 2015.

[11] U. Kuckartz, Qualitative Text Analysis: A Guide to Methods, Practice & Using Software, Londres: SAGE Publications Ltd, 2014.

[12] H. Kerzner, Strategic Planning for Project Management Using a Project Management Maturity Model, J. W. &. Sons, Ed., 2002, p. 272.

[13] L. M. Gonzalez Amaral and J. Pascoal Faria, "A Gap Analysis Methodology for the Team Software Process," Seventh International Conference on the Quality of Information and Communications Technology, pp. 424-429, 2010.

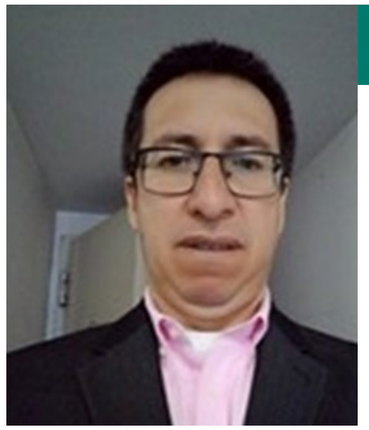
[14] T. Addagada, "Do We Need a Mature GAP Analysis?," 2012. [En línea]. Disponible: https://archive.is/20130120053323/http://clients.criticalimpact.com/go.cfm#selection-495.13-497.16

[15] R. Depietro, E. Wiarda and M. Fleischer, "The context for change: Organization, technology and environment," The processes of technological innovation, vol. 199, pp. 151-175, 1990.

[3]

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

AUTHORS



Irving Reascos

Irving Reascos es ingeniero en Sistemas Computacionales, Magíster en informática y PHD en Tecnologías & Sistemas de Información. Docente de las carreras de Ingeniería en Sistemas Computacionales e Ingeniería de Software en la Universidad Técnica del Norte (Ecuador). Profesor investigador, su área de interés está centrada en los Sistemas de Información Empresarial, lo que incluyen metodologías para la implantación de Aplicaciones Informáticas Empresariales (AIE) en las pymes e implantación de Entornos Virtuales de Aprendizaje en instituciones de educación superior

Elvis Moreta

Elvis Moreta es estudiante de la carrera de Ingeniería en Sistemas Computacionales. Actualmente realiza su trabajo de titulación en el área de Migración de Datos.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Síntesis de Sistemas de Conmutación Mediante Permutación de Tablas de Código Gray (Método PGC)

Switching Systems Synthesis Method Using Permuted Gray Code Tables (PGC Method)

ARTICLE HISTORY

Received 06 August 2020
Accepted 02 November 2020

César Troya-Sherdek
Faculty of Applied Science
International University of Ecuador
Quito, Ecuador
cesartroyasherdek@gmail.com
https://orcid.org/0000-0002-4274-2649

Valentin Salgado-Fuentes
Department of Mechanical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark
vasafu@mek.dtu.dk

Jaime Molina
Department of Mechanical Science
Kachariy Higher Technical Institute
Quito, Ecuador
jaime.molina@itk.edu.ec

Gustavo Moreno
Department of Electronic Science
Kachariy Higher Technical Institute
Quito, Ecuador
gustavo.moreno@itk.edu.ec

Síntesis de Sistemas de Conmutación Mediante Permutación de Tablas de Código Gray (Método PGC)

Switching Systems Synthesis Method Using Permuted Gray Code Tables (PGC Method)

César Troya-Sherdek
Faculty of Applied Science
International University of Ecuador
Quito, Ecuador
cesartroyasherdek@gmail.com

Valentin Salgado-Fuentes
Department of Mechanical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark
vasafu@mek.dtu.dk

Jaime Molina
Department of Mechanical Science
Kachariy Higher Technical Institute
Quito, Ecuador
jaime.molina@itk.edu.ec

Gustavo Moreno
Department of Electronic Science
Kachariy Higher Technical Institute
Quito, Ecuador
gustavo.moreno@itk.edu.ec

Resumen— Encontrar la función más corta en los sistemas de conmutación es una necesidad para el desarrollo de sistemas automáticos eficientes. Actualmente, existen varias metodologías que tienen como objetivo solucionar esta necesidad con diferentes técnicas. Este artículo propone una nueva metodología para encontrar una fórmula proposicional que describa un problema de un sistema de conmutación utilizando varias tablas de verdad que se basan en una original, estas tablas se generan utilizando los principios y permutaciones del Código Gray. Como se mostrará, el código utilizado tiene una relación directa con los caminos hamiltonianos, donde cada permutación es una conexión diferente en un hipervolumen y cada nodo se representa como una combinación de bits. Para verificar y validar el método, se desarrolló un algoritmo utilizando el MATLAB y se comparó con las soluciones del software Boole-Deusto. Finalmente, se presentan ejemplos de ejecución, comparación de costos computacionales y propuestas de trabajos futuros.

Palabras Clave— *Caminos hamiltonianos, código Gray, funciones booleanas, hipercubo, problemas discretos, sistemas de conmutación*

Abstract— Finding the shortest function on switching systems is a necessity for the development of efficient automatic systems. Currently, several methodologies aim to solve this need with different techniques. This article proposes a new methodology to find a propositional formula that describes a switching system problem using several truth tables which are based on an original one; these tables are generated using Gray Code principles and permutations. As it will be shown, the used code has a direct relation to the Hamiltonian paths, where each permutation is a different connection in a hypervolume, and each node is represented as a bit combination. An algorithm was developed using MATLAB and compared with the solutions from the software Boole-Deusto to verify and validate the applicability and implementation of the method. Finally, examples of execution, computational cost comparison and future work proposals are presented.

Keywords— *Boolean functions, discrete problems, Gray Code, Hamiltonian Paths, hypercube, switching systems.*

I. INTRODUCTION

The solution of switching systems problems is of increasing importance in the development of modern technologies as well as in the implementation of automated control strategies. Thus, some authors like P. Roth [1], Quine [2] or Karnaugh [3] made much effort to improve the efficiency of these solutions. In propositional logic, a truth table defines a problem and a propositional formula (also

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

known as truth function or Boolean function), can be obtained to describe any given truth table. Several methods can be used to obtain this formula, e.g. Veitch chart, Karnaugh map [4], minterms, maxterms [5] or Boolean algebra. However, as recognized by Quine [2], "The quest is to descry a technique to find the shortest truth-function formula" or by Veitch [6] "The problem is how to depict a Boolean function of "n" variables so the human eye can quickly see how to simplify the function". Veitch and Karnaugh used graphical methods, but higher-order problems present severe difficulties since the method involves human inspection.

This work proposes a new method to find a propositional formula that describes a switching system problem. Based on the original truth table, Gray code principles and permutations [7] are used to generate multiple truth tables. Gray codes are named after Frank Gray who in 1947 patented the idea of generating a binary codification that is used in applications that depends essentially on the bits looping. They are represented as a function G(i) where the consequential G(i+1) differ in exactly one bit [8], it is essential to note that the permuted tables that start with Gray code structure will maintain that property in all permutations. The advantage of using Gray codes to rearrange the truth tables, as will be explained in the method, lies on generating clusters that can be grouped and simplified using Boolean algebra theorems like identities and complements.

To verify and validate the proposed method, an algorithm using MATLAB 2014b is developed and tested with 2 to 7 bits logic tables. The computations are performed on a 64-bit architecture Intel XEON E3-1505M 2.80 GHz with 32 Gb RAM personal workstation capable of achieving a propositional formula that adequately solves the original switching system truth table. The solutions from the current method are compared with the solutions from the software Boole-Deusto [9].

