International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:11, No:10, 2017

Analysis of the Topics of Research of Brazilian Researchers Acting in the Areas of Engineering

Jether Gomes, Thiago M. R. Dias, Gray F. Moita

***Abstract*—**The production and publication of scientific works have increased significantly in the last years, being the Internet the main factor of access and diffusion of these. In view of this, researchers from several areas of knowledge have carried out several studies on scientific production data in order to analyze phenomena and trends about science. The understanding of how research has evolved can, for example, serve as a basis for building scientific policies for further advances in science and stimulating research groups to become more productive. In this context, the objective of this work is to analyze the main research topics investigated along the trajectory of the Brazilian science of researchers working in the areas of engineering, in order to map scientific knowledge and identify topics in highlights. To this end, studies are carried out on the frequency and relationship of the keywords of the set of scientific articles registered in the existing curricula in the Lattes Platform of each one of the selected researchers, counting with the aid of bibliometric analysis features.

***Keywords*—**Research topics, bibliometrics, topics of interest, Lattes Platform.

I. Introduction

THE large number of information available on the Internet and the socialization of scientific activity by networks of researchers are essential factors for the current development of science [1]. Services such as digital libraries, relationship networks, bibliographic repositories, and individual scientific production sites are examples of how the Internet has contributed significantly to the number of published works, allowing users not only to access available content, but also to its technical and scientific production from its interaction with this medium. In this way, published and available works can be accessed instantly contributing to the expansion of knowledge [2].

Knowledge is the fundamental element for the generation of development. In addition to scientific dissemination contributing to the democratization of knowledge, it brings the ordinary citizen closer to the benefits that he has the right to claim for the improvement of social welfare, giving him a clearer vision about the true causes and effects of the problems he faces in the day by day [3].

In a globally competitive society, implementing technical and scientific knowledge is an indispensable task for economic and social development, especially for developing countries that are consumers of this knowledge in order to identify their needs and peculiarities [4]. However, often its implementation plan is the existence of limited resources and an increasing requirement of rationality and objectivity in the application of the few available resources.

As important as having investments is having the ability to control, understand, and measure the scientific footing of nations and individualized groups, businesses, and foundations that must decide their scientific priorities. Therefore, it is essential to study, to know and measure, scientific production for the implementation of this knowledge, overcoming the difficulties and reaching the presuppositions of rationality and objectivity [4].

According to Yi and Choi [5], the understanding of scientific production can promote new advances in science. In Brito et. al. [6], the authors point out that works of this nature are considered urgent in Brazil and can portray what is developed and published in science, technology and innovation, making it possible to generate parameters to guide efforts and investments in order to boost research results.

A growing interest by researchers around the world in the most diverse areas regarding the extraction of knowledge in scientific databases has been revealed [7]-[10], [5]. However, the work uses conventionally international scientific databases. Therefore, because they are international, they may not represent what is produced in Brazil [6]. Thus, analyzing a data source that encompasses several types of publication, especially in national and multi-area vehicles, becomes a relevant task for understanding Brazilian science.

For Pritchard [11], bibliometrics stands out as one of the main metric sciences of content analysis, and can be applied to scientific data sources in order to obtain quantitative information about publications. Dias [2], points out that with the use of bibliometrics it is possible to identify the trends and the growth of scientific knowledge in several areas, observe the dispersion of scientific knowledge, support investment policies and understand how scientific evolution happens. Thus, this work aims to perform an analysis of keywords of scientific publications extracted from the curricula of individuals who work in the areas of engineering. The study contemplates bibliometric analyzes, in order to map the research topics of the Brazilian researchers, highlighting them that stand out. With this, a detailed view is presented on the main topics of research that the researchers that work in the areas of engineering have been producing.

II. Background

Efforts to identify research topics are a way of improving understanding of what has been produced about science. These

Jether Gomes, Thiago M. R. Dias, and Gray F. Moita are with the CEFET-MG - Federal Center for Technological Education of Minas Gerais Av. Amazonas, 7675, Nova Gameleira, 30510-000 Belo Horizonte, MG, Brazil (e-mail: jethergomes@yahoo.com.br, thiago@div.cefetmg.br, gray@dppg.cefetmg.br).

studies can be based either on word counts extracted from the titles of the publications or on the keywords of bibliographic productions [12]. For example, Trucolo and Digiampietri [13] developed and applied linear and nonlinear regressions of the indexes of importance based on frequency of terms extracted from titles of scientific publications of a historical basis, in order to identify trends of subjects and branches of research in science short, medium and long term.

