World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:2, 2014

Identification of Coauthors in Scientific Database

Thiago M. R Dias, Gray F. Moita

***Abstract*—**The analysis of scientific collaboration networks has contributed significantly to improving the understanding of how does the process of collaboration between researchers and also to understand how the evolution of scientific production of researchers or research groups occurs. However, the identification of collaborations in large scientific databases is not a trivial task given the high computational cost of the methods commonly used. This paper proposes a method for identifying collaboration in large data base of curriculum researchers. The proposed method has low computational cost with satisfactory results, proving to be an interesting alternative for the modeling and characterization of large scientific collaboration networks.

***Keywords*—**Extraction and data integration, Information Retrieval, Scientific Collaboration.

I. 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.

A scientific collaboration network is a network where the vertices are the authors of scientific papers, and there is an edge between two authors if they published together, i.e., collaborated in the production of a scientific article [4].

For Freire [3], collaborative networks are different citation networks, in which nodes are documents and edges there is a publication quoted the other. The intensity of relationships among researchers can be measured by the total number of publications together, for example, adding weight one (1) for each publication to the edge by a pair of authors. Thus, the more these two publications the authors have set in, the greater the strength of the relationship, or the edge weight.

To analyze a collaboration network can discover many topological properties of the network, as the number of authors, the number of publications, the number of reviewers per article, the probability that two authors have a collaborator in common, the shortest path between two authors distant network and the number of connected components. You can also identify other important features that make the ranking of researchers according to their importance to a research group, country or world possible, or identify which groups of individuals in a network are more important [2], [4], [5].

For this work, the main motivation is to use a network based collaboration and publications of scientific work of researchers, identifying links with the aid of techniques for social network analysis and thus model a network characterized by collaboration scientific peer researchers can thus understand your pattern, structure, characteristics and perform the ranking of individuals within the studied collaboration network, and other analyzes.

II. Related Work

Collaboration networks have been studied in Spain and the United States in an attempt to form relationships scientific cooperation in network format, from individuals, groups and institutions, nationally or internationally [6].

A network of ancient collaboration, which is still in academic reference, is the collaboration network of the great Hungarian mathematician Paul Erdõs. Through it, one obtains the number of Erdõs of each researcher [7].

The number represents the distance Erdõs collaboration between a person and Paul Erdõs measured by authorship of academic papers.

A co-author to be awarded a number of Erdõs, must write an academic document with an author that has a finite number of Erdõs. Erdõs Paul is the only person who has a number of Erdõs zero. For any other author, the fewer Erdõs of all employees is k, then its number of Erdõs is k + 1.

According to Grossman and Ion [7] Erdõs wrote more than 1,416 scientific articles, mainly in collaboration. He had 504 direct employees. These are people with equal number of Erdõs 1. Authors who have collaborated with them (but not with Erdõs own) have a number of Erdõs 2 (6,593 people), those who have worked with people who have a number of Erdõs 2 (but not with Erdõs or anyone who owns a number of

Thiago M. R. Dias is with the Federal Centre for Technological Educationof Minas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil; (e-mail: thiagomagela@gmail.com).

Gray F. Moita is with the Federal Centre for Technological Education ofMinas Gerais, Av. Amazonas, 7675, Nova Gameleira, 30510-000, Belo Horizonte, MG, Brazil; (e-mail: gray@ppg.cefetmg.edu.br).

Erdõs 1) have a number of Erdõs equal to 3 (33,605 authors), and so on a person who does not have a path to Erdõs in collaboration network has a number of Erdõs infinity.

Newman [8], using networks with researchers from the fields of biology, physics and mathematics and trying to answer a variety of questions about collaboration patterns, found multiple results across studies of these networks. Among them, found that the number of employees in research in the area of network biology is much higher than in mathematics because of the way research (biology working with laboratory experiments with many people, and mathematics is more theoretical, few people working in a survey). Also concluded that, in recent years the number of collaborations between mathematicians has increased due to changing social organizations in the mathematical community, the emergence of better communication systems, and possible changes in the types of research questions and approaches used.

