Automatic Patent Clustering using SOM and Bibliographic Coupling

Magali R. G. Meireles[1], Juan R. S. Carvalho[2],

Zenilton K. G. do Patrocínio Júnior[1], Paulo E. M. de Almeida[3]

[1] Institute of Mathematical Sciences and Informatics,
Pontifical Catholic University of Minas Gerais
Rua Walter Ianni, 255, São Gabriel, Belo Horizonte, Minas Gerais, Brazil,
CEP 31.980-110

[2]Computer Engineering, Pontifical Catholic University of Minas Gerais

[3]Computer Department, Federal Center for Technological Education of Minas Gerais

{magali,zenilton}@pucminas.br,juanrequeijo41@gmail.com,

pema@lsi.cefetmg.br

Abstract. Patents are usually organized in classes generated by the offices responsible for patents protection, to create a useful format to the information retrieval process. The complexity of patent taxonomies is a challenge for the automation of patent classification. Beside this, the high numbers of subgroups makes the classification in deeper levels more difficult. This work proposes a method to cluster patents using Self Organizing Maps (SOM) networks and bibliographic coupling. To validate the proposed method, an empirical experiment used a patent database from a specific classification system. The obtained results show that patents clusters were successfully identified by SOM through their cited references, and that SOM results were similar to k-Means algorithm results to perform this task. This study can contribute to the development of the knowledge organization systems by evaluating the use of citation analysis in the automatic clustering of patents in a constrained knowledge domain, at the subgroup level of current patent classification systems.

1. Introduction

Nowadays, with the growth of digital patent collections, demanding higher level of computer support, the need to automatically organize the information available has increasing priority. To create an alternative to the information retrieval process, the patents are usually presented in classes generated by the offices responsible for patents protection. According Baeza-Yates and Ribeiro-Neto (2011), a natural solution to solve the problem of finding documents on a restricted domain of knowledge is to group documents by common topics and name each group with one or more meaningful labels. Each labelled group is a set in which we can insert documents whose contents

can be described by its label. The classification process provides a mean to organize and to manage information, which allows better understanding and interpretation of the data. One compelling argument for classification systems is that there is an innate tendency for humans to compartmentalize information [Smith 2002]. Patent offices organize patent applications into very large topic taxonomies. The vocabulary is quite diverse and to avoid narrowing the scope of the invention, the applicants prefer use general terms. Because the patents describe new inventions, usually they are different at a semantic level [Tikk et al. 2007]. The complexity of patent taxonomies is a challenge for the automation of patent classification.

Even though numerous attempts are found in the literature for building automatic classification systems, some shortcomings can be identified, such as limited subclass level accuracy. This problem arises from the granularity of large patent classification systems, such as the Unites States Patent Classification (USPC) and International Patent Classification (IPC). The high number of subgroups makes the classification in subgroups level more difficult. If an error, for example, is made at class level, the error is propagated to subclass and group level. According Smith (2002), the use of clustering software was investigated as a potential tool for the reclassification process. The reclassification process is the process by which patent categories are grouped together in larger ones, or broken down in smaller ones, as well as the subsequent process of re-tagging some patents that were classified under the modified categories. This process can be further subdivided in two subtasks. The first one can suggest new categories and the second one is the process of automatic re-tagging of the patents according to new patent categories [Benzineb and Guyot 2011]. The idea is to subdivide large, fast growing subclasses into smaller ones that could be more efficiently browsed during a prior art search. This research aims to reclassify patents, suggesting new categories, using cited patents as attribute of the categorization process. The method here proposed is particularly useful for constrained domains of knowledge, in which keywords of the documents are similar among each other, as the subgroups of a patent classification system. In this case, it becomes important to find another attribute to identify in-between categories. The clustering process is made by Self Organizing Maps (SOM) Artificial Neural Network (ANN), and the attributes used are the presence or the absence of the cited patents.

