Journal of Physics:
Conference Series

PURPOSE-LED PUBLISHING™

PAPER • **OPEN ACCESS**

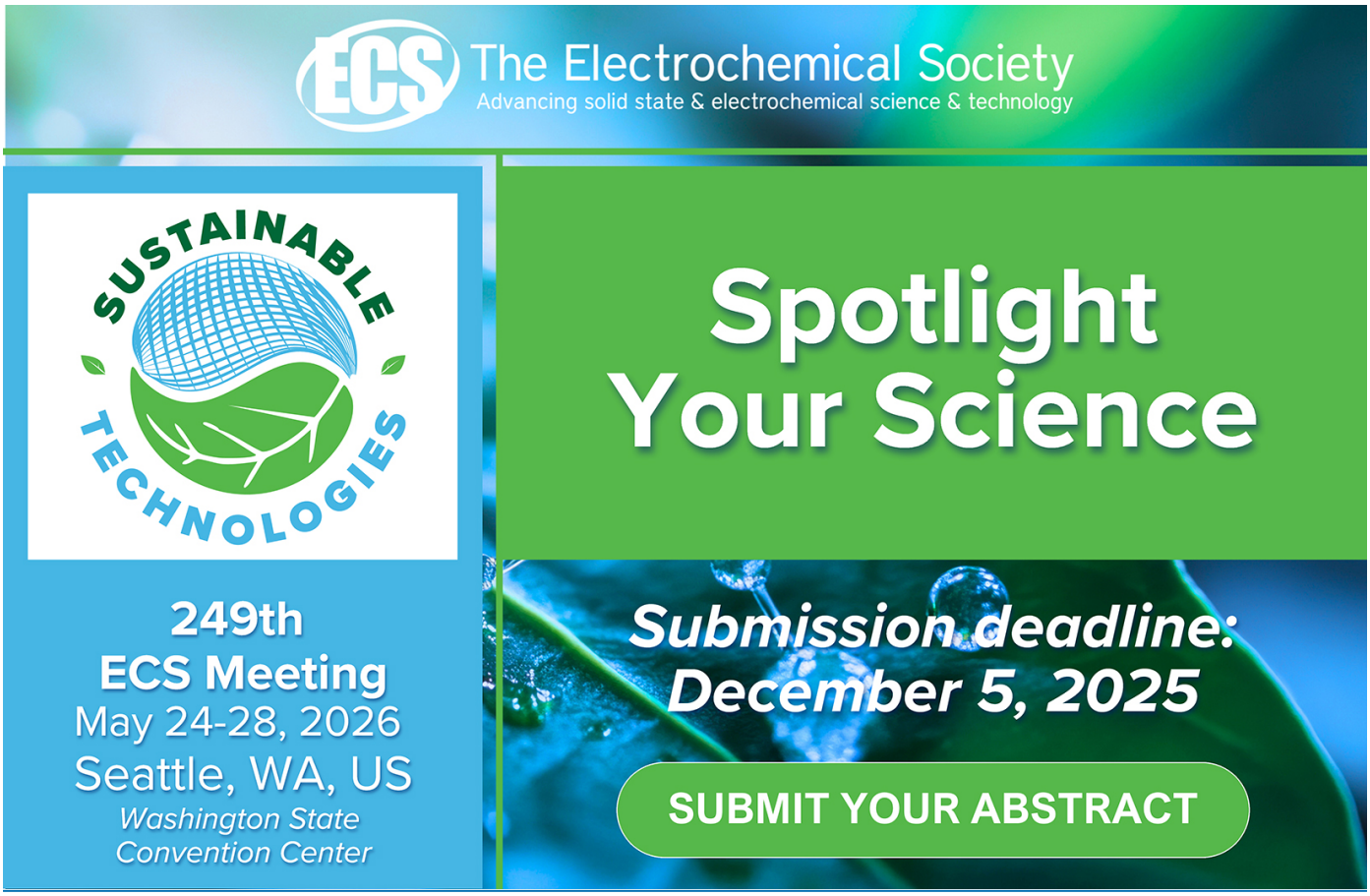
Application of data mining as a tool in computer science

To cite this article: T Velásquez-Pérez *et al* 2020 *J. Phys.: Conf. Ser.* **1587** 012018

View the article online for updates and enhancements.