II. METHOD

The Gray code tables present a single bit variation in each row, allowing to identify groups (2n members) of bits with a 'true' logic output that can be simplified. By permuting the columns of the Gray table, all possible clusters appear. The approach used has a direct relation with the Hamiltonian paths in hypercubes [10] where each permutation is a different path, and each node (vertex) is represented as a row (bit combination) of the truth table as can be seen in Fig. 1 [11]; transforming a multidimensional analysis into a unidimensional one. The right side of Fig. 1 shows a hypercube, whose vertices represent all possible combinations of the permutation. The vertices with a filled circle are the returned true outputs, and the vertices with the empty circle are the false outputs. Also, the arrows show the Hamiltonian path for the given Gray code truth table permutations that are represented in the left side of Fig. 1.

A	B	C	D	
0	0	0	0	1
1	0	0	0	1
1	0	0	1	1
0	0	0	1	1
0	1	0	1	1
1	1	0	1	1
1	1	0	0	1
0	1	0	0	1
0	1	1	0	0
1	1	1	0	0
1	1	1	1	1
0	1	1	1	1
0	0	1	1	1
1	0	1	1	1
1	0	1	0	0
0	0	1	0	0

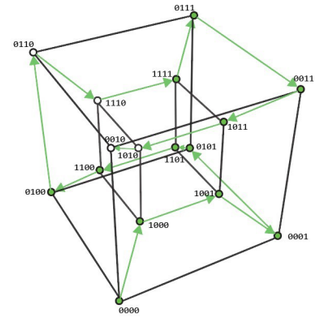


Fig. 1 Multidimensional representation of a Gray code table permutation.

The algorithm of the proposed method consists of four basic steps: preparation, generation, depuration and output. To illustrate the operation, a three-bit logic table that corresponds to the unidimensional output array XT = (0,1,0,1,1,0,0,1) will be used. For the first step, the original 'Gray table' must be generated along with all the possible permuted tables and each one of the eight-element logic output 'X' must be assigned to the corresponding input combination. The columns from the original 'Gray table' are named alphabetically using capital letters and used as independent arrays, for the example they will be: AT = (0,0,0,0,1,1,1,1); BT=(0,0,1,1,1,1,0,0); CT = (0,1,1,0,0,1,1,0). The number of permutations can be determined using (1) as described by Benavides [12], where the interchangeable values 'r' are the same than the number of bits 'n' so it can be expressed only as the factorial of 'n!'.

$$NTables = \frac{n!}{(n-r)!} \quad (1)$$

The permuted tables are assembled by concatenating the 'n' columns in the order given by the permutations and reassigning the labels of the columns to its original order; each row has also assigned its corresponding logic value of 'X' based on the input bits configuration. With the 'n!' tables filled, the generation stage begins, the objective is to mark the pairs of input sets combinations that give a 'true' logic output. Moreover, is important to keep track of the input sets that give a 'true' logic output

but never get in pairs with other combination through all the permutations, these sets are going to be known as 'elusive sets'. Fig. 2 shows the assembled permuted tables corresponding to the example; the dashed lines point out the selected input sets, the continuous vertical line indicates the corresponding true values and the continuous horizontal lines keep track of the 'elusive sets'.

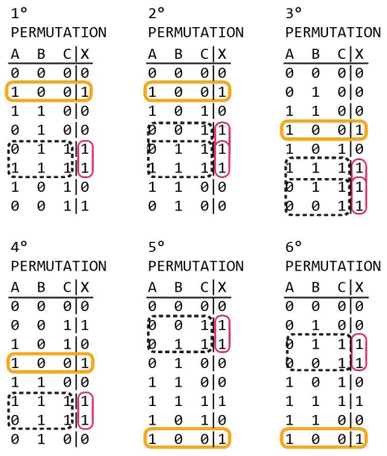


Fig. 2 Assembled permuted tables.

The pairs of input sets selected are reorganized in a two-column array called 'global combinations' while the 'elusive sets' occupy a namesake unidimensional array, as shown in Fig. 3a. In the depuration stage, the repeated sets must be deleted regardless of the order that the combinations were found (i.e., 011 111 is the equivalent of 111 011). The result of this operation is shown in Fig. 3b.

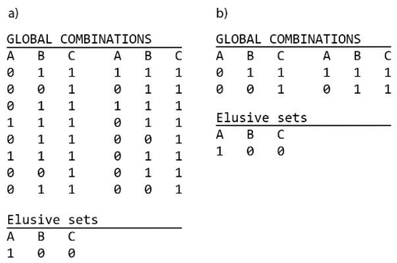


Fig. 3 a) Global combinations and Elusive sets. b) Depurated Global combinations and Elusive sets.

The output stage is required to translate the outcome of the depuration stage into the traditional Boolean algebra notation using the cleaned 'global combinations' array where the unchanged bits are subjected to a logical 'AND' whereas each row of the array to a logical 'OR'. On the 'elusive sets', the bits of each row are subjected to a logical 'AND' while the rows to a logical 'OR'. In both arrays, those bits with a 'false' logical state take the denied label (~), and bits with a 'true' logical state take only the label of the bit. The first combination changes bit 'A' while bits 'B' and 'C' remain the same, so the rule for this combination gives the output (B·C). In the second row, the bit 'B' changes, the bit 'A' remains with '0' and the bit 'C' remains with '1' so the logical output is (~ A · C). In the row from the 'elusive sets', the output is (A · ~ B · ~ C). The complete output is the logical OR of the previous rules: (~ A · C) + (B · C) + (A · ~ B ·~ C).

In some cases, there are associations involving two configurations sets that have already been marked; these are called 'ghost sets' and most commonly appear with four or larger numbers of input bits such as XT= (1,1,0,1,1,0,0,1,1,1,0,1,1,0,0,1) which corresponds to a four-bit function. Fig. 4a shows the 'global combinations' array for this logical output after removing the repeated sets. To remove the 'ghosts sets' is necessary to find clusters of logical bits that repeat their groupings. In Fig. 4a, the binary combinations that are used more than once are marked with '1'. On the other hand, the first appearance of each combination and those that appear only once are marker with '0' (β column in Fig. 4a). To mark a combination of sets as selected at least one of the two sets must be marked with a zero. Continuous lines in Fig. 4a note the 'cleaned global combinations'. A clearer representation of this can be seen in Fig. 4b, where the global combinations are placed in the Karnaugh map format. The continuous lines indicate the 'cleaned global combinations' and the shaded sections represent the 'ghost sets'. It is important to note that in order to optimize the cleaning of 'ghosts sets' it is necessary to reorder the table of 'global combinations' by placing the 'elusive sets' (dashed lines) at the end of the table, these are represented by pairs of identical combinations in Fig 4a. The final output will be: (A·~C~D) + (A·C·D) + (~B·~C·D) + (~A·C·D) + (~A·~C·~D).

