Em Questão
ISSN: 1807-8893
emquestao@ufrgs.br
Universidade Federal do Rio Grande do Sul
Brasil

Rodrigues Dias, Thiago Magela; Farias Moita, Gray
A method for the identification of collaboration in large scientific databases
Em Questão, vol. 21, núm. 2, mayo-agosto, 2015, pp. 140-161
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brasil

em questão
E-ISSN 1808-5245

A method for the identification of collaboration in large scientific databases

Thiago Magela Rodrigues Dias
Doutorando; Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG);
thiagomagela@gmail.com

Gray Farias Moita
Doutor; Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG);
gray@dpg.cefetmg.br

Resumo: The analysis of scientific collaboration networks has contributed significantly to improve the understanding of the collaboration process between researchers. Additionally, it has helped to understand how scientific productions by researchers and research groups evolve. However, the identification of collaborations in large scientific databases is not a trivial task, given the high computational cost of the prevalent methods. This paper proposes a method for identifying collaborations in large scientific databases, namely, ISColl – Identification of Scientific Collaboration. Unlike methods that use techniques such as exhaustive comparisons of publication pairs, the proposed method produces satisfactory results with a low computational cost, thus providing an interesting alternative for the modelling and characterization of large scientific collaboration networks. To demonstrate the potential of the proposed technique, tests were conducted using scientific publications data registered in the Lattes Platform of CNPq, with the obtained results yielding excellent accuracy during the identification of scientific collaborations.

Palavras-chave: Extraction and data integration. Information retrieval. Identification of collaboration.

1 Introdução

The production and publication of scientific papers have increased considerably in recent years. The rapid proliferation of research publications on the Internet can be considered as the primary factor accelerating the distribution of this class of publications. Services such as digital libraries, social networks, websites and bibliographic repositories that act as a personal storehouse for an individual's scholarly or scientific productions are some examples of how the Internet has contributed considerably to the expansion of the number of published works.

Users can not only access available content but also record the technical and scientific outputs of their interactions on the Web.

In recent years, in addition to scholarly or scientific production, there has been a steady growth in the study of networks in relation to various disciplines ranging from computer science and communication 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.

Besides describing a network as a collection of vertices and edges, a network can also be characterized according to its structural and topological properties, which are mostly derived from graph theory. These structural and topological properties help to explain the structure of a network. They include the centrality of certain network nodes, the density of their relationships, their ability to interconnect and communicate, among others. One of the main uses of social network analysis is to analyze such properties. With social network analysis, we seek to understand the relationships and the flow of information between people, groups and organizations. The unit in social network analysis is not the individual, but the collection of individuals and the relationships between them (REVOREDO et al., 2012).

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 (DING, 2011).

Scientific social networks are specific types of social networks that represent the social interactions among academic. There are different motivations for the study of scientific collaboration networks, such as recommending new collaborators, or ranking and rating researchers and research groups. Collaborations among researchers is often perceived to reflect a social relation (LEE et al., 2010).

From the data available in scientific publications, it is possible to build collaborative networks in which the authors of published works are represented as network nodes, and the publications that these authors have collaborated on are represented as edges (links between the nodes).

Gayen and Chandra (2014) note that the strength of the relation is often determined by the number of publications being made by the collaborating researchers. Thus this relation among the collaborators can be represented by a weighted network, where the nodes represent the authors and the links' weights are determined by the strength of the relation.

The identification of these collaborations in large databases is not an easy task due to a number of factors, such as ambiguity in the names of authors that are inserted in the published works, the lack of citations of a particular author, typos and grammatical errors in citations, and so on.

There are several methods for identifying scientific collaborations, the most common of which are the methods that use techniques such as exhaustive comparison of publication pairs (polynomial of order 2). In exhaustive comparison of publication pairs, the titles of articles by a particular author are compared with the titles of all the articles from the other authors in the database. However, this technique has a high computational cost which prevents its use in repositories with vast amounts of data.

