A Computational Model to Textual Extraction And Construction of Social And Complex Networks

Patrícia Freitas Braga
Programa de Modelagem Computacional,
SENAI Cimatec,
Salvador, BA, Brazil
patyfb04@gmail.com

Hernane Borges de Barros Pereira
Programa de Modelagem Computacional,
SENAI Cimatec,
Salvador, BA, Brazil
Departamento de Ciências Exatas, UEFS,
Feira de Santana, BA, Brazil
hbbpereira@gmail.com

Macelo A. Moret
Programa de Modelagem Computacional,
SENAI Cimatec,
Salvador, BA, Brazil
Departamento de Física, UEFS,
Feira de Santana, BA, Brazil
mamoret@gmail.com

Abstract - **This work aims presenting a computational modeling to extract specific data from textual repository, in order to build social and complex networks. These networks structures are implicit in texts. This paper presents the model process, which involves text mining by regular expressions, and the construction of networks. To validate the model, an experimental procedure was applied to build scientific collaboration networks in the context of post-graduation programs.**

Keywords-component; Text Mining; Regular Expressions; Socia; Complex Networks; Scientific Collaboration Networks;

I. INTRODUCTION

Social and complex networks have topological characteristics which allow the understanding of their dynamic. The networks behavior could reflect aspects such structure composition, weaker links, centrality points, structure vulnerabilities, expansion capacity, clusters presence, and many other features that configure complex networks. In this research, the objective is to build scientific collaboration networks in productions of from post-graduation programs. The information needed to compose these networks structures was (the information) in digital texts, and due to this fact, text mining and complex networks were the two main points to the development of this model. The data to be extracted from the textual structure is very specific, and considering this, regular expressions were chosen as a text mining form.

There are some text mining and complex networks related works (e.g. [5]). However, the major part of these existent researches is grounded on semantic structures to build complex networks. Semantic structures are based on existent relationships among words. In this research, the networks structures are implicit in the texts, and they are not semantic-related. The relations are built according to a logical meaning. The motivation to the development of this model is reasoned by an academic necessity to measure researcher's relations in co-authorship in post-graduation programs. One significant work, with the same subject as related in this research [8], where he focused the scientific collaboration networks. However, he used some publication databases in a variety of fields, to obtain information about publications and authors, and their collaboration networks. Evaluating the proposed model results, it was possible to analyze the scientific productivity of researchers and higher education institutions, and understand the scientific collaboration networks dynamic, besides other aspects.

II. TEXT MINING AND REGULAR EXPRESSIONS

According to [2], Text Mining consists in extracting regularities and recognizing patterns or tendencies from large text volumes. Text mining uses non-structured information, dependent of natural language. The general absence of a well defined structure in the texts makes necessary the pre-processing of texts to normalize the data. This process is called Natural Language Processing (NLP). In a simplified form, in text mining process, the text is processed by NLP where the texts have their dimensionality reduced, and then are submitted to a statistical analyzer which will index the most frequent terms. In this research, regular expressions were used as a text mining procedure, which were based on pattern recognition in texts. Regular Expressions are based on the existence of formal patterns in texts, such as words formatting, characters arrangement, among other aspects. The problem is that not every text has well defined

978-1-4577-1133-6/11/$26.00 ©2011 IEEE

formatting structure and this is the point where the text mining complexity appears.

According to [7], regular expressions are like mathematical expressions which operates in characters sequences instead numbers. The mathematical bases of regular expressions are the Kleene's Algebra, more specifically Kleene Star, and Finite Automaton Theory. According to [9], a language is considered regular, if and only if, it is possible to build a finite automaton which recognizes the language. The Kleene star consists of a unary operation accomplished over strings or symbols. Consider *k* a set of strings, *k** the smallest subset of *k*, that contains 0 or more strings and ends with the concatenation operation between them, according to the following model:

{“a”, “c”} * = {∅, “a”, “c”, “abab”, “abc”, ...}

where ∅ represents the empty set. In this model, the regular expressions were selected as a way to extract data, whereas that the relevant data is contained in digital texts and has its willing and formatting defined.

