Analyzing Keyword Networks for the Identification of Correlated Research Topics

Thiago M. R. Dias, Patrícia M. 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 distribution of these works. Faced with this, there is a growing interest in understanding how scientific research has evolved, in order to explore this knowledge to encourage research groups to become more productive. Therefore, the objective of this work is to explore repositories containing data from scientific publications and to characterize keyword networks of these publications, in order to identify the most relevant keywords, and to highlight those that have the greatest impact on the network. To do this, each article in the study repository has its keywords extracted and in this way the network is characterized, after which several metrics for social network analysis are applied for the identification of the highlighted keywords.

Keywords**—**Extraction and data integration, bibliometrics, scientometrics.

I. Introduction

In recent years, in addition to scientific production, there has been a steady growth in the study of networks in relation to various disciplines ranging from computer science to areas such as sociology and epidemiology. The network can be characterized as a graph, consisting of a set of nodes (vertices) and links (edges) between nodes. These bonds may be, directed or non-directed, and may optionally have an associated weight. Many, perhaps almost all-natural phenomena can usually be described in terms of a network. The brain can be characterized as a network of neurons bound by synapses. The Internet is also an example of an important network these days.

The topics cited have already been objects of study by several researchers; however, it has only recently been that network analysis has become an important area of research [1]. This is partly due to the advancement of computers. Computers have aided in the empirical study of real networks and allowed researchers in different areas to apply their techniques in large networks.

The strong relationship between the scientific and socio-economic domain has provided a growing interest in understanding the mechanisms involved in scientific activities [2]. The cooperative relationship between researchers was also analyzed [1].

II. Related Work

With the ever-increasing competition among research organizations and institutions, it becomes important for its members to discover potential collaborators in order to leverage scientific production. Recent studies show that research groups with a well-connected network tend to be more productive [3], [4].

Community co-authoring networks may reveal interesting facts about them, such as the groups that collaborate best, the intensity of relationships between authors or authors working with a greater degree of collaboration. The study of co-authoring networks can also be used to compare patterns of collaboration among different scientific communities [5].

In [6], Canibano and Bozeman suggest that curriculum analysis can be used as a sufficiently comprehensive source of information in academic research, and that its usefulness has been widely explored since 2000.

In the work of Petersen [7], important factors for academic success in scientific networks are highlighted. Among them are the abundance of scientific production that enhances the attraction of future opportunities and size of the team of collaborators, co-authors and collaboration network. In view of this, it is evident the importance of new studies in order to understand and analyze how scientific collaboration happens, as well as to propose new tools that aim to boost scientific production.

In the research conducted by Cataldi et al. [8] the authors recognize the important potential of Twitter and propose a technique for detecting emerging themes, which allows to recover in real time most of the emerging topics expressed by the community. First, it extracts the content (set of terms) from the tweets and models the life cycle of the terms according to an aging technique with the intention of exploring the emerging terms. A term can be defined as emergent if it often occurs in the specified time interval and was relatively rare in the past. In addition, considering that the importance of a content also depends on its source, the social relations in the network with the Page Rank algorithm are analyzed in order to determine the users' authority. Finally, topic graphs are generated that link emerging terms with other semantically related keywords, allowing for the detection of emerging topics, under time constraints specified by the user. They offer different case studies that show the validity of the proposed approach (as shown in Fig. 1).

Thiago M. R. Dias, Patrícia M. Dias, and Gray F. Moita are with the Federal Centre for Technological Education of Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil; (e-mail: thiagomagela@gmail.com, patriciamdias@gmail.com, gray@dppg.cefetmg.edu.br).

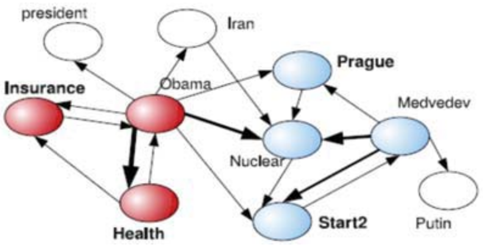


Fig. 1 Network analyzed [9]

In Zhu et al. [9], based on the network composed of 111,444 key words from articles in the area of information science extracted from the Scopus repository, the average distance and cluster coefficient metrics are evaluated. The authors, with the application of complex network techniques and through calculations, reveal the small world effect of the keyword network. At the core of the keyword network, the degree of centrality is used to conduct a preliminary study on how to detect the hotspots of a research discipline.

