Bibliometric Knowledge Organization: A Domain Analytic Method Using Artificial Neural Networks†

Magali Rezende Gouvêa Meireles*, Beatriz Valadares Cendón**,
Paulo Eduardo Maciel de Almeida***

*Institute of Mathematical Sciences and Informatics, Pontifical Catholic University of Minas Gerais,
Av. Dom José Gaspar, 500, Coração Eucarístico, Belo Horizonte, MG, Brazil, CEP 30.535-901
<magali@pucminas.br>
**School of Information Science, Federal University of Minas Gerais, Av. Antônio Carlos, 6627,
Pampulha, Belo Horizonte, MG, Brazil, CEP 31.270-901 <cendon@eci.ufmg.br>
***Computer Department, Federal Center for Technological Education of Minas Gerais, Av. Amazonas,
7675, Nova Gameleira, Belo Horizonte, Minas Gerais, Brazil, CEP 30.510-000 <pema@lsi.cefetmg.br>

Magali Rezende Gouvêa Meireles is adjunct professor at the Institute of Mathematical Sciences and Informatics of the Pontifical Catholic University of Minas Gerais (PUC Minas) in Brazil. She holds a Dr.Sc. in Information Science from Federal University of Minas Gerais (UFMG), a M.Sc. from Federal Center for Technological Education of Minas Gerais (CEFET-MG) and a B.E. in Electrical Engineering from UFMG. Her research interests are related to knowledge organization, information systems and applied computational intelligence. In 2013-2014, she was a Visiting Professor with the Faculty of Science and Engineering at Queensland University of Technology (QUT) in Australia.

Beatriz Valadares Cendón holds a Ph.D. and a MLIS from the School of Information at the University of Texas at Austin, USA, and a B.E. in Civil Engineering from the Federal University of Minas Gerais (UFMG), Brazil. Currently she is professor at the School of Information Science, UFMG. She taught at the Division of Library and Information Science of the University of South Florida, USA (1994). A researcher for the National Council for Scientific and Technological Development, Brazil, her interests are in the areas of information retrieval systems, use of electronic journals, information systems evaluation and information behavior.

Paulo E. M. de Almeida received B.E. and M.Sc. degrees in electrical engineering from Federal University of Minas Gerais, Belo Horizonte, Brazil, in 1992 and 1996, and a Dr. Eng. degree from São Paulo University, São Paulo, Brazil in 2002. He is an associate professor at the Federal Center for Technological Education of Minas Gerais (CEFET-MG), Belo Horizonte, Brazil. His research interests include computational intelligence applied to information systems, transportation, control systems, and renewable resources. In 2000–2001, he was a Visiting Scholar at Colorado School of Mines, USA. In 2013-2014, Dr. Almeida was a Visiting Professor at QUT, Australia

Meireles, Magali Rezende Gouvêa, Cendón, Beatriz Valadares, Almeida, Paulo Eduardo Maciel de. **Bibliometric Knowledge Organization: A Domain Analytic Method Using Artificial Neural Networks.** *Knowledge Organization. 41(2)*, 145-159. 35 references.

Abstract: The organization of large collections of documents has become more important with the increase in the amount of digital information available. In certain constricted domains of knowledge, keywords and subject descriptors tend to be similar and therefore insufficient to differentiate documents. In this context, instead of relying only on the presence of common terms, the identification of common cited references can be useful to define semantic relationship among documents. The purpose of this work is to add another instance on the research linking information retrieval and bibliometric techniques aided by information technology. A domain analytic method was developed to generate clusters of documents, which uses self-organizing maps, in the scope of artificial neural networks, to categorize documents. The results obtained show that this approach successfully identified clusters of authors and documents through their cited references. In addition, further qualitative analysis of these clusters demonstrates the existence of semantic relationships between the documents. This study can contribute to the development of the field of knowledge organization by evaluating the use of artificial neural networks in the automatic categorization of documents in a constricted knowledge domain based on the analysis of the references cited by these documents.

† The authors would like to thank Prof. Maurício B. Almeida for the valuable suggestions and the financial support of Pontifical Catholic University of Minas Gerais (PUC Minas), Federal University of Minas Gerais (UFMG) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES–Process BEX 18208-12-6).

Received 5 September 2013; Revised 11 December 2013; Accepted 18 December 2013

Keywords: articles, documents, categorization, references, network

1.0 Introduction

The human brain is constantly looking for patterns and similarities in the world around in a permanent effort to sort all that interacts with it. Human beings have a natural tendency to group objects by selecting them by their common properties, and thus understand more clearly the surrounding context. The processes of categorization and classification are naturally performed in searching for the interpretation and understanding of the world. With the growth of digital document collections, the need to automatically organize the information available has increased. The enormity of the collections of electronic documents has motivated the development of tools and techniques that assist an user in the organization of these collections (Hussin and Kamel 2003). Citations are recognized as an important source for the indication of research groups of authors that relate to each other and define the growth in a particular area of expertise. Starting from the assumption that the documents could be grouped, using the cited references as attributes and that the categories generated could gather closely related documents, the current work proposes a method for the categorization of articles that uses artificial neural network (ANN). Next, the context of this research and its relationship with information science (IS) is presented.

Hjorland (2002) discussed eleven approaches that have been used in IS to produce domain-specific knowledge. Three of those approaches are of special interest to this research. The first one discusses the bibliometrical studies. The author stated that they can be used as a tool in domain analysis because it is empirical and based on detailed analysis of connections between individual documents. The second approach is about constructing special classifications. According to the author, these classification systems organize the logical structures of categories and concepts in a domain as well as the semantic relationship between the concepts. The third approach is the knowledge gathered by means of domain analytic studies about scientific cognition, expert knowledge and artificial intelligence. Hjorland understands that these fields offer useful techniques, providing mental models of a domain for knowledge elicitation in order to produce expert systems, which may supplement other approaches to domain analysis in IS. According to Smiraglia (2013), domain analysis provides a set of techniques for extracting and analyzing the semantic intellectual content of coherent groups which share a set of common hypotheses, epistemological consensus on methodological approaches, and social semantics. Domain analytic methods draw out the concepts that form these components of domain coherence.

Hjorland (2008) stated that "there exists no closed "universe of knowledge" that can be studied by knowledge organization in isolation from all the other sciences' study of reality." Many authors have used more than one approach to produce knowledge. Some of them combined citation analysis with ANN (Campanario 1995, Moya-Anegón et al. 2006; Chen and Chang 2010). Gnoli (2008) proposed ten questions to be addressed by research in the 21st century. Two of them relate to this work. The first one discusses if the knowledge organization principles can be extended to a broader scope. Gnoli stated that "the field was evolving from its documentary origins to embrace a much broader range of disciplines." The second question discussed by Gnoli is about how knowledge organization systems can represent all the layers of knowledge, "developing systems more efficient in representing all the relevant dimensions of the content of the documents." López-Huertas (2008) emphasized that a bibliometric approach is a broader perspective, which make relationships that are not based on document content.