This paper presents an efficient method for the identification of collaboration in large amounts of scientific data. By using our method, so called ISColl (Identification of Scientific Collaboration), it is possible to achieve the

modeling and characterization of collaboration networks in repositories with large amounts of data. In our strategy, each title is parsed only once, as it is not necessary to compare with other titles for the identification of collaborations. Thus, the proposed method has a low computational cost and achieves better results than the exhaustive comparison of publication pairs in the analysis of large data repositories.

2 Related Work

In recent years, there has been considerable research into scientific collaboration networks (NEWMAN, 2001a; 2001b; 2001c; 2004). The structure of scientific collaboration networks can reveal several interesting aspects of academic communities (NEWMAN, 2004).

With increasingly fierce competition among organizations and research institutions, it becomes important for members to discover potential collaborators in order to leverage the scholarly or scientific production. Recent studies show that research groups with a well-connected social network science tend to be more productive (LOPES et al., 2011).

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 (PROCOPIO; LAENDER; MORO, 2011).

Cañibano and Bozeman (2009) 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 the year 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 (DIGIAMPIETRI; MUGNAINI; ALVES, 2013).

The study by Petersen et al. (2012) 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 collaborators, 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 scholarly or scientific production.

Other proposals for social network analysis research can be seen in Dias et al. (2013) and in Mena-Chalco and Cesar-Junior (2009). These proposals are based on the potential for mining, visualization and structure analysis of social networks of researchers, institutions and groups – especially using data from scholarly articles produced by the researchers.

Identification of contributions is a complex task mainly due to the nature of the data to be analyzed. This data usually does not have a well-defined pattern; it presents misspellings and lacks uniformity in the various ways in which an author can cite a collaborator in his work. Therefore, an efficient method to perform the identification of scientific collaborations, particularly in large databases, is required.

After the characterization of the network with all the vertices and edges, various techniques of social network analysis can be applied in order to understand how the network is structured (i.e., its topology) and how it behaves when nodes or edges are removed.

The proposed technique/system makes it possible to eliminate the necessity for human intuition in the social network analysis, which can help to automate the process of selecting potential new collaborators. Additionally, it enhances existing collaborations generated by network analysis. Our method also identifies thematic areas of research and enhances this identification with the analysis of keywords reported in published works.

In this paper, a method to identify and characterize the collaborative relationships among a set of researchers is proposed. Our method involves analyzing all the studies published by each researcher in the base to be analyzed. The publications can be of different types such as books, abstracts, and articles that are published together.

3 Methodology

Scientific collaboration networks have been the subject of many studies due to the wealth of data available in various formats and repositories. A major challenge in the analysis of these collaborative networks is the diversity in the structure and format of production repositories (LOPES, 2012; LAENDER et al., 2011).

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

The choice of the Lattes Platform for extraction is related to the fact that it is extremely rich in data. The platform integrates the scientific data Curricula (CVs) and the research areas of S&T institutions, and records academic data and scholarly or scientific production of researchers and institutions. The data is updated by the researchers themselves. Currently, the Lattes Platform has approximately 3.5 million registered CVs.

Several papers for scientific data analysis have explored the Lattes Platform as a primary source of information (DIAS et al., 2013; ALVES; YANASSE; SOMA, 2011; FERNANDES; SAMPAIO; SOUZA, 2011).

Although the data from the Lattes curricula is freely available, the CVs are accessed by a query that displays them individually via an interface. Hence, techniques and tools for extracting and integrating the data with other scientific databases that complement the information are necessary.

The integration with external sources, such as repositories of journals and reviews of postgraduate courses, is important to complement the data reported in the curricula. Since the goal of this study is the identification of collaboration among the authors that belong to a given network, the framework for extraction and integration of scientific data developed by Dias et al. (2013) was used to obtain the data to be analyzed.

In Dias et al. (2013), the whole process of extraction and data integration is divided into three main parts: Extraction, Processing and Visualization. However, for the purpose of our current study, only the results of the extraction step that have the details of the curriculum and the subject of the research study were used.

The data extraction process in the framework begins with the acquisition of identifiers for the Lattes curricula that have been obtained with a requisition to 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 eXtensible Markup Language (XML) format.