III. Social and Complex Networks

Complex Networks have non-trivial topological characteristics which describe their behavior. Newman [6] defines a network as a set of objects called vertices, which has connections between them, called edges. By the mathematical definition, a network is represented by a graph, which is constituted by a pair of sets $G = (V, E)$, where V is a set of n vertices ($n = |N|$) and E is a set of edges that connects the elements of V (Gross;Yellen, 2004). There are some theories that explain networks behavior, but well known complex network models are random networks, proposed by Erdös and Renyi [10], small-world networks, defined by Wattz and Strogatz [3], and scale free networks, presented by Barabási and Albert [1].

In random networks, every vertex has the same probability to receive new links, i.e., the average degree <k> is fixed, and the degree distribution decays very fast. Wattz and Strogatz demonstrated the presence of clusters in the networks, characterizing the small-worlds networks [3]. In real networks, there is no uniformity in their dynamic, which is demonstrated by Barabasi and Albert, and the scale free network model were proposed [1]. In this model, the connectivity follows power law, i.e. the more connected a vertex is, more links it receives. There are some properties that characterize complex networks behavior, as following:

- **Transitivity** or Clustering – is the presence of three vertices closed cycles in a network, given by the coefficient clustering

$$Ci = \frac{2E}{k_i(k_i - 1)}$$

  where k_i é is the connectivity degree of the vertice i and $|E|$ is the total of edges between its neighbors;

- **Connectivity** – is the number of edges incident on a vertex, defined by

$$k_i = \sum_{j \in N} a_{ij}$$

  where k_i is the vertex degree, i is the vertex and a_{ij} are the edges linked to this vertex;
- **Degree Centrality** – refers to the most connected or the most influent vertices;

There are other topological properties which characterize complex networks. Some of these topological indexes were obtained from built networks, and subsidized the analysis of scientific collaboration networks and their behavior.

IV. Model General Description

This model proposes the identification of implicit networks contained into digital documents, using regular expressions as text mining form. In text mining common process, the relevant content are based in words frequency, however, in this model, the data of interest are very specific and not dependent of terms frequency. Three main steps are defined to build the networks: Text Mining, Data Insertion, and Network Building. The functional process is related in the following sequence:

1. Text Mining
   a) Selection of documents set
   b) Patterns identification
   c) Generation of data lists to insert in the database
2. Data Insertion
   a) Primary data insertion
   b) Relationships insertion
3. Network Building
   a) Filters definition
   b) Networks building and recording to .NET (Pajek) format

A. Text Mining

How to extract only specific data from the document? Regular expression uses patterns notation to be identified in the texts. This kind of information extraction enabled more specific results in the query, than the text mining common application, which is very used in content summarization processes. For instance, in the accomplished experiment, the author's names have specific formatting, like in “**BRAGA, P. F (Student)”.** In the pattern creation, it is fundamental to observe the characters of the strings, which will compound the query pattern. In this case, one possible regular expression could be:

```
[A-Z]+\,[A-Z\.\s]+\([A-Za-z\s]+\)
```

The regular expressions are based on consecutive Boolean tests taking into account the set of strings. While the condition is been satisfied, the character is added to the results of the query. This way, we could obtain only the information we need, such as author's names, paper's title, etc.

B. Data Insertion

This step contains two sub processes: the primary data insertion and the relationships insertion. The model architecture is based on researcher's social networks. The primary data defines the basic network entities (the researchers, publications, projects, etc). These data are organized as lists and taken into account for network building. After the insertion of all primary data, the next step is to relate these data lists. The network links are represented by co-authorship relations, in papers, conference proceedings, or both, or other contexts.

C. Network Building

The network building process uses a set of researchers related to scientific productions. This model contains a module to define the network scope, providing filters to allow multiple selections. The network could have a wider or narrower scope, depending on the defined filters. There is an additional module to evaluate a post-graduate program based on its scientific productivity. This module contains:

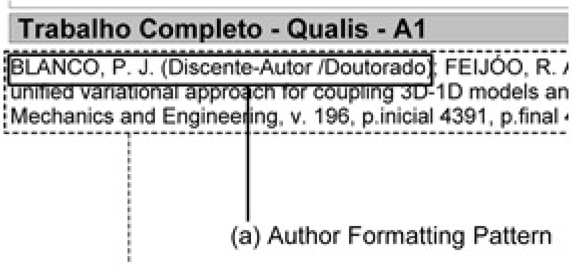
- Researchers data, which could be filtered by year, program and institution;
- Amount of researcher's publications, by qualification level (Brazilian Qualis);
- Measurement of institution productivity, calculated from the quantity of publications by year and triennium, and based on pre-defined qualifications weights, in the database management module.

