Qualitative and Quantitative Methods in Libraries (QQML) Special Issue Bibliometrics and Scientometrics: 71-79, 2015

Citation-Based Document Categorization: An Approach Using Artificial Neural Networks

Magali Rezende Gouvêa Meireles[1] and Beatriz Valadares Cendón[2]

[1]Associate Professor, Institute of Mathematical Sciences and Informatics, Pontific Catholic University of Minas Gerais
[2]Associate Professor, School of Information Science, Federal University of Minas Gerais

Abstract: The automatic organization of large collections of documents becomes more important with the growth of the amount of information available in digital form. This study contributes to this issue evaluating the use of Artificial Neural Networks (ANNs) to automatically categorize documents through the analysis of the references cited in these documents. The article describes the method developed to generate clusters of documents based on bibliometric concepts. The method is grounded on the premise that the presence of common citations is indicative of relationships among documents and thus publications are categorized using citations as the main input information. ANNs are typically used to solve problems related to approximation, prediction, classification, categorization and optimization. Many of the experiments reported in the literature describe the use of SOM networks, Self Organizing Maps, in the organization of documents for information retrieval. SOM networks are used in this work in order to categorize documents in a test database. In this categorization process, the semantic relationships among documents are defined not by the identification of terms in common, but by the presence of common cited references and their years of publication. After validation of the method, through the use of a prototype, a database was created, containing the references cited in 200 articles published in the *IEEE Transactions on Neural Networks Journal*, between years of 2001 and 2010. The publications were categorized by the ANN and presented in groups organized by their common citations. The results obtained show that the ANN successfully identified clusters of authors and texts, through their cited references. These clusters, formed through automatic classification of documents, evidence the existence of semantic relationships between the documents. They can be useful, for example, to automatically identify groups of researchers working in related fields or for identifying research trends in specific domains of knowledge. Another application would be in the process of information retrieval, where they could assist users in the development or reformulation of their queries.

Received: 23.1.2014 Accepted: 30.11.2014 ISSN 2241-1925

Keywords: Bibliometrics, Document Clustering, Information Science, Classification, Information Retrieval Systems, Artificial Neural Networks

1. Introduction

The brain is constantly looking for patterns and similarities in the world in a permanent effort to sort all that interacts with it. There is a natural tendency to group objects by selecting them by their common properties, and thus understand more clearly the world. The processes of categorization and classification are performed naturally by the human being in searching for the interpretation and understanding of the world.

The citations are recognized as an important source for the indication of groups that define and relate the growth of research in a particular area of expertise. According to Alvarenga (1998), the cited references in papers can reveal semantic relations among them, even without using words as thematic representations for these documents.

The purpose of this paper is to categorize documents without using words as units of representation of documents. The attributes used for the categorization of articles were the references listed in the papers analyzed and their year of publication. Artificial Neural Networks were used to categorize the papers.

The following three sections present concepts related to bibliometric techniques, to the process of categorization, and to Artificial Neural Networks. Section 5 presents the methodology and, in the section 6, the results are discussed.

2. Bibliometric Techniques

Bibliometrics offers a set of methods and measures to study the structure and the process of the scholarly communication. Many studies are devoted to statistical analysis of the digital content and try to develop quantitative assessments of the information flow.

Borgman and Furner (2002) point out that citation analysis is the most popular bibliometric approach and that it can be used to identify relationships among document regardless of the presence of equal terms in the documents evaluated. According to Guedes and Borschiver (2005), citation analysis identifies the research front of a particular scientific field through the set of authors who are cited in recent literature and estimate the immediacy factor of a published article based on the hypothesis that, in certain scientific fields, papers cited most frequently are more relevant than less cited papers.

Cronin (1984) says that the habit of quoting shows conformity and consistency in the act of the intellectual production, often governed by tacit and internalized norms. The facts cited in the text gain credibility when the literature used is referenced, which connects the reader to other sources of information on the subject. The citations denote a special relationship between the citing and the cited papers. Accordingly, Leal (2005) states that the quote can be understood as a social process when considering all the previous experience of the author, his network of knowledge and his own considerations.

3. Categorization

Historically, the notion of category has been addressed in different ways. According to Xavier (2008), the concept of categories, as discussed today, was born with Aristotle, who lived between 384 and 322 BC. "Categories" is the first of the five treatises composing the "Organon", the work that exposes the Aristotelian logic. It is assumed that this is the treaty that introduces the content of all the other four, On Interpretation, Prior Analytics, Posterior Analytics and Topics.

