11th World Congress on Computational Mechanics (WCCM XI)
5th European Conference on Computational Mechanics (ECCM V)
6th European Conference on Computational Fluid Dynamics (ECFD VI)
E. Oñate, J. Oliver and A. Huerta(Eds)

IDENTIFYING RELEVANT KEYWORDS IN SCIENTIFIC COLLABORATION NETWORKS

THIAGO M. R. DIAS [1], GRAY F. MOITA[1]

[1] CEFET-MG - Federal Center for Technological Education of Minas Gerais
Av. Amazonas, 7675, Nova Gameleira, 30510-000
Belo Horizonte, MG, Brazil
e-mail: thiagomagela@gmail.com and gray@dppg.cefetmg.br

Key Words: *Extraction and data integration, Information Retrieval, Scientific Collaboration Networks.*

Abstract. The analysis of titles and abstracts of scientific publications has been the focus of studies of several works that aim to understand what are the main interests for research groups and certain areas of research. This analysis becomes important because it is possible to analyze research topics being studied and which are the focus of interest. In this work, a proposal for analysis of keywords from scientific publications using techniques of social network analysis is performed. For this, all keywords of the publications under review are inserted into a graph with a click, and after the construction of the entire graph and application of metrics, diverse information like connected words, the relationship between words and distance between words can be obtained .

1 INTRODUCTION

The graphs or networks are powerful tools that allow abstractions encode relationships between pairs of objects, in which vertices represent objects and edges the relationships. In some cases the vertices and edges correspond to physical objects in the real world, in others, the vertices are real objects while edges correspond to intangible relationships, and there are still cases where vertices and edges are pure abstractions [1].

In transport networks, for example, the route map used by an air carrier naturally forms a graph where the vertices are airports, and there is an edge between two vertices if there is a direct flight between two airports. Already in communication networks, a set of computers connected by a communication network can be modeled as a graph, where each vertex represents a computer and edges represent physical connections between them [1].

Among the various types of networks, there are social networks. A social network is a set of people or groups who have some kind of relationship between them [2].

In Freire [3], discloses that relationships between people can be friendship, kinship or collaboration (e.g., in an article co-authors). In a social network of friendship, the relationship between two people can represent a friendship between them. In a network of kinship relationships between people can indicate that two people belong to the same family.

In the scientific domain, an example of a social network is the scientific collaboration network that is observed as a graph in which the vertices correspond to the authors of specific scientific publications and the lines or edges correspond to co-authorship relationship. In this type of network, the edges mayor may not be weighted. The addition of weight represents the number of joint papers in which the authors connected by the edge under analysis have participated.

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 and communications to sociology and epidemiology.

A network can be characterized as a graph that consists of a set of nodes (vertices) and links (edges) between the nodes. These links can be either directed or not directed, and can optionally have an associated weight. Many, perhaps most, natural phenomena can usually be described in terms of a network. The brain may be characterized as a network of neurons connected by synapses. The Internet is also an example of an important network for society today.

The abovementioned topics have been studied by several researchers; however, it was only recently that the analysis of networks has become an important area of research. This is partly due to the advancement of computers. Computers have aided in the empirical study of real networks, and have enabled researchers from different fields to conduct technical analyses of large networks.

The strong relationship between the scientific and the socio-economic domain has led to a growing interest in understanding the mechanisms involved in scientific activities. Furthermore, it has resulted in many studies that analyze the elements of its construction and the characteristics of language and discourse used in scientific communication. The ratio of collaboration between researchers has also been analyzed [4].

For Stroele et al. [5], scientific social networks are specific types of social networks that represent social interactions originating in the academic environment. These interactions usually occur through the publication of scientific articles, academic guidelines and the development of research projects. Various goals can lead to the study of scientific collaboration networks, such as recommendation of new collaborators, intensifying the collaboration, ranking of groups or individuals, or identifying groups and their characteristics.