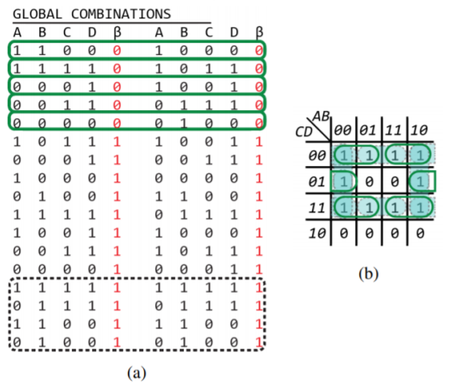


Fig. 4 a) Ghost sets depuration. b) Ghosts sets equivalent in Karnaugh maps.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

The importance of placing the 'elusive sets' at the end of the 'global combinations' table before executing the last depuration stage is to achieve a higher degree of simplification in the results, avoiding that the bits are selected individually generating functions equally equivalent but with more significant extension and complexity. To assess the effectiveness of the method in the treatment of the 'ghost sets', more cases with a different number of input bits were tested. Besides, unbalance cases with logic outputs containing more true outputs (1's) than false outputs (0's) and conversely, were used to make sure that proper simplification is accomplished. In these cases, sparse, random or uniform location of the true outputs (1's) was also considered.

Two examples corresponding to 3-bit combinations are presented to verify the entire method and its operation. The first example has an input XT= (0,0,0,1,1,0,0,1) and the second example is defined by the input XT= (1,0,0,1,1,0,0,1). Furthermore, the computational effort of the algorithm was analyzed to know how efficient the method could be in comparison with other fore-mentioned methods. However, even when the number of bits of the test cases is constant, the location of the true outputs (1's) changes the procedure performed in the depuration step and the running time needed. Therefore, to have a measurement that can be used as a benchmark, the input bit (1,0) was used in different computations but increasing with the number of bits; e.g. (1,0,1,0,1,0,1,0) for 3 bits, (1,0,1,0,1,0,1,0,1,0) for 4 bits, etc. The reason to use this logic output combination is due to the result in all the cases is the denied last variable (~ C and ~ D respectively). With only one variable as an answer, the eliminations of the repeated combinations in the depuration step increase accordingly with the number of bits and the running time. Finally, to increase the efficiency of the method, a more in-depth analysis was carried out to find a way to qualify the yielded permutations from the generation step based on better simplification perspectives.

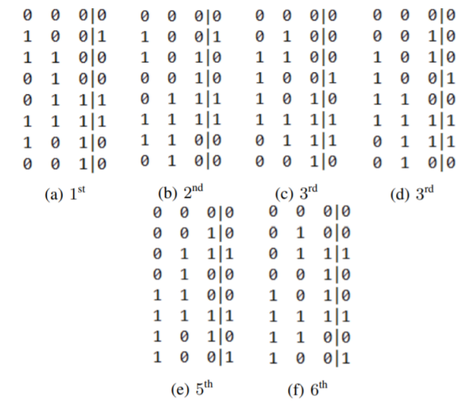


Fig. 5 Permutations - example 1.

From each of these permutations the local combinations are extracted and grouped in the table of 'global results' in Fig. 6, it is crucial to note the presence in the table of six identical rows with the combination [1 0 0 - 1 0 0] indicating a recursive 'elusive set'.

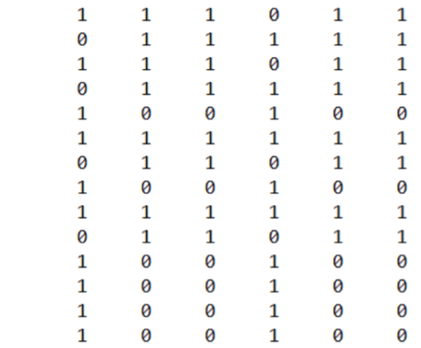


Fig. 6 Global Results - example 1.

Fig. 7a presents the results of performing the purification of repeated combinations on the global results, only four of the fourteen originals rows remained after the first filtration stage. Three of these results will remain in the form of 'elusive sets', however, at the time of performing the procedure described previously to eliminate 'ghost sets' the last two rows are discarded because they present redundant combinations that do not provide direct information to the solution, this final depuration result is presented in Fig. 7b.

III. RESULTS

A. Example number 1

From the vector XT= (0,0,0,1,1,0,0,1), six existing permutations are generated and presented in Fig. 5.

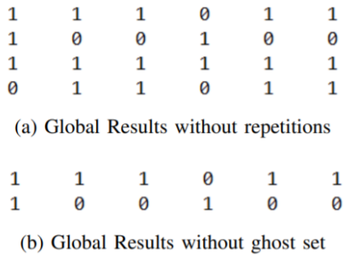


Fig. 7 Depurations results - example 1.

Finally, the translation to traditional algebra notation is done by analyzing the bits that do not change in each row; the first combination changes only in the first bit while the second one remains constant in all of them, applying a logical OR (+) connection between the rows, the final Boolean function solution is (B·C) + (A·~B·~C).

B.Example number 2

From the vector XT= (1,0,0,1,1,0,0,1), six existing permutations are generated and presented in Fig. 8. It is interesting to note in these permutations that all true outputs (logic 1s) can be grouped in one or some of the tables, therefore, in the filtered results it will be observed that all 'elusive set' were eliminated.

Fig. 8 Permutations - example 2.

In Fig. 9 the 'local results' were grouped into the 'global results' array and sorted by placing the 'elusive sets' at the end (criterion presented in Fig 4), those being the last six rows of the table before the purification stage.

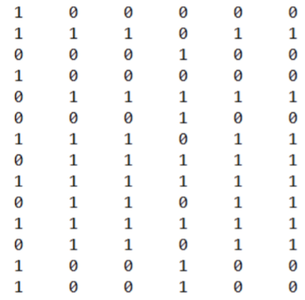


Fig. 9 Global Results - example 2.

The first filtering step presented in Fig. 10a eliminates the repeated combinations between rows moving from fourteen combinations to only five, of which three remain to be 'elusive sets', then in Fig. 10b the results of eliminating 'ghost sets' are presented, only two final combinations were conserved, and all remaining elusive sets were eliminated.

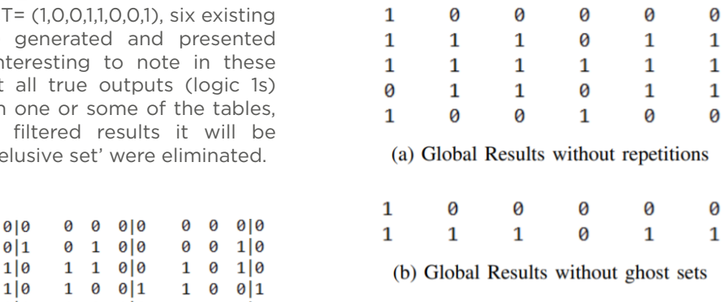


Fig. 10 Depurations results - example 2.

Restructuring the results into the Boolean algebra format yields to the function (~B·~C) + (B·C). Both presented examples were compared with the Boole-Deusto software, and the solutions achieved the same function, confirming the operation and effectiveness of the methodology.

C. Computational effort & qualification study

Table I compiles the average running time of the algorithm, measured for 2, 3, 4, 5, 6 and 7 bits. As can be seen, the running time increases exponentially with the increment of inputs bits 'n' due to the augment of tables (n!). This behaviour can be represented with (2) obtained through a regression method. With the equation, a simplified expression of the computational cost of the algorithm is achieved.

Bits	Time [s]
2	0.0828
3	0.1115
4	0.1804
5	1.1327
6	23.2123
7	1996.5976

Table I Computational time cost.

$$t = (9.8338294 \cdot 10^{-9}) \cdot e^{(3{,}7191 \cdot n)} \quad \textbf{(2)}$$

The qualification study yielded multiple truth tables sorted using Gray code (permutations), each one of them with unique 'true output' (logic 1's) clusters as shown with the continuous vertical lines in Fig. 11.