As highlighted by Barite (2000), Ranganathan took the concept of categorization from the field of philosophy to the classification of knowledge, and to prove that the categories are the foundation of any system of organization of knowledge, he built a system of classification, the "Colon Classification", from its theoretical postulates. According to Jacob (2004), categorization is the process of dividing the world into groups of entities whose members have similarities between them within a given context. When the individual aggregates entities into categories he perceives order in the world that surrounds him.

According to Lima (2010), Eleanor Rosch transformed categorization in a research question. Rosch has developed her work in the 70's and created the prototype model. According to this model, concepts are represented by a group of characteristics and not by the use of the definitions. The grouping of concepts in a given category would be by the similarity with the prototypes, that is, those members of the category that most reflects the redundancy of the category's structure as a whole.

4. Artificial Neural Network

An artificial neuron can be represented by a simplified mathematical model of the processes in a biological neuron. An Artificial Neural Network, ANN, can be defined as a topology of interconnected artificial neurons, in which typically input neurons, internal neurons and output neurons can be identified. The way the neurons are organized and connected depends on the network architecture. Neural networks implement algorithms which try to achieve a desired performance approaching natural neural systems through techniques such as learning experience and by generalizing from similar situations.

ANNs are used primarily in problems of approximation, prediction, classification, categorization and optimization. Meireles, Almeida and Simões (2003) point out that the vast majority of applications reported in the literature focuses on the industrial area. Souza (2006) points out that in an Information

Retrieval System, SRI, the ANNs can be used to perform pattern matching between queries and documents of the system's collection.

Networks Self Organizing Maps, SOM, used in this study, are self-organizing maps developed by Kohonen in the 80s. They are structures based on topological maps present in the cerebral cortex. SOM networks work basically building a topological map where nodes that are topologically close respond similarly to similar input patterns. Each input neuron is connected to each output neuron with its respective association weight. Braga, Carvalho and Ludermir (2000) point out that in the literature there are examples of data categorization processes using SOM networks.

5. Methodology

After validation of the method, through the use of a prototype containing 10 papers, 200 papers were selected, whose publication occurred between 2001 and 2010, using the IEEE Xplore digital library searching tool. In this retrieval process, the goal was to find the term "neural network" appearing in the titles of the papers. The papers were chosen preferably in the *IEEE Transactions on Neural Networks Journal*, since their database allowed retrieval of all data required for the composition of the database. The citations of these papers amounted to a total of 6015 references.

Two databases were prepared. The first one was composed of titles of articles, keywords, publication year, number of references cited for each paper and the numeric codes associated with each of the references cited by the paper. These cited references for each paper of the first database were recorded in the second database, which listed its numeric code, title, authors and publication year for each citation.

Once the database was built, a comprehensive preprocessing phase was accomplished to avoid inconsistencies in the database. Durães (2009) states that eighty percent of the time consumed in the pre-processing of real applications are spent on data cleaning. Missing information, incorrect or inconsistent records in the database should be corrected to avoid compromising data quality. In addition to typografical errors identified in the data, some semantic discrepancy could also be found after this process.

The process of developing a test base and the categorization of the papers can be viewed in Figure 1.

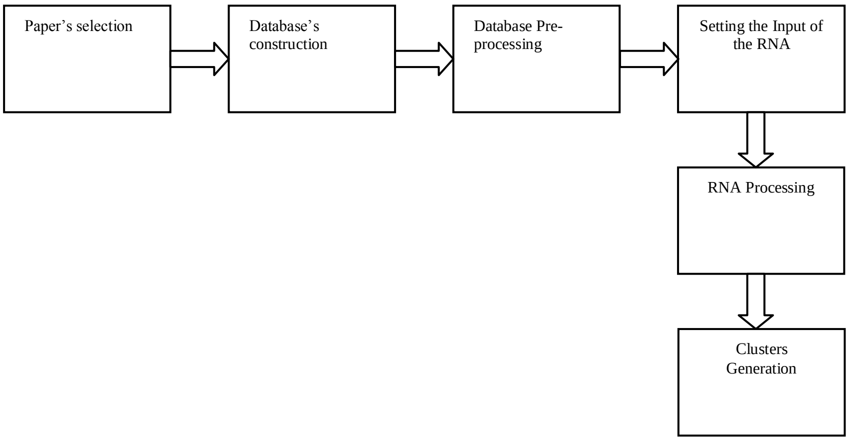

Figure 1: Representation of the methodological steps

During their recording, each one of the citations received a numeric code. As the citations were repeated in several articles, it was essential that each one was registered with a unique numeric code, even if it repeated in different parts of the database associated with different papers. To ensure this condition and assign a unique code number to the same publication, even if referenced by separate papers, a computer program written in Java was designed to find duplicates and properly conciliate them into the final database.