In [9], the authors analyzed the scientific literature in three different regions of the world, Brazil, North America and Europe, through collaborative networks obtained from a database of publications in Science computation, DBLP. The results obtained for different metrics indicate that the process of production of knowledge has changed differently in each region. Research is increasingly done in collaboration in different sub-areas of Computer Science. The size of the giant connected component indicates the existence of isolated collaboration groups in the European network, unlike the degree of connectivity found in Brazil and North America. Was also analyzed the temporal evolution of the social networks representing the three regions. The number of authors per article has increased over a period of 12 years. It was observed that the number of collaborations between authors grows faster than the number of authors.

Newman [4] approached metrics to measure the intensity of the relationship on scientific collaboration networks. This intensity is represented by weight the edges of the collaborative network. First showed up a simple metric that is to add weight to an edge 1 for each article that has a couple of authors together. That is, the weight of the edge is the number of authors wrote that two articles together.

In the same work in question, Newman introduced a new metric for measuring the intensity of the relationship networks of scientific collaboration, hereafter called Metrics Newman. It works as follows: each article contributed by a number of authors 1/n-1 adds to the intensity of the collaboration, or the weight of the edge, where n is the number of authors of the article.

For the present work, we used some metric of social network analysis for understanding the structure and behavior of the network as the degree of vertices, minimum paths and groupings except the process for the extraction and identification of scientific collaboration.

As the main purpose is to understand the structure and pattern of collaboration, the links (edges) were generated without assessing weights, i.e., identified a collaboration, this will not depend on the number of jobs that this pair of search features together.

III. Methodology

For the extraction and modeling of the network to be evaluated, the drilling platform and characterization proposed by Dias et al. was used [10]. The platform extracts a set of scientific data and integrates several repositories in order to generate data networks scientific publications and digital libraries.

Data were used from the CNPq Lattes Platform for this work. The Lattes Platform was designed to integrate the information systems of federal agencies in Brazil, streamlining the management process of Science and Technology (S&T), both from the point of view of the user and of the promotion and teaching and research institutions agencies.

The choice of the Lattes Platform for extraction is related to the fact that she is extremely rich, because dealing with the integration of scientific data curricula (CVs) and the area of S&T institutions, recording the academic data and scientific production of researchers and institutions, allowing the update of the data is performed by the researchers themselves. Currently the Lattes Platform has approximately 3.5 million registered CVs.

Among their data repositories is possible to find information about the higher level courses and postgraduate recognized with their respective concepts of periodic various areas knowledge with assessment level and even a bank containing information on theses and dissertations from 1987 in postgraduate programs in the country.

The whole process of extraction and integration of data is divided into three main parts called Extraction, Processing and Visualization. However, for this work only the results of the extraction step that has the curricula, study object were used.

All data extracted by the framework begins with the acquisition of identifiers for Lattes curricula obtained by the query interface of the platform later to be stored locally. The acquisition strategy begins with a consultation in order to get all the identification codes of the curricula for these in search platform interface provided by all researchers, listed in order of updating the curricula are located. With this, you can get all registered curricula, which in January 2014 totaled approximately 3,500,000.

After the query, a crawler makes all the collection of identifiers and generates a list of codes that will allow access to individual curriculum for each registered researcher. This list is stored in a file for later use extractor curricula.

Armed with a list of the file containing all the identifiers of the curriculum, the curriculum extractor allows you to select a quantity of resumes to extract, extract resumes from a certain date update, or the entire repository.

All resumes are stored on disk in XML format one by one until the end of the list sent to the extractor. In the extraction stage, several structural flaws that are in the curriculum are addressed. These very mistakes that hinder the extraction process. Examples of these problems are post- graduate courses and are reported as completed but have no year of completion, postgraduate courses in progress that do not have the starting year, published works that have no year of publication, among others. Importantly, in this case, handling exceptions were incorporated to circumvent these problems and allow files to be saved normally and with maximum consistency possible.

In addition to the curriculum, the framework also extracts data from other repositories such groups and lines of research platform, as well as data from other digital libraries in order to supplement the data curricula. Were used only however, the curriculum for this job data.

The processing steps and display responsible for carrying out the integration of the extracted sources and provide an interface for viewing the data were not used.

With all of the resumes stored in a standard format, the method is applied to the identification of scientific collaborations.