III. Methodology

In this work, data from the Lattes Platform of CNPq were used. The Lattes Platform was designed to integrate the information systems of the Brazilian federal agencies, optimizing the Science and Technology (S&T) management process from the user's point of view, as well as development agencies and teaching and research institutions [10].

The choice of the Lattes Platform for extraction is related to the fact that it is extremely rich, since it deals with the integration of scientific data from curricula and institutions of the S&T area, recording the academic data and the scientific productions of the researchers and institutions, allowing that the updating of the data is carried out by the researchers themselves. Currently Lattes Platform has approximately 4.9 million registered curricula.

Several articles for the analysis of scientific data have explored the Lattes Platform as a primary source of information [10]. Although data from the Lattes Platform curricula are freely available, they are visualized through a query interface that presents the curricula individually. In view of this, techniques and tools for the extraction and integration of data with other scientific data bases to complement the information become necessary.

To extract the curricula to be analyzed, the framework presented in Fig. 2 was used.

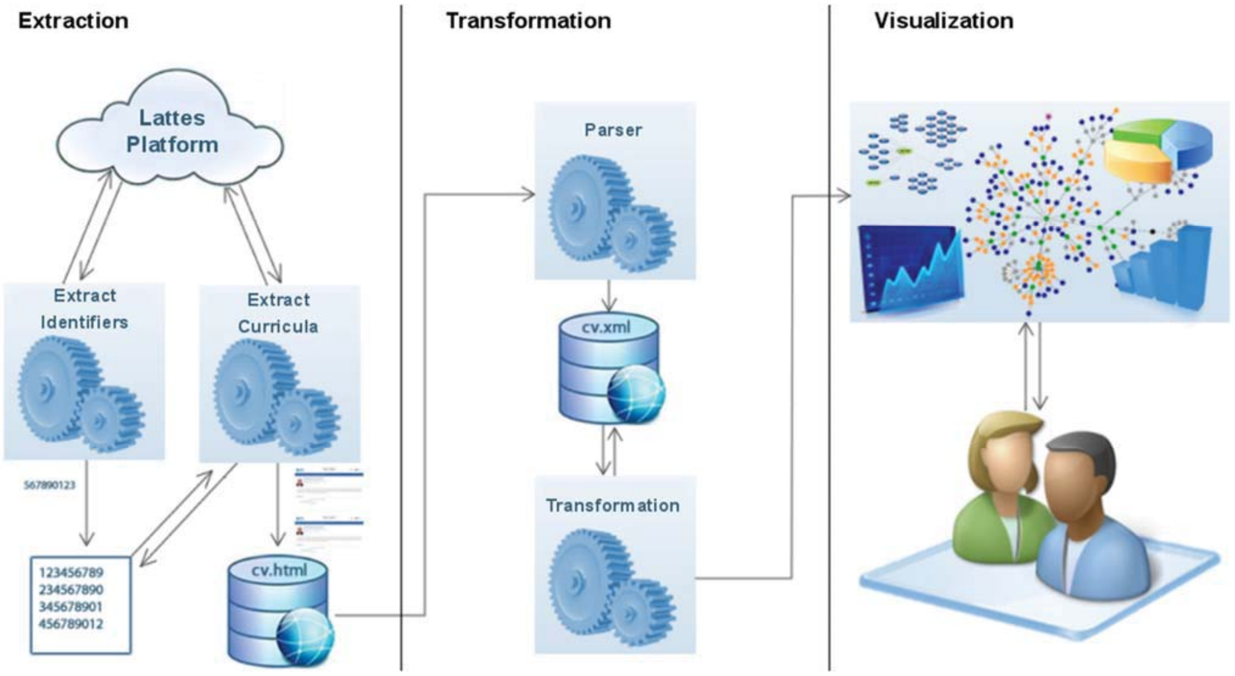


Fig. 2 Framework for Extraction and Integration of Scientific Data [10]

The framework extracts data from the Lattes Platform through the individual identifiers of each of the curricula, which are extracted and stored in XML (eXtensible Markup Language) format for further analysis. In the framework, several other functionalities are incorporated, such as the extraction of research groups and researchers that make up the platform, however, for the purpose of this work, only the resume extractor was used.