Scientific collaboration can be measured in several ways, e.g., published articles, research projects or guidelines. In this work only papers published in collaboration were considered.

After all the resumes are stored in a standard format, the proposed method is applied to identify scientific collaborations. In this method, all the titles of the articles registered in the curriculum of each author are analyzed, and they become the basis of the entire construction of the collaboration network. The steps involved in the identification process are listed in 1.

Algorithm 1 - Algorithm for identification of collaboration

```
Identification-Collaboration (list-of-publications)

  // Each publication have an id_author
 // Co-author bound have an id.
//Each publication is concatenated with the year of publication

1.  n ← number of articles author
2.  for i ← 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.    if hash[x] in dictionary  // checks whether x is in the dictionary
8.      dictionary[x] ← id_author
9.    else dictionary  ← x, id_author
10.   return: Adjacency_matriz
```

Source: the author.

As shown in the algorithm for identification of collaboration, each registered title of a study in a particular curriculum undergoes a transformation process that strips the title of accentuation, spaces and words that have no semantic value. The strategic objective of the algorithm is to minimize typos and grammatical errors that may be present in the titles of the articles. Consequently, all the text is standardized in lowercase. The resulting string is concatenated with the year of publication and is subsequently transformed into a key that represents the work under review (lines 2-6 of the algorithm). An example of this transformation is shown in Table 1.

Table 1 - Transformation of titles into keys

Line	Result
3	Modeling and Characterization of Scientific Networks: A Study of the Lattes Platform 2013
4	Modeling Characterization Scientific Networks: A Study the Lattes Platform 2013
5	ModelingCharacterizationScientificNetworksAStudytheLattesPlatform2013
6	modelingcharacterizationscientificnetworksastudythelattesplatform2013

Source: the author.

After transformation, the key is inserted in the dictionary that is used for the characterization of the collaboration network. If the key already exists in the dictionary, the identifier of the originator of the curriculum in question is linked to the key; otherwise, the key is inserted and becomes index in the dictionary. An example of a dictionary is given in Table 2.

Table 2 - Example dictionary

Key	Author
modelingcharacterizationstudyscientificnetworkslattesplatform2013	Id01, Id25
studyaboutinfluenceacademicperformancestudentsuserssocialnetworks2013	Id25, Id145, Id98
analysiscollaborationnetworksscientificpublications2013	Id01, Id25, Id85
…. ….	
identificationprocessreviewersscientificnetworks2013	Id01, Id25, Id174

Source: the author.

Importantly, each of the registered Lattes CVs owns an unique identifier. This identifier is used to tag the user and the platform, and also to allow access to each individual user's CV.

A problem encountered by the authors in the registration of a publication is a difficulty related to the registration of the collaborators. Due to a lack of standardization and the possibility of ambiguous citations in which two or more

collaborators can use the same citation name, one cannot ensure that the name entered in the collaborators list belongs to only one user of the platform. This makes the characterization of the collaboration network by the names in citations difficult.

To overcome this problem, the Lattes Platform allows an author to link his co-authors to their identifiers. Instead of simply entering the name of a collaborator, the unique code attached to the name of this collaborator is entered. Hence, in addition to the citation name of a particular contributor, the identification code is also embedded in the registered publication.

However, these links are not always possible to establish because the updating process of the curricula is not automatic, i.e., the author must manually bind his co-authors to their identifiers. But, when the co-authors' identifiers are registered, ISColl displays even better performance than the other methods that are usually used to identify collaboration. That is because collaboration can be identified even if only one of the authors has registered a particular title, in contrast to methods that use exhaustive comparison of publication pairs.

4 Results

As proof of concept, groups of researchers who have registered in the platform were selected in order to verify the efficiency of ISColl in identifying collaborations.

The metrics adopted to verify the results were based on the principle of precision and recall. These metrics are commonly used in Information Retrieval (IR) systems.