Medeiros and Mena-Chalco [14] analyzed more than 650,000 Lattes Platform curricula in order to study the social network composed of all the people who declared to work in at least one of the following major areas: Humanities, Applied Social Sciences or Linguistics, Literature and Arts. In addition, they analyzed the frequency of the words of the titles of the publications to identify which are being used more by period of time and that, somehow, are the most important for these areas. They emphasized that network usage for viewing search interactions for large groups is impracticable due to the large number of nodes and connections that appear. Therefore, as support for the analysis, they used the concepts of word maps for the 200 most frequent words of each area, for periods of time studied.

In the work of Cataldi et al. [15], the authors recognize the important role of *Twitter* in the dissemination of information and proposes a technique of real-time detection of emerging topics expressed by the community under time constraints specified by the user. First, they extracted the content of the tweets and modeled the set of terms according to an aging metric to highlight the emerging terms. In addition, we analyzed the social relationships in the network with the Page Rank algorithm, in order to determine the users' authority. Finally, a directed graph was generated that links the emergent terms with other semantically related words.

Souza et al. [16] analyzed semantic networks constructed from terms extracted from the call titles of an electronic attendance system with the use of social network analysis techniques. The authors highlighted three types of centrality for analysis of the built network. The degree centrality, which treats the importance of a vertex in the connections it establishes with neighboring vertices. The centrality of proximity, which shows the importance of a word in relation to the closest neighbors and its importance in relation to the entire network of words. Finally, the centrality of intermediation, which quantifies the number of times a word acts as a bridge along the shortest path between two other words. The presented results allowed visualizing the recurrent problems in order to facilitate the decision-making and the definition of tendencies in the requests of certain users and sectors.

In Zhu et al. [10], based on the network composed of 111,444 key words from publications in the area of information science, social network analysis metrics were applied and identified the small world effect, showing that words are close to each other. In addition, it was also identified with the calculation of the degree of centrality, that some words have a high number of links with others and that this demonstrates their importance in the network. Faced with this, the authors carried out a preliminary study on how to detect the most relevant terms of a line of research. This method was also compared with the frequency identification strategy of the words, justifying that the analysis based on degree of centrality tends to be more efficient.

Khan and Wood [17] analyzed the networks of keywords and words extracted from the titles of scientific articles referring to the area of Information Technology Management (GIT) through social network analysis techniques and the burst detection algorithm with the proposal of Map the scientific knowledge and highlight the emerging topics that have been produced on the subject, given the importance of the theme for industry and its constant evolution (Fig. 1). The study information was extracted from the Web of Science library query interface. The networks were constructed based on approximately 2000 words of the 893 articles published between 1995 and 2014 referring to 40 magazines, 64 countries, 914 institutions and 1914 researchers. The metrics used for analysis were number of components, diameter, density, clustering coefficient, mean grade, intermediation and centrality measures. However, despite the validity of the results, the authors suggest limitations in the study because the data analyzed are not sufficient to generalize the conclusions about the GIT area.

III. Development

The data source used for this work was the Lattes Platform of the National Council for Scientific and Technological Development (CNPq). This platform aims to integrate the information systems of the Brazilian federal agencies, optimizing the Science and Technology (S & T) management process [9].

The choice of platform is related to the fact that it deals with the integration of scientific data from curricula and institutions of all S & T areas that exist throughout the Brazilian science trajectory. Currently, the platform has more than 4.6 million curricula [2]. These freely available data on the Internet have not yet been widely analyzed, although they are typically used to evaluate or verify data from researchers and groups [7]. It is also used as a source of information for bodies that evaluate Brazil's National Graduate System, and for funding and research funding agencies and scholarship offers.

Despite the data available, these are only visualized through a query interface that presents each curriculum individually, and thus does not allow a more comprehensive analysis. In view of this, for a more detailed analysis of groups, institutions or even the entire nation of Brazilian scientists, techniques and tools for the extraction and analysis of these scientific data are necessary.