After the extraction and storage, each of the publications registered in a curriculum are analyzed and their keywords form a click, which is inserted in the graph by juxtaposition. Therefore, all the articles registered are analyzed and all the words of these articles that can represent research topics are inserted in the graph, processing is terminated when all data is analyzed. (Fig. 3).

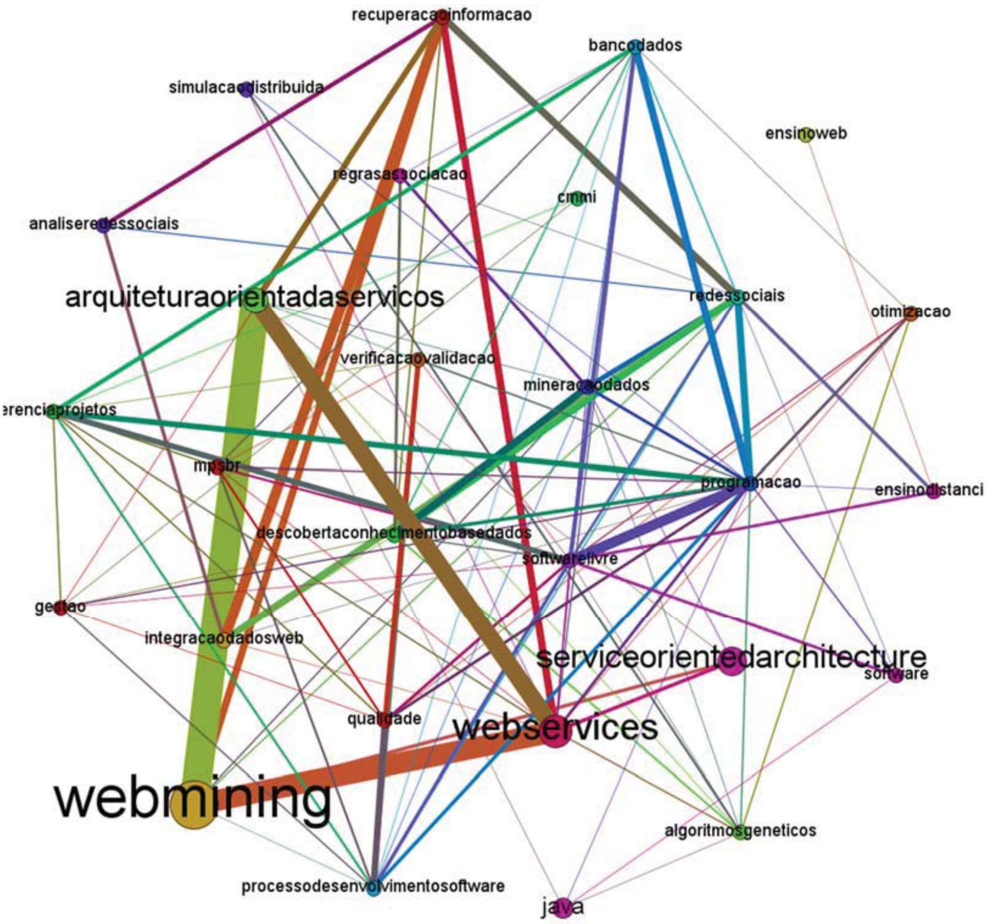


Fig. 3 Keyword network example

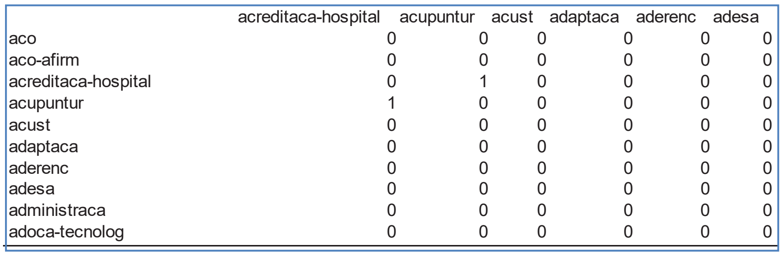
After the construction of the graph, in which the keywords that were used in the same publication, are connected. So, we can look at those keywords that were most commonly used together. Given this, it is possible to visually identify the most relevant words and their links to each group of curricula analyzed. Thus, it is easily possible to extract which topics (keywords) have the greatest influence on the networks analyzed. In addition, it is possible to identify words that are more related to a set of other words. This type of analysis presents itself as an excellent mechanism for the identification of related issues. That is, words that have a sparse edge with other words means that they have a high degree of collaboration (edge with weight greater than average). In the example of Fig. 3 the words "webmining" and "arquiteturaorientadaservicos" have a high degree of collaboration.