You may also like

- Workshop on Intakes of Radionuclides: Occupational and Public Exposure, Avignon, 15-18 September 1997
G Etherington, A W Phipps, J D Harrison et al.
- Perspectives on biologically inspired design: introduction to the collected contributions
Jeannette Yen and Marc Weissburg
- Prototype virtual assistant as a tool to support research activities at CORPOSUCRE
C M Pineda-Pertuz, H J Romero, J A Vergara et al.



This content was downloaded from IP address 8.243.64.131 on 06/10/2025 at 20:58

Application of data mining as a tool in computer science

T Velásquez-Pérez[1], **J A Camargo-Pérez**[1], **and E L Quintero-Quintero**[1]
[1] Grupo de Investigación en Tecnología y Desarrollo en Ingenierías, Universidad Francisco de Paula Santander, Seccional Ocaña, Colombia

E-mail: tvelasquezp@ufpso.edu.co, jacamargop@ufpso.edu.co

Abstract. Computational science involves a systemic approach for adequate management of both information and knowledge; through theories of information and computation. Today trends such as big data and data mining emerge, with which behavior patterns previously only visible to the human eye can be established. This article, part of the analysis of big data, based on research found on the science websites, whose objective was based on knowing the conceptual evolution of Big data as well as the search for statistically regular behaviors over time in different elements related to the production and consumption of scientific information. For this study, data mining will be specified for Big data analysis, resulting in instructions to identify and understand trends and patterns within the different research areas.

1. Introduction

Business intelligence aims to analyze consolidated data in an environment called a data warehouse, understood as a unified repository for all the data collected by the various systems of a company, normally, it is hosted on a corporate server or increasingly, in cloud. Being a data warehouse architecture that enables business executives to organize, understand and use their data to make strategic decisions. Storage in big data is done by distributed files, allowing agile management in large volumes of data. Bearing in mind that the ultimate objective of big-data is to find the knowledge contained in the data and then apply it to improve any type of process, therefore big-data is not a substitute concept for the business intelligence concept, on the contrary, big-data is a tool that facilitates the development of business intelligence processes; therefore, it can be said that these are complementary since Big data offers deep analysis and a global vision of the data, Business Intelligence provides the user with a more structured experience [1].

Bibliometric indicators are defined as numerical data on the phenomena of scientific activity related to the production, transmission, and consumption of information [2]. These allow quantifying the number of publications, subjects in which it is published, languages and countries with more production, the volume of authorship, etc. They synthesize the bibliographic characteristics using a numerical value that compares with observations of other documentary sets and studies the evolution over time. The science, technology, and innovation indicators are contained in data sources, manuals and other indicators [3].

Bibliometric implies the statistical analysis on the descriptive information, contributions, and citation of the scientific productions of a bibliographic nature, or also from a thematic perspective, which allows detecting the research fronts from the topics treated there so that can get useful summary information to establish analytical relationships. This identification of topics can be done using the traditional content analysis method, generally used in qualitative studies on small numbers of documents, or using statistical

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

modeling, which allows to extract and analyze information from large text collections, applying unsupervised algorithms [4].

The bibliometric analysis carried out in the projects on scientific publications constitutes a very important link in the research, therefore it has become a tool that allows us to qualify the quality of the dynamics of knowledge generation and the impact of this process on the environment. This project has allowed monitoring of volume, evolution, visibility and knowledge structure around big data. In this way, scientific activity and the impact of both research and sources can be assessed. There are different techniques for analyzing scientific activities and one of them is bibliometric, which is a discipline for the quantitative evaluation, and in some qualitative way, of scientific production [5].

With the increasing advance in information technology, it brings the increase in the digitization of data, which has an impact in the same way in the studies that focus on evaluating tools and techniques of help in these subjects, that is why trends take importance as data mining, business intelligence, big data among others, significantly support decision-making with "business intelligence".

The rapid proliferation of information and communication technology has resulted in the rapid growth of digitized data and has also attracted considerable attention to research opportunities in big data analysis and business intelligence in management, social sciences, and humanity. The big data and analysis trend for business intelligence provides great resources and a powerful methodology to support the data-based decision-making process, which is the core of "business intelligence" [6].

Big data analysis associated with a database search, extraction and analysis can be seen as an innovative information technology (IT) capability that can improve the performance of the company. Although some leading companies are actively adopting big data analysis to strengthen market competition and open up new business opportunities, many companies are still in the initial stage of the adoption curve due to lack of understanding and experience with big data [7].