The three approaches presented by Hjorland (2002) form the theoretical basis of the method here proposed: bibliometrical studies, categorization and artificial intelligence. Also, the nature of this work comes to meet the discussion proposed by Gnoli (2008) because this is an interdisciplinary research that uses concepts of IS, tools and concepts from artificial intelligence and the field of cognitive science by means of the categorization process. The domain analytic method here proposed does not use words as units of knowledge representation. It seeks other layers of knowledge to establish relationships between documents. It explores the relationship between the citing and the cited documents. The method is particularly useful for constricted domains of knowledge in which the keywords of the documents are similar and it becomes important to find another attribute to identify categories between them. To categorize a group of documents retrieved using the same keywords, specific vocabularies would need to be used to find similarities between these documents. The proposed method obtains categories which represent groups of documents semantically related without the necessity of definition of new words in a new query, avoiding, thus, all the language problems related to the use of words such as thematic representation of documents (Smeaton 1991). The attributes used by the ANN for categorization are the presence or the absence of their references and the year of publication. To validate the proposed method an empirical experiment used a database containing the references cited by 200 articles published between 2001 and 2010 on a specific research subject. The results obtained show that the ANN successfully identified clusters of authors and documents, through their cited references. This study contributes to some issues in knowledge organization evaluating the use of ANN to automatically categorize documents in a constricted knowledge domain through the analysis of the references cited by these documents.

The remaining part of the article is organized as follows. Section 2 presents some important concepts of the field of bibliometry, of the process of categorization and gives a brief background on ANN. At section 3, selected works which use ANN for document categorization are reviewed. Details of the proposed method, our findings, discussion and conclusions are presented in final sections.

2.0 Background

This section presents the theoretical basis of the research here presented, which are bibliometric techniques, constricted domain categorization process and the ANN. Self-organizing maps (SOM) networks, chosen to be used in this research, are artificial neuron maps developed by Teuvo Kohonen in the 80s. They are responsible for execution of the categorization process. The attributes used to perform this process are originated on the bibliographic coupling existing between the documents of a collection.

2.1 Bibliometric techniques

Bibliometrics offers a set of methods and metrics to study the structure and the process of scholarly communication. In the current era of digital information, many bibliometrical studies are devoted to statistical analysis of the digital content and try to develop quantitative assessments of the information flow. Among bibliometric techniques, citation analysis is the most popular bibliometric approach and can be used to identify relationships among document regardless of the presence of equal terms in the evaluated documents (Borgman and Furner 2002).

In bibliometrics, bibliographic coupling and co-citation are examples of studies on the assessment of document similarities as showed by Figure 1. For bibliographic coupling, citing documents are the subject for the analysis. The degree of bibliographic coupling for documents A and B is reflected in the frequency of the documents that are cited by both A and B. The focus of the co-citation analysis is on the cited documents, by calculating the frequency of A and B that are co-cited by specific documents (Lai and Wu 2003).

According to Hjorland (2008), bibliographic coupling was introduced by Kessler, in 1963, and co-citation analysis was suggested independently by Marshakova and Small in 1973. Kessler (1963), in his experiments, found a

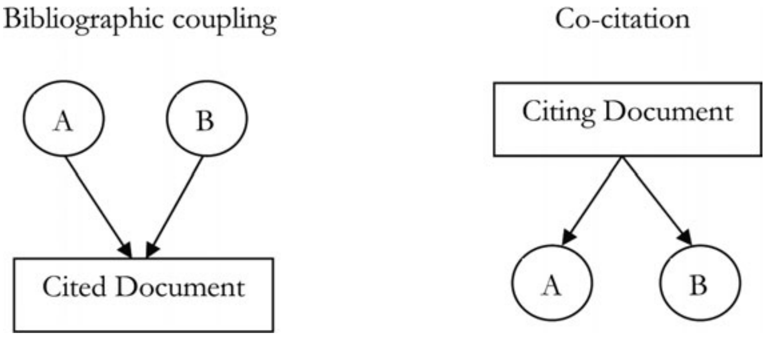

Figure 1. Examples of bibliographic coupling and co-citation
Adaptaded from Lai and Wu (2003).

high degree of logical correlation between the papers grouped using criterion of bibliographic coupling. Marshakova (1973) understood that co-citation is the logical opposite of bibliographic coupling. Small (1973) concluded that through the study of the structures created, co-citation provides a tool for monitoring the development of scientific fields and for assessing the degree of interrelationships among specialties.

The habit of citing gives conformity and consistency to the act of intellectual production, often governed by tacit and internalized norms (Cronin 1984). Statements cited in the text gain credibility when this text informs sources that support it, connecting the reader to other sources of information on that subject. The citations denote a special relationship between the citing and the cited papers and it is possible to identify, by the author's considerations, his previous experience in that subject and his network of knowledge. All those arguments strengthen the statement that bibliometrics is a strong approach because it shows real connections between individual documents, representing the authors' explicit acknowledgment dependency between papers, researchers, research fields and geographical regions (Hjorland 2002).

2.2 Categorization

"The notion of category, from Aristotle to the present time, has been used as a basic intellectual tool for analysis of the existence and how the things change" (Barite 2000). The concept of categories 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 presents 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 knowledge organization, he built a classification system, the "Colon Classification", from his 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.

Particularly relevant to the current research is the work developed by Eleanor Rosch in the 70s. The author created the prototype model, which represents concepts by a group of their characteristics and not by the use of their definitions. The grouping of concepts in a given category would be defined by their similarity with the prototypes, which are those members of the category that most reflect the redundancy of the category's structure as a whole (Rosch and Mervis 1975). Khoo and Na (2006), when discussing semantic relations as a meaningful association between two or more entities, cited binarity as one of the property of lexical- semantic relations. The binary relation was used in this work to represent each document by the presence or the absence of all 200 articles' citations of the database created. Hjorland (2002) emphasized the high practical value of the classification research and the benefits of co-operating classification with other approaches to domain analysis as bibliometric studies.

2.3 Artificial neural networks

An artificial neuron can be understood as a simplified mathematical model of the processes that happens in a biological neuron. An ANN can be defined as a structure of interconnected artificial neurons, in which typically input neurons, internal neurons and output neurons can be identified. The way neurons are organized and connected depends on the network architecture. The ANN architecture defines the number of network layers, the number of nodes in each layer, the type of connection between the nodes and the network topology. The topology of an ANN is associated with the number of neurons in the existing layers. ANNs implement algorithms which try to achieve a desired performance through techniques such as learning by experience and generalizing from similar situations, a process which is called training of the ANN. In a categorization process, after the training of the ANN, each group of elements related by their common characteristics corresponds to the activation of one neuron in the output layer of this network. ANN are primarily used in problems of approximation, prediction, classification, categorization and optimization. The vast majority of applications reported in the literature focuses on the industrial area (Meireles et al. 2003). However, they have also been used in information retrieval systems (IRS) and in categorization processes as presented in the next section.

The SOM networks are structures based on topological maps present in the cerebral cortex. Each input neuron is connected to each output neuron through its respective association weight. Here, a topological map is understood as a map which presents information related to the displayed points, not taking into consideration the distance between or the location of the points. SOM networks work basically building a map where nodes that are topologically close respond similarly to similar input patterns. In the literature there are examples of data categorization processes using SOM networks (Haykin 1994).