The IR metrics used are as follows (BAEZA-YATES; RIBEIRO-NETO, 2013):

Recall ratio: The fraction of relevant documents retrieved.

$$Recall = \frac{|R \cap A|}{|R|} \quad (1)$$

Precision ration: The fraction of documents retrieved that are relevant.

$$Precision = \frac{|R \cap A|}{|A|} \quad (2)$$

When performed an X query, R is the set of documents relevant to X, and A is the answer set for a query in X obtained by a recovery algorithm, where R ∩ A is the intersection between the sets R and A, i.e., the relevant documents (R) that were recovered.

Equations (1) and (2) are used to identify the set of scientific collaborations to be analyzed; Recall evaluates the fraction of actual collaborations identified, and Precision assesses whether the fraction of collaborations identified are true. Therefore, true identifications feature collaborations that really exist. The actual data for comparison were manually calculated.

In order to evaluate our method, four groups of researchers affiliated to Federal Post Graduate Programs in Brazil were selected. These researchers are faculty members who keep their CVs updated. The choice for such groups was based upon their expertise fields, recognition in these fields and high publication rate. The groups were selected to evaluate researchers in different areas based on the assumption that the titles of articles in different fields of research have different standards.

In order to obtain a baseline for comparison of results, the same groups of researchers were also analyzed with another method that identifies collaborations. Thus, it was possible to compare the results between the two different methods of identification. The Levenshtein distance was chosen for the comparative analysis.

The Levenshtein distance determines the similarity between two strings. The Levenshtein distance is calculated as the minimum number of insertions, deletions, or substitutions of characters required to transforming one string into another. In the method, two publications are considered equal if the percentage of similarity between their titles is greater than a threshold value. The values used in this study were 80% and 90%.

Several studies have used Levensthein distance to calculate similarities between titles of publications, and thereby characterize scientific collaborations. In Digiampietri et al. (2014), the authors used a rate of 90% similarity to consider two titles as a publication; although more restrictive, the authors aimed at increasing the accuracy rates. Already in Digiampietri et al. (2012), the proposed approach uses Levenstheim distance combined with the specific characteristics of publications, such as year of publication and number of pages in order to increase their accuracy rates.

The evaluation metrics (Precision and Recall) for the four chosen research groups are shown in Table 3.

Table 3 - Comparative analysis for identifying collaboration

	GROUP 01 – AREA: AGRONOMY – 12 researchers		
Metric	Levenshtein 80%	Levenshtein 90%	ISColl
Precision	100%	100%	100%
Recall	94.73%	92.5%	94.73%
	GRUPO 02 – AREA: MEDICINE – 13 researchers		
Metric	Levenshtein 80%	Levenshtein 90%	ISColl
Precision	100%	100%	100%
Recall	92.10%	91.3%	100%
	GRUPO 03 – AREA: SOCIAL SCIENCES – 38 – researchers		
Metric	Levenshtein 80%	Levenshtein 90%	ISColl
Precision	100%	100%	100%
Recall	94.44%	90.48%	100%
	GRUPO 04 – AREA: INTERDISCIPLINARY – 25 – researchers		
Metric	Levenshtein 80%	Levenshtein 90%	ISColl
Precision	100%	100%	100%
Recall	95.23%	92.37%	97.61%
	AVERAGE		
Metric	Levenshtein 80%	Levenshtein 90%	ISColl
Precision	100%	100%	100%
Recall	94.12%	91.66%	98.08%

Source: the author.

From the results in Table 3, it can be seen that, in the four networks, both methods achieved 100% accuracy for the Precision metric. This indicates that the collaborations identified were actual collaborations. In other words, all methods did not identify collaborations that were false, i.e., all the collaborations indicated do actually exist.

When converting the title of an article into a dictionary key, there is a possibility that two authors that have no cooperation are linked because they have different papers with the same title. In order to minimize this problem, ISColl inserts the year of publication at the end of the key. Therefore, the key consists of the transformed title (transformed as previously described) concatenated with the year of publication. This is the new key that is entered into the dictionary. Although this strategy does not definitively solve the problem, it significantly reduces the possibility of erroneous linking of authors who have not collaborated, since it is highly unlikely that two authors would have published articles with the same title in the same year.