One of the problems faced by researchers in the technological area is associated with the proper management of information, this implies the incursion into trends of intelligent systems that use techniques or tools such as machine learning, web 3.0 or semantics, computational intelligence, the data mining used many times in social networks. There are frameworks or tools for massive data processing based on the MapReduce paradigm, as well as big data, Apache Hadoop, Spark, which has allowed data mining methods to be managed efficiently, as well as learning algorithms for machines in different domains [8]. In order to comply with the investigation, the following phases were carried out systematically.

1.1. Phase 1. Identification of the topic and tools

This phase starts from the need that exists to determine the evolution that big data has had in recent years, it is opted to use mining through the bibliometric for the analysis of metadata or set of data generated in specialized databases as is the science website.

1.2. Phase 2. Information search

In this phase, a search equation is defined that adjusts to the needs, making use of the search equations that the science website has integrated. Certain elements are delimited in it, which allowed obtaining better results in terms of reviewing the scientific production.

1.3. Phase 3. Debugging

In this phase, duplicate elements that were debugged through the Vosviewer platform are eliminated in order to obtain cleaner results.

1.4. Phase 4. Analysis

This phase refers to the analysis of the data sets through the Vosviewer tool, in which correlation and density maps were generated through the use of clustering or grouping using unsupervised learning techniques.

2. Methodology

The development of the present investigation was worked under a descriptive investigation with a quantitative approach, the descriptive investigation according to [8] is oriented to describe in detail and exhaustively the phenomena in one or more moments. For this investigation, the scientific database, the web of science, is used, with the results of the search between article titles, their abstracts and the keywords identified based on the combination of terms; Once the data set was obtained from the results export function offered by the science website in .txt format, which is a database format that supports the software used for the analysis, we proceeded to carry out the analysis of the base information obtained using Vosviewer bibliometric data analysis software using graphical representations. The data within the software was analyzed under unsupervised learning, which is based on searching groupings based on similarities, these analyzes are shown to the user through diagrams that allow further analysis within the object of study.

3. Results

The development of this project was based on an analysis of 305 articles, which corresponds to the most cited on the science website. 33 attributes and 305 instances were obtained. The method of consultation on the web of science was as follows: (ts = big data), English language, and types of documents: article or book, refined by Best Essential Science Indicators (ESI), articles: very cited in the field period: 2007-2019. The database used was the science website, once the results obtained were obtained, the files were exported to a .txt file which would allow it to be uploaded to the VOSviewer software for its respective analysis according to certain criteria selected in the tool, such as the level of co-occurrence, the number of citations per documents and country.

Figure 1 shows the number of publications that mention big data, it can be determined that as of 2009, research began to be done in the area, precisely in those years it was where social networks, business intelligence boomed, etc. Stated that in 2010, information was scarce, expensive and difficult to obtain, but right now, the collection and storage of massive data has been simplified to the point that the amount of information is doubled at least every two years [10].

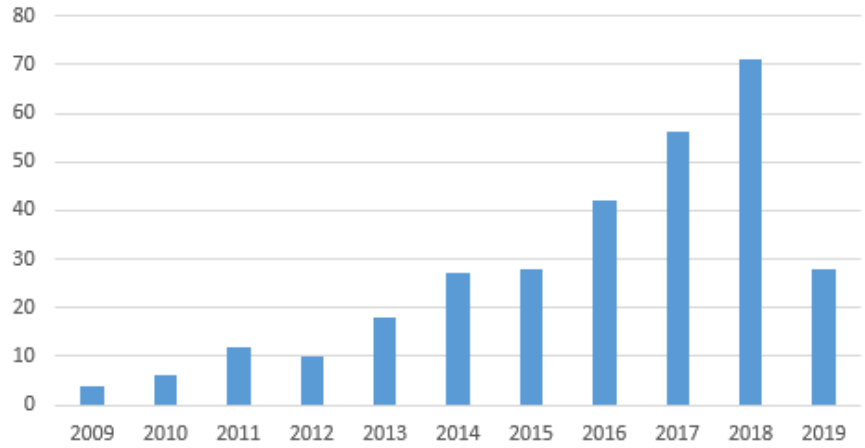


Figure 1. Number of publications per year [11].