In order to acquire the curricula to be analyzed, the scientific data extraction framework developed by Dias et al. [9] Fig. 2.

The process of extracting the data starts by acquiring a list of the curriculum codes obtained through the request in the query interface of the Lattes Platform, so that the identifiers can then be stored locally. With the identifiers in hand, the framework downloads the files, storing them in XML (eXtensible Markup Language) format. Although HyperText Markup Language (HTML) can be extracted, the XML version is the most suitable for automatic processing, since it has all

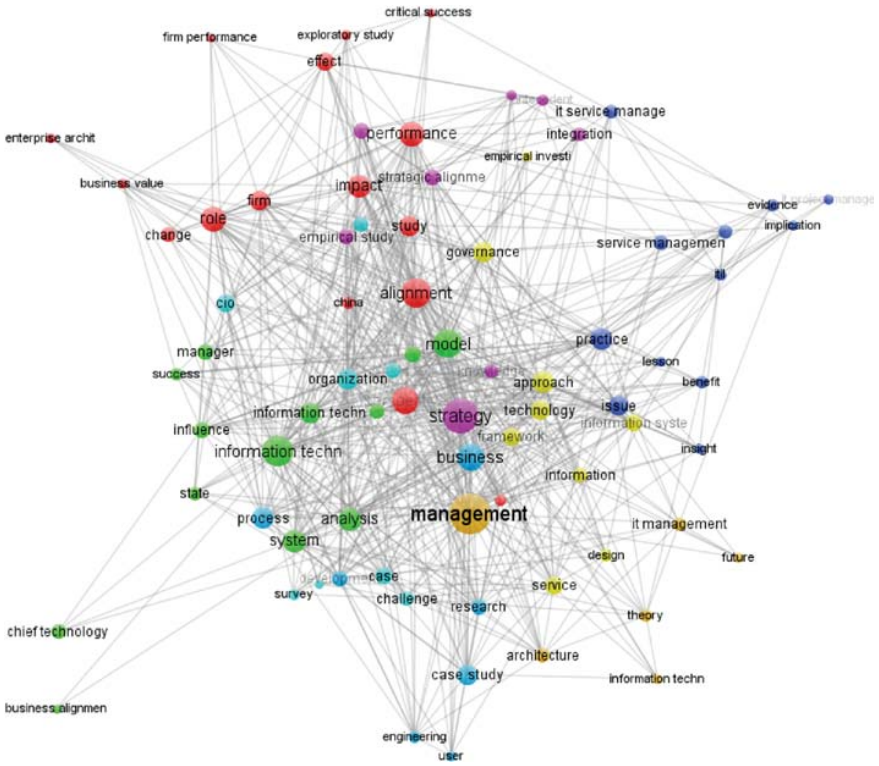

Fig. 1 Networks of words extracted from titles of scientific productions [17]

the sections and fields well delimited, as well as contain the keywords of scientific articles, main object of study of this work.

The justification for the study of keywords in detriment of the titles of the publications, these widely used in researches in the literature, is given by the fact that the keywords of a certain work have as main objective to describe the themes of search that guide the content described without worrying about its semantics.

The data collection took place in March 2016, totaling 208,079 curricula of individuals who declared that their main area of activity was Engineering. Then, the mining of XML files was carried out in order to extract information from individuals and their scientific articles. This step is important because it extracts the information that really needs to be processed and analyzed, and with that, it decreases the computational processing time.

Ferraz et al. [18] points out that analyzing keywords from articles registered in the Lattes Platform is not a trivial task, given that the choice of words does not follow a predefined pattern. As a result of this, one usually has a very large collection and no pattern. In an attempt to overcome this problem, a method was developed that performs the keyword processing in order to exclude possible words with noises or that do not represent a research topic.

The method starts by getting the number of keywords extracted from each article and its references to it. Therefore, each word passes through a process of language detection, information used in the *stemming* step. In continuation, in the *lowercase* stage, the words are converted to lowercase with the proposal to standardize the set.