Alternatively, this paper aims to explore the keywords of scientific publications for generating networks enabling thus apply techniques for analysis of social networks for knowledge extraction.

Techniques like to find under way and us with a greater degree of influence are used in order to identify those most impactful and influential among the words that make up the network.

2 RELATED WORK

With increasingly fierce competition among organizations and research institutions, it becomes important for members to discover potential collaborators in order to leverage the scientific production. Recent studies show that research groups with a well-connected social network science tend to be more productive [6,7].

Networks of co-authorship of a community can reveal interesting facts about them such as which groups collaborate better, the intensity of relationships between authors, or which authors work with a greater degree of collaboration. The study of networks of co-authorship can also be used to compare the patterns of collaboration between different scientific communities [8].

Canibano and Bozeman [9] have suggested that the curriculum vitae method can be used as a sufficiently comprehensive source of information in academic research, and that its usefulness has been widely explored from 2000. However, few studies have investigated the use of curricula for conducting social network analysis, whereas several others have analyzed co-authorship and the effects of scientific collaboration on the career of the researcher [10].

The study by Petersen et al. [11] highlights factors that are of great importance to academic success in scientific networks. These factors include the abundance of scientific literature that enhances the attractiveness and the size of future opportunities for employees, and the co-author collaboration network. In view of this, it is evident that further study is necessary in order to understand and analyze how scientific collaboration happens as well as to design new tools aimed at boosting scientific production.

Other proposals for social network analysis research can be seen in [12-17]. These proposals are based on the potential for mining, visualization, and structure analysis of social networks of researchers, institutions, groups, and thematic research in a particular area, from their scientific productions — especially scholarly articles produced by the researchers.

In the work of Cataldi et al. [18], they recognize this primary role of Twitter and we propose a novel topic detection technique that permits to retrieve in real-time the most emergent topics expressed by the community. First, we extract the contents (set of terms) of the tweets and model the term life cycle according to a novel aging theory intended to mine the emerging ones. A term can be defined as emerging if it frequently occurs in the specified time interval and it was relatively rare in the past. Moreover, considering that the importance of a content also depends on its source, we analyze the social relationships in the network with the well-known Page Rank algorithm in order to determine the authority of the users. Finally, they leverage a navigable topic graph which connects the emerging terms with other semantically related keywords, allowing the detection of the emerging topics, under user-specified time constraints. They provide different case studies which show the validity of the proposed approach. Figure 1.

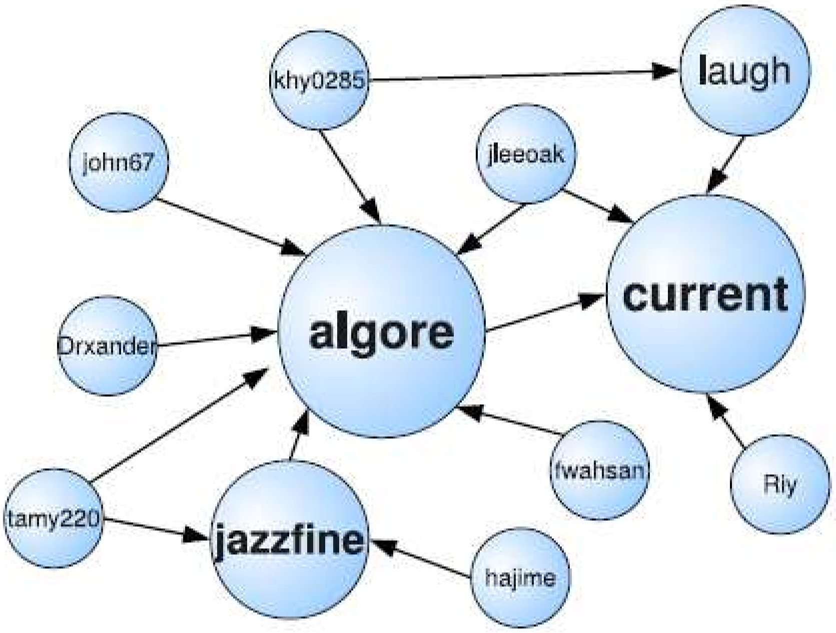

Figure 1: The size of the nodes highlights their importance in the considered community. [18]