3.1. Co-occurrence analysis

Distance-based maps are maps in which the distance between two elements reflects the strength of the relationship between the elements. A shorter distance generally indicates a stronger relationship. In many cases, the elements are distributed quite unevenly on distance-based maps [12].

Figure 2 shows the analysis regarding the level of co-occurrence on the 305 articles analyzed, the tool yielded 7 clusters, the approach between clusters indicates the level of closeness and co-occurrence that is the relationship between terms the functionality of VOSviewer is especially useful for easily displaying large bibliometric maps to interpret [13].

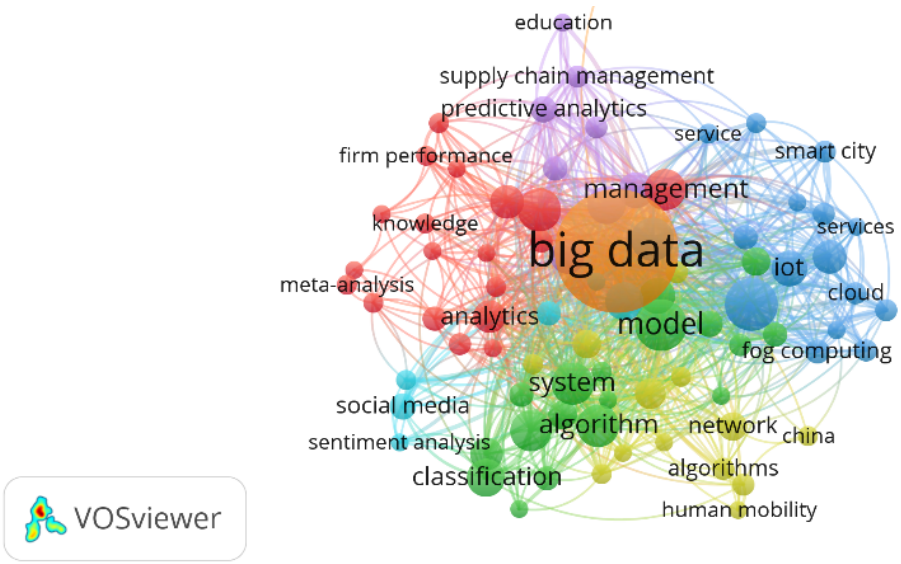


Figure 2. Co-occurrence analysis.

Figure 3 shows that although the main search was about big data, the figure shows us a fairly demarcated cluster in terms of internet of things (IoT) which is highlighted within a red circle , what we would call the centroid for cluster number 5, this leads us to think that the internet of things through the use of sensors and different electronic devices are generating large volumes of information that are being stored, other elements with considerable force were security and cloud computing, in terms of security, it is evident that this is an element inherent to the information since this is defined as the main asset, it turns out to be sensitive, therefore security has become a topic of continuous relevance, so in recent years the study of the theoretical bases and practical implementations to ensure confidentiality in information management. As a second element, we have cloud computing also known as cloud services, it is a paradigm that allows us to offer computing services through a network, which is usually the internet. Social networks such as Facebook, e-commerce and 5G networks are directly related to the handling of large volumes of information.

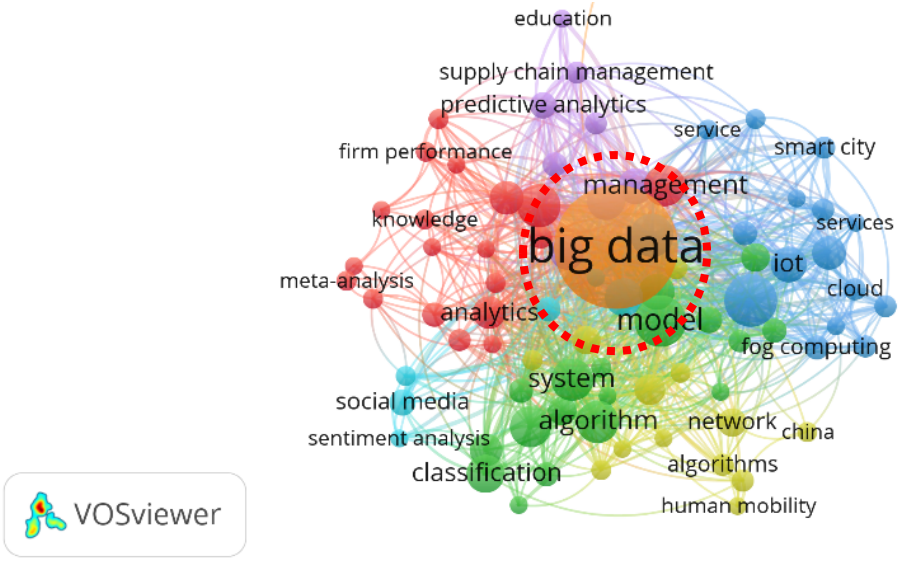


Figure 3. Density diagram.