Note that this problem also occurs in methods that use exhaustive comparison of publication pairs. By using techniques such as the Levenshtein distance, identical titles would be considered collaborations, even if the collaboration does not exist.

However, as it can be seen in Table 3, this situation does not occur in any of the groups. Moreover, there were no false positives identified even after conducting tests with even larger groups. Therefore, both methods achieve good accuracy results as papers with the same title by different researchers are not common in the base of CV´s used.

In the analysis of Recall, which indicates how much collaboration was identified from the actual collaborations obtained, the ISColl achieved better overall results than the Levenshtein distances method. For methods using Levenshtein distances, similarity rates indicate the proximity percentage between two strings to be considered. The higher the rate of similarity becomes, the more restrictive is the method. As shown in the Recall in Table 3, the results obtained for the 90% rate were compared to the lower percentage of 80%. Since the 80% rate presented better results in the comparison, the comparative analysis of the methods was done using the results of Levenshtein distances with 80% similarity.

Levenshtein distances achieved an average of 94.12% Recall while ISColl obtained a percentage of 98.08%. In the analysis of the individual

networks in each area, ISColl achieved better results in all groups, except for Group 01, in which the methods are equivalent.

When looking at the results of Group 01, which consisted of researchers in the area Agronomy, both methods achieved a recall of 94.73%. The methods evaluated were not able to identify two edges within all group members. ISColl was not able to identify collaborations because bonds were not formed between different authors; hence, the generated keys were similar and did not indicate collaboration.

Levenshtein distances can identify papers with divergence in titles as collaboration because of the calculation to determine the similarity between two text strings. However, two edges were not identified because only one of the authors of the paper provided the information about this publication. Hence, the Levenshtein distances method was not able to identify the collaboration.

This limitation is not a problem for ISColl. Even if only one of the authors has registered his article, ISColl links his co-authors to the key of that article and their names are linked to the article identifier.

This strategy of linking the identifiers of co-authors enabled ISColl to achieve 100% Recall in groups 02 and 03. Levenshtein distances cannot achieve comparable results due to the fact that the authors are not used in the comparison, i.e., only the actual titles are used. For instance, for medicine (Group 02), the titles of the papers are mostly textual and often very lengthy. Therefore, they are prone to typing mistakes, which makes the task of the Levenshtein algorithm even more difficult. Similar situations occurred in Group 04, where the same articles were registered with different titles and no method was able to identify collaboration. An example of a registered article with two divergent titles is given in Table 4.

Table 4 - An article registered with differing titles

A Methodology for Supporting Discourse Analysis of Multimedia Content Television
An Approach Based on Data Mining and Image Analysis to Support the Analysis of Speech Videos Televised Programs

Source: the author.

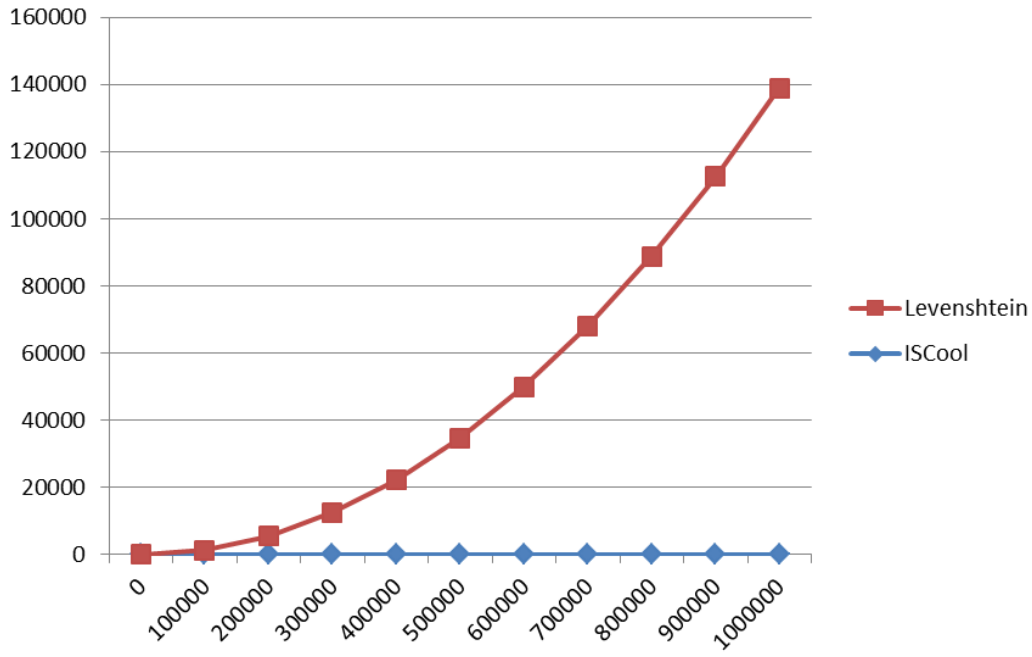
When analyzing groups of authors who work in different areas of knowledge and, consequently, have distinct patterns in the titles of their articles, ISColl produces satisfactory results when compared with other strategies that use exhaustive comparison of publication pairs.