Alternatively, by constructing keyword networks from individual curricula (Fig. 3), it is possible to perform similarity analysis of networks. That is, networks that resemble are excellent indications that two researchers have similar profiles and recommendation of researchers who can work together can be performed.

IV. Results

With the adoption of social network analysis metrics, it is possible to identify characteristics that are not visually identified. These characteristics are important because they can reveal valuable information in order to foster research based on emerging issues. For the analysis of neighbors in common, an n × n matrix is created, where n is the number of words and in this way values are inserted that represent the amount of neighbors (words) in common that each keyword has with another (Fig. 4). These matrices are important because their results allow the performance of works in several areas of research such as classification and recommendation systems, which aim to classify words or recommend words that can work together with new words that a researcher can consider for future research.

For the present work, the matrix of neighbors in common allows to identify the keywords that have a certain amount of neighbors in common, characterizing them as very close in the analyzed network. Therefore, the number of neighbors in common allows to indicate the proximity of two keywords in the network. In addition to matrices with different metrics, several other graphs can be generated for the analysis of keyword networks (Fig. 5).



	acreditaca-hospital	acupuntur	acust	adaptaca	aderenc	adesa
aco	0	0	0	0	0	0
aco-afirm	0	0	0	0	0	0
acreditaca-hospital	0	1	0	0	0	0
acupuntur	1	0	0	0	0	0
acust	0	0	0	0	0	0
adaptaca	0	0	0	0	0	0
aderenc	0	0	0	0	0	0
adesa	0	0	0	0	0	0
administraca	0	0	0	0	0	0
adoca-tecnolog	0	0	0	0	0	0

Fig. 4 Matrix of neighbors in common

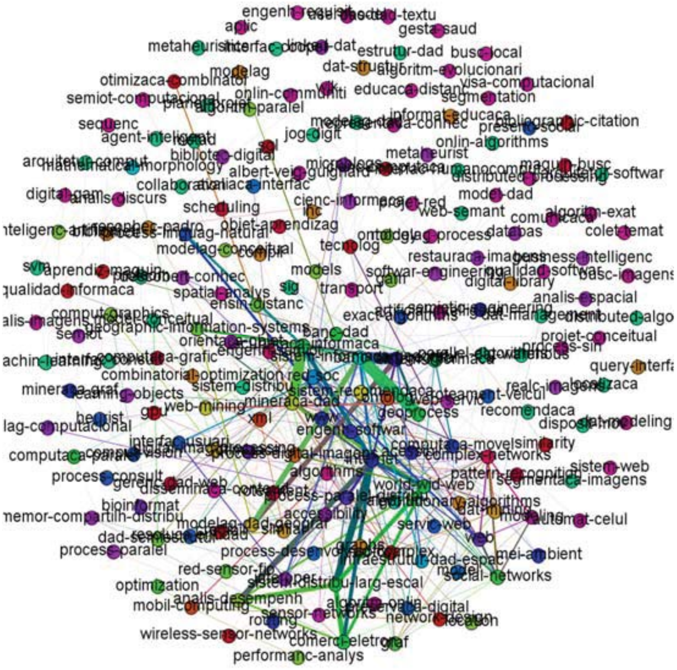


Fig. 5 Network of words with edge weight and type of publication

Fig. 5 shows a network of keywords of curricula of the group of researchers of the Graduate Program in Mathematical and Computational Modeling of CEFET-MG, and their relations, these relations are represented by the words that were used in the same article. The thickness of the edges indicates the number of publications in which the words appear in the same publication and the color of the nodes indicates the different types of publications in which the words were published. Given this, it is possible to observe the words used together with greater frequency and the words that are not used with others in chapters of books, articles in congress annals, magazines, among others. That is, sparser edges mean a greater degree of co-occurrence.

Another example of a network that can be generated is the network in Fig. 6. In this network, we represent keywords and edges when two or more keywords appeared in the same job. However, the color of the nodes represents the research area in which these words were used. Given this, it is possible to identify the keywords that are being searched in a certain area of knowledge, as well as how these words are related.

The network of Fig. 7 indicates keywords and their relationships where the isolated nodes of the network represent words that were used in isolation. Therefore, it is possible to identify underused words (minor nodes) and words that have no correlation with any other words. Therefore, in an analysis of words that most correlate, the isolated words could be isolated, reducing in this way the number of nodes to be analyzed.