For this, all the titles of the articles registered in the curriculum of each author are analyzed and consequently become the basis of the entire construction of the collaboration network. All stages of the identification method can be seen in Fig.1.

```
Identification-Collaboration

1. n ← number of articles author
2. fori ← 1 to n
3. x ← string[i]  // x is article title [i]
4. x ← stopword[x]  // removes token without semantic value
5. x ← normalization[x]  // remove whitespace and accentuation
6. x ← lowercase[x]
7. ifhash[x] indictionary  // checks whether x is in the dictionary
8. dictionary[x] ← id_author
9. elsedictionary  ← x, id_author
```

Fig. 1 Algorithm for identification of collaboration

As shown by the Fig.1 algorithm, each under a certain curriculum undergoes a transformation that aims to achieve the same title without words with any semantic value, without any stress and without spaces. After, all text is standardized in lowercase and the resulting string is concatenated with the year of publication and this is subsequently transformed into key, steps 2-6 of algorithm.

TABLE I
Transformation of Titles in Keys

Line	Result
3	Modeling and Characterization of Scientific Networks: A Study of the Lattes Platform 2013
4	Modeling Characterization Scientific Networks Study Lattes Platform 2013
5	ModelingCharacterizationScientificNetworksStudyLattesPlatform2013
6	Modeling Characterization Scientific Networks Study Lattes Platform2013

After transformation of the title, a check is performed in order to check if this key already exists in the dictionary used for identifying collaborations. If the key already exists in the dictionary, the identifier of the originator of the curriculum in question is inserted into the switch position, otherwise the key is inserted and the handle in that dictionary. An example of dictionary can be seen in Table II.

TABLE II
Example of Built Dictionary

Key	Identifier
Modelingcharacterizationstudyscientificnetworkslattesplatform2013	Id01, Id25
Studyaboutinfluenceacademicperformancestudentsuserssocialnetworks2013	Id25, Id145, ID98
Analysiscollaborationnetworksscientificpublications2013	Id01, Id25, Id85
….	
….	
Identificationprocessreviewersscientificnetworks 2013	Id01, Id25, Id174

Importantly, the curricula registered in the Lattes Platform have a unique identifier linked to its owner. Possession of identifiers for employees of a particular author, which can be found and linked to a name for a search interface at the time of registration of a production it is possible to inquire what their co authors are with the insertion of a identification. Given this, the construction of the network, when an author to insert the identifiers of its employees into the database of a particular job, the insertion of the key that represents this work, usually happens in the dictionary; however, beyond the author's own identifier, also identifiers of its employees, which gives better results by the proposed method are inserted .

However, are not always registered identifiers of employees, this is mostly because it is not an automatic process to update the curricula, i.e., the author must bind his co-authors to their identifiers manually, which complicates this process. However, when it is possible to register the bond makes the algorithm has better performance, because even if only one of the authors has registered a particular title, you can identify collaboration, unlike the methods that work with cross- validations.

IV. Results

The network of collaboration for this work has published data curricula that have papers published in editions of CILAMCE - Iberian Latin American Congress on Computational Methods in Engineering, which is in its 34th edition, these editions in the period 1977-2013.

The CILAMCE aims to create a forum where engineers, researchers and students can exchange ideas and information about the computational methods available systems and improvements in computing technology to solve various complex problems of practical and theoretical engineering.

In all its editions, CILAMCE has played an important role in the dissemination of the latest applications and computational developments in engineering among professionals, researchers and students of Latin American Latin community.

In addition to technical sessions and tutorials, the CILAMCE also includes tutorials and invited lectures by renowned researchers from national and international community.

The base contains data as crafted article, authors of articles

International Science Index, Computer and Information Engineering Vol:8, No:2, 2014 waset.org/Publication/9997687

researchers, and institutions of origin for researchers and year of publication. Using the platform of extracting Dias et al. [10] were unable to extract the network of scientific collaboration of all authors of articles published in CILAMCE on some issue and who registered this article in your Lattes. Given this, the identification of collaboration between the authors and the general network with data extracted from the Lattes Platform was held was generated. The network has 3,442 authors and can be viewed in Fig. 2.