In Zhu et al. [19], based on the network comprised of 111,444 keywords of library and information science that are extracted from Scopus, and taken into consideration the major properties of average distance and clustering coefficients, the present authors, with the knowledge of complex network and by means of calculation, reveal the small-world effect of the keywords network. On the basis of the keywords network, the betweenness centrality is used to carry out a preliminary study on how to detect the research hotspots of a discipline. This method is also compared with that of detecting research hotspots by word frequency.

3 DEVELOPMENT

Data from the CNPq Lattes Platform were used to produce this work. The Lattes Platform was conceived to integrate the information systems of Brazilian federal agencies, optimizing the Science and Technology (S&T) management process from the standpoint of both the user as well as promotion agencies and institutions of education and research. [20]

The selection of the Lattes Platform for extraction is related to it being extremely rich. This is because it deals with the integration of scientific data, both curricular and institutional in the S&T field, recording academic data and scientific production from researchers and institutions, allowing the researchers themselves to update the information. Currently, the Lattes Platform includes approximately 3 million recorded curricula vitae.

Several papers for scientific data analysis have explored the Lattes Platform as a primary source of information [21-27]

In Dias et al. [20] paper, the whole process of extraction and data integration is divided into three main parts: Extraction, Processing, and Visualization (Figure 2). However, for the purpose of our study, only the results of the extraction step that have the details of the curriculum and the subject of the research study were used.3.1 Title

The title should be written centered, in 14pt, boldface Roman, all capital letters. It should be single spaced if the title is more than one line long.