In the *Stop Words* step, words that have no semantic value are removed. Next, the normalization process is performed to extract the accented letters and replace them with their unstressed equivalent. In the *stemming* stage, each word is reduced to its radical, and with that, it avoids the inclusion of words with the same meaning in different ways. In the case of compound words, the process runs on each word individually, and then are concatenated by forming a single word.

Each word resulting from the whole process is inserted into the dictionary keeping its original format and reference through the article code. If the word is already present in the dictionary, a corresponding counter is incremented. Table I shows an example of a word transformation. While Table II illustrates an example of the dictionary.

TABLE I
EXAMPLE OF TRANSFORMATION OF A WORD EXTRACTED FROM A SCIENTIFIC ARTICLE

Step Algorithm	Original Word
1	Gerência de Dados, 1000
2	Gerência de Dados, 1000, Portuguese
3	gerência de dados, 1000, Portuguese
4	gerência dados, 1000, Portuguese
5	gerencia dados, 1000, Portuguese
6	gerenc dad, 1000, Portuguese
7	gerenc-dad, 1000

The dictionary is composed of the number of words that have been reduced to a given radical, by the radical itself, the codes referring to the articles they appeared and the original

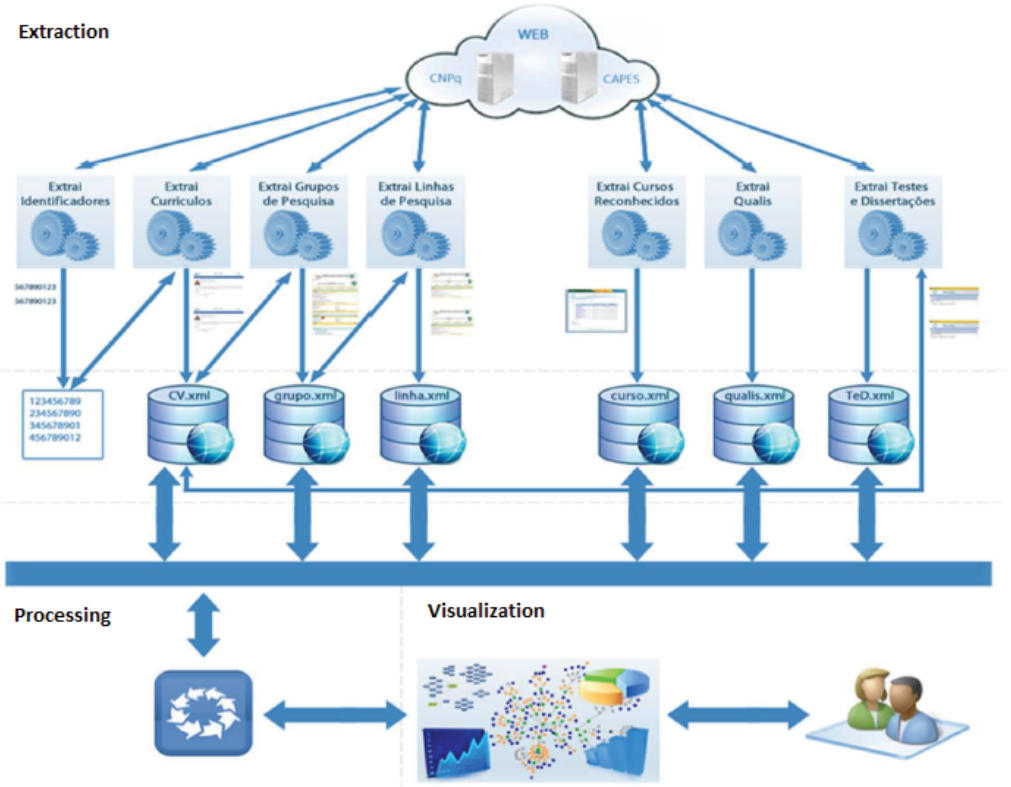

Fig. 2 Framework for data extraction from the Lattes Platform [9]

TABLE II
DICTIONARY EXAMPLE WHERE THE TRANSFORMED WORD IS INSERTED

Frequency	Radical	C. Article	Word 1	C. Article	Word 2	...
2	gerenc-dad	1000	Gerência de dados	500	Gerência de dados	...
25	mineraca-dad	1000	Mineração de Dados	100	Mineração de Dados	...
5	banc-dad	30	Banco de Dados	1000	Banco de Dados	...

keywords. This strategy was adopted so that all radicalized keywords could be traceable to their respective publication. For example, in this case, it may be noted that in the code article 1000 the following keywords in Portuguese appeared *Gerência de Dados*, *Mineração de Dados* and *Banco de Dados*.