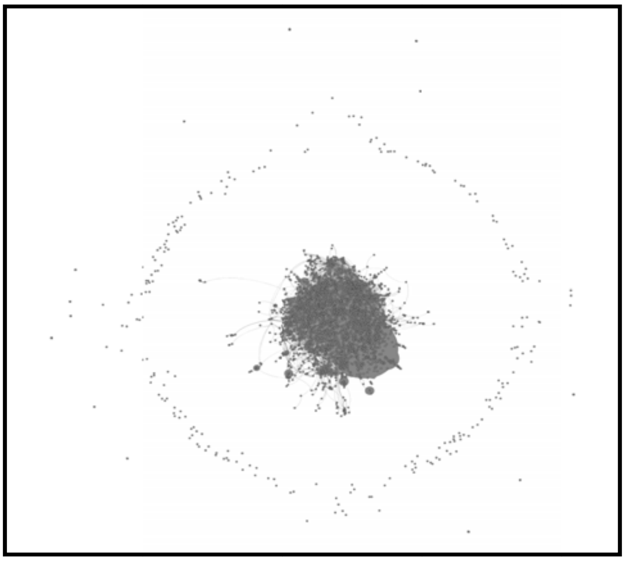

Fig. 2 Network Authors CILAMCE

You can verify that the network has a number of individual authors, featuring works that were produced individually or in partnership with other authors who are not registered on the platform for some reason, given that the network was generated only with authors with registered Lattes curricula.

However, there is a high degree of concentration among curricula that are characterized by links between nodes. Several papers have multiple authors and institutions providing different collaboration not only among researchers, but also between institutions and research groups. The average number of authors per article is 2.80 whichmake the network become well connected with over the years. In 1999 the average number of authors per article was 2.08.

The average degree is the number of employees that a given node has. Before it is possible to evaluate how an author is connected to another within this network. Although some authors not be linked to any other (isolated vertices in the network), which gives you a grade 0, others stand out for their high degree of collaboration.

In the analyzed network exists six authors with more than 60 employees. These authors are mostly titration doctors with more than 10 years, following in event publications and guidelines have completed MSc and PhD. Moreover, the authors cited only 6 1 is not linked to UFRJ and has no training as a Civil Engineering. Important to emphasize that these four authors are fellows of CNPq productivity and that together account for a total of 139 articles published in all editions of CILAMCE. The author with a greater degree (70) has a total of 27 articles, one single paper with 10 co-authors.

V.Conclusion

Relationships in scientific collaboration networks are characterized by different intensities. To perform the calculation of intensities, there are various metrics such as number of publications or number of collaborations, each one focusing on specific. The application of these metrics is possible to define how a participant or group of participants in the network is important or influential, performing a score for these criteria the following metric.

When analyzing the results after application of the network, it is possible to identify the number of publications is increasing and together they characterized the network has a high level of collaboration.

You can also identify that there are researchers who are extremely influential in the network by having a high number of publications and reviewers and thus become extremely important authors for group/institution to which it belongs, besides the network as a whole. Moreover, it is noticed that in recent years the percentage of collaboration between the authors of CILAMCE has increased considerably.

New features can be considered and aggregated in order to obtain more effective results and consider the interaction of age (age of the publication), geographic location of authors to define future events and even check how the spread of information occurs, since intuitively the information spreads faster among the vertices of higher intensity.

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] M. E. J. Newman. Co-Authorship Networks and Patterns of Scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1):5200, 2004.

[5] P. J. L. Alvarenga, M. A. Gonçalves and D. R. Figueiredo. Ranqueamento Supervisionado de Autores em redes de Colaboração Científica. Simpósio Brasileiro de Banco de dados – SBBD 2012. São Paulo – SP. 2012.

[6] J. P. M. Oliveira, G. R. Lopes and M. M. Moro. Academic Social Networks. In: ER Workshops 2011.

[7] J. Grossman, P. Ion, and R. D. Castro. "The Erdösnumberproject (2007)." Acessado em Outubro de (2007).

[8] Newman, M. E. J. centrality. Physical Review E. 64, 016132, 2001.

[9] V. S. A. Menezes, G. Z. Silva and J. M. Souza. Análise de Redes Socias Científicas: Modelagem Multi-relacional. In: CSBC 2012 - XXXII Congresso da Sociedade Brasileira de Computação, 2012, Curitiba.

[10] T. M. R. Dias and G. F. Moita,Extração e Modelagem de Redes de Colaboração Científica. In: Conferência IADIS Ibero-Americana WWW/Internet, 2013, Porto Alegre. Brazil.