3.0 Artificial neural networks for categorization

Many of the experiments reported in the literature describe the use of SOM for categorization of documents in order to organize them in an alternative format for information retrieval (Luo and Zincir-Heywood 2003; Yen and Wu 2006; Yu et al. 2008). These are works found in journals and conferences in which researchers from fields such as engineering, computer science and information science discuss the use of ANN associated with the process of categorization. The applications discussed in this section use different types of network architectures and learning algorithms. The architecture of an ANN defines how its elements are interconnected. The learning algorithms relate to ANN's ability to learn using examples without having been programmed. Supervised learning networks use the information provided by an external supervisor and present desired responses on its output to the provided input patterns. This type of network may be used only when desired outputs are known and previously recorded data relating inputs and outputs exist. Thus, the supervisor adjusts the parameters of the network to find relationships between existing inputs and provided outputs. Knowing these relationships, it is possible to find outputs for new inputs submitted to the network. In unsupervised learning, there is no supervisor to monitor the learning process. From the instant in which the network identifies regularities among the input data, it generates internal representations to encode the input characteristics and automatically creates new groups.

3.1 Categorization based on textual content

Some selected studies are briefly reviewed here to show how different variables can be used as attributes for the categorization process. They used as attributes in their experiments: word repetition, use of similar words, textual similarities and context similarities found in the existing documents. Sharma and others (1994) joined supervised and unsupervised learning approaches in their experiments. While supervised learning techniques presented high performance when dealing with patterns similar to those existing on training sets, they did not perform well to recognize new categories of patterns. On the other hand, ANN based on unsupervised learning paradigm presented better recognition performance and also required lower self-training time, when compared to supervised learning networks. During the research process, they explored hardware implementation possibilities for both paradigms and highlighted the improvement of hardware ANN performance.

Lensu and Koikkalainen (1999) presented a method that can be used in document retrieval using queries. The process found similar words in the documents and further more categorized the documents based on the contexts in which these words were inserted. To evaluate the experiment, they used 18,937 questionnaires completed by students of Finnish schools and identified 115,474 words and 73,608 contexts. The textual analysis procedure was able to find similar documents even when these documents contained words different from the query words, e.g. with different endings and with spelling errors. For instance, the method was able to identify two documents with phrases like "listen to the teacher and do the homework" and "pay attention and do the exercises" as belonging to the same context and to group them in the same cluster.

Kohonen and others (2000) described the implementation of a system capable of organizing a vast collection of documents according to textual similarities. The authors state that the interpretation of the search results would become easier if these results were already presented according to the content similarities. In their work, the articles were represented as points in a two-dimensional structure and the geometric relations between these points represented the similarity relations between the articles, forming maps. The purpose of the document map was to add value to the text retrieval, providing a meaningful visual basis for presenting the search results and providing clues to select the most relevant texts. The maps would be especially useful when the user did not know well the domain or when the user had only a vague idea of the content of the texts that were being examined. According to the authors, organized presentations of data provide the users the possibility to retrieve relevant information that was not explicitly defined in their searches.

Bakus and others (2002) highlighted the growth of the researchers' interest in studies that explore methods and tools for organizing electronically available data. In their work, they defined an approach for categorizing documents that identified some of the contexts in which the words were inserted, using phrases rather than words. SOM was used with an algorithm for extracting phrases. The corpus used was composed of 21,578 articles from REUTERS texts base. From this corpus, 1,000 articles were selected to test the proposed categorization method. Among the remaining articles, 10,000 documents were chosen for the training in the sentences extraction phase. The authors showed that there was an improvement in the performance of the categorization process when using phrases rather than words.

In a later work, Hussin and Kamel (2003) used a hierarchically organized network, which was built from a SOM network and from an adaptive reasonance theory (ART) network, called SOMART by the authors. SOM

https://doi.org/10.5771/0943-7444-2014-2-145 - am 13.01.2026, 12:21:10. https://www.inlibra.com/de/agb - Open Access -

network was used to separate the collection of documents in groups and ART network was used to increase the quality of the clusters. The results showed that the experiment presented better quality of categorization and that the combination of networks was efficient in terms of runtime. The corpus was composed also by 1,000 articles collected from the same set of 21,578 documents provided by REUTERS in the previous example.

Wang and Yu (2009) proposed a model for text categorization based on the use of an ANN and the latent semantic analysis (LSA) method. This method is able to index texts for information retrieval and to establish a vector space, where each dimension corresponds to a term and each text is represented as a sum of its components. The goal of the LSA method is to reduce the number of dimensions of the produced vector space with the least possible loss of information. Besides reducing the number of dimensions, the method presents an important relationship between the terms. The training algorithm used, called back-propagation, is a supervised algorithm as described earlier in this section. The ANN is a multilayer perceptron (MLP) network which uses back-propagation modified training algorithm. The MLP is one of the most used architectures and is characterized by having at least one intermediate neurons layer. This modification has been proposed to increase the speed of network training. LSA method, originally proposed as a method of information retrieval, was used in text categorization to improve the accuracy and efficiency of the process.

In the work developed by Phuc and Hung (2008), a categorization system that used graphs was presented to group similar documents and to extract the main ideas of the documents. The model was able to indicate, according to the authors, the structural information of documents, as well as the semantic relationship between the words used in the representation, the position of the words in the documents and some implicit concepts in the documents. After the categorization was performed, SOM output was used to identify the words that helped to define the main ideas of the set of 500 documents.

3.2 Categorization based on citations

The next works illustrate ANN which use citations as attribute for the categorization process. Morris and others (2001) emphasized the fact that most of the methods described in the literature for document categorization use word frequency histograms as an attribute of the categorization process. Their method of article visualization used the connections contained in the document's citations and, according to the authors, was able to identify innovations in the area investigated and the influence of these innovations on apparently unrelated technologies, providing timelines of technological trends. The practical application of the method used 118 documents.

He and Hui (2001) described a publication retrieval system based on citations. This system indexed the scientific publications available on some websites and stored them in a database. The article described two categorization processes, which generated groups of documents and groups of authors. The categorization method grouped authors based on analysis of citations of their works. In this process, it was assumed that if the frequency with which two authors were cited by the same researchers was high, then, these two authors belonged to the same research field. This an example of the use of the co-citation concepts described on section 2. For the categorization of documents, two techniques were used, Kohonen's SOM networks and Fuzzy ART networks. The system extracted words from the titles of the references cited by the documents and used them in the categorization process. The system architecture used a citation indexing agent, which located the articles on sites specified by users or sites containing the specified keywords, converted the text in the articles, identified the reference section and saved the references in a database. Tests to validate the method were made with publications classified under the subject "information retrieval," in the site of the Institute for Scientific Information (ISI). The journals selected were from the area of "Information Science" or "Library Science." The authors selected 1,466 articles, published from 1987 to 1997, from the 367 periodicals, generating a total of 44,836 citations. Figure 2 represents the documents categorization process proposed by the authors of that work.