The CAPES (Coordination for the Improvement of Higher Level -or Education- Personnel) uses some indexes to evaluate an institution scientific productivity, using as parameters, the number of researchers publications, which post-graduate programs have the major part of these productive researchers, the level of these publications, among others. If an institution doesn't present good results, its quality classification is lowered by CAPES. These data will favor the improvement of higher education quality, once that this quality classification could raise or lower a post-graduation program qualification.

V. Experimental Procedure – Scientific Collaboration Networks

To validate the model, CAPES indicators notes were selected as textual source, which contain information about publications during 2007 and 2008 (Fig.1.0).

Data lists were created from the text mining of these documents, such as: researchers list, papers list, conference proceedings list, chapters list, programs list and Qualis list. One aspect observed was the author's names replications in the documents. For example, BRAGA, P. F. in a document could also be written like BRAGA, P. in another one. To solve this problem, the model provides a verifying function to evaluate the author's names by a determined similarity degree. From a defined similarity percentage, the verifier function sends an acceptation or rejection message, depending of the similarity degree among the words. All lists were inserted into the database.



Trabalho Completo - Qualis - A1

BLANCO, P. J. (Discente-Autor /Doutorado); FEIJÓO, R.
unified variational approach for coupling 3D-1D models an
Mechanics and Engineering, v. 196, p.inicial 4391, p.final

Fig. 1.0 – A part of CAPES indicators note.

After observing the information contained in the texts, regular expressions were created, as the following example (Table I):

TABLE I. PATTERNS CREATED FROM THE CAPES INDICATORS NOTES

<table>
<tr><th>Use</th><th>Textual Pattern Model</th><th>Regular Expression</th></tr>
<tr><td>Pattern Authors/Vinculation</td><td>BLANCO, P. J.(Student Author/ Doctorate)</td><td>[A-ZÁ-Ú Ã-ÕÇÂ-Ô \s\-]+\,\s(?(\()|[A-Z\.\s]+) \([A-ZÁ-Ú Ã-Õ Â-ÔÇa-zá-úã-õâç0-9\.\/\-\s]+\)</td></tr>
<tr><td>Pattern Program, Institution e Year.</td><td>COMPUTACIONAL MODELING - INSTITUTION - 2007</td><td>[A-Z A-ZÁ-Ú Ã-ÕÇÂ-Ô Ç Â- Ô]+\s[A-ZÁ-Ú Ã-ÕÇÂ-Ô]+\s\/\s[A-ZÁ-Ú Ã-ÕÇÂ-Ô]+ \s\-\s[0-9]+</td></tr>
</table>

From the text mining process, the resulting lists were related to compound network structures.

- **Researcher/Program**: using the Researchers and Program lists;
- **Researcher/Paper**: using the Researchers and Papers lists;
- **Researcher/Conference Proceedings**: using the Researchers and Conference Proceedings lists;
- **Researcher/Chapter**: using the Researchers and Chapters lists;
- **Paper/Qualis**: using the Papers and Qualis lists;
- **Conference Proceedings/Qualis**: using the Conference Proceedings and Qualis lists.

After the data relationship, the networks were built based on the defined filters, by publication type, associating all the publications types and delimitating publications by quails. The Fig. 2.0 shows the co-authorship network of a post-graduate program in 2007.

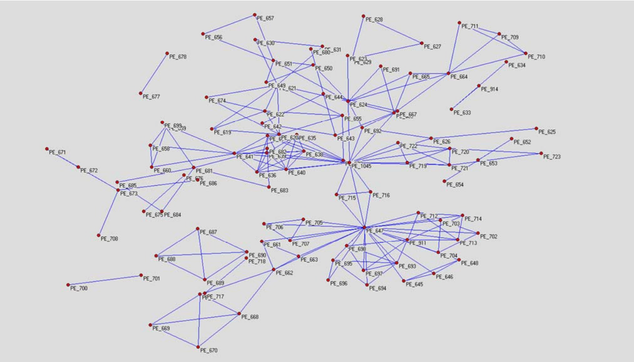


Fig. 2.0 – Co-authorship network in 2007.