To generate a data sequence to be used in the input ANN, the program created a logical attribute for each paper containing the information on the presence or absence of each one of the 6015 references registered. The presence of each reference in a paper was represented by the value one in the position relative to the numeric code assigned to that citation and the absence by a value zero.

The unsupervised ANN was fed only with data relating to the citations of each paper and their respective year of publication.

6. Experiments and Discussion

During the preprocessing, several inconsistencies were identified in different instances of the titles and of the authors of the citations. As noted by MacRoberts and MacRoberts (1989), the editing errors are a technical limitation of indexes related to citations, hindering the implementation of citation analysis.

Some words were found in the singular and in the plural (e.g., inequality and inequalities) in the titles of the papers that make up the references or with the addition of a hyphen or a quotation, which were identified by the program as a difference between the titles of the works. There were also many typographic errors, as the case of two words not separated by a space or titles without one or more words (e.g., for solving monotone variational ..., or for monotone... or for solvingmonotone ...). There were also variations in the form of registering the authors' names (e.g., X. Hu and X. L. Hu) or even the absence of one or some of the authors. After identifying these inconsistencies, the titles of these papers and the authors' names were manually changed according to the information provided by the IEEE Xplore digital library, thus allowing the program to identify 100% of the similarity between same references, as expected.

After adjusting the parameters of the ANN, some tests were performed to get 4, 9, 10, 12, 16, 25 e 36 clusters of papers. The number of clusters is directly related to the topology of each ANN in the experiments. The purpose of these tests was to verify whether the topologies would be able to group the papers into similar clusters. In four of these tests, a significant part of the papers was grouped by the SOM algorithm in seven clusters, as shown in the example maps in Figure 2. These four topologies generated clusters more consistently than the other three and, therefore, only their output results were selected for analysis.

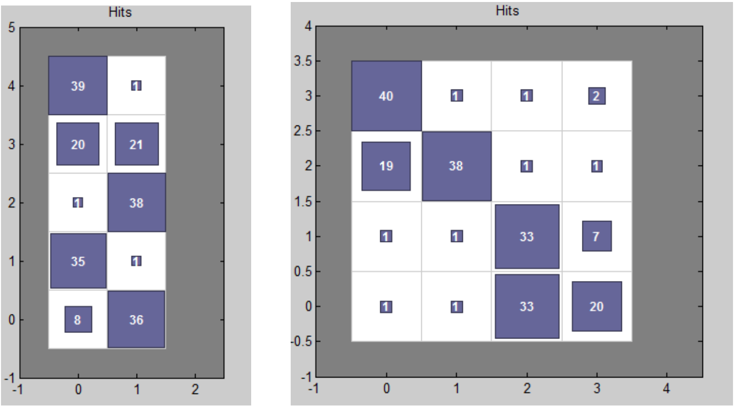

Figure 2: Categorization Maps

Each of the four topologies selected, presented a group with six or more papers which was called "GroupOne". These were coded as A, B, C and D, respectively. This GroupOne presented 8 papers in the topology of 10 clusters, 5 papers in the topology of 12 clusters, 6 papers in the topology of 16 clusters and 6 papers in the topology of 25 clusters. Table 1 presents the papers of GroupOne in each one of these topologies.

TABLE 1. The same GoupOne categorized on 4 Topologies

Number of clusters	GroupOne	Number of papers	Papers
10	A	8	1, 11, 18, 35, 47, 48, 49, 50
12	B	5	1, 47, 48, 49, 50
16	C	6	1, 6, 18, 47, 48, 49, 50
25	D	6	1, 18, 35, 47, 48, 49, 50

Papers 1, 47, 48, 49 e 50 showed up in each one of the four selected topologies and paper 18 appeared in three of the four topologies. To evaluate the results, these six papers were selected and year of publication, their citations and keywords were compared in Table 2.

TABLE 2. The relation between the six papers of the GroupOne

Papers	Citations common to 3 or more papers	Year	keyworks
1	1, 3, 4, 5, 6, 9, 11, 12, 13, 15, 16, 17, 18, 19, 20, 24, 26, 30, 33	2009	Asymptotic stability, k-winners-take-all (WTA), linear programming, neural network, quadratic programming
18	1, 3, 4, 5, 6, 9, 11, 12, 13, 15, 18, 19, 20, 26, 30	2010	Convergence, linear and quadratic programming neural network, stability
47	1, 3, 4, 5, 9, 12, 13, 15, 16, 17, 19, 20, 24, 26, 30	2008	winners-take-all (k-WTA), Global asymptotic stability, optimization, quadratic programming (QP), recurrent neural network
48	1, 4, 5, 9, 12, 18, 19, 24, 26	2008	Differential inclusion, Lyapunov stability, global convergence , hard-limiting activation function, nonlinear programming, quadratic programming,

			recurrent neural network
49	4, 6, 9, 13, 17, 18, 19, 24, 26, 33	2008	Constrained optimization, convergence, convex and nonconvex problems, recurrent neural networks
50	1, 3, 5, 9, 11, 12, 16, 17, 18, 19, 20, 24, 26, 33	2007	Global convergence, linear programming, linear variational inequality (LVI), quadratic programming, recurrent neural network