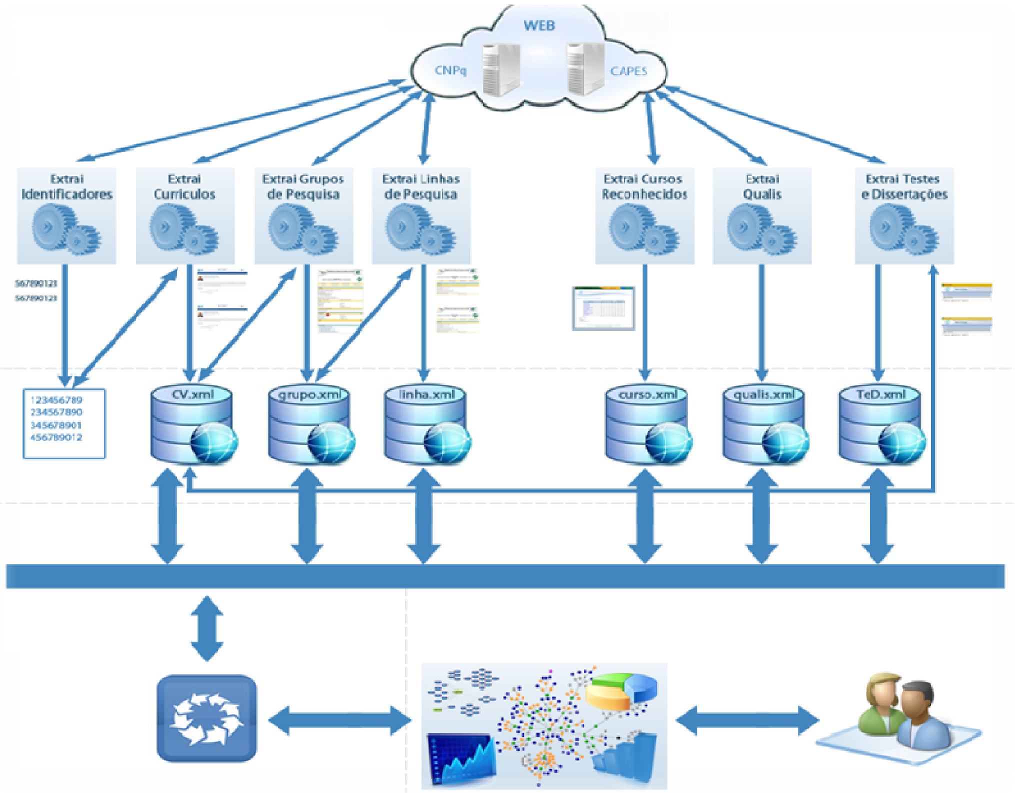

Figure 2: Framework for Extraction Lattes Platform. [20].

The data extraction process in the framework begins with the acquisition of identifiers for the Lattes curricula that have been obtained with a requisition on the platform. These identifiers are then stored locally. The acquisition strategy begins with a request that results in a list containing all the identification codes of the registered curricula.

Subsequently, a crawler collects the identifiers and generates a list of codes that will allow access to the curriculum of each individual researcher for extraction. All the extracted Lattes curricula are stored in the eXtensible Markup language (XML) format.

Subsequently, each publication registered in a curricula are analyzed and their keywords form a clique that is inserted in the graph by the juxtaposition of key words. Given this, every published study it is then inserted into the graph until all titles are analyzed and the graph is finalized. Figure 3.

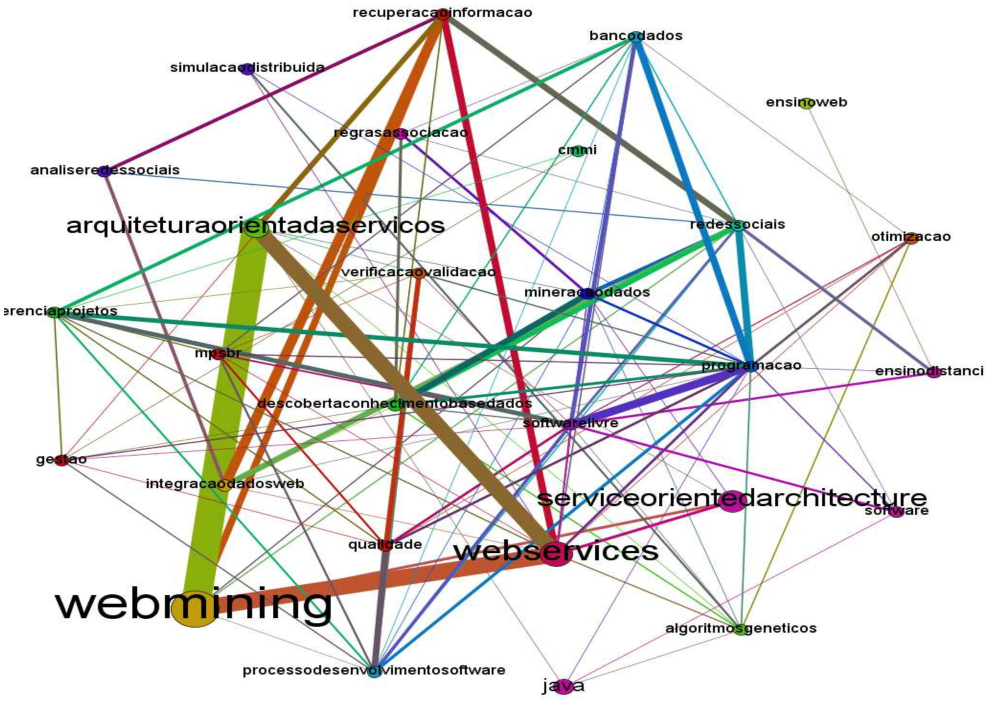

Figure 3: Example Network Keywords

With the constructed graph, where nodes represent words and edges correspond to occurrence of two or more words in the same publication, it is possible to observe those most frequent words (larger nodes) and the words appear together more often (more sparse edges).