Most studies about categorization found in the literature use keywords as attributes for categorization or group the documents by similarities found in their contents and contexts in which they are inserted. In the work described in the preceding paragraph, the categorization of articles was performed using as attributes the words extracted from the titles of the references cited by these articles as attributes for categorization. The work presented in the following sections also uses the cited references as attributes and proposes a categorization method that uses SOM networks. Nevertheless, it does not use words during the categorization process. The current work generates an input file for ANN using only the information on the presence or the absence of each one of the references for the articles of a specific collection.

4.0 Proposed method

The main contribution of this work is to propose a method, depicted in three phases, to implement the cate-

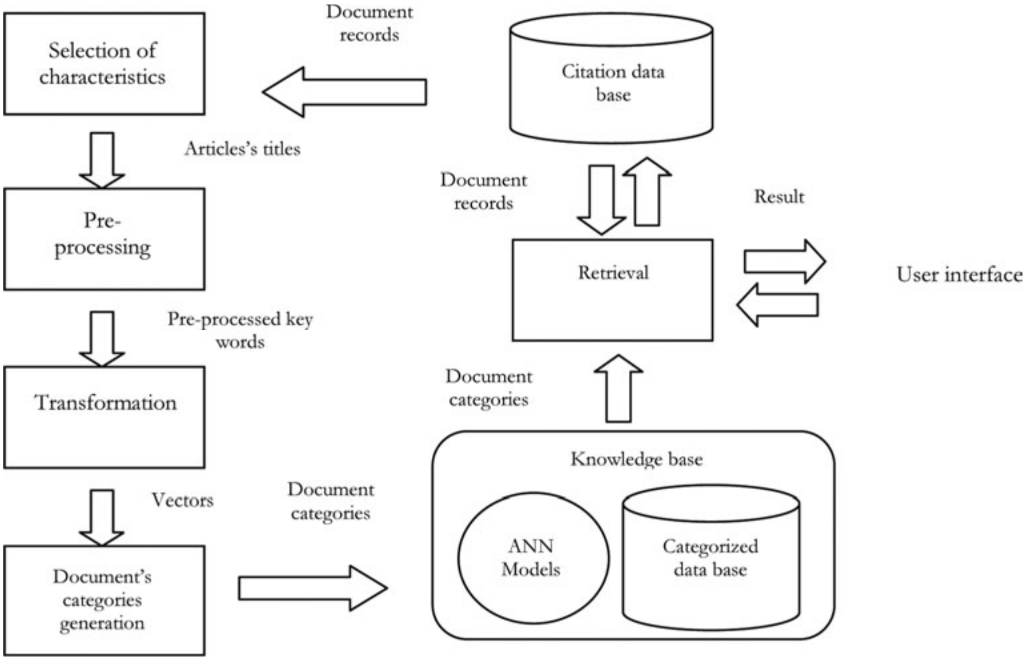

Figure 2. The process of categorization proposed by He and Hui (2001).
Adaptaded from He e Hui (2001).

gorization of documents based on citations. Hjorland (2007), summarizing the theory of classification for information retrieval, stated that "different domains develop specific languages and may need different descriptions." The concepts related with bibliographic coupling were used here as alternative descriptions and were applied in an alternative categorization process, provided by an ensemble of SOM networks. The three phases of the proposed method are described in Figure 3. In the first one, composed by the first four blocks, the test base for a constricted domain of knowledge is developed. To generate the data sequence to be used as the ANN input, the database is preprocessed and prepared. The second phase is characterized by the ANN processing and generation of articles' categories using different ANN topologies. The third phase defines the criteria to select the best topologies and to find the groups with the largest number of references in common.

Figure 4 summarizes the proposed method. A database, comprised by a collection of articles and their references, is preprocessed (Phase I) and the ANN input file is generated. Each ANN topology outputs a set of categories of articles, using the references cited by these articles as primary attributes of categorization (Phase II). These set of categories feed the decision-making process, which defines a final set of categories which has documents with strong semantic relationship based on their citations (Phase III).

5.0 The empirical experiment

A practical experiment with the three phases described in the last section was designed to validate the proposed method and to verify its effectiveness in finding relationships among papers using solely their cited references and their publication year.

5.1 Phase I of the experiment

In phase I, a representative group of documents was selected, in a specific area of knowledge. Data concerning the articles and references cited in each article are registered in two databases. After organizing the cited references and correcting inconsistencies, this data is used to create the ANN input file. In phase II, the ANN processes the input data and gathers the articles in groups.

To create a test collection, a domain with 19 articles from the *Journal of the American Society for Information Science and Technology* (JASIST) was selected. The criteria used to select the corpus were publication year from 2009 to 2010 and the presence of the expression "artificial neural network" in their titles. These 19 articles had 662 cited references which were used to create the cited references database. After organization of the 662 references, it was observed that only 19 of them were cited by more than one article and that each one of these 19 references was cited at most two times. As the proposed categorization method performed by the ANN used the cited references

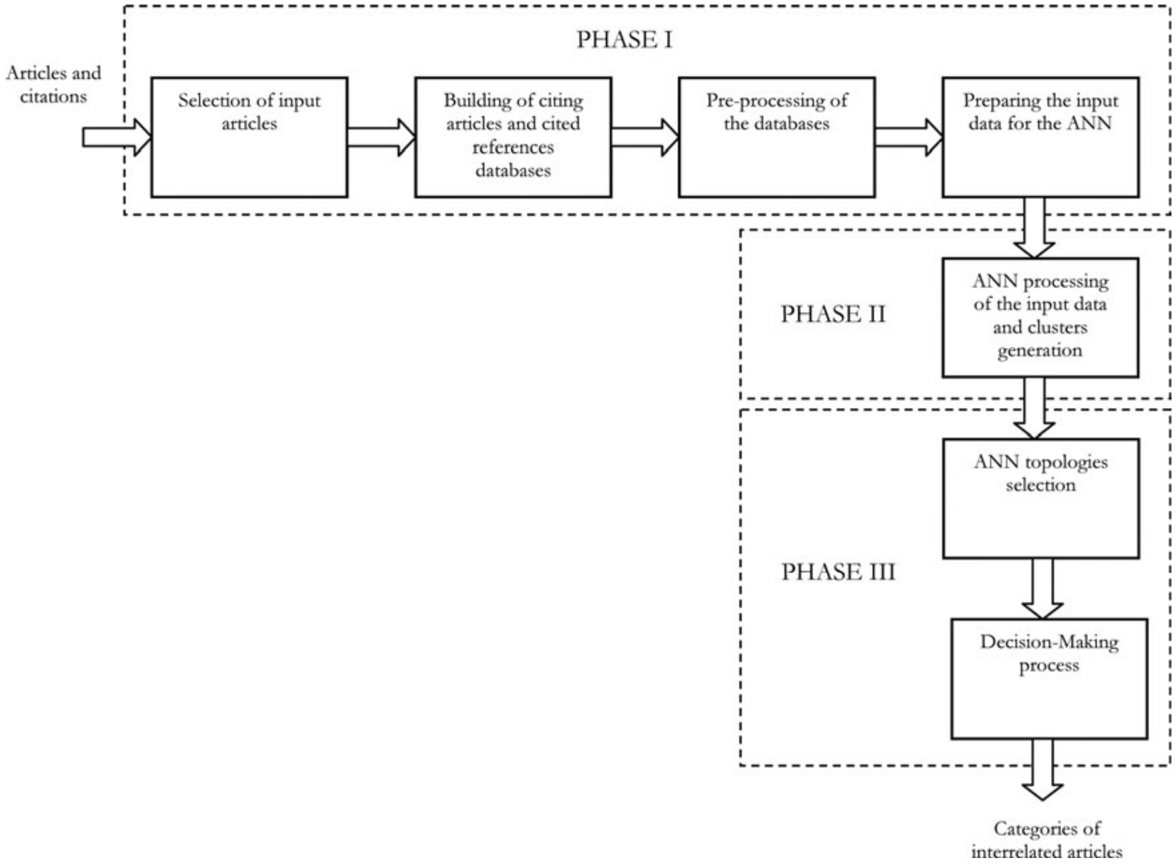
PHASE I
Articles and citations
Selection of input articles
Building of citing articles and cited references databases
Pre-processing of the databases
Preparing the input data for the ANN
PHASE II
ANN processing of the input data and clusters generation
PHASE III
ANN topologies selection
Decision-Making process
Categories of interrelated articles