Fig. 11 Rating equation applied to 3 permutation tables.

From this clustering, (3) can be developed where: 'm' represents the number of groups found on each table and 'n' is the number of true values in each group, the result of such qualification is exemplified in Fig. 11

$$\alpha = \sum_{i=0}^{i=m} (3^{n_i}) \qquad \textbf{(3)}$$

The second permutation has higher simplification potential with an alpha value of 810; it should be noted that, in Fig 11, only 3 of the 24 permutations were used.

IV. DISCUSSION

In the case of the method effectiveness, some cases were detected where the degree of simplification obtained was not fully accomplished compared to other methods. For instance, in the 4-bit function previously presented (Fig. 4a), the solution with the proposed method deduced from the 'global combination' table was: (A·~ C·~D) + (A·C·D) + (~B·~C·D) + (~A·C·D) + (~A·~C·~D) while the equivalent solution obtained by Boole-Deusto was: (~C·~D) + (C·D) + (~B·~C). It can be deduced that the method yields accurate results but not as effective as other methods, so an improvement in the depuration stage is needed.

Regarding the performance of the method, the time used for the programmed algorithm to solve the functions is longer than the one needed by Boole-Deusto. However, it is considered that significant improvements can be achieved and these results can be reduced considerably by improving steps taken on the algorithm, e.g. analyzing only the truth table with the best perspectives of simplification as explained below. Also, (2) and the results of the datasets were compared to well-known growth rate models and datasets presented by [13], this comparison confirms that the proposed model behaviour should be similar to an exponential model.

The qualification procedure could allow reducing the computational time by discarding unnecessary analyzes on inconvenient permutations. During the various testing stages, a characteristic behaviour has been observed in the way in which the 'local results' are organized in the 'global results' array. Therefore, it is considered that by applying (3) to qualify the permutations would be possible to establish a combination order that allows a more profound simplification by eliminating redundant tables that do not contribute new information to the resolution. Nonetheless, the improvement of the output stage would be covered in further studies.

V. CONCLUSION

This paper proposes the so-called PGC method to find a propositional formula of a switching system problem by using Gray code principles and Boolean algebra. The main advantage of the proposed method is that it does not require in-depth knowledge of Boolean algebra, and unlike graphical methods, the outcome does not require visual inspection. Moreover, the

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

method is simple to implement and deploy in any programming tool since it does not require complex development techniques or advanced levels of analysis.

The proper operation of the method was demonstrated by comparing solutions obtained using the implementation described with manual methods and Boole-Deusto software. Although in some cases, the solution obtained did not represent the best possible result, an adequate degree of simplification was achieved, and all outputs obtained by the different methods are correspondingly equivalent. Furthermore, an equation that correlates the amount of time that the algorithm needs to solve a problem based on the number of logical inputs helps to estimate the computational time of the analyzed system before its deployment. Even though the computational time of the algorithm might be more significant than other methods, a possible step of the implementation has been identified as the future step of optimization for future developments of the PGC method.

Finally, the described method could also be extended to solve sequential logic problems, decision trees, route optimization, reduction of logic circuits or even for the academic purpose of using a flat interpretation of Hamiltonian hypercubes.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of M.Sc. Richard Andrade and M.Sc. Jose Beltran for their useful and valuable advice and mentorship. Their comments and suggestions during the performance of the project have helped the team to direct the focus of the analysis in the right track.

REFERENCES

[1] J. P. Roth, "Algebraic topological methods for the synthesis of switching systems. I.," Transactions of the American Mathematical Society 88.2, pp. 301-326, 1958.

[2] W.V. Quine "A way to simplify truth functions." The American mathematical monthly 62.9, pp. 627-631, 1955.

[3] M. Karnaugh, "The map method for synthesis of combinational logic circuits," Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics 72.5, pp. 593-599, 1963.

[4] R.B Hurley, "Probability maps," IEEE Transactions on Reliability 12.3 pp. 39-44, 1963.

[5] Astola, Jaakko, and R. Stankovic. “Fundamentals of Switching Theory and Logic Design A Hands-on Approach,” Springer, 2006.

[6] E.W. Veitch, "A chart method for simplifying truth functions," Proceedings of the ACM national meeting (Pittsburgh), 1952.

[7] Bitner, R James., Gideon Ehrlich, and E.M. Reingold. "Efficient generation of the binary reflected Gray code and its applications," Communications of the ACM 19.9, pp. 517-521, 1976.

[8] W. Press, et al. Numerical recipes in Fortran 77: volume 1, volume 1 of Fortran numerical recipes: the art of scientific computing. Cambridge university press. 1992.

[9] J. Zubia, J. García, Sanz Martínez, and S. Borja, "BOOLE-DEUSTO, la aplicación para sistemas digitales," 2001.

[10] J. Dybizbanski, and A. Szepietowski, "Hamiltonian paths in hypercubes with local traps," Information Sciences 375, pp. 258-270, 2007.

[11] K. Sankar, V. Jaya, M. Pandharipande, and P. S. Moharir. "Generalized gray codes," Proceedings of 2004 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS IEEE. 2004.

[12] D. Benavides, “Diseño, implementación y evaluación de unidades didácticas de matemáticas en MAD 2,” pp. 265. 2019.

[13] L. Egghe and I. Ravichandra Rao, "Classification of growth models based on growth rates and its applications," Scientometrics 25.1, pp. 5-46, 1992.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

AUTHORS

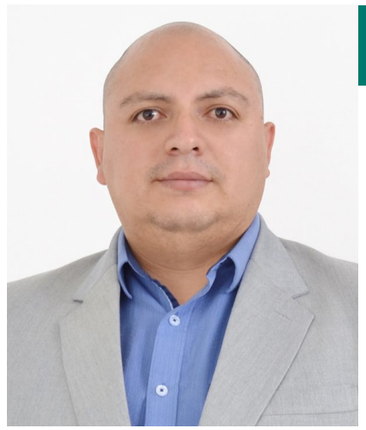
César Troya-Sherdek

César David Troya Sherdek, Marketing / Sales operations / Business Intelligence Analyst in General Motors Ecuador, Associate professor at ITK Instituto Tecnológico Kachary for the electronics department, Cum Laude in Mechatronic Engineering from the International University of Ecuador, MBAc from ADEN University, has worked in research groups in the aeronautics area, aerospace, mathematics and computing, also lectured on data science and artificial intelligence.

Valentin Salgado-Fuentes

Valentin Salgado Fuentes- Born in Quito - Ecuador in 1991 and Graduated from the SEK International University of Ecuador in 2014 as a Mechanical Engineer with a specialization in Energy and Control processes. In 2016 he moved to Denmark to study a Master degree in Engineering Design and Applied Mechanics at the Technical University of Denmark (DTU). Since 2018, he is a PhD student at the Section of Thermal Energy at DTU developing advance numerical models of complex thermal systems.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021



Jaime Molina

Jaime Vinicio Molina Osejos, Professor at SEK International University Auxiliary Investigator by the Senescyt, Master in Design, Production and Industrial Automation. Leader of the Laboratory of Metallography and Industrial, Coordinator of the Master's Degree in Mechanical Design, Manufacturing of Vehicle Autoparts (2016 - 2018). Coordinator of the Careers of Mechanical Engineering in: Design and Materials (2012 - 2014), Member of the Research Committee of the UISEK (2011) Professor since 2010 of in SEK International University. and Kachariy Higher Technological Institute

Gustavo Moreno

Gustavo Adolfo Moreno Jimenez is Professor at ITK Instituto Tecnologico Kachary director of Electronics area. He received his Master of Science in Technology Management from Marshall University (United States), his Master in Pedagogy and University Management from SEK University (Chile), and his Bachelor in Science at Electronic Engineering from ESPE University (Ecuador). He is a Senescyt certified Investigator, winner of “Ideas Bank” Senescyt Award in 2015, and winner of “Teaching Best Practices” SEK University Award in 2017.

LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC), Vol VIII, Issue 1, January 2021

Vehículo Eléctrico con Algoritmo de Control de Velocidad y Freno Regenerativo y Diseño de una Aplicación Web Móvil Basada en IoT

Electric Vehicle with Speed Control Algorithm and Regenerative Braking and the Design of a Mobile Web App based on IoT

ARTICLE HISTORY

Received 13 October 2020
Accepted 02 November 2020

Alex Pulamarin
Department of Electrical and Electronic
National Polytechnic University
Quito, Ecuador
alexdiaz11_corp@hotmail.com

Vehículo Eléctrico con Algoritmo de Control de Velocidad y Freno Regenerativo y Diseño de una Aplicación Web Móvil Basada en IoT

Electric Vehicle with Speed Control Algorithm and Regenerative Braking and the Design of a Mobile Web App based on IoT

Alex Pulamarin
Department of Electrical
and Electronic
National Polytechnic
University
Quito, Ecuador
alexdiaz11_corp@hotmail.com

Resumen— Actualmente, la mayoría de los vehículos eléctricos para la movilidad personal no mantienen una velocidad constante y no frenan automáticamente cuando el sistema lo requiere. Además, esta energía de frenado no se utiliza para alimentar el propio sistema. Para superar estos problemas, este trabajo presenta la implementación de algoritmos de control y una aplicación web para un vehículo eléctrico de movilidad personal. Se implementa un algoritmo de control en cascada donde el lazo interno es el par y el lazo externo es la velocidad. Además, se diseñó una aplicación web móvil basada en el internet de las cosas (IoT) para mostrar en tiempo real, información importante sobre el estado del sistema. Finalmente, se realizaron simulaciones y pruebas reales que indicaron una respuesta rápida de par y velocidad cuando el sistema está sujeto a diferentes escenarios de movilidad.

Palabras clave— *controlador PI, motor BLDC, algoritmo de velocidad, control adaptativo de referencia de modelo (MRAC), frenado regenerativo, base de fuego, interfaz de programación de aplicaciones (API).*

Abstract— Currently, most electric vehicles for personal mobility do not keep a constant speed and they do not brake automatically when the system requires. Moreover, this braking energy is not used to power the system itself. To overcome these problems, this work presents the implementation of control algorithms and a web app for an electric vehicle for personal mobility. A cascade control algorithm is implemented where the internal loop is torque and the external loop is speed. Additionally, a mobile web app based on IoT was designed to display in real time some important information about the system status. Eventually, simulations and real test were performed indicating fast torque and speed response when the system is subjected to different mobility scenarios.

Keywords— *PI controller, BLDC motor, speed algorithm, model reference adaptive control (MRAC), regenerative braking, firebase, application programming interface (API).*

I. INTRODUCTION

Electric vehicles are having great impact on society because they are more efficient than fossil fuel vehicles. Furthermore, electric mobility is sustainable, innovative and it generates no emission of greenhouse gases which directly benefits the environment [16].

This project focuses on an electric vehicle for personal mobility that will help to solve environmental problems and decrease traffic rush in big cities, as well as reducing destination time and road space. Moreover, this project presents solutions to improve the use of battery time and brakes without any mechanical devices. These advantages save cost and physical space. Additionally, a mobile web app based on IoT has been designed to indicate important variable information about the vehicle status [1].

To keep a constant speed and to brake automatically in electric vehicles this work focuses on the control algorithm design and the design of a web app based on IoT technologies. Eventually, tests are carried out, and results are analysed indicating energy efficiency and performance.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

II. BACKGROUND AND RELATED WORK

Each year there are innovative steps in the electric design of vehicles, where new technologies and technics in software and hardware are included in their manufacturing [11][12]. This research will be presenting important features in the software design of an electric vehicle for personal mobility.

Current electric vehicles for personal mobility do not maintain a fixed speed while riding. Therefore, it reduces their performance in speed and torque, when they are expose too many disturbances due to the unexpected road characteristics, and load [13]. One of the advances implemented is that the system brakes electrically and automatically without any mechanical component and without any frictional waste of energy. Instead, this energy from braking is tapping to power the entire system and recharge the battery pack. Furthermore, there are benefits such as saving space due to the nonuse of mechanical brakes and gears required to stop the vehicle.

To complement the system, a mobile web app has been designed in which important variables can be visualized. This information is crucial to optimize the life spam from the system. As a result, the user can monitor speed, temperature, battery level and other parameters. Besides that, the system can be locked and unlocked from the web for more security. Personal electric vehicles do not have these features like in the case of scooters, unicycles, skateboards, among others [14][15]. This research will start first with an analysis of a permanent brushless DC motor or simply BLDC motor in order to design the corresponding algorithms that will be executed on a chip to control this machine properly. These algorithms are fundamental to process and control variables such as current and speed. On the other side, these signals of current and speed have noise, thus both need to be filtered by a Butterworth filter. This filter improves substantially the output signal reducing the margin error less than 2%. Before the speed is processed, it is important to determine first the high and low states from the speed sensor. This information is used to determine the signal period and thus the synchronous speed.

In regard to the control algorithms, proportional and integral algorithms (PI) are used for both torque and speed. The current controller algorithm is based on the magnitude optimum criterion where the entire closed transfer functions become approximately one [4]. Then, the speed controller utilizes the criteria of the cancellation of poles and zeros to determine its gains. When the vehicle is subjected to many disturbances, the system slows down in speed response. Therefore, a type of a model reference adaptive algorithm (MRAC) is implemented in the speed loop, which updates its gains every 10 rpm.

In the interest of having more comfort when riding the vehicle, a web application based on the internet of things (IoT) has been developed. This app displays useful information such as speed, distance travelled, electronic board temperature, and power level. These data are first transferred from the chip to the firebase and then to the web app in real time.

III. CHARACTERISTICS OF A BLDC MOTOR

The principal part that provides force to the electric vehicle is a BLDC motor as shown in the Fig. 1. The principal parts of this motor are the stator and the rotor. The stator contains the coils and the rotor the permanent magnets.

Electrical and mechanical parameters of the BLDC motor are illustrated in Table I [1], which will be used to later to determine the controller parameters.

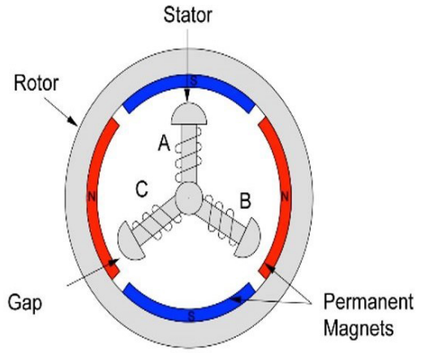


Fig. 1. Three-phase BLDC motor.

Table I. Electrical and mechanical parameters[1].