IV. Results and Discussions

Of all the curricula registered in the Lattes Platform in March 2016 (4,457,260) a total of 208,079 reported as their main large area, the Engineering. Such large area is the fifth with the largest amount of curricula. Besides the large area of activity of a given individual, the curricula also have information about the areas of action, so it is possible to see which areas are most representative (Fig. 3).

As can be observed, the area of Electrical Engineering holds approximately 24% of the curricula, followed by Civil Engineering and Mechanical Engineering. Consequently, it is to be expected that such areas will also be responsible for the greater number of articles published and thus have a direct influence on the main keywords. It should be noted that only 0.36% of the curricula linked to a large area of Engineering did

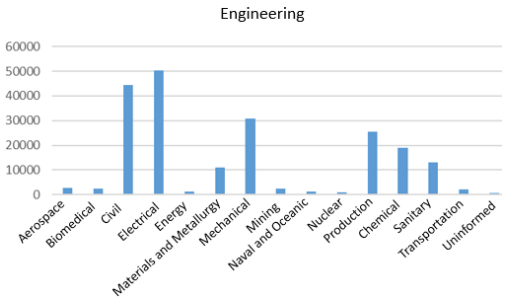

Fig. 3 Distribution of curricula by area in the major area of Engineering

not specify the area of activity. Although most of the curricula have been updated for at least two years (76%), a large number do not have registered scientific publications (Table III).

When verifying the curricula with scientific production registered, it is possible to identify the main keywords used and that represent the main topics of studies (Table IV). The result presented had as ranking criterion the frequency of the main words.

It can be noticed that for the set of 20 most frequent

TABLE III
CURRICULA WITHOUT INFORMED SCIENTIFIC PRODUCTION

Big Area	Total Curricula	Articles in Annals of Congress		Articles in Journal	
		No Production	%	No Production	%
Engineering	208.079	153.833	73,93	177.538	85,32

TABLE IV
MAIN TOPICS OF STUDY USED BY INDIVIDUALS WHO WORK IN ENGINEERING

Keyword	Frequency
Simulação	1.713
Ergonomia	1.712
Otimização	1.670
Meio Ambiente	1.618
Qualidade	1.596
Sustentabilidade	1.479
Biodiesel	1.440
Modelagem	1.331
Geoprocessamento	1.310
Redes Neurais	1.195
Corrosão	1.158
Elementos Finitos	1.109
Arquitetura	1.091
Ensino	1.068
Sensoriamento Remoto	1.032
Reciclagem	1.031
Design	1.006
Inteligência Artificial	993
Concreto	960
Irrigação	938

keywords, several are related to the areas of activity with the largest number of individuals with registered Lattes curricula, such as *Meio Ambiente*, frequent concern of Civil Engineering; *Ergonomia* highly referenced by those who work in Production Engineering; As well as *Arquitetura* often investigated by Civil Engineering researchers. However, it is important to highlight keywords that are generic research themes that are frequently applied in works from different areas such as *Simulação*, *Otimização*, *Qualidade*, *Modelagem* and *Redes Neurais*. In view of this, it is observed that among the most frequent topics when considering the keywords used in the publications registered in the Lattes curricula of individuals who have worked in the areas of engineering, these are directly influenced by themes linked to the areas with the greatest number of curricula, as well as topics adopted by researchers from different areas. Therefore, analyzes that consider criteria other than frequency, such as those based on social network analysis metrics, can provide greater precision in identifying the most relevant or impacting topics in the set analyzed.

V. CONCLUSION

Considering the study carried out here, it is emphasized that this type of analysis is characterized as an important mechanism, since it allows identifying the most impacting research topics within a research community. When analyzing the keywords of Lattes curricula, it is possible to consider publications made in annals of congresses, which would not be feasible to verify in other international data sources. With this, you can get an accurate view of the most researched topics.

