MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media

MediaTIC: Una plataforma analítica de medios digitales costarricenses en redes sociales

Cristina Soto-Rojas[1], Carlos Gamboa-Venegas[2], Adriana Céspedes-Vindas[3]

Soto-Rojas, C; Gamboa-Venegas, C; Céspedes-Vindas, A. MediaTIC: A Social Media Analytics Framework For the Costa Rican News Media. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 18-24.

https://doi.org/10.18845/tm.v33i5.5070

1 University of Costa Rica (UCR), National Center of High Technology (CeNAT).
2 National Center of High Technology (CeNAT). Email: cgamboa@cenat.ac.cr.
3 Universidad Estatal a Distancia (UNED). Email: acespedesv@uned.ac.cr.

Keywords

Social Media; Big Data; MongoDB; Spark; R; Facebook.

Abstract

Social media sites such as Facebook are tools that democratize contents. These tools facilitate the creation of news and ideas. In Costa Rica, we designed MediaTIC framework to track social media outlets communities by accessing their posts and user interactions. In order to do this, we created a big data store and designed a software architecture that uses high computing hardware (cluster) to manage software escalation, data volume, and future requirements. We used MongoDB, Spark and R technologies to build the platform and manage the information. This platform, allowed us to run faster queries, improving the performance time in 110%-180% approximately. The main objective is to provide visualizations and information for Costa Rican social media analysts that can give value to Social Communication Science in Costa Rica by using high technology underneath.

Palabras clave

Redes sociales; Big Data; MongoDB; Spark; R; Facebook.

Resumen

Las redes sociales como Facebook son herramientas democratizadoras de contenido. Estas herramientas facilitan la creación y visualización de ideas y noticias, favoreciendo la creación de comunidades virtuales. En Costa Rica, diseñamos MediaTIC como una plataforma para examinar estas comunidades y su comportamiento utilizando la información de las publicaciones y las interacciones entre usuarios. Para lograr esto, implementamos un repositorio *big data*, el cual trabaja sobre una infraestructura de alto rendimiento (cluster computacional) lo que permite que esta plataforma sea un sistema escalable que puede trabajar con grandes volúmenes de datos. En la implementación de esta plataforma y el manejo de los datos se utilizaron MongoDB, Spark y R. Esta plataforma nos permite ejecutar consultas a mayor velocidad, mejorando el tiempo de respuesta en un 110%-180% aproximadamente. El objetivo principal es proveer visualizaciones y estadísticas que brinden información de valor a la investigación en Comunicación Social en Costa Rica.

Introduction

The usage of social media has increased the generation and collection of data. This data contains more information that can be analyzed and used to respond to different questions. This project focuses on social media outlets with facebook pages. We record posts, comments and interactions between users in these pages. This information is stored in a platform that will allow users to determine media behavior in Costa Rica during certain months, seasons or major events.

MediaTIC is a computational platform for the analysis and visualization of big data produced by the principal digital media outlets in Costa Rica on facebook. The main objective of the project is to develop a computer system that not only collect information but also applies information retrieval algorithms and social network analysis to visualize the most relevant information through a web interface. The working group is composed of professionals in the area of computing and communication from three departments: Laboratory of Research and Technological Innovation (LIIT) at UNED, Communication Research Center (CICOM) at UCR, and Advanced Computing Laboratory (CNCA) at CeNAT (National Center for High Technology).

Social Media Analytics Framework

There are many studies related to social media analytics, with data collection and analysis being some of the most important steps. Social media analytics framework is seen as a guide to find and resolve conflicts [1] and it helps to identify challenges and design solutions [1], [2], [3]. In Figure 1 we present an analog framework for our project.

First, consider the tracking step. The tracking is a fundamental step that explains the process to recollect the data [1]. In our case, the data was collected manually through Netvizz tool, which allows data to be downloaded two weeks after the date they were published. The selected media are: La Nación, CR Hoy, Telenoticias, Repretel, Semanario Universidad, El Financiero, Noticias Monumental, Prensa Libre, Diario La Extra and Amelia Rueda.