R	Phase AB resistance	0.5 ohms
L	Phase AB inductance	0.4 mH
τ_e	Electric time constant	0.0008 s
r	Wheel radius	0.21 m
T_{PWM}	PWM Time	$6.41 * 10^{-5} s$
τ_{mech}	Mechanical time constant	0.0678
B	Coefficient of viscosity	0.02336 Nms
J	Inertia	0.1375 Kgm^2
K_E	Torque Constant	1 Nm/A

IV. CURRENT AND SPEED MEASURMENT

A.Current Measurement

Three Hall Effect sensors of 0-200A input and 0-50mA output are connected per phase as seen in Fig. 2. These current signals are conditioned by a differential amplifier circuit and then processed, filtered, and analysed by a microcontroller that gives feedback to the control current closed loop.

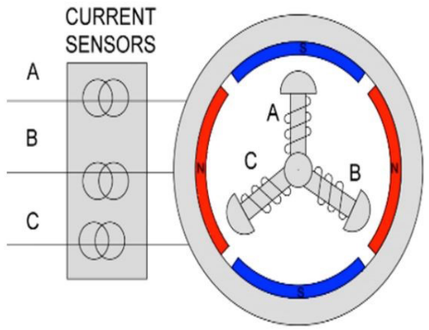


Fig. 2. Current sensors- connection per phase.

To have a better current signal a Butterworth filter is used, which eliminates noise produced by electromagnetic interference. In (1), u(s) is the filtered current signal and e(s) is the current signal nonfiltered.

$$H(s) = \frac{u(s)}{e(s)} = \frac{1}{(s+1)(s^2+s+a)} \quad (1)$$

Equation (2) is in the frequency domain and it cannot be coded in a microcontroller therefore a differential equation from (6) is implemented as follows:

$$u(n) = Te[n-3] + u[n-1] - u[n-2] \quad (2)$$
$$+u[n-3] - Tu[n-3]$$

B. Speed Measurement

To measure speed, hall effect sensors are embedded in the wheel and distributed in the coils 120 electrical degrees. These sensors can be high or low state (see Fig. 3).

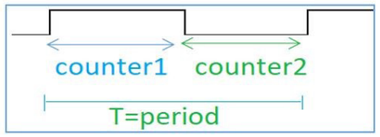


Fig. 3. Hall effect sensors high and low state signal.

In the microcontroller a timer is set to compute both high and low state to finally add up and obtain the signal period(T). Now, the synchronous speed can be easily calculated from (3) where p is the number of poles (p=56) and T is the period.

$$\Omega = \frac{120}{p*T} \quad (3)$$

C. Temperature Measurement

The electronic board has three temperature sensors, which are located in each branch of the inverter. They alert the rider of overload and overcurrents, avoiding serious hardware damage. To calculate this temperature, it is known from the sensor datasheet that 1^o C/10mV, the supply voltage sensor is 3.3V and the microcontroller is 10bit-ADC(Analogue Digital Converter 1024) as indicated in (4).

$$Temperature = ADC * \frac{3.3}{1024} * 100 \quad (4)$$

V. CONTROLLERS ALGORITHM

To design the current and speed controllers a PI (proportional and integral) algorithm has been implemented. A PI algorithm has been chosen because the proportional response changes the output for a given change in the error and the integral response eliminates the steady state error that happens due to the proportional algorithm [10].

A.Current Controller Algorithm

Due to the electromagnetic torque is directly proportional to the torque constant and current, a current algorithm is implemented to control torque [2][3]. This algorithm is proportional and integral (PI) as indicated in (5), where k_{pi} is the proportional gain, and τ_i

$$PI(s) = k_{pi}\left(\frac{1+\tau_i s}{\tau_i s}\right) \quad (5)$$

Fig. 4 illustrates a closed loop control. It is made up of a PI controller and the power electronics G_p and the BLDC motor transfer function G_M[5].

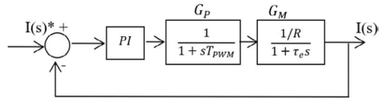


Fig. 4. Closed loop current.

Thus, the open-loop transfer function is:

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

$$G_{iaT} = k_{pi}\frac{1+\tau_i s}{\tau_i s}\frac{1/R}{1+\tau_e s}\frac{1}{1+sT_{PWM}} \quad (6)$$

Based on (6) and on the magnitude optimum criterion k_{pi} and τ_i are expressed as [4]:

$$k_{pi} = \frac{\tau_e R}{2T_{PWM}} = 3.12 \quad (7)$$

$$\tau_i = \tau_e = 8 * 10^{-4} \quad (8)$$

Fig. 5 is the closed loop current subroutine where errors are compute based on the current set point and the machine coil current then a control action is taken to generate a pulse width modulation or PWM signal.

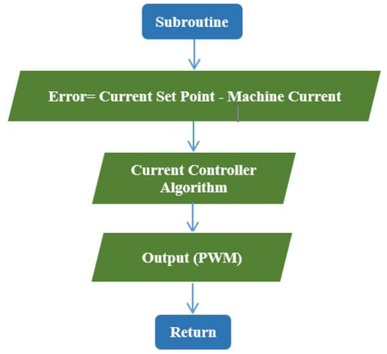


Fig. 5. Closed loop current subroutine.

B. Speed Controller Algorithm

The closed loop speed is shown in Fig. 6. It contains a PI algorithm, K_T is the torque constant and the mechanical transfer function $G\Omega$.

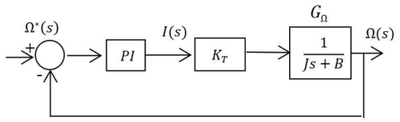


Fig. 6. Closed-loop speed.

The PI control algorithm is expressed as:

$$PI = k_{p\Omega}\frac{1+\tau_\Omega s}{\tau_\Omega s} \quad (9)$$

The open-loop transfer function from Fig. 6 is:

$$G_{\Omega a}(s) = k_{p\Omega}\frac{1+\tau_\Omega s}{\tau_\Omega s} * \frac{1/B}{1+\frac{J}{B}s} \quad (10)$$

Constants $k_p\Omega$ and $\tau\Omega$ are determined by the cancellation of poles and zeros and using Table I as follows:

$$\tau_\Omega = \frac{J}{B} = 5.88 \quad (11)$$

$$k_{p\Omega} = \tau_\Omega B = 0.137 \quad (12)$$

Fig. 7 is the closed loop speed subroutine where errors are calculated based on the speed set point and the machine mechanical speed. Then a control action is run, which generates an output equal to the current set point.

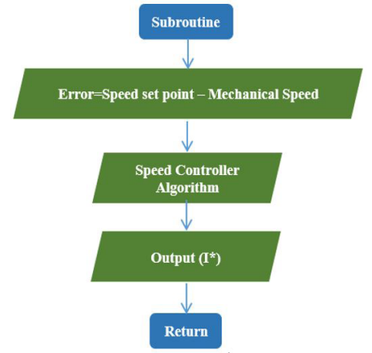


Fig. 7. Closed loop speed subroutine.

These constants are used to simulate the speed response in closed loop. In Fig. 8, the overshoot is 22% and the settling time is 0.422.

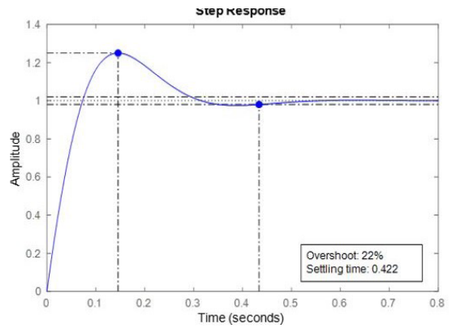


Fig. 8. Closed loop speed response simulation.

