A System for Discovery of Knowledge in Data Repository Education

Emerson de Sousa Costa, Thiago Magela Rodrigues Dias, and Patrícia Mascarenhas Dias

***Abstract*—Studies on scientific production date have received attention from Researchers in various fields to gain insight into the evolution of research in general. Such studies enable the analysis of scientific production for various purposes and one of the challenges in this type of analysis lies in the diversity of repositories containing data in different formats and structures. Currently, Bibliometric and scientometric Analyzes are the Characterized the main scientific analysis techniques on the production of a set of Researchers. These Analyzes aim to carry out the measurement of scientific communication, including the study of sciences to identify its structure, evolution and connections. The proposal of this project is to Develop an information system to perform scientificometric Analyzes of sets of Individuals, having as a source of data repositories of scientific publications. As a case study will be Analyzed the Brazilian Researchers who have acted in the Interdisciplinary area in Brazil. These Analyzes intend to present an overview of the area under study in Brazil, having the main source of data for the Analyzes, date of scientific publications of this set, extracted from its curricula of the Lattes Platform. Therefore, it will be possible to identify the profile of These Researchers, where They are acting, how the collaboration between Them Occurs, and Also, the qualitative analysis of Their research, Thus enabling the accomplishment of Analyzes comparative with the different areas of knowledge. After validation with the case study carried out Several other sets of Individuals Can Be Analyzed.**

***Index Terms*—Bibliometric, scientific collaboration, lattes platform, scientific collaboration network.**

I. Problem Description

A new generation of services mainly in the Web is changing the way to disclose and make available the scientific and technological production. There is currently a trend that strengthens the exchange of information and collaboration between people. The strong relationship between the scientific field and socioeconomic has provided a growing interest in understanding the mechanisms that involve scientific activities, and can point to several studies that analyze elements of its construction as the characteristics of language and discourse used in scientific communication [1] or a collaborative relationship between researchers [2].

With competition increasingly fierce between research institutions, it is important to its members find potential employees in order to carry out work in collaboration to boost its scientific production and achieve better results in their research. We can point to recent works where it is shown that research groups with a well-connected scientific social network generally tend to be more productive [3], [4].

Allied to this, the presence of data releases available in different formats and in different repositories difficult consultations by users who require a unified view of these data or the identification of groups of individuals who are working with a particular theme in different institutions or regions. The growth and evolution of the Web, for example, created a lot of textual data with little structure, heterogeneous, generated with no concern for standardization, and especially fragmented in different repositories. Despite this disruption, often data that Web users seek are not present in a single source and mechanisms for automatic integration of data is desirable.

Semi-structured data from the web often can not be integrated with the use of transactions provided by DBMSs (Database Management Systems) traditional due to the lack of standardization in the format and structure of the data, inconsistencies in the records as incomplete and with errors grammar, among others. When these data are in text format, you can use information retrieval techniques and word processing to establish a similarity function in braces or other fields of the records. For this, they are usually extracted scientific data repositories available on the web and integrated for further analysis,

However, it is often difficult to identify relevant information from these data sets, which creates need for methods and tools that can transform them, automatically, into information that bring relevant knowledge. Thus, automated techniques for knowledge discovery can be applied to these collections in order to obtain useful information about the vehicle combination.

The bibliometric study, especially in large bibliographic repositories, is not a trivial task given the amount of data to be analyzed and characteristics of these repositories, which, in most cases, do not have a set pattern. Therefore, in this research, techniques to allow obtain scientometric information on the entire set of data from researchers at the Interdisciplinary area with registered Lattes curriculum in the Lattes Platform are implemented in a system that aims to conduct studies on sets of researchers formed by informed criteria, obtaining thus an unprecedented picture of the scientific production of the analyzed sets.

In addition to the scientometric information in the scientific field, the collaboration networks have been widely studied. To [5], evidence points that way researchers collaborate has a strong impact on their productivity. At work, the authors analyze the correlation between analysis metrics of social

Manuscript received February 12, 2019; revised May 24, 2019.

The authors are with Centro Federal de Educação Tecnológica de Minas Gerais – CEFET-MG, Brazil (e-mail: emerson@cefetmg.br, thiagomagela@cefetmg.br, patriciamdias@gmail.com).

doi: 10.18178/ijiet.2019.9.8.1261

networks of researchers collaborative networks of computer science area and its scientific performance. The results show that researchers establish strong collaborative ties and have an intermediary role within the network, in general, have a higher performance index. Before that, a collaboration network-based analysis will also be held.

This study brings as differential the adoption of specific computational techniques for the recovery, treatment and analysis of large volume of curricular data. Such techniques make it possible to analyze large groups of individuals to obtain a priori knowledge not yet discovered.