The tool provides two plain-text files, one that contains all the posts with the publication date, text, and reactions (like, love, sadness, wow, angry). The other file contains the comments for each post with the publication date, text, reactions and the order chain between the comments.

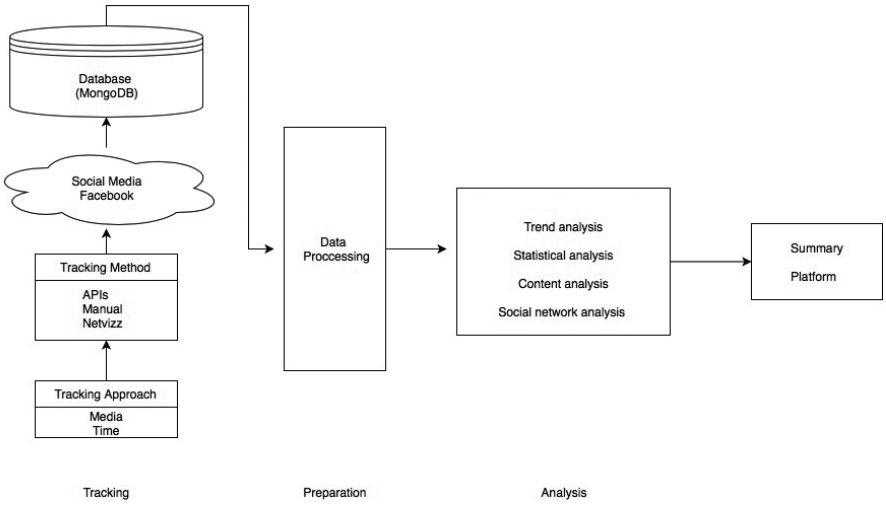

Figure 1. Social Media Analytic Framework Scheme

Preparation

Once we have the data collection step, the next step is the data storage [1]. We have selected MongoDB [4], a document-based database chosen in virtue of its facility to store data with different structures and its efficiency in queries. This database model makes possible to have different collections in the same storage space. We created a collection for posts and another collection for posts comments.

This Mongo database is located in the cluster Kabré. Hosted at CeNAT, Kabré is a multidisciplinary supercomputer, with 32 Intel KNL nodes dedicated to bioinformatics, simulation, and others. 4 k40 NVIDIA GPU nodes assigned to machine learning projects, and 3 nodes for big data applications. Each one of the big data nodes has 64 GB of memory and two Intel Xeon E5-2650 processors at 2.20 GHz, all together add 70 TB of storage and are connected through a 10 Gigabit Ethernet network which makes possible fast communication and low latency in data transmission [5][6].

To access the database a web platform was implemented, which we called “Mediatic-CMS” for Control Management System. The main goal of this platform is to provide a tool for queries and

the manual assignment of post categories (politics, sports, nationals, international). There is an automatic alternative to do this category assignment as an option. Users can utilize the platform to upload netviz data and enrich the collection. There are two main modules, one for posts, and one for comments. In the post section, users can look up for information using different criteria such as: media name, dates, category (main topic), tone, format, reporter, if it is a public or nonpublic affair, and keywords. In the comment section, users can filter information by media, dates, public or non-public news and keywords. In both sections, users have the option to download the information as Comma-Separated-Value files for complementary analysis.

Processing and analysis

The next step is analysis [1]. To compute the queries for the analysis we used Mongo Spark Connector [7], which allows us to optimize the queries as we can see in the comparative Table 1. Once the data had the structure as need it after the queries the next step is the visualization of these results.

As we can see in Table 1, the performance using mongodb and spark shows an increase in performance for Spark in 110%-180% depending on the query type. Spark offers a considerably shorter response time in this first phase of the project.