Given this, it is possible to visually identify the most relevant words and their links to each group analyzed curricula. Thus, it is easily possible to extract what are the topics (keywords) that have major influences in the analyzed networks.

4 RESULTS

With the adoption of the metric social network analysis, you can identify features that are not visually identified. These features are important because they can reveal valuable information with the aim of boosting research based on emerging themes.

Sample results of these metrics are matrix presenting the analysis of the frequency of a particular word for each author, the distance matrix that represents the distance between a given word and another in the graph under consideration and also the matrix of neighbors in common that allows see the words that are linked with the plumb set of words. Table 1.

Table 1: Matrix of neighbors in common

	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

These matrix are important before their results allow performing work on various research area as classification and recommendation systems, which aims to classify the words have or even recommend words that can work together or new words that a specific researcher might consider in their future research.

Beyond the arrays with different metrics, other graphs can be generated spara network analysis considering key words. Figure 4.

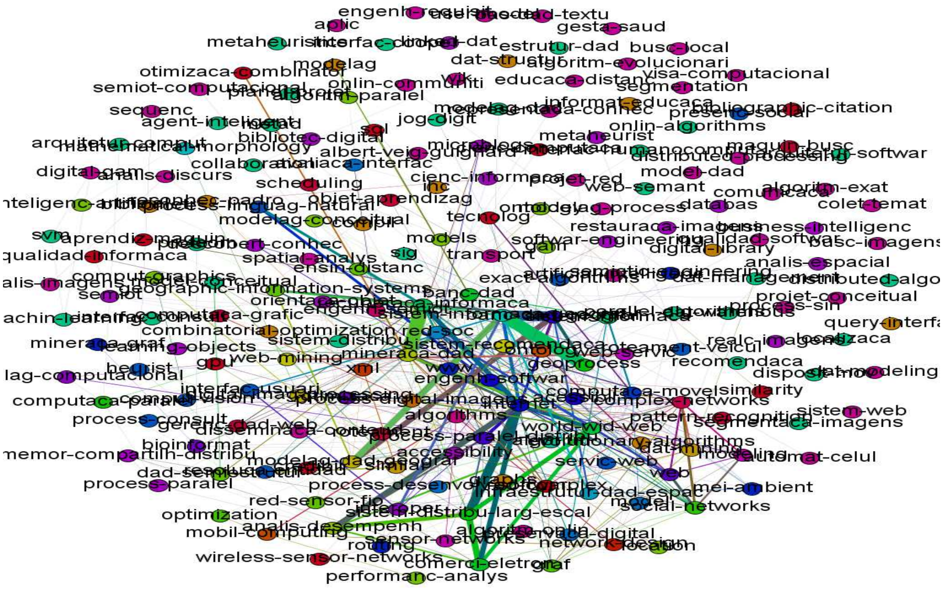

Figure 4: keyword network

Figure 4 shows a network of key words and their relationships, these relationships which are represented by words that were used in a same article. The thickness of the edges indicate the number of publications in which the words appeared in the same publication. Given this, one realizes what are the words used most often and which words are not used together in the same work.

Another example of a network that can be generated is the network of Figure 5. In this network, nodes represent authors and edges between these authors indicate keywords that the authors used in common. Therefore, it is possible to identify researchers who have worked with the same words.