The general motivation of the research, as mentioned, is to explore the publications of researchers in various fields of knowledge with the adoption of scientometric analysis and based on social network analysis.

II. Goals

Seeking greater understanding of how does the Brazilian scientific research in various fields of knowledge, this study has the general objective to propose an information system able to perform bibliometric analysis and based on social network analysis in scientific publications data to knowledge discovery on sets analyzed. Getting this way, knowledge about how their research has evolved based on scientometric and social network analysis on data available in the curricula of the Lattes Platform.

Therefore, the following specific objectives should be addressed:

- Identification of the researchers of the areas to be analyzed;
- Extraction of Lattes Platform curricula of researchers to build a data repository;
- General characterization of the identified set;
- Scientometric application metrics in publications of data (quantitative and qualitative methods);
- Identification collaborations;
- Modeling and characterization of collaborative networks;
- Application metrics of social network analysis in networks characterized;
- comparative analysis with other areas of knowledge.

Specifically, the challenges are explored in:

- Extraction and data integration;
- scientometric analysis for the determination of indicators of scientific research;
- scientific collaboration identification;
- Modeling and characterization of scientific collaboration networks;
- Adoption of social network analysis metrics for extraction of knowledge about the modeled networks.

III. Related Work

In [6] the authors point out an increase in the number of shared scientific works, driven, mainly, by the ease in the way of distributing the works, allowing a greater speed in the exchange of information. In addition, the authors emphasize that collaborative work saves time and financial and material resources, resulting in a process increasingly stimulated by institutions and agencies that fund research. For [7], the interaction between two or more scientists who, in a given context, aims to share activities to achieve results and achieve common goals, is defined as scientific collaboration and is characterized as a complex phenomenon.

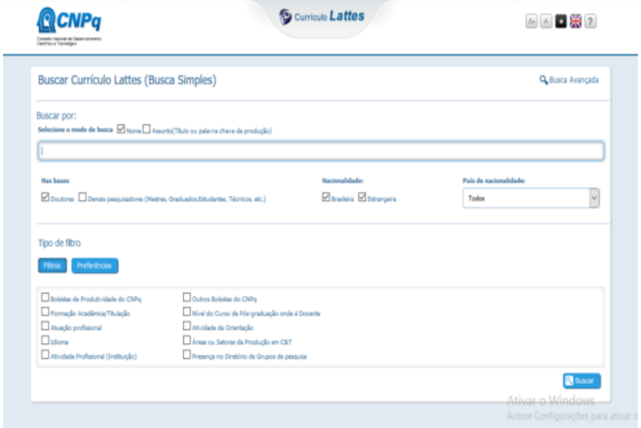
In his doctoral research, [8] analyzed Brazilian scientific production in the field of health. The survey included 3,066 journals, corresponding to a total of 38,349 articles published between 1990 and 2002 extracted from the Web of Knowledge. Bibliometric methods and co-occurrence of terms were the instruments used to measure scientific activity. The descriptors and identifiers of the articles were related to other variables, such as authors, year of publication and affiliation of the publication. From there, it was possible to identify networks of cooperation between researchers and institutions, as well as the descriptors of the field of knowledge studied.

In the research [9], unlike the works found in the literature, presents a comprehensive study on Brazilian scientific production. For this, the author developed LattesDataXplorer, a framework responsible for extracting the entire data set from the curricula. In this framework, techniques were also implemented for bibliometric and metric analyzes based on social network analysis to study scientific collaboration networks. Thus, it was possible to characterize the entire curriculum repository of the Lattes Platform and also to present a detailed view on Brazilian scientific production. According to the author, the results found are unprecedented in view of the comprehensiveness of the analyzes performed and the large number of individuals considered. As one of the future works, it suggests the study of the content of the publications to understand which topics were and are studied by Brazilian science.

Already in the work of [10] the authors make an evaluation of the performance of the main researchers who work in Computer Science. Data extracted from the Lattes Platform curricula of 406 researchers in CNPq research productivity in the area of Computer Science are used in the five types of scholarship (1A, 1B, 1C, 1D and 2). The evaluation considered three central dimensions, being: the researcher's career time, in which the number of years after the conclusion of the doctorate was considered; number of targeted students; and scientific productivity, referring to the volume of publications and quotations. Regarding the researchers' career time, it is observed that those with the highest levels of scholarship also have a longer career. Regarding the guidelines, there is a distinction between the number of master's and doctoral degrees, in which there is approximately a doctorate orientation for each year of career among all the fellows. When considering the master's guidelines, this figure falls to less than 0.5 guidance for each year upon completion of the doctorate. When evaluating the scientific production of the researchers, it is observed that the volume of publications increases with the level of scholarships, except for level 1A, which is similar to 1C, because it includes older researchers, these were not linked to programs with certain level of maturity. These evaluations demonstrate consistency between the evaluated items and the scholarship modalities in which the researchers fall into the area of Computer Science.