V. Conclusion

With keyword analysis of scientific publications, a number of relevant information can be extracted that can help in understanding which research topics are evolving and therefore should receive more attention.

The method proposed in this work analyzes all the keywords that compose a publication and the construction of a keyword graph is performed by insertion of clicks, and after the construction of the graph, social network analysis metrics are applied to identify information Relevant to the understanding of these networks. In addition, it is possible to obtain knowledge about research topics in which certain groups of researchers have directed their efforts and how these topics have been investigated in several areas of research.

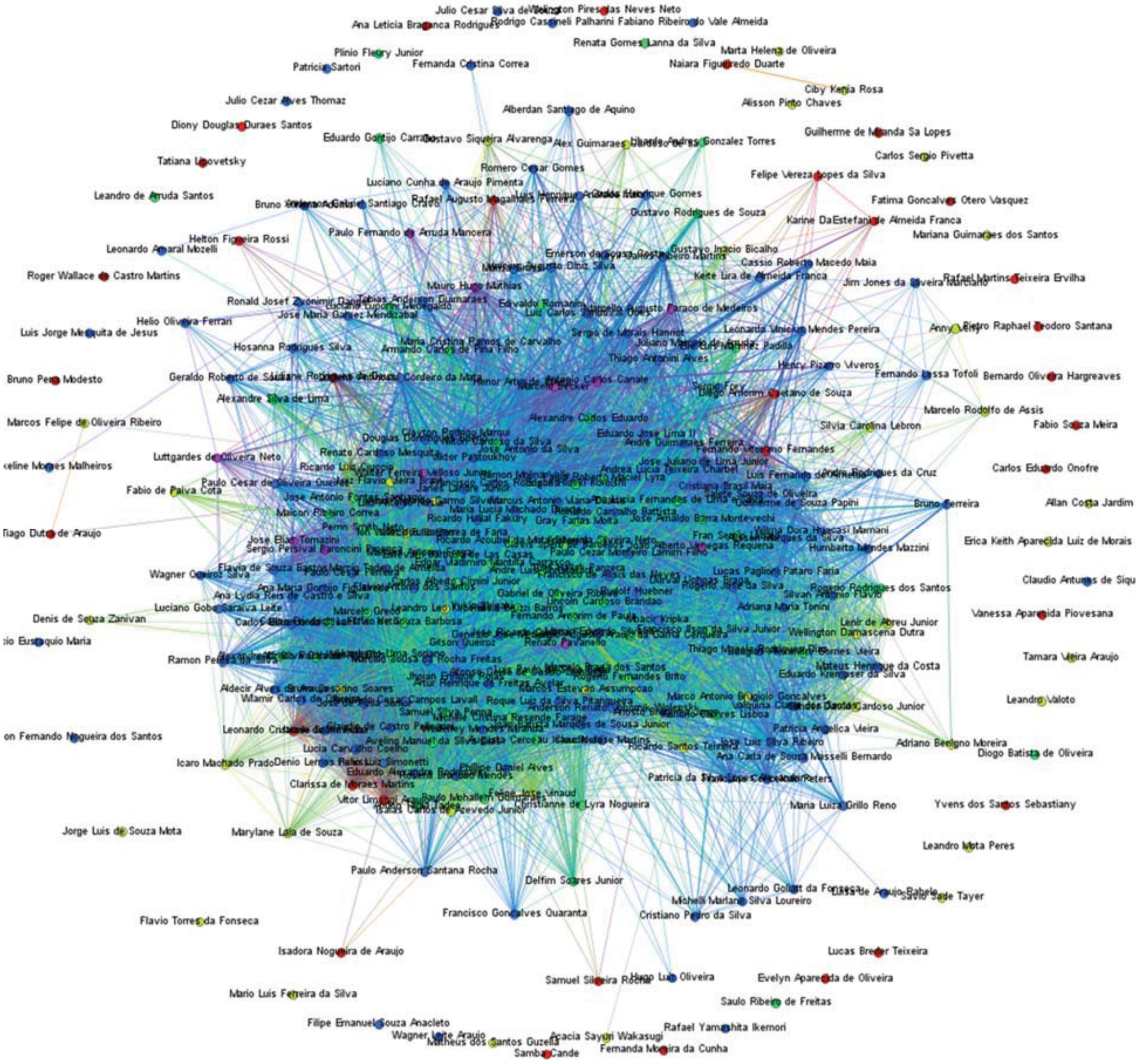

Julio Cesar Silva de Souza
Wellington Pires das Neves Neto
Ana Leticia Braganca Rodrigues
Rodrigo Casaneli Palharini Fabiano Ribeiro do Vale Almeida
Renata Gomes Lanna da Silva
Marta Helena de Oliveira
Plinio Fleury Junior
Naiara Figueiredo Duarte
Patricia Sartori
Fernanda Cristina Correa
Ciby Kenia Rosa
Julio Cezar Alves Thomaz
Alisson Pinto Chaves
Aberdan Santiago de Aquino
Diony Douglas Duraes Santos
Guilherme de Miranda Sa Lopes
Eduardo Gontijo Carrano
Gustavo Siqueira Alvarenga
Alex Guimaraes
Gonzalez Torres
Carlos Sergio Pivetta
Tatiana Lipovetsky
Romero Cesar Gomes
Felipe Vereza Lopes da Silva
Luciano Cunha de Araujo Pimenta
Leandro de Arruda Santos
Fatima Goncalves Otero Vasquez
Gustavo Rodrigues de Souza
Karine DaEstefania de Almeida Franca
Mariana Guimaraes dos Santos
Paulo Fernando de Arruda Mancera
Helton Figueira Rossi
Leonardo Amaral Mozelli
Gustavo Inacio Bicalho
Cassio Roberto Macedo Maia
Roger Wallace de Castro Martins
Keite Lira de Almeida Franca
Jim Jones da Silveira Marciano
Rafael Martins Teixeira Ervilha
Mauro Hugo Mathias