The main advantage found in ISColl is that it can identify the collaboration even when only one of the authors registers the paper in his CV. In these cases, the collaborators' names must be tied to their identifier in the platform. However, it is important to note that even in databases with distinct characteristics, where this type of binding cannot be performed, ISColl can achieve good results at a low computational cost.

To evaluate the computational performance of Levenshtein distances, the number of comparisons required to identify co-authorship in a list of n publications is in the order of $\theta(1/2\ (n2 + n)) = O(n2)$. In practice, if n = 10,000, and each comparison is performed in 0.001 seconds, then 13.89 hours will be required to process all the group comparisons (MENA-CHALCO; DIGIAMPIETRI; CESAR-JUNIOR, 2012). This is due to the fact that Levenshtein distances detect inconsistencies in the titles with a quadratic computational cost.

ISColl has a computational cost of $\theta(n)$ comparisons, since the only comparison performed is that which is made to check the existence of the key in the dictionary. Therefore, to compare n = 10,000 publications, where each comparison takes 0.001 seconds, 0.002 hours would be necessary for the identification process, i.e., only 10 seconds. A comparative analysis of the computational cost of Levenshtein distances and ISColl for analyzing large quantities of securities is shown in Figure 1.

Figure 1 - Comparison of Computational Cost for Levenshtein distances and ISColl



Source: the author.

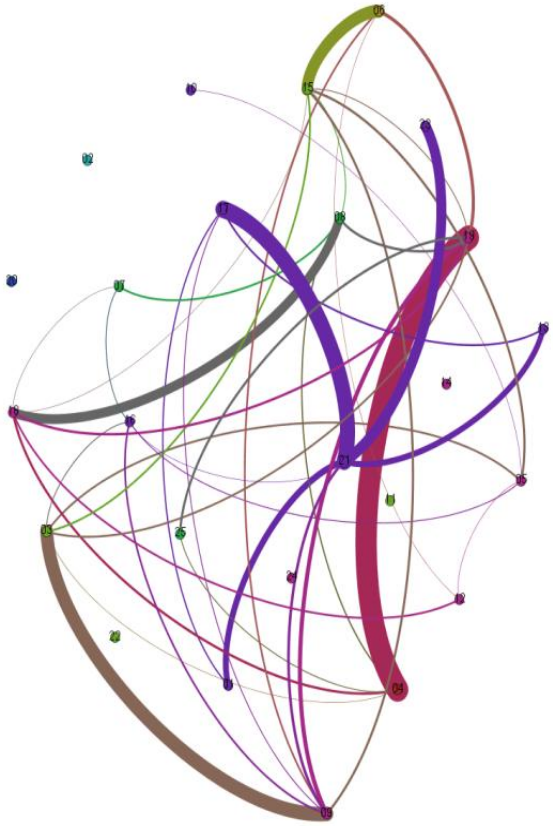
Figure 1 shows the number of hours required to perform the entire process of collaboration identification. In a database with n = 300,000 publications, where each comparison is performed in 0.001 seconds, Levenshtein distances requires 12,500 hours to perform all the comparisons. By contrast, ISColl requires only 0.08 hours, i.e., 5 minutes. Thus, we observe an increase in the computational cost as the amount of work to be scanned increases. As exhaustive comparison methods have polynomial costs, it is infeasible for them to be adopted for the analysis of large amounts of data.