Table 1. Comparison time between MongoDB and Spark in the following operations: Aggregation, Sort and Group

Time in ms		
Query	MongoDB	Spark
1: Aggregation	339	3
2: Sort	332	3
3: Group	183	1

Data visualization

Data Visualization has been worked using R programming language. We are constructing an interactive site with Shiny that allows users to change the graphics in real-time. The type of graphics are word clouds, heat maps and time series. The user can filter by month, year and media.

Word Clouds of Post Contents

Word clouds are considered a text-mining technique. They work by highlighting the frequency of words in an input. This process requires clean data, that means that punctuation, numbers and special characters are removed. Words that match the stop-words language list (such as: articles, prepositions, and others) are also removed. Finally, the text, once clean, goes through a stemming-process that allow counting the words with the same root. The most frequent words are graphed. The bigger the word, the most frequent it is in the text.

In MediaTIC words clouds are used to visualize the most frequent words that a media outlet used during a specific period. For example, we can identify the topic of the month by the media. In figure 2, there is a word cloud sample with December's posts most frequent content.

Heat Map by Category

Heatmaps are a graphic tool to visualize clusters. In this kind of representations, numerical data is substituted by colors. The stronger the color, the bigger the number. The color scheme helps with the identification of those categories that are more prevalent or more subtle in the objective of analysis.

In MediaTIC heat maps are used to help with the determination of the most frequent categories (or topics) that each media published during a period. The most frequent the topic, the darker the color. For example, we can easily identify how popular a category like “Sports” or “International” is for media on its Facebook page during a specific month of the year. Figure 3 shows a heatmap of December 2018 categories post-publication by media.

Team Series by Media

Time series allows us to visualize the behavior of a variable in a period and is possible to analyze the tendencies of this variable. In MediaTIC, since we have a temporary variable, one of the considered analyzes consisted of visualizing the behavior of the media as a series of time. The behaviors considered were the engagement and the number of publications over the data from the year 2018 classified by media. These can be seen in Figure 4.

Figure 2. December, 2018. Word Cloud with data from La Nacion, CR Hoy, Semanario Universidad and Telenoticias.

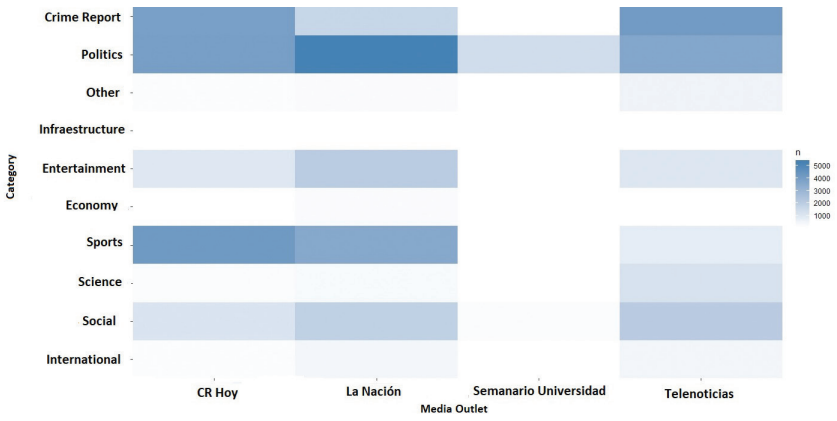

Figure 3. .December, 2018. Heat map by category with the number of news per outlet and category-subject.

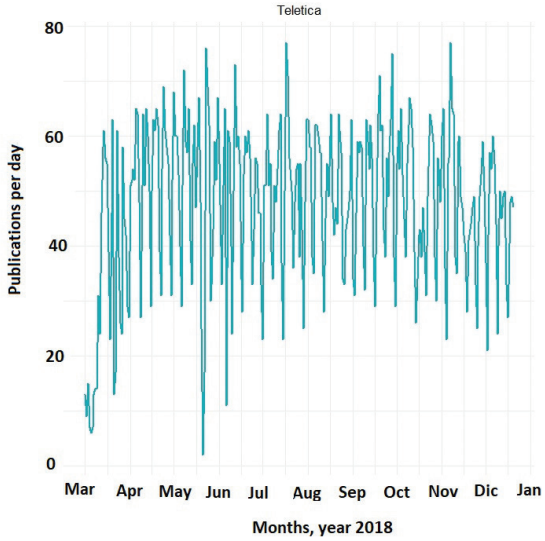

Figure 4. Time Series of Telenoticias, 2018 and the number of publications per day.