Ronald Josef Zvonimir Dantes
Helio Oliveira Ferrari
Luis Jorge Mesquita de Jesus
Anny Vieira
Pietro Raphael Teodoro Santana
Hosanna Rodrigues Silva
Henry Pizarro Viveros
Bruno Pena Modesto
Geraldo Roberto de Sousa
Fernando Lessa Tofoli
Bernardo Oliveira Hargreaves
Alexandre Silva de Lima
Marcelo Rodolfo de Assis
Marcos Felipe de Oliveira Ribeiro
Silvia Carolina Lebron
Fabio Souza Meira
Alexandre Carlos Eduardo
Luttgardes de Oliveira Neto
Carlos Eduardo Onofre
Fabio de Paiva Cota
Luis Fernando
Allan Costa Jardim
Bruno Ferreira
Tiago Dutra de Araujo
Erica Keith Aparecida Luiz de Morais
Humberto Mendes Mazzini
Wagner Queiroz Silva
Claudio Antunes de Siqu
Lucas Paglioni Pataro Faria
Rogerio Rodrigues dos Santos
Vanessa Aparecida Piovesana
Denis de Souza Zanivan
Luciano Gobo Saraiva Leite
Lenir de Abreu Junior
Wellington Damascena Dutra
Ramon Pereira
Tamara Vieira Araujo
Mateus Henrique da Costa
Aldecir Alves
Leandro Valoto
Leonardo Cristiano
on Fernando Nogueira dos Santos
Patricia Angelica Vieira
Adriano Benigno Moreira
Diogo Batista de Oliveira
Icaro Machado Prado
Denio Lemos Panisset
Ana Carla de Souza Masselli Bernardo
Clarissa de Moraes Martins
Patricia da Silva
Maria Luiza Grillo Reno
Yvens dos Santos Sebastiany
Jorge Luis de Souza Mota
Marylane Lala de Souza
Christianne de Lyra Nogueira
Leandro Mota Peres
Paulo Anderson Santana Rocha
Delfim Soares Junior
Leonardo Goliatt da Fonseca
Micheli Mariane Silva Loureiro
Savio Sade Tayer
Francisco Goncalves Quaranta
Cristiano Pedro da Silva
Flavio Torres da Fonseca
Lucas Breder Teixeira
Isadora Nogueira de Araujo
Samuel Silveira Rocha
Hugo Luis Oliveira
Evelyn Aparecida de Oliveira
Saulo Ribeiro de Freitas
Mario Luis Ferreira da Silva
Rafael Yamashita Ikemori
Filipe Emanuel Souza Anacleto
Wagner Leite Araujo
Matheus dos Santos Guzella
Acacia Sayuri Wakasugi
Fernanda Moreira da Cunha
Samba Cande