It is noticed that among the analysis of frequency of words, the most representative areas influence it, as well as the topics adopted in several areas. As a result, it is expected that as future work, we incorporate analyzes that consider temporal factors to determine the relevance of a topic, as well as analyzes based on social network metrics to determine those more central words and, consequently, with a greater degree of importance.

ACKNOWLEDGMENT

The authors thank Federal Center for Technological Education of Minas Gerais (CEFET-MG) and Foundation for Research Support of the State of Minas Gerais (FAPEMIG) for their assistance in the research.

REFERENCES

[1] M. Castells, "A sociedade em rede," *São Paulo: Paz e Terra*, vol. 8, 1999.
[2] T. M. R. Dias, "Um estudo da produção científica brasileira a partir de dados da plataforma lattes," Ph.D. dissertation, CEFET-MG, 2016.
[3] D. Carneiro. (2003) C&T em prol da cidadania. (Online). Available: http://memoria.ebc.com.br/agenciabrasil/noticia/2003-10-03/ct-em-prol-da-cidadania
[4] S. G. SAES, "Aplicação de métodos bibliométricos e da "co-word analysis" na avaliação da literatura científica brasileira em ciências da saúde de 1990 a 2002." Ph.D. dissertation, Universidade de São Paulo, 2005.
[5] S. Yi and J. Choi, "The organization of scientific knowledge: the structural characteristics of keyword networks," *Scientometrics*, vol. 90, no. 3, pp. 1015–1026, 2011.
[6] A. G. C. de Brito, L. Quoniam, and J. P. Mena-Chalco, "Exploração da plataforma lattes por assunto: proposta de metodologia," *Transinformação-ISSN 2318-0889*, vol. 28, no. 1, 2016.
[7] L. Digiampietri, "Análise da rede social brasileira," Ph.D. dissertation, School of Arts, Sciences and Humanity, University of São Paulo (USP), 2015.
[8] J. P. Mena-Chalco, L. A. Digiampietri, F. M. Lopes, and R. M. Cesar, "Brazilian bibliometric coauthorship networks," *Journal of the Association for Information Science and Technology*, vol. 65, no. 7, pp. 1424–1445, 2014.
[9] T. M. R. Dias, G. F. Moita, P. M. Dias, and T. H. J. Moreira, "Identificação e caracterização de redes científicas de dados curriculares," *iSys-Revista Brasileira de Sistemas de Informação*, vol. 7, no. 3, pp. 5–18, 2014.
[10] D. Zhu, D. Wang, S.-U. Hassan, and P. Haddawy, "Small-world phenomenon of keywords network based on complex network," *Scientometrics*, vol. 97, no. 2, pp. 435–442, 2013.
[11] A. Pritchard, "Statistical bibliography or bibliometrics," *Journal of documentation*, vol. 25, p. 348, 1969.
[12] J. Choi, S. Yi, and K. C. Lee, "Analysis of keyword networks in mis research and implications for predicting knowledge evolution," *Information & Management*, vol. 48, no. 8, pp. 371–381, 2011.
[13] C. Trucolo and L. Digiampietri, "Análise de tendências da produção científica nacional da área de ciência da computação," *Revista de Sistemas de Informação da FSMA*, vol. 14, pp. 2–9, 2014.
[14] C. B. Medeiros and J. Mena-Chalco, "The dynamics of multidisciplinary research networks-mining a public repository of scientists cvs," in *World Social Science Forum*, 2013, pp. 1–17.
[15] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. ACM, 2010, p. 4.
[16] J. Souza, D. Lyra, J. Cavalcanti, R. Simão, Z. Cézar, A. N. Duarte, and A. V. Brito, "Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento."

[17] G. F. Khan and J. Wood, "Information technology management domain: Emerging themes and keyword analysis," *Scientometrics*, vol. 105, no. 2, pp. 959–972, 2015.
[18] R. R. N. Ferraz, L. M. Quoniam, and E. A. Maccari, "The use of scriptlattes tool for extraction and on line availability of academic production from a department of stricto sensu in management," in *11th International Conference on Information Systems and Technology Management–CONTECSI*, vol. 17, 2014.