IV. Research Methodology

In this research, initially was defined as subject the analysis of the data contained in the curricula of the Lattes Platform set of researchers. The choice of the Lattes Platform for data extraction is related to the fact that it has a vast amount of data as it deals with the integration of scientific data curricula and institutions of S&T area, recording the academic, technical data and productions scientific, still allowing the update of personal data is carried out by the researchers themselves. Currently the Lattes Platform has approximately 5 million registered CVs, resumes these that have information about academic, research areas, professional activities, academic guidance, and technical and scientific production. The Lattes Platform search page can be viewed in Fig. 1.


CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico
Currículo Lattes
A+ A ?

Buscar Currículo Lattes (Busca Simples)

Busca Avançada

Buscar por:

Selecione o modo de busca ☑ Nome ☐ Assunto(Título ou palavra chave da produção)

Nas bases

☑ Doutores ☐ Demais pesquisadores (Mestres, Graduados,Estudantes, Técnicos, etc.)

Nacionalidade:

☑ Brasileira ☑ Estrangeira

País de nacionalidade: Todos

Tipo de filtro

Filtros Preferências

☐ Bolsistas de Produtividade do CNPq
☐ Formação Acadêmica/Titulação
☐ Atuação profissional
☐ Idioma
☐ Atividade Profissional (Instituição)
☐ Outros Bolsistas do CNPq
☐ Nível do Curso de Pós-graduação onde é Docente
☐ Atividade de Orientação
☐ Áreas ou Setores da Produção em C&T
☐ Presença no Diretório de Grupos de pesquisa

Buscar

Ativar o Windows
Acesse Configurações para ativar o

Fig. 1. Lattes platform search page.

Curricula that are part of the Lattes Platform became a national standard for individual assessment of the scientific and academic activities, since they aggregate data from researchers from all areas of knowledge, making the Lattes Platform an extremely extensive and valuable source for analyzing and understand the behavior of research groups. For the analysis presented here, a set of computational components (Fig. 2) was used for the collection and processing of data.

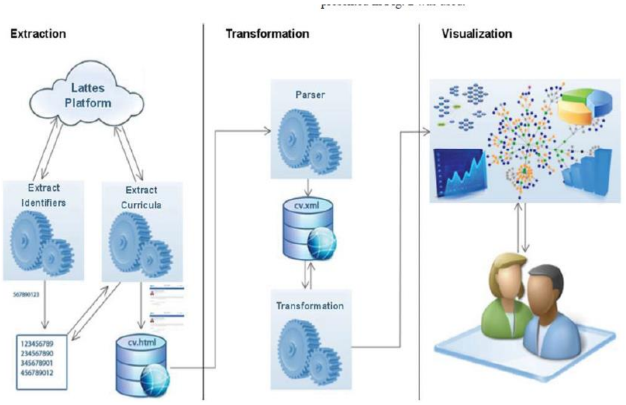


Fig. 2. Framework for extraction and integration of scientific data [11].

Although the data from the Lattes Platform curricula are freely available, these are displayed in query interface provided by CNPq presenting curricula individually, without the possibility of analyzes and comparisons with other curricula. In view of this, techniques and tools for the extraction of curricular data are needed.

Initially, an extensive literature review of the work on the topic was held. This review showed that many relevant studies have explored the curriculum data Lattes Platform in recent years, with significant results for understanding of certain areas of knowledge.

Thus, methods for the extraction of data in the Lattes Platform of Interdisciplinary area researchers will be proposed. Such methods will be implemented and therefore are extracted every set data to be analyzed. Given the amount of data to be analyzed, several studies will be performed to identify which tests can be applied.

The adoption of metrics for scientometric analysis of the scientific production of a particular area of research allows measuring and understanding significantly occur as the research and what the study trends of its researchers. With this, various metrics to scientometric analysis will be applied to the extracted data (Fig. 3).

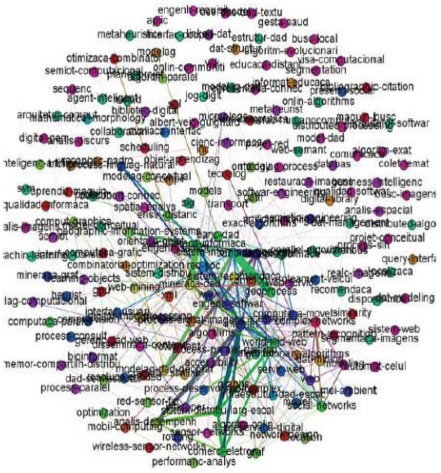


Fig. 3. Scientific collaboration.