Six of these citations (6,11,15,16,30,33) were used in three papers, four citations (3,13,17,20) in four papers, six citations (1,4,5,12,18,24) in five papers and three citations (9,19,26) in six papers. These papers were published in four different years (2007, 2008, 2009 and 2010) while the totality of the papers in the database had 10 different publication years, indicating that this wasn't a predominant attribute used by the ANN in the categorization process.

The papers in Table 2 had an average of 6 keywords. The comparison of these reveal that "quadratic programming" was repeated, showing up in five of the papers, "recurrent neural network" was repeated in four papers and "neural network" in two papers.

7. Conclusions

At the point of the study, the attributes used by the ANN in the clustering process are still not clear. However, the results obtained from the analysis of the six papers more frequent in the GroupOne showed a large number of citations common to three or more of them. This can be taken as an indication of the presence of a semantic relationship among the papers although a categorization process that used keywords as the attribute for grouping wouldn't necessarily find the same result. The method can become an important alternative to information retrieval processes in many contexts, insuring more diversity into results retrieved by those classes of information systems.

The results are preliminary results of a research project which is underway.

Acknowledgments

The authors would like to thank the financial support of FAPEMIG and Pontific Catholic University of Minas Gerais (PUC Minas), and the Federal University of Minas Gerais (UFMG) for the opportunity to develop this research as well as the valuable contributions of Prof. Dr. Paulo E. M. de Almeida from CEFET-MG.

References

Alvarenga, Lídia., (1998). Bibliometria e arqueologia do saber de Michel Foucault – traços de identidade teórico-metodológica. Ciência da Informação, Brasília, Vol. 27, No. 3.

Barite, M. G., (2000). The Notion of "Category": Its Implications in Subject Analysis and in the Construction and Evaluation of Indexing Languages. Knowledge organization, Vol. 27, No. 1/2, 4-10.

Borgman, C. L. and Furner, J., (2002). Scholarly Communication and Bibliometrics. Annual Review of Information Science and Technology, Vol. 36, No. 1, 2-72.

Braga, A. P.; Carvalho, A. C. P. L. F.; Ludermir, T. B., (2000). Redes neurais artificiais: teoria e aplicações. Rio de Janeiro: LTC.

Cronin, B., (1984). The citation process. London: Taylor Graham, 103 p.

DURÃES, Rodrigo Leite., (2009). Validação de Modelos Baseados em RNA Utilizando Análise Estatística de Dados e Lógica Fuzzy. 123f. Dissertação (Mestrado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais. Belo Horizonte.

Guedes, V. L. S and Borschiver, S., (2005). Bibliometria: uma ferramenta estatística para a Gestão da informação e do conhecimento, em Sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: *Proceedings* CINFORM – Encontro Nacional de Ciência da Informação VI, Salvador, Bahia.

Jacob, E., (2004). Classification and categorization: a difference that makes a difference. Library trends, Vol. 52, No. 3, 515-540.

Leal, I. C., (2005). Análise de citações da produção científica de uma comunidade: a construção de uma ferramenta e sua aplicação em um acervo de teses e dissertações do PPGCI - UFMG. 94f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Minas Gerais. Belo Horizonte.

Lima, G. A. B. O., (2010). Modelos de categorização: apresentando o modelo clássico e o modelo de protótipos. Perspectivas em Ciência da Informação, Vol. 15, No. 2, 108-122.

Mac Roberts, M. H. and Mac Roberts, B. R., (1989). Problems of citation analysis: a critical review. Journal of the American Society for Information Science, Vol. 40, No. 5, 342-349.

Meireles, M. R. G.; Almeida, P. E. M. and Simões, M. G., (2003). A Comprehensive Review for Industrial Applicability of Artificial Neural Networks. IEEE Transactions on Industrial Electronics. EUA, Vol. 50, No. 3, 1-18.

Souza, R. R., (2006). Sistemas de Recuperação de Informações e Mecanismos de Busca na *web*: panorama atual e tendências. Perspectivas em ciência da informação, Belo Horizonte, Vol. 11, No. 2, 161-173.

Xavier, B. R., (2008). As categorias de Aristóteles e o conhecimento científico. Pensar, Fortaleza, Vol. 13, No. 1, 57-64.