Conclusions and future work

Data provided by Netviz is enough to obtain a good variable set for analysis. Many permutations of data can be done to provide value to final users using advanced computing techniques.

Different visualizations are possible due to library packages that an environment like R provides. These visualizations enable diffusion and understanding of the data collected. High Computing infrastructure establishes a good set of resources that helps with the extraction and manipulation of big data with a very good performance.

Data retrieval is optimized by the use of the big data cluster infrastructure and MongoDB-Spark Connector. The time used to run the query is considerably better with spark. The performance is expected to be more notorious as the volume of data increases.

Since we have temporal and text variables, the analysis of temporal series and sentimental analysis are interesting future approaches applying studies from Rodrigues [8] and Nodarakis [9]. In temporal series, the next step is to analyze the stationarity and tendencies to make predictions. In sentimental analysis, since the language is Spanish there need to build the dictionary, and then proceed to analyze the posts and the comments of the gathered news.

At the current state of the project, we are working on the connection of our front-end (R-Shiny) and back-end (MongoDB) using Spark as a sort of middleware to allow us with the volume and processing time. It is important to consider that Netviz tool for data extraction from Facebook has the risk to stop working, and we need to accomplish the collecting step with other set of tools and sources different from facebook.

References

[1] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. “Social media analytics – Challenges in topic discovery, data collection, and data preparation”. *International Journal of Information Management*, Vol. 39, no. 2018, pp 156–168, April, 2018. https://doi.org/10.1016/j.ijinfomgt.2017.12.002.

[2] Labrinidis, A., Papakonstantinou, Y., Patel, J. M., & Ramakrishnan, R. “Big Data and Its Technical Challenges”. *Communications of the ACM*. Vol. 57, no 7, pp 86-94, July, 2014. https://doi.org/10.1145/2611567.

[3] Verma, J. P., Agrawal, S., Patel, B., & Patel, A. “Big Data Analytics: challenges and applications for text, audio, video, and social media data”. *International Journal on Soft Computing, Artificial Intelligence and Applications*, Vol 5, No 1, pp 41–51, February, 2016. https://doi.org/10.5121/ijdps.2017.8101.

[4] Chodorow, K and Dirolf, M. “MongoDB: The Definitive Guide”, O’Reilly Media, Inc. 2010

[5] Intel. Hadoop* Clusters Built on 10 Gigabit Ethernet, 2012, [Online]. Available: https://www.arista.com/assets/data/pdf/Whitepapers/Hadoop_WP_final.pdf

[6] Arista. 10 Gigabit Ethernet: Enabling Storage Networking for Big Data, 2016, [Online].Available:https://www.arista.com/assets/data/pdf/Whitepapers/AristaStorageNetworkingWhitepaper_v.6.0_GF.pdf

[7] Shoro, A. G., & Rahim, S. T. “Big Data Analysis: Ap Spark Perspective”. *Global Journal of Computer Science and Technology: C Software & Data Engineering*, Vol 15, No 1, 2015. Vol 15. No 1. pp 7-14. Retrieved from http://www.computerresearch.org/index.php/computer/article/viewFile/1137/1124.

[8] Rodrigues, A. P., Chiplunkar, N. N., & Rao, A. “Sentiment analysis of social media data using Big Data Processing Techniques”. *International Journal of Computer Applications*, Vol 22, No. 6, pp 56, 2016.

[9] Nodarakis, N., Tsakalidis, A., Sioutas, S., & Tzimas, G. (2016). “Large scale sentiment analysis on Twitter with Spark*”. CEUR Workshop Proceedings*, Bordeaux, France, 2016,Vol 1558.