Considering these results and the cost to generate large collaboration networks from the curricula of scholarly or scientific production data, ISColl presents itself as an interesting alternative for the analysis of large amounts of data with satisfactory results.

After analyzing all the article titles, a network is generated by building networks of cliques. A set of vertices is called a clique when all the vertices are interconnected. Therefore, for each dictionary key that corresponds to an article,

the elements representing the authors of the article are inserted into the network as a clique. Since a vertex to be inserted may already be present in the collaboration network, the cliques are linked by the juxtaposition of the common vertex (see Figure 2).

Figure 2 - Collaboration network, group 4



Source: the author.

After modeling the network, it is possible to identify the vertices that have greater intensity of collaboration. These vertices are characterized by thicker edges. The vertices are then sorted according to their research areas and several other features that can be drawn from the author information. It is possible to apply various metrics of social network analysis to a network

constructed in this way in order to better understand the specific characteristics of each vertex as well as the topological characteristics of the network.

5 Conclusions

The proposed method, ISColl, is used to obtain a dictionary consisting of keys, which are processed article titles linked to the identifiers of the authors of the paper. Using our technique, the network construction is performed by the juxtaposition of cliques on a collaboration graph.

ISColl presents excellent results in what concerns the Precision of the values, and also very satisfactory results when evaluated with regard to Recall. These results were obtained with four groups of researchers of distinct areas, which, in turn, reveal different characteristics in the titles of their papers.

The limitation of the proposed method is the possibility of existing two distinct publications with the same titles published in the same year. ISColl cannot identify the distinction between the publications in that case. Finally, the method has a transformation process to avoid typing erros in the registrations of the publications, but if the number of errors is large, it cannot carry the identification.

The main advantage of adopting our method is the savings made in the computational cost. As only one comparison is performed for each article title, it is possible to characterize the collaboration network with a linear complexity of value (O(n)). In contrast, methods that use other techniques, such as exhaustive comparison, have a polynomial complexity of value (O(n2)). Thus, ISColl can be used to build networks that have a very large number of authors and publications to be analyzed, making it an excellent alternative for the identification of collaborations in large amounts of data.

Referências

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. LattesMiner: a multilingual DSL for information extraction from lattes platform. In: CONFERENCE ON SYSTEMS, PROGRAMMING, AND APPLICATIONS: SOFTWARE FOR HUMANITY, 2011, Portland. **Proceedings…** New York: ACM 2011. p. 85-92.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Recuperação de informação**: conceitos e tecnologia das máquinas de busca. 2. ed. Porto Alegre: Bookman, 2013.

CAÑIBANO, C.; BOZEMAN, B. Curriculum vitae method in science policy and research evaluation: the state-of-the-art. **Research Evaluation**, Oxford, v. 18, n. 2, p. 86-94, 2009.

DIAS, T. M. R. et al. Modelagem e caracterização de redes científicas: um estudo sobre a Plataforma Lattes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2., 2013, Maceió. **Anais...** [S.l.]: UFMG, UFRJ, 2013.

DIGIAMPIETRI, L. A.; et al. BraX-Ray: an x-ray of the brazilian computer science graduate programs. **PLoS One**, San Francisco, v. 9, p. e94541, 2014.

DIGIAMPIETRI, L.; 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, 2., 2013, Maceió. **Anais...** [S.l.]: UFMG, UFRJ, 2013.

DIGIAMPIETRI, L. et al. Dinâmica das relações de coautoria nos programas de Pós-Graduação em Computação no Brasil. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2012, Curitiba. **Anais...** Curitiba: UFPR, 2012.

DING, Y. Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. **Journal of Informetrics**, Amsterdam, v. 5, n. 1, p. 187-203, 2011.