In addition to the scientometric analyzes will also be carried out several studies on scientific collaboration networks. Given the number of studies of possibilities to be carried out with the subjects studied, methods for the characterization of networks will be implemented for data analysis.

To assess the main techniques implemented initial analysis will be performed, resulting in a set of information that enables preliminary identification an overview of the analyzed data. Based on this initial analysis will be presented preliminary results that show a macro view on how to have occurred Brazilian Interdisciplinary research in the area. Also, verify that new studies may be obtained from the initial results of this project.

The presented project is multidisciplinary since for its construction will involve techniques from various areas such as statistics, Big Data, as it will be necessary to process a large volume of data, use of Artificial Intelligence techniques such as natural language processing, and further, concept of graph theory to application of metrics, and also the design concepts and Analysis of Algorithms and complexity analysis, rows and cells.

V. Results

It is expected that after implementation of the system proposed here and its validation with a case study, this research result in a complete application that allows performing bibliometric and scientometric analysis of a group of individuals with registered resumes in the Lattes Platform, resulting in a tool that allows to understand the basis of the scientific publication of the data set publication profile.

In addition, the proposed system will incorporate techniques based on social network analysis in the scientific collaboration networks, enabling trace as sets of researchers collaborate. Such analyzes provide important knowledge to be proposed policies to encourage collaboration.

A potential advantage of the tool proposed here compared to other tools is that it is being developed with the aim of using as little computational resources as possible so that the analyzes to be performed can contemplate large data sets and serve as tools for studies in several areas of application. All the tools built will be made available in order that it can be used in other studies and even be adapted for analysis of other sources of scientific data.

Thanks

The authors thank FAPEMIG and CEFET-MG for their assistance in the research.

References

[1] HOFFNAGEL, "JC practice citation in academic papers (the practice of citation in academic works.)," *Language books and Society*, vol. 10, no. 1, p. 71, 2009.

[2] Y. Ding, "Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks," *Informetrics J.*, vol. 5, no. 1, pp. 187-203, 2011.

[3] G. R. Lopes *et al.*, "Ranking strategy for graduate programs evaluation," in *Proc. International Conference on Information Technology and Applications*, pp. 59-64, 2011.

[4] M. A. Brandao *et al.*, "Using link semantics to recommend collaborations in academic social networks," in *Proc. the International Conference on World Wide Web Companion*, XXII, Rio de Janeiro, pp. 833-840, 2013.

[5] A. J. Wanderley *et al.*, "Identifying correlations between metrics of social network analysis and performance index for computer science researchers (identifying correlations between social network analysis metrics and the performance index of computer science researchers)," Brazilian Workshop on Social Network Analysis and Mining, Brasilia, 2014.

[6] A. J. Meadows and A. A. B. Lemos, "A comunicação científica," *Brasília: Briquet de Lemos/Livros*, 1999.

[7] D. H. Sonnenwald, "Scientific collaboration," *Annual Review of Information Science and Technology*, vol. 41, no. 1, pp. 643-681, 2007.

[8] 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. 2005. 183. (Tese de Doutorado)," Programa de Pós-Graduação em Saúde Pública, USP, São Paulo, 2005.

[9] T. M. R. DIAS, "Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes. 2016. 181 (Tese de Doutorado)," Programa de Pós-Graduação em Modelagem Matemática e Computacional, PPGMMC/CEFET-MG, 2016.

[10] H. LIMA *et al.*, "Assessing the profile of top Brazilian computer science researchers," *Scientometrics*, vol. 103, pp. 879- 896, 2015.

[11] T. M. R. Dias *et al.*, "Modelagem e caracterização de redes científicas: Um estudo sobre a plataforma lattes," Brazilian Workshop on Social Network Analysis and Mining, Anais, Maceió, 2013.

Emerson S. Costa is professor at the Federal Center of Technological Education of Minas Gerais - CEFET-MG. He got the PhD in mechanical engineering, the master in mathematical and computational modeling, graduate in mathematics.

Thiago M. R. Dias is professor of the Federal Center for Technological Education of Minas Gerais - CEFET-MG. He got the PhD in mathematical and computational modeling, the master in mathematical and computational modeling, graduate in computer science.

Patrícia M. Dias is PhD student at the Federal Center of Technological Education of Minas Gerais - CEFET-MG. she got the master in mathematical and computational modeling, graduate in mathematics.