Fig. 6 Network of authors (vertices) by words (edges)

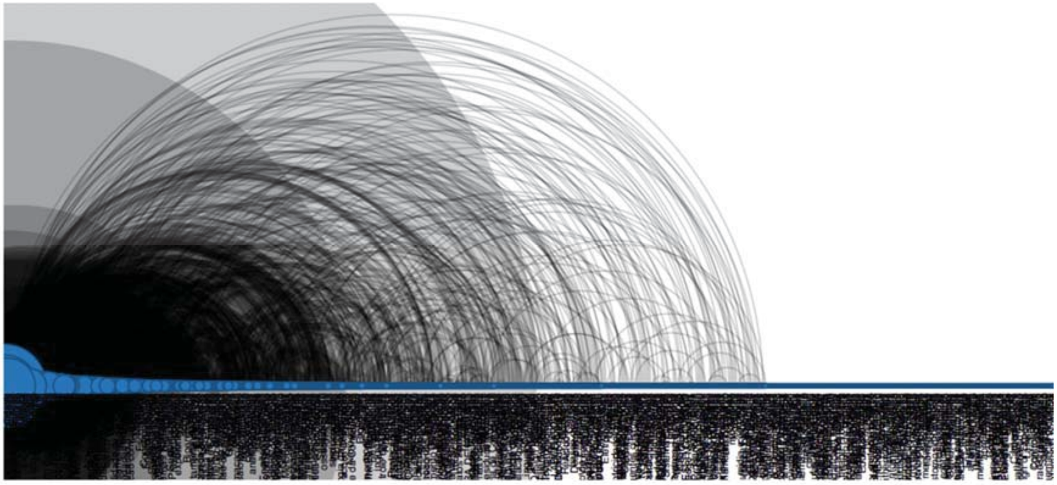


Fig. 7 Keyword network with isolated nodes

Acknowledgment

The authors are grateful for the support of CEFET-MG, for this work.

References

[1] Ding, Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. J. Informetrics, v. 5, n. 1, p. 187-203, 2011. In: < http://dblp.uni-trier.de/db/journals/joi/joi5.html#Ding11 >.

[2] Alves, A. D.; Yanasse, H. H.; Soma, N. Y. Perfil dos bolsistas pq das áreas de engenharia de produçao e de transportes do cnpq: enfoque na subárea de pesquisa operacional. XLIII Simpósio Brasileiro de Pesquisa Operacional, 2011a, Ubatuba, SP, Brasil.

[3] Brandão, M. A.; Moro, M. M. Recomendação de Colaboração em Redes Sociais Acadêmicas Baseada na Afiliação dos Pesquisadores. SBBD - Simpósio Brasileiro de Bancos de Dados, 2012, São Paulo, Brasil.

[4] Lopes, G. R. et al. Ranking Strategy for Graduate Programs Evaluation. ICITA 2011 Journal of Information Technology and Applications, 2011, Sydney, Australia.

[5] Júnior, P. S. P.; Laender, A. H. F.; Moro, M. M. Analysis of Network Co-authoring the Brazilian Symposium on Databases. SBBD - Simpósio Brasileiro de Banco de Dados, 2011, Florianópolis, Brasil.

[6] Cañibano, C.; Bozeman, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. Research Evaluation, v. 18, n. 2, p. 86-94, 2009. ISSN 0958-2029.

[7] Petersen, A. M. et al. Persistence and uncertainty in the academic career. Proceedings of the National Academy of Sciences, v. 109, n. 14, p. 5213-5218, 2012. ISSN 0027-8424.

[8] Cataldi, M.; Caro, L. D.; Schifanella, C. Emerging topic detection on Twitter based on temporal and social terms evaluation. Proceedings of the Tenth International Workshop on Multimedia Data Mining. Washington, D.C.: ACM: 1-10 p. 2010.

[9] Zhu, D. et al. Small-world phenomenon of keywords network based on complex network. Scientometrics, v. 97, n. 2, p. 435-442, 2013/11/01 2013. ISSN 0138-9130. In: < http://dx.doi.org/10.1007/s11192-013-1019-3 >.

[10] Dias, T. M. R. et al. Modeling and Characterization of Scientific Networks: A Study of the Lattes Platform. Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2013, Maceió, Brasil.