3.2. Production analysis and citation of scientific production

An analysis was made on the level of citation of articles between different countries, it is evident that countries such as the People's Republic of China followed by the United States, England, Australia, Germany, Spain, and France, are the states that more research They have done about big data, it is to emphasize that these are developed countries and that invest in technological development.

Countries like the people's republic of China followed by the United States are the ones that are doing more research in areas such as big data, the internet of things and cloud computing, nothing strange is done since they are industrialized countries that are constantly pointing to the development of new technologies. Over the years, big data will take on more strength, the stored data will grow vertiginously and the internet of things will generate an exponential increase in the data collected in the databases, but not only that, the analysis also indicates that countries should focus on security as a fundamental aspect to maintain the principles of integrity, confidentiality, and availability of information, countries should be concerned with maintaining data protection since as well as there is information growth, growth will occur directly proportional on computer attacks.

The identification of the research lines and their interconnections, based on the information contained in the databases, is essential to understand the knowledge structure of a scientific domain. In this sense, the research lines are similar to the thematic groups that make up an area of knowledge [14].

4. Conclusions

The significance of this study lies in the realization of a first approach to the early history of big data as an emerging technology through the use of data mining techniques, the history can be described as early, due to the lack and recently documented the emergence of scientific studies in the field with just 10 years of having been mentioned and known as it is known today and having begun to study in-depth by researchers.

As it was found in the review carried out the 305 articles found on the science website and submitted on the Vosviewer platform as a whole, confirms the prevalence of the term big data, mainly due to the emergence of cloud computing, the flow of social networks and devices that store information through objects interconnected through the internet. This research allows us to understand the universe or the context in which big data has evolved, it becomes increasingly complex because it is related to many technologies, and in each of them special attention must be paid, especially on the subject of security.

References

[1] Gill H and Rao P 1996 *Data Warehousing: La Integración de Información para la Mejor Toma de Decisiones* (México: Prentice-Hall)
[2] Sotolongo G, María A, Guzmán Sánchez V 2001 *Revista Ciencias de la Información* **1** 27
[3] Schab E, Rivera R, Bracco L, Coto F, Cristaldo P, Ramos L, Rapesta N, Pablo Núñez J, Retamar S, Casanova C, De Battista A, Herrera N E 2018 Minería de datos y visualización de información *XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018)* (Argentina: RedUNCI)
[4] González C M, Varela S and Miguel S E 2018 Uso de R en bibliometría. Exploración de técnicas para la detección de temas locales de investigación *Conferencia Latinoamericana sobre el Uso de R en Investigación+Desarrollo (LatinR)* (CABA: Universidad de Palermo)
[5] Rosselli D 2019 *Acta Neurologica Colombiana* **35(1)** 1
[6] Liang T P, Liu Y H 2018 *Expert Systems with Applications* **111** 2
[7] Kwon O, Lee N, and Shin B 2014 *International Journal of Information Management* **34** 387
[8] Bello-Orgaz G, Jung J J, and Camacho D 2016 *Information Fusion* **28** 45
[9] Castaño Garrido C, and Quecedo Lecanda M 2002 *Revista de Psicodidáctica* **14** 5
[10] Cukier K, and Mayer-Shonberguer V 2012 *Big Data la Revolución de los Datos Masivos* (España: Houghton Mifflin Harcourt)
[11] Garfield E 2019 *Web of Science* (Boston: Clarivate Analytics)
[12] Van Eck N J, Waltman L 2010 Software survey: VOSviewer, a computer program for bibliometric mapping *Scientometrics* **84** 523
[13] Hu Y, Sun Z and Wu D 2019 *IOP Conference Series: Earth and Environmental Science* **300** 032098:1
[14] Gálvez C 2016 *Revista Española de Salud Publica* **90** 1