Figure 3. Representation of the methodological phases.

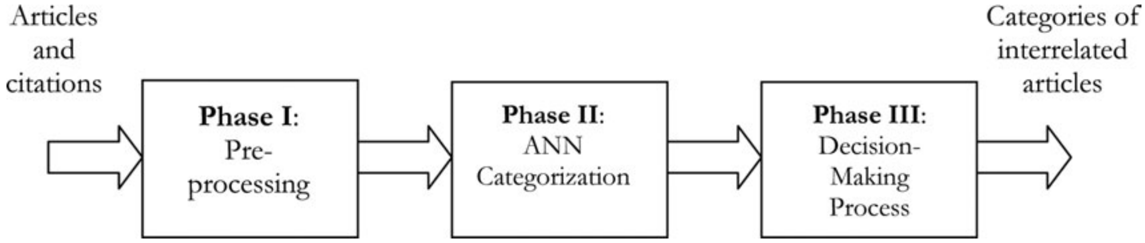
Articles and citations
Phase I: Pre-processing
Phase II: ANN Categorization
Phase III: Decision-Making Process
Categories of interrelated articles

Figure 4. Proposed method

and the year of publication of the article as attributes for categorization, the existence of a large number of references in common among different articles of the database was essential for the success of the categorization process. The characteristic of this test collection, in which the authors of the articles cited few references in common, would hinder the work of the ANN in finding similarities among the articles that could justify creating groups. For this reason, this test collection was discarded.

For the next test collection, the domain of knowledge was constricted in an attempt to obtain a greater number of common cited references. A specialized journal was chosen to constrain the domain and to ensure that the retrieved articles could present a stronger semantic relationship. To create this test collection, the IEEE Xplore digital library searching tool was used to select 200 documents with publication year ranging from 2001 to 2010. The search criterion was the presence of the term "neural network" in the document title. The papers were chosen preferably from the *IEEE Transactions on Neural Networks Journal*, which allowed retrieval of all required data for the composition of a database composed of all the authors and all the cited references. The citations in these papers amounted to a total of 6015 references. Using this data, two databases were prepared. The first one contained the 200 citing articles. The database included for each document its titles, keywords, publication year, number of cited references and an unique numeric code associated with each of the references cited by the document. The cited references in each citing article of the first database were recorded in a second database, which listed for each cited reference its numeric code, title, authors and publication year.

Some words in the titles of the same cited references were found in the singular and in the plural (e.g., inequality and inequalities) or other words had the addition of a hyphen or a quotation mark in them, which created a difference between the titles of the documents. There were also many typographic errors, such as 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 documents and the authors' names were manually changed according to the correct citation provided by the IEEE Xplore digital library.

Although the citations were repeated in several articles, cited by different authors, each one should be registered with a unique numeric code, even if it showed up several times in the database, in association with different papers. To ensure this condition and to assign a unique code number to each publication, even if referenced by different documents, a computer program written in Java language was developed to find duplicates and properly conciliate them into the final database. To generate the data sequence to be used as the ANN input for each citing article, this program was also used to create a logical (binary) attribute containing the information on the publication year and the presence or absence of each one of the 6015 cited references. In this logical attribute, the first information was the year of the article's publication and the other 6015 positions were filled by binary elements ("0" or "1"). If the reference related to the position of a code numeric had been cited by the article, this position was represented by the value "1." If the reference had not been cited, that position was filled with "0." Table 1 shows an example of the resulting data after the described processing.

5.2 Phase II of the experiment

In this phase, some tests were performed with the 200 document attributes to get 4, 9, 10, 12, 16, 25 and 36 clusters of papers. The number of clusters is directly related to the topology of each ANN in the experiment. The purpose of these tests was to verify whether the topologies would be able to gather similar documents into the same clusters.

5.3 Phase III of the experiment

In this phase, it was observed that four SOM topologies, with 10, 12, 16 and 25 clusters, grouped most of the papers into seven clusters as shown in Figure 5. These four SOM topologies were selected for analysis of the resulting clusters, as they had strong similarity between them and could generate more consistent results.

A decision making process was implemented in order to identify groups of articles representative of similar clusters across topologies, that is, those that had a large percentage

Year and references Articles	Year of publication	Reference 1	Reference 2	Reference 3	…	Reference 6015
Article 1	2009	1	1	1	…	0
Article 2	2009	0	0	1	…	0
Article 3	2010	0	0	1	…	1
…	…	…	…	…	…	…
Article 200	2004	0	1	0	…	1

Table 1. Example of the graphical scheme of the database.