FERNANDES, G. O.; SAMPAIO, J. O.; SOUZA, J. M. XMLattes: a tool for importing and exporting curricula data. In: WORLD CONGRESS IN COMPUTER SCIENCE, COMPUTER ENGINEERING, AND APPLIED COMPUTING, 2011, Las Vegas. **Proceedings...** Las Vegas: WORLDCOMP, 2011.

GAYEN, A.; CHANDRA, J. Role of trust in evolution of scientific collaboration networks. In: INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING, 7., 2014, Beijing. **Proceedings...** New York: ACM, 2014.

LAENDER, A. et al. Ciência Brasil - the brazilian portal of science and technology. In: SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 38., 2011, Natal. **Anais eletrônicos...** Natal, 2011.

LEE, D. et al. Complete trails of coauthorship network evolution. **Physical Review E**, New York, v. 82, 026112, 2010.

LOPES, G. R. **Avaliação e recomendação de colaborações em redes sociais acadêmicas.** 2012. Tese (Doutorado em Ciência da Computação) – Curso de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

LOPES, G. R. et al. Ranking strategy for graduate programs evaluation. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY AND APPLICATIONS , 7., 2011, Sydney. **Proceedings...** Sydney: ICITA, 2011. p. 59-64,

MENA-CHALCO, J. P.; CESAR-JUNIOR, R. M. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, Porto Alegre, v. 15, n. 4, p. 31-39, 2009.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; CESAR-JUNIOR, R. M.. Caracterizando as redes de coautoria de currículos Lattes. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2012, Curitiba. **Anais...** Curitiba: UFPR, 2012.

NEWMAN, M. E. J. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences,** Washington, v. 98, n. 2, p. 404-409, 2001a.

NEWMAN, M. E. J. Scientific collaboration networks. I. Network construction and fundamental results. **Physical Review E**, New York, v. 64, n. 1, p. 016131_1-016131_8, 2001b.

NEWMAN, M. E. J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. **Physical Review E**, New York, v. 64, n. 1, p. 016132_1-016132_7, 2001c.

NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. **Proceedings of the National Academy of Sciences,** Washington, v. 101, suppl. 1, p. 5200-5205, 2004.

PETERSEN, A. M. et al. Persistence and uncertainty in the academic career. **Proceedings of the National Academy of Sciences,** Washington, v. 109, n. 14, p. 5213-5218, 2012.

PROCOPIO, S. P., LAENDER, A. H. F., MORO, M. M. Analysis of network co-authoring the brazilian symposium on databases. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 26., 2011, Florianópolis. **Anais...** Porto Alegre: SBC, 2011.

REVOREDO, K. et al. Mining scientific literature for analysis of collaboration in research communities. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2012, Curitiba. **Anais...** Curitiba: UFPR, 2012.

Um método para identificação de colaborações em grandes bases de dados científicos

Abstract: A análise de redes de colaboração científica tem contribuído significativamente para melhorar a compreensão do processo de colaboração entre os pesquisadores. Além disso, tem ajudado a compreender como as produções científicas de pesquisadores e grupos de pesquisa têm evoluído. No entanto, a identificação de colaborações em grandes repositórios de dados científicos não é uma tarefa trivial, tendo em vista o alto custo computacional dos métodos frequentemente utilizados. Este artigo propõe um método para identificar colaborações em grandes repositórios de dados científicos, denominado ISColl – Identificação de Colaboração Científica. Ao contrário dos métodos que utilizam técnicas como a validação cruzada, o método proposto produz resultados satisfatórios com um baixo custo computacional, proporcionando, assim, uma alternativa interessante para a modelagem e caracterização de grandes redes de colaboração científica. Para comprovar todo

o potencial do método proposto, são realizados testes com dados de publicações científicas da Plataforma Lattes do CNPq, obtendo excelentes resultados para o processo de identificação de colaborações científicas.

Keywords: Extração e integração de dados. Recuperação de informações. Identificação de colaboração.

Recebido: 03/02/2015
Aceito: 05/06/2015