C. Adaptive Speed Control Algorithm

To improve the speed response a type of a model reference adaptive algorithm (MRAC) has been design (See Fig.9) [6].

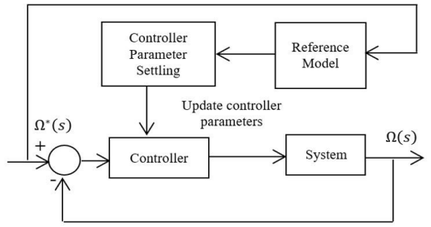


Fig. 9. Type of MRAC.

To implement this controller, speed is divided into intervals of 10 rpm and then speed constants change during this interval as indicated in Table II. $k_{p\Omega}$ and $k_{i\Omega}$ are determined using (10) and the root locus when K constant moves to the left [See Fig. 10(a)] [5].

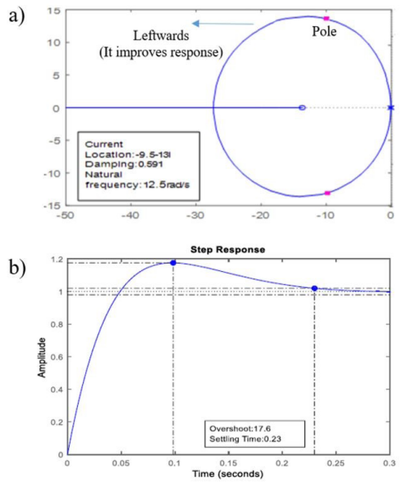


Fig. 10. a) Root Locus with K=1 b) Speed Response MRAC simulations.

In Fig. 10(b), there are improvements in speed response because the overshoot and the settling time go down to 17% and 0.23. Thus, speed response becomes faster.

Table II $k_p\Omega$ And $k_i\Omega$Speed Constants

Speed (rpm)	$k_{p\Omega}$	τ_Ω
0 a 10	0.137	5.88
10 a 20	0.28368	4.95
20 a 30	0.32256	4.23
30 a 40	0.40188	3.49
40 a 50	0.41472	3.30
50 a 60	0.4728	2.96
60 a 70	0.50688	2.69
70 a 80	0.56736	2.47
80 a 90	0.576	2.37
90 a 100	0.61464	2.28
100 a 110	0.64512	2.11
110 a 120	0.68556	2.04
120 a 130	0.6912	1.980
130 a 140	0.75648	1.85
140 a 150	0.8064	1.69

D. Difference Equations Algorithm

To implement the PI control algorithm into the microcontroller, it is necessary to transform from the frequency domain to Z domain and then into a difference equation. To discretize the controller, forward Euler method is used as shown in (13), where T is the sample time [7].

$$S \rightarrow \frac{z-1}{T} \qquad (13)$$

Thus, PI(z)=u(z)/e(z) where u(z) is the output and e(z) is the input or error:

$$u(z) = z^{-1}u(z) + k_{pi}e(z) \qquad (14)$$
$$+(\tau^{-1}T - kp_i)z^{-1}e(z)s$$

The inverse Z-transform is applied to (14) obtaining (15), which is programmed in the microcontroller.

$$u(n) = u(n-1) + kp_i\, e(n) \qquad (15)$$
$$+(\tau^{-1}\, T - kp_i\,)e(n-1)$$

VI. MOBILE WEB APP BASED ON IOT

The web app developed is based on the internet of things or IoT. It evolves multiple technologies such as, real-time analytics, sensors, wireless networks and embedded systems.

The web app components are classified in three main stages, microcontroller, database, and web development as shown in Fig. 11.

Fig. 11. Web app components.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

A. Microcontroller

The microcontroller board has got a Wi-Fi chip that allows connectivity to the internet and it is compatible with TCP / IP protocol. To set up the microcontroller, libraries are attached to receive and send data to the Firebase. In addition, an application programming interface or API key is necessary to exchange data. These data are speed, distance traveled, battery level and temperature.

B.Firebase Realtime Database

The firebase real time database is a cloud-hosted database. Data is stored and synchronized in real time to the microcontroller.

The firebase real time database allows secure access to the database directly from the microcontroller. Data is persisted locally, and when offline, real time events continue. When the microcontroller is connected, the real time database synchronizes the local data changes with the remote updates that occurred while the microcontroller was offline, solving any trouble automatically [8]

C. Web Design

The web app was design by using Hypertext Markup Language (HTML) is the standard markup language for documents designed to be displayed in a web browser. It displayed images and other objects such as interactive forms that may be embedded into the page. HTML is assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript. CSS brings style to the sheet by giving color, layout, and fonts. Java script instead manipulates data from the page itself and from the firebase. Java establishes connectivity with and the firebase through an API key along with other parameters such as domain, uniform resource locator (URL) and an identifier (ID). To run API in Java, three main components are required Java compiler, Java virtual machine and Java API as illustrated in Fig. 12 [9].

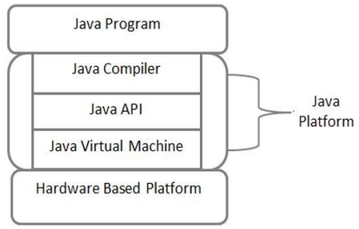


Fig. 12. Basic Java structure.

VII. RESULTS

The vehicle was tested under different conditions of speed, current and disturbances and which results are presented below.

A. Torque Controller Response

Tests were performed for a reference value of 15Nm torque, which results are presented in Fig. 13, presenting an overshoot of Mp<18% and an steady state of error<3%.

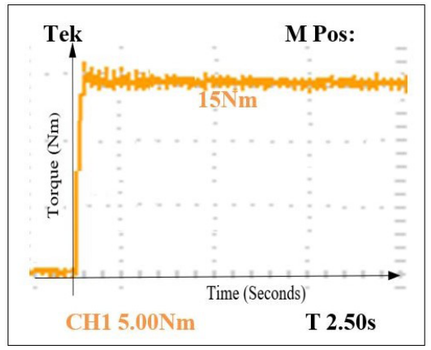


Fig. 13. Current controller response in closed loop for 15 Nm.

B. Speed controller response

Tests were performed for speed references of149 rpm and the results obtained are shown in Fig. 14 with an overshoot Mp<20% and an error<4%.

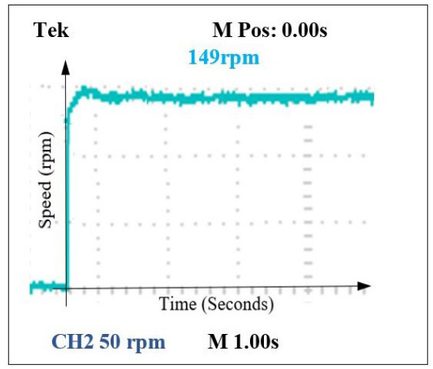


Fig. 14. Speed controller response in closed loop for149 rpm.

Under disturbances, the controller also responds property and which its results are shown in Fig. 15 for a reference speed of 70rpm.