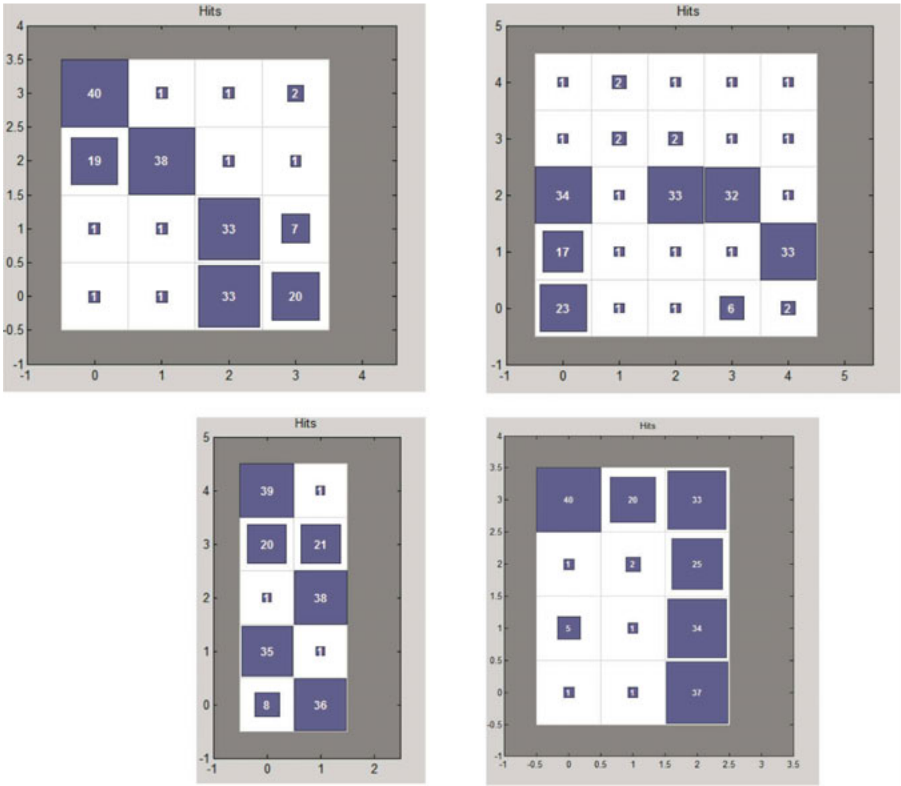

Figure 5. Categorization Maps.

of articles in common. The set of articles selected to represent the seven clusters of the four topologies were composed by the articles in common to all of them. The seven groups were called GroupOne, GroupTwo, GroupThree, GroupFour, GroupFive, GroupSix and GroupSeven. In the categorization process, the ANN used as attributes the publication year and the information on the presence or the absence of cited references for each citing article. GroupOne was the group of articles that had the largest number of references in common and was somewhat influenced, in its composition, by the publication year of the component articles. In the formation of the other groups, the publication year was also the predominant attribute in the categorization process. As the goal of the current work is to verify if the ANN is able to find, within a given collection, a group of articles with an expressive number of common references and if these article have semantic relationships, this group, called GroupOne, was chosen to demonstrate the process for definition of articles' clusters and to present the complete analysis of semantic similarity of the articles done for each group.

In each one of the four selected topologies in phase 3 (TP10, TP12, TP16 and TP25) GroupOne was renamed according to the number of the clusters of each topology. So they were called "GroupOne_10," GroupOne_12," GroupOne_16" and GroupOne_25." GroupOne presented 8 papers in the topology of 10 clusters, 5 papers in the topology of 12 clusters, 7 papers in the topology of 16 clusters and 7 papers in the topology of 25 clusters. Table 2 presents the papers belonging to GroupOne in each one of these topologies. The articles are denoted by the letter A followed by the numeric code they received in the citing articles data base.

Articles A1, A47, A48, A49 and A50 showed up in each one of the four selected topologies and the paper A18 appeared in three of the four topologies. To evaluate this occurrence, all six papers attributes were depicted and analyzed as presented in Table 3. The cited references were denoted, in Table 3, by the letter R followed by the numeric code they received in the cited references database.

The articles in Table 3 have an average of six keywords each. A direct comparison of these keywords reveals that "quadratic programming" showed up in five of the articles, "recurrent neural network" in four articles, "linear programming" in three articles and "neural network," "convergence," "asymptotic stability" and "k-winners-take-all" in two articles. These articles were published in four different years (2007, 2008, 2009 and 2010) while the totality of the articles in the database had 10 different publication years, which can indicate that this was not a

GroupOne	Number of Articles	Articles								
GroupOne_10	8	A1		A11	A18	A35	A47	A48	A49	A50
GroupOne_12	5	A1					A47	A48	A49	A50
GroupOne_16	7	A1	A6		A18		A47	A48	A49	A50
GroupOne_25	7	A1			A18	A35	A47	A48	A49	A50

Table 2. The same GroupOne categorized on 4 Topologies.

Articles	References common to 3 or more articles	Year	keywords
A1	R1, R3, R4, R5, R6, R9, R11, R12, R13, R15, R16, R17, R18, R19, R20, R24, R26, R30, R33	2009	Asymptotic stability, k-winners-take-all (WTA), linear programming, neural network, quadratic programming
A18	R1, R3, R4, R5, R6, R9, R11, R12, R13, R15, R18, R19, R20, R26, R30, R55, R802, R806	2010	Convergence, linear and quadratic programming, neural network, stability
A47	R1, R3, R4, R5, R9, R12, R13, R15, R16, R17, R19, R20, R24, R26, R30, R802, R806, R1351	2008	Winners-take-all (k-WTA), Global asymptotic stability, optimization, quadratic programming (QP), recurrent neural network
A48	R1, R4, R5, R9, R12, R18, R19, R24, R26, R55, R802, R806, R1351, R1843	2008	Differential inclusion, Lyapunov stability, global convergence, hard-limiting activation function, nonlinear programming, quadratic programming, recurrent neural network
A49	R4, R6, R9, R13, R17, R18, R19, R24, R26, R33, R55, R806, R1351, R1843	2008	Constrained optimization, convergence, convex and nonconvex problems, recurrent neural networks
A50	R1, R3, R5, R9, R11, R12, R16, R17, R18, R19, R20, R24, R26, R33, R55, R1351, R1843	2007	Global convergence, linear programming, linear variational inequality (LVI), quadratic programming, recurrent neural network

Table 3. The six articles of the GroupOne: comparison of characteristics.

predominant attribute used by the ANN in the categorization process.

To evaluate the similarity existing among citing papers of GroupOne, the common references, their keywords and their authors were also analyzed. Eight of these references (R6, R11, R15, R16, R30, R33, R802, R1843) were used in three articles, seven references (R3, R13, R17, R20, R55, R806, R1351) in four articles, six references (R1, R4, R5, R12, R18, R24) in five articles and three references (R9, R19, R26) in all the six articles. Each two citing paper of GroupOne had, at least, seven references in common. This number reached twenty one references in common, as in the case of the articles A1 and A18. These facts make clear the high degree of bibliographic coupling that exists between the articles of GroupOne. Figure 6 shows the number of common references between the articles of this group.

Table 4 presents the authors of the six articles in GroupOne associated with their respective articles and publication years. The authors' names were replaced to preserve their identity. Of the eight authors of the six articles of GroupOne, six work in China and two in Greece. Among them there is similarity of research interests, as revealed by their curricula and in the large number of common references they use in their publications.

Authors	Articles (Publication Year)
A	A1(2009)
B	A47(2008),A48(2008),A50(2007)
C	A18(2010)
D	A49(2008)
E	A49(2008)
F	A48(2008)
G	A18(2010)
H	A1(2009), A47(2008),A50(2007)

Table 4. Articles's Authors.