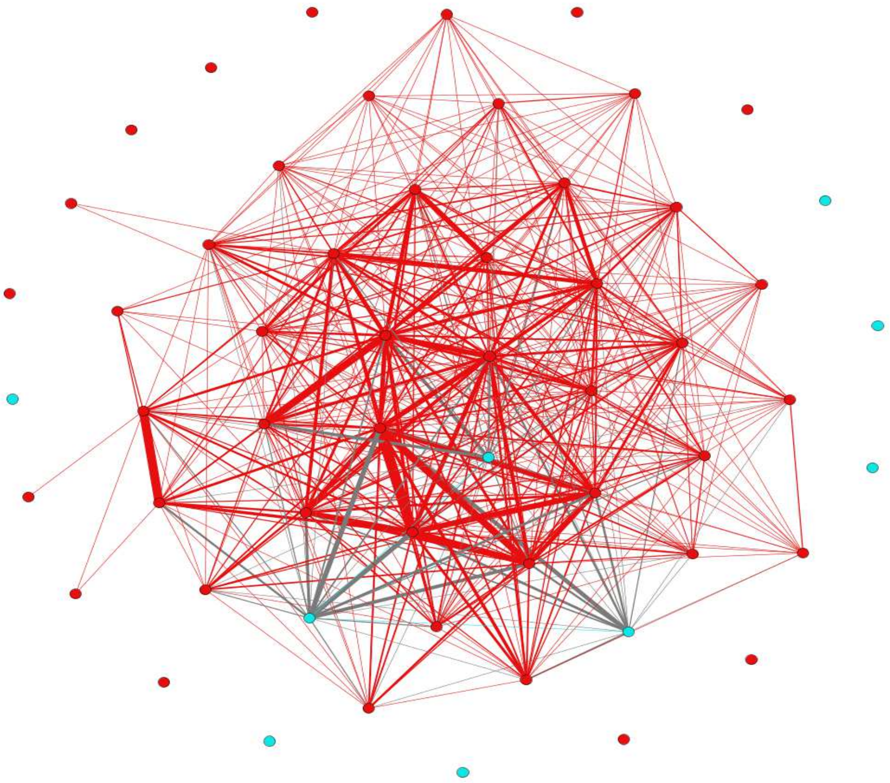

Figure 5: Network of authors per keywords

5 CONCLUSIONS

With analysis of keywords from scientific publications, one can extract various relevant information that may assist in the understanding of which research topics are evolving and thus direct research to topics that are evolving.

The method proposed in this paper analyzes all key words that make up a publication constructing a graph of keywords, and after the construction of the metrics graph of the social network analysis are applied and information relevant to understanding these networks can be obtained.

Given this, it is possible to obtain knowledge about research topics in which certain groups of researchers have directed their efforts and how these themes have been investigated in several research areas.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support provided by the CEFET-MG, FAPEMIG and CAPES during this work.

REFERENCES

[1] L. D. Nowell. and J. M. Kleinberg. The Link Prediction Problem for Social Networks. In CIKM. New Orleans,USA, pp. 556–559, 2003.

[2] M. E. J. Newman. The Structure of Scientific Collaboration Networks. Proceedings of the National Academy of Sciences of the United States of America, 98(2):404, 2001.

[3] V. Freire ano D. R. Figueiredo. Ranking in Collaboration Networks Using a Group Based Metric. Journal of the Brazilian Computer Society, 17:255-266. 2011.

[4] Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. J. Informetrics, 5(1), 187-203.

[5] Ströele, V., Zimbrão, G., & Souza, J. M. Análise de redes sociais científicas: modelagem multi-relacional. In Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), Curitiba, Brasil, 2012

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

[7] Lopes, G. R., Moro, M. M., da Silva, R., Barbosa, E. M., & de Oliveira, J. P. M. Ranking Strategy for Graduate Programs Evaluation. In ICITA - 7th International Conference on Information Technology and Application, Sydney, Austrália, 2011 (Vol. 1, pp. 253-260)

[8] Júnior, P. S. P., Laender, A. H. F., & Moro, M. M. Analysis of Network Co-authoring the Brazilian Symposium on Databases. In SBBD - Simpósio Brasileiro de Banco de Dados, Florianópolis, Brasil, 2011 (Vol. 1)

[9] Cañibano, C., & Bozeman, B. (2009). Curriculum vitae method in science policy and research evaluation: the state-of-the-art. Research Evaluation, 18(2), 86-94.