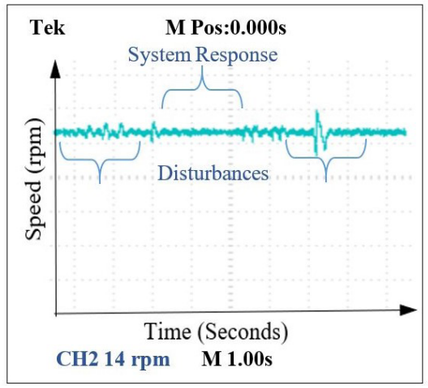


Fig. 15. Speed controller in closed loop for 70 rpm under disturbances.

C. System Response in Regenerative Braking Mode

The system is ridden in different scenarios where A is flat terrain and B, C are descending slopes as indicated in Fig. 16.

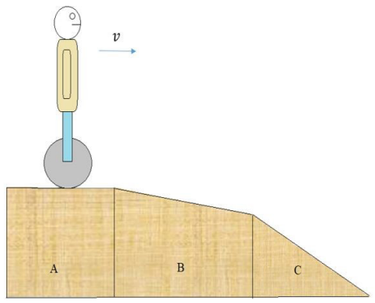


Fig. 16. Mobility scenarios.

In Fig. 17 can be observed that when the vehicle is ridden on flat terrain the torque positive , therefore there is energy consumption but when the vehicle is ridden on descending slopes the torque is negative therefore there is energy regeneration. This energy can be used to supply power to the electronic board and recharge the pack of batteries.

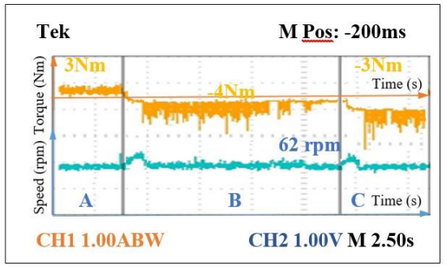


Fig. 17. Energy regeneration test to 62 rpm.

D. Mobile Web App

Fig. 18 is the web app design which indicates speed in Km/h, distance travelled , battery level, MOSFET temperature, the led can change from red to green indicating whether the vehicle saves energy or consumes energy and also allows us the option to lock the vehicle.

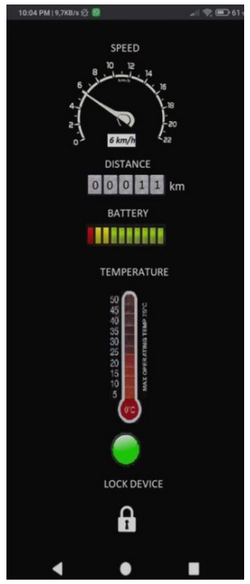


Fig. 18. Mobile web app.

VIII. CONCLUSION AND RECOMMENDATIONS

The software for an electric vehicle for personal mobility was designed and implemented with torque, speed, and energy regeneration.

Additionally, tests were performed under minimum, normal, and maximum operating conditions demonstrating a good system performance. The torque response has an overshoot less than 18% and a steady state error less than 3% while the speed response has an overshoot less than 20% and an error less than 4%, thus, meeting the design criteria. Under disturbances the system reacts fast and smoothly. On the other hand, the mobile web app becomes handy when riding because it presents valuable information of the vehicle status. Additionally, it helps to increase the lifetime of the vehicle when it warns high temperature values.

In regard to the regenerative braking, the energy produced can supply energy to the electronic board. Moreover, it can recharge the battery pack y some time intervals. In this stage, the time of the battery can last

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

for 20 to 30 minutes and it can regenerate an average energy of 35 Wh. Furthermore, when the system is in regenerative braking mode, it brakes automatically, keeping a fixed speed and eliminating mechanical brakes and frictional waste, which are great economic and space saving benefits.

Other control algorithms can be implemented to improve speed response and energy efficiency. In regard to the web app parameters such as speed record, graphs and GPS can incorporated.

REFERENCES

[1] P. Alex, "Design and implementation of speed and torque control with regenerative braking for a platform prototype of an electric unicycle," February 2019.
[2] T. Miller, "Brushless permanent magnet and reluctance Motor Drives," Oxford University, USA, 1989.
[3] Mohan Ned, "Electric drives, an integrate approach," USA, MNPERE Minneapolis, 2001.
[4] G. Papadopoulos, "PID controller tuning using the magnitude optimun criterion," Springer, Suiza, 2015.
[5] K. Ogata, "Modern Control Engineering," PEARSON, 5th Ed., Madrid, 2010.
[6] Embetion, "Adaptive algorithms," [Online]. Available: https://www.embention.com [Accessed: Sep.11,2020].
[7] MIT, "Discrete Approximation of Continuous Time Systems," [Online]. Available: https://ocw.mit.edu/terms/ [Accessed: Sep.11,2020]
[8] Firebase, "Firebase Realtime Database," [Online]. Available: https://firebase.google.com/docs/database [Accessed: Nov.05, 2020]
[9] Firebase, "App development platform," [Online] Available: https://firebase.google.com/docs/database [Accessed: Nov.06, 2020]
[10] C. Kuo "Automatic control systems," PRENTICE HALL, 7th Ed, New York, 2010
[11] D. Stefano, C. Pablo, S. Aldo, G. Patrik, P. Pietro and V. Fabio " Torque-fill control and energy management for a four-wheel-drive electric vehicle layout With two-speed transmissions," " IEEE Transactions on Industry Applications, October 2016.
[12] M. Arash, S. Aldo, G. Patrick, F. Saber and S. Jasper "A fast and parametric torque distribution strategy for four-wheel-drive energy-efficient electric vehicles" IEEE Transactions on Industry Applications," IEEE Transactions on Industrial Electronics, July 2016.
[13] R. Desna, H. Eko, S. Randen and D. Dadet "PENS-Wheel (One-Wheeled Self Balancing Vehicle) Balancing Control using PID Controller," Conference International Electronics Symposium 2016.
[14] R. Jiageng Ruan and S. Qiang "A Novel Dual-Motor Two-Speed Direct Drive Battery Electric Vehicle Drivetrain," IEEE Access, April 2019.
[15] R. Bakhtiar, H.Eko, S. Raden and P. Dadet "PENS-Wheel (Self Balancing One-Wheel Vehicle) Mechanical Design and Sensor System," Conference International Electronics Symposium 2016.
[16] NewScientist, "Electric cars really are a greener option than fossil fuel vehicles," [Online]. Available: www.newscientist.com [Accessed: Nov. 11, 2020]

AUTHOR

Alex Pulamarin

Alex Pulamarin he was born in Cayambe-Ecuador. He carried out his studies at National Polytechnic University. His major is in Electronic and Control Engineering. He has participated in several research projects in electric vehicles in the department of power electronics at National Polytechnic University and at Massachusetts Institute of Technology MIT. Areas of interest: Electrical Machines, Power Electronics, Embedded Systems, Instrumentation, and Industrial Control.

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

LATIN-AMERICAN JOURNAL OF COMPUTING

Published by

Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Quito-Ecuador

https://lajc.epn.edu.ec/
lajc@epn.edu.ec

January 2021

ESCUELA POLITÉCNICA NACIONAL

FIS

CEDIA

ISSN:1390-9266 e-ISSN:1390-9134 LAJC 2021

Vol VIII, Issue 1, January 2021

https://www.epn.edu.ec

QUITO - ECUADOR

LAJC

LATIN-AMERICAN JOURNAL OF COMPUTING

Vol VIII, Issue 1, January 2021

latindex catálogo 2.0

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

CiteFactor Academic Scientific Journals