The results obtained from the analysis of the six papers more frequent in GroupOne_10, GroupOne_12, GroupOne_16 and GroupOne_25, showed that the ANN successfully grouped semantically similar articles in the same

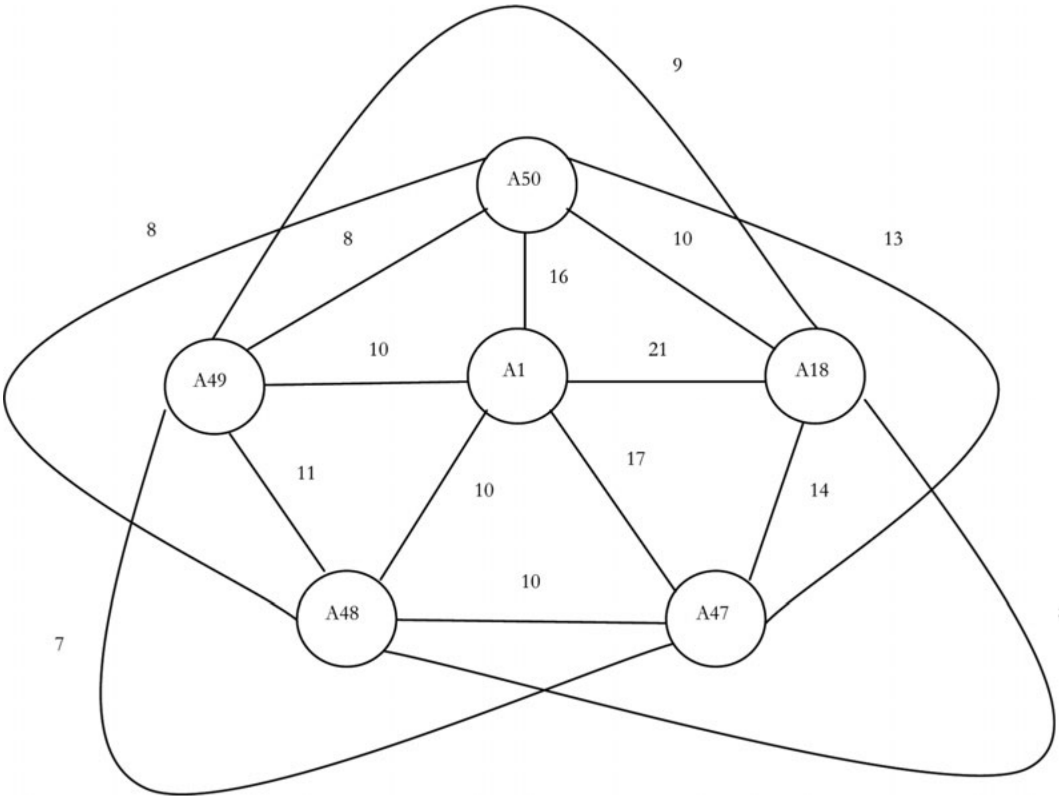

Figure 6. Number of common references among articles in GroupOne.

cluster. The articles in the same group had a large number of cited references common to three or more of them. There is an expressive number of keywords in common and all papers discuss the same subject. This can be taken as an indication of the presence of a semantic relationship among the papers.

These results were compared with others obtained in a categorization algorithm that uses keywords as attributes. In that experiment, the software used was Tanagra. The input file used in Tanagra had the information on the presence or absence of 661 keywords selected by the authors of the 200 articles. The database used for this experiment constricted the field of research and limited the choice of the articles in a specialized journal to those which had the term "neural network" in their titles. As consequence the algorithm created a category with all the documents which presented this expression as keyword. The other groups, significantly smaller, were characterized by keywords derived from that main expression (e.g., recurrent neural network and radial basis function neural network). Members of GroupOne found by SOM were not grouped together in the same cluster. The categorization process using keywords proved to be not efficient to produce groups of articles semantically related in a constricted content domain as used in this work.

6.0 Discussion

Editing errors are a technical limitation of citation databases, hindering the implementation of citation analysis (MacRoberts and MacRoberts 1989). To assure accuracy and consistency, the test base was preprocessed in phase I of the method. Once the database was built, an exhaustive preprocessing phase was accomplished to correct several inconsistencies which were identified in different instances of the titles and of the authors in the cited references. Missing information, incorrect or inconsistent records in the database were corrected to improve data quality. Through this process, in addition to typographical errors identified in the data, some semantic discrepancy was also found. At the end of the preprocessing, all the inconsistent records were manually changed according to the correct citation provided by the IEEE Xplore digital library. During phase I, a computer program written in Java was developed to find duplicates between the numeric code associated with the cited documents to properly conciliate them into a consistent and a reliable database to be used by the SOM network.

In phase II, some tests were performed with the SOM networks ensemble, to verify if the topologies of each ANN would be able to join similar documents into the same clusters. In phase III, four SOM topologies, with 10, 12, 16 and 25 clusters, were selected because they grouped most of the papers into seven clusters, showing similarity between them. A specific group, named GroupOne, was chosen to demonstrate the process for definition of articles' clusters and to present the complete analysis of semantic similarity of the articles done for each group. During the experiment, it was observed that the articles in GroupOne had a large number of cited references in common, as well as an expressive number of common keywords. All papers of that group discussed the same subject.

All these similarities identified in the documents of GroupOne showed that the proposed method successfully identified, in a constricted knowledge domain, a cluster of documents with a strong semantic relationship among them. The method is able to identify similarities in this kind of domain of knowledge where keywords and subject descriptors tend to be similar and therefore insufficient to differentiate documents.

7.0 Conclusions

The main contribution of this work is the domain analytic method proposed, particularly useful for constricted domains of knowledge in which keywords are similar and where it is necessary to determine additional attributes to identify similarities among the documents. This method can be an important alternative to domain analysis, bringing result sets different than usual retrieval by the use of keywords and producing domain specific knowledge. The method used ANN ensembles to automatically categorize documents through the analysis of the references cited by these documents. The theoretical bases of this work are bibliometric thechniques, categorization and ANN, coming to meet the discussion proposed by Hjorland (2002) and Gnoli (2008) and highlighting the importance of interdisciplinary studies in knowledge organization. The proposed method is divided in three phases, which were implemented by an empirical experimental presented and discussed here in details.

Hjorland (2002) stated that the field of AI had historically been related to and dominated by individualistic rather than social ways of thinking and such research had mostly been done with a mechanical view of human thinking, neglecting the historical and cultural aspects of human cognition. The method proposed here overcame that fact supplementing the use of ANN with bibliometric concepts. Citation analysis and co-citation concepts, used to categorize documents, denoted special relationships between citing and cited papers and revealed many social aspects related to the authors, as their previous experience in that subject and their network of knowledge.

The results of the categorization process which uses citations as attributes permit the known advantages of citation databases such as the identification of groups of researchers working in related fields and the identification of research trends in specific domains of knowledge. Another application of the identification of articles' clusters by their common references is to use the keywords of the articles in the categories for the formulation or reformulation of a query to a database in the process of information retrieval. Using the proposed method, a user can employ these keywords to formulate a new query and this is particularly useful, for example, when the user has only a vague idea about the content of the texts that will be examined. This combination of methods for information retrieval could improve final search results. The proposed method uses bibliographic coupling concepts, from the bibliometric studies, as an objective measure of the semantic relationship between scientific documents. This research strengthened the importance of using citations for categorization and for information retrieval and confirmed some of the difficulties encountered in these studies during the development stage of the testing base. As ANN used the common citations as the attribute of the categorization process, the database must be composed by articles which have a large number of common citations. Whilst that fact is a limitation for this work, this method can be applied when the research is in an advanced phase and the first query already has retrieved some related documents with some references in common.