[10] Digiampietri, L. A., Mugnaini, R., & Alves, C. Analysis of Participation in supervised production of Advisors: A Case Study in Computer Science. In Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), Maceio, Brasil, 2013 (Vol. 1)

[11] Petersen, A. M., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2012). Persistence and uncertainty in the academic career. Proceedings of the National Academy of Sciences, 109(14), 5213-5218.

[12] Dias, T. M. R., & Moita, G. F. Extraction and Modeling of Scientific Collaboration Networks. In Conferência IADIS Ibero-Americana WWW/Internet, Porto Alegre, Brasil, 2013 (Vol. 1)

[13] Alvarenga, P. J. L., Gonçalves, M. A., & Figueiredo, D. R. Ranqueamento Supervisionado de Autores em Redes de Colaboração Científica. In SBBD - Simpósio Brasileiro de Banco de Dados, São Paulo, Brasil, 2012 (Vol. 1)

[14] Freire, V. P., & Figueiredo, D. R. (2011). Ranking in collaboration networks using a group based metric. Journal of the Brazilian Computer Society, 17(4), 255-266.

[15] Oliveira, J. P., Lopes, G. R., & Moro, M. M. Academic Social Networks. In International Conference on Conceptual Modeling (ER 2011), Bruxelas, Belgica, 2011 (Vol. 1)

[16] Mena-Chalco, J. P., Junior, C., & Marcondes, R. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. Journal of the Brazilian Computer Society, 15(4), 31-39.

[17] Reijers, H. A., Song, M., Romero, H., Dayal, U., Eder, J., & Koehler, J. (2009). A collaboration and productiveness analysis of the BPM community. In Business Process Management (pp. 1-14): Springer.

[18] Cataldi, Mario and Di Caro, Luigi and Schifanella, Claudio. (2010). Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. Proceedings of the Tenth International Workshop on Multimedia Data Mining. MDMKDD '10. Washington, D.C.

[19] Zhu, D., Wang, D., Hassan, S. U., & Haddawy, P. (2013). Small-world phenomenon of keywords network based on complex network. Scientometrics, 97(2), 435-442.

[20] Dias, T. M. R., Moita, G. F., Dias, P. M., Moreira, T., & Santos, L. Modeling and Characterization of Scientific Networks: A Study of the Lattes Platform. In Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), Maceió, Brasil, 2013 (Vol. 1)

[21] Fadigas, I., & Pereira, H. (2013). A network approach based on cliques. Physica A: Statistical Mechanics and its Applications, 392(10), 2576-2587.

[22] Mena-Chalco, J. P., Digiampietri, L. A., & Cesar-Jr, R. M. Caracterizando as redes de coautoria de currículos Lattes. In Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), Curitiba, Brasil, 2012 (pp. 1-12)

[23] Alves, A., Yanasse, H., & Soma, N. Extração de Informação na plataforma Lattes para identificação de redes sociais acadêmicas. In Workshop dos Cursos de Computação Aplicada do INPE, São José dos Campos, Brasil, 2009 (Vol. 9)

[24] 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. In XLIII Simpósio Brasileiro de Pesquisa Operacional, Ubatuba, SP, Brasil, 2011a (Vol. 8)

[25] Alves, A. D., Yanasse, H. H., & Soma, N. Y. SUCUPIRA: Um Sistema de Extração de Informações da Plataforma Lattes para Identificação de Redes Sociais Acadêmicas. In CISTI'2011 (6ª Conferência Ibérica de Sistemas e Tecnologias de Informação), Chaves, Portugal, 2011b

[26] Fernandes, G. O., Sampaio, J. O., & Souza, J. M. XMLattes - A Tool for Importing and Exporting Curricula Data. In WORLDCOMP'11 - The 2011 World Congress in Computer Science, Computer Engineering, and Applied Computing, Las Vegas, Nevada, USA, 2011.

[27] Arkin, A., Askary, S., Bloch, B., Curbera, F., Goland, Y., Kartha, N., et al. (2004). Web services business process execution language version 2.0. Working Draft, December.