The scientific collaboration networks were constructed and verified to validate the model reliability. All them were evaluated as correct. From the built networks, some topological properties were obtained, as showed in the following example (Table II):

TABLE II. INDICES OF CO-AUTHORSHIP NETWORKS IN 2007-2008. (BRAGA, 2010)

Properties	2007	2008
Vertices	107	163
Average Shortest Paths	2,0713	1,801
Clustering Coefficient	0,751	0,399
Density	0,0308	0,029
Degree Centralization	0,199	0,250

These indexes allow us to analyze the behavior of the constructed networks. Among some observations, we could notice that the knowledge diffusion occurs more easily in 2007 networks than in 2008. This is justified by some aspects as the average shortest paths and coefficient clustering. Average shortest paths indicate more articulation between the researchers in the network. That is, if there is no long distance between the researchers, the sharing of ideas in publications is facilitated, for instance. The higher coefficient clustering in 2007 infers that the researcher's networks in 2007 are more collaborative from a co-authorship perspective. Comparing the built networks with random networks with the same *n* vertices and <*k*> (average degree), it was observed that the co-authorship networks presented small-world behavior, because their coefficient clustering is higher than the coefficient clustering of the correspondent random network (C_{RD} = 0,362092 and co-authorship network C = 0,751) and average shortest path is similar (L_{RD} = 1,248 and co-authorship networks L = 2,0713).

VI. RESULTS EVALUATION AND CONCLUSION

The model critical point lies on patterns creation, in text mining step. The query precision to get better results is associated to the pattern refinement. The error in the results floats due to the regular expression dependency. Despite of the critical point, the model showed to be efficient, minimizing fully manual work on the process of networks building using texts sources. To minimize this task complexity, the model provides a module to help the user in creating regular expressions. This module is composed of buttons, each one containing embedded regular expression syntax. This means that, the user does not need to write regular expressions fully hand made. She/he does not need to know all about regular expression, she/he just has to define text formating rules and click the buttons. Because the network structures are implicit in the texts, the relationships are manually created. Other relevant aspect of the model is that the processes are centralized. Commonly, they are found separately in computational models: text mining, data management and network building.

These data from the constructed networks will subsidize the study and comprehension of the behavioral dynamic of these social networks. For instance, some information obtained from the constructed network properties revealed tendecy in collaboration between researchers from the same institutions, which means that, the knowledge sharing tends to be limited by the academic institution context. The information storage in the model database could generate other information, not only the researcher's relationships. Thus, as proposal of future activities, the use of data mining will be applied in the model to extract other implicit information from the database, such as the most recurrent subjects or other networks, such as relationships among papers, not only between researchers. From the academic assessment perspective, these data obtained by the model will favor the promoting and funding researches by research funding agencies.

REFERENCES

[1] A. Barabasi; , R. Albert, R. Emergence of scaling in random networks. Science, v. 286, 1999.

[2] C. Aranha; E. Passos . "A Tecnologia de Mineração de Textos. RESI-Revista Eletrônica de Sistemas de Informação", RESI-Revista Eletrônica de Sistemas de Informação, n. 2, 2006.

[3] D. Wattz,; S. Strogatz." Collective dynamics of small-world networks". Nature,v. 393, 1998.

[4] J. L , Gross.; J. Yellen. "Handbook of Graph Theory". New York: CRC PRESS, 2004.

[5] L. Antiqueira, et al., "Modelando Textos como Redes Complexas", XXV Congresso da Sociedade Brasileira de Computação, Jul. 2005.

[6] M. Newman, "The structure and function of complex networks", SIAM Review,v. 45, n. 2, p. 167-256, 2003.

[7] N. Good, "Regular Expressions Recipes for Windows Developers: A Problem Solution Approach", United States, Ed. Apress, 2005.

[8] M. Newman. "The structure of scientific collaboration networks". Proc. Natl. Acad. Sci. USA, v. 98, jan. 2001.

[9] P. Meneses, "Linguagens Formais e Autômatos", UFRGS, n.3, 4º Edition, Ed. Sagra, 2000.

[10] P. Erdös; A. Rényi, On the evolution of random graphs, Publications of the Mathematical Institute of the Hungarian Academy of Sciences no. 5, pp. 17-61, 1960