Further research could determine a possible threshold for the number of common citations below which the method is not viable. It would be also important to evaluate whether there is any article that must be together with the generated clusters but it is not. The categorization process is a natural process for human beings, who seek to create groups as a way to organize information. The gigantism of collections of documents makes it a challenge for the users to find the ones that actually meet their needs. The proposed domain analytic method provides an alternative to search for relevant information in the process of information retrieval.

References

Bakus, J., Hussin, M. F. and Kamel M. 2002. A SOM-based document clustering using phrases. In Wang, L., Rajapakse, J.C., Fukushima, K., Lee, S.-Y. and Yao, X., eds., ICONIP'02: Proceedings of the 9th International Conference on Neural Information Processing: computational intelligence for the E-age: November 18-22, 2002, Orchid Country Club, Singapore. Singapore: Nanyang Technological University, pp. 2212-6.

Barite, M. G. 2000. The notion of "category": its implications in subject analysis and in the construction and evaluation of indexing languages. *Knowledge organization* 27: 4-10.

Borgman, Christine L. Borgman and Furner, Jonathan. 2002. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology* 36: 2-72.

Campanario, J. M. 1995. Using neural networks to study networks of scientific journals. *Scientometrics* 33: 23-40.

Chen, Yu-Shan and Chang, Ke-Chiun. 2010. Analyzing the nonlinear effects of firm size, profitability and employee productivity on patent citations of the US pharmaceutical companies by using artificial neural network. *Scientometrics* 82: 75-82.

Cronin, Blaise. 1984. *The citation process*. London: Taylor Graham.

Gnoli, Claudio. 2008. Ten long-term research questions in knowledge organization. *Knowledge organization* 35: 137-49.

Haykin, Simon S. 1994. *Neural networks: a comprehensive foundation*. New Jersey: Prentice Hall.

He, Yulan and Hui, Siu Cheung. 2001. PubSearch: a web citation-based retrieval system. *Library hi tech* 19: 274-85.

Hjørland, Birger. 2002. Domain analysis in information science: eleven approaches – traditional as well as innovative. *Journal of documentation* 58: 422-62.

Hjørland, Birger. 2007. Semantics and knowledge organization. *Annual Review of Information Science and Technology* 41: 367-405.

Hjørland, Birger. 2008. What is knowledge organization (KO)? *Knowledge organization* 35: 86-101.

Hussin, M. F. and Kamel, M. 2003. Document clustering using hierarchical SOMART neural network. In *Proceedings of the 2003 International Joint Conference on Neural Network*. Portland, Oregon: Institute of Electrical and Electronics Engineers Inc., pp. 2238-2242.

Jacob, Elin K. 2004. Classification and categorization: a difference that makes a difference. *Library trends* 52: 515-40.

Kessler, Myer Mike. 1963. Bibliographic coupling between scientific papers. *American documentation* 14: 10-25.

Khoo, Christopher S. G. and Na, Jin-Cheon. 2006. Semantic relations in information science. *Annual Review of Information Science and Technology* 40: 157-228.

Kohonen, T. et al. 2000. Self organization of a massive document collection. *IEEE transactions on neural networks* 11: 574-85.

Lai, Kuei-Kuei and Wu, Shiao-Jun. 2003. Using the patent co-citation approach to establish a new patent classification system. *Information processing and management* 41: 313-30.

Lensu, A. and Koikkalainen, P. 1999. Similar document detection using self-organizing maps. In Jain, L.C., ed., *Third International Conference on Knowledge-Based Intelligent Information & Engineering Systems,* Adelaide, Australia: Institute of Electrical and Electronics Engineers Inc., pp 174-7.

López-Huertas, María J. 2008. Some current research questions in the field of knowledge organization. *Knowledge organization* 35: 113-36.

Luo, X. and Zincir-Heywood, A. N. 2003. A Comparison of SOM Based Document Categorization Systems. In *Proceedings of the 2003 International Joint Conference on Neural Network*. Portland, Oregon: Institute of Electrical and Electronics Engineers Inc., pp. 1786-1791.

Mac Roberts, Michael H. and Mac Roberts, Barbara R. 1989. Problems of citation analysis: a critical review. *Journal of the American Society for Information Science* 40: 342-9.

Marshakova, Irina V. 1973. A system of document connection based on references. *Scientific and technical information serial of VINITI* 6 no.2: 3-8.

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* 50 n.3: 1-18.

Morris, S. A., Wu, Z. and Yen, G. 2001. A SOM mapping technique for visualizing documents in a database. In *IJCNN'01: proceedings: International Joint Conference on Neural Networks : Washington, D.C., July 15-19, 2001*. Piscataway, N.J.: Institute of Electrical and Electronics Engineers Inc., pp. 1914-9.

Moya-Anegón, Félix, Herrero-Solana, Víctor and Jiménez-Contreras, Evaristo. 2006. A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research. *Journal of information science* 32: 63–77.

Phuc, Do and Hung, Mai Xuan. 2008. Using SOM based graph clustering for extracting main ideas from documents. In *RIVF 2008 2008 IEEE International Conference on Research, Innovation and Vision for the Future in Computing & Communication Technologies University of Science Vietnam National University Ho Chi Minh City, July 13 17, 2008*, pp. 209-14.

Rosch, Eleanor and Mervis, Carolyn B. 1975. Family resemblances: studies in the internal structure of categories. *Cognitive psychology* 7: 573-605.

Sharma, Anil K., Sheikh, Sohail, Pelczer, Istvan and Levy, George C. 1994. Classification and clustering: using neural networks. *Journal of chemical information and computer sciences* 34: 1130-9.

Small, Henry. 1973. Co-citation in the scientific literature: a new measurement of the relationship between two documents. *Journal of the American Society of Information Science* 24: 265-9.

Smeaton, Alan F. 1991. Prospects for intelligent, language-based information retrieval. *Online review* 15: 373-82.

Smiraglia, Richard P. 2013. Is FRBR a domain? domain analysis applied to the literature of the FRBR family of conceptual models. In Proceedings from North American Symposium on Knowledge Organization, Vol. 4. University of Wisconsin-Milwaukee. Available http://iskocus.org/NASKO2013proceedings/Smiraglia_IsFRBRaDomain.pdf

Wang, Wei. and Yu, Bo. 2009. Text categorization based on combination of modified back propagation neural network and latent semantic analysis. *Neural computing and applications 18*: 875-81.

Yen, G. G. and Wu, Zheng. 2006. A self-organizing map based approach for document clustering and visualization. *IJCNN '06. International Joint Conference on Neural Networks, 2006*. United States: Institute of Electrical and Electronics Engineers Inc., pp. 3279-86.

Yu, Yan, He, Pilian, Bai, Yushan and Yang, Zhenlei. 2008. A document clustering method based on one-dimensional SOM. In Lee R., ed., *Proceedings 7th IEEE/ACIS International Conference on Computer and Information Science (IEEE/ACIS ICIS 2008) In conjunction with 2nd IEEE/ACIS International Workshop on e-Activity (IEEE/ACIS IWEA 2008),* pp. 295-300.