Usually, automated search service works creating a word list to conduct a query, extracted from the title, abstract and brief summary portions of the patent application. But, according Meireles et al. (2016), there is no agreement in the literature about the best attributes to use in patent representation. The method proposed here 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. To cluster a group of documents retrieved using the same keywords, specific vocabularies would need to be used to find similarities between these documents. An empirical experiment using a patent database, containing references cited by 117 patents, is proposed here to validate the method. These patents were chosen among four specific subgroups of a classification system. The objective is to show that the proposed method is able to identify new subgroups in these four subgroups and so suggest a new redistribution of patents in this classification system. The experiments here discussed have revisited the method developed by Meireles et al. (2014), implementing another application and another algorithm, using a different

constrained knowledge domain. The results obtained show that SOM successfully identified clusters of patents, through their cited references, and that K-Means results were similar to SOM results, showing consistency of the proposed method. The measure of similarity included in this paper proves that there is similarity between the algorithms' output.

The remainder of this article is organized as follows. Section 2 presents some concepts related to clustering process and similarity metrics. Section 3 introduces automatic patent clustering systems using citation information. The methodology, results, discussion and conclusions are presented in final sections.

2. Clustering Techniques and Similarity Metrics

The steps to cluster and to classify documents, used by machine learning algorithms, are inspired by the described human behaviour. As described by Croft, Metzler and Strohman (2010), document clustering is the task of grouping related documents together while, classification is the task of automatically applying labels to data, for example, labels to documents. Both have been studied for many years by information retrieval researchers, with the aim of improving the effectiveness and the efficiency of search applications. In machine learning, learning algorithms are typically characterized as supervised or unsupervised. In supervised learning, a model is learned using a set of fully labelled documents, called the training set. Once a model is learned, it can be applied to a set of unlabelled documents, called the test set. Classification is a supervised learning problem. Clustering is the most common example of unsupervised learning. It takes a set of unlabelled data objects as input and then groups the objects using some notion of similarity. The first step is to identify a number of important features in the documents, which will help to properly distinguish them among the possible labels. In the second step, these features are extracted from each document. In the third step, the evidence from the extracted features is combined to find appropriate labels or groups (clusters).

Two important clustering algorithms are Hierarchical Clustering and K-Means. They start from some initial clustering of the data and then iteratively improve the existing clusters, by optimizing some objective function. Some authors have compared the performance of these algorithms [Widodo, Budi 2011; Kukolj et al. 2012] using different attributes to group patents. They used datasets from the fields of Information and Communication Technology (ICT) and of consumer electronics, respectively. In other algorithm, the K Nearest Neighbour-Clustering, a cluster is formed around every input instance. For input instance x, the K points that are nearest to x according to some distance metric and x itself form a cluster [Croft, Metzler, Strohman 2010]. In the literature, there are also examples of data clustering processes using SOM networks [Haykin 1994]. These 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. SOM networks work basically building a map where nodes that are topologically close to each other respond similarly to similar input patterns.

To quantitatively express the extent to which the clusters of each algortithm agree with the created groups, a clustering similarity measure called Measure of Concordance (MoC) can be used [Pfitzner, Leibbrandt, Powers 2009]. To provide a

measure of the degree of concordance between clustering S, created by one method, and clustering M, generated another method, MoC is defined as

$$MoC\ (S,M) = \begin{cases} 1,\ if\ I = J = 1; \\ \frac{1}{\sqrt{IJ}-1}\left(\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\|F_{ij}\|^2}{\|S_i\|\|M_j\|} - 1\right), otherwise, \end{cases} \quad (1)$$

in which the norm operator $\|\,.\,\|$ represents the size (or the number of compounding instances) of common fragments among clusters, $\boldsymbol{F_{ij}}$, the size of clusters $\boldsymbol{S_i}$ and the size of clusters $\boldsymbol{M_j}$. There are $\boldsymbol{I}$ clusters in $\boldsymbol{S}$ and $\boldsymbol{J}$ clusters in $\boldsymbol{M}$. Each individual cluster in $\boldsymbol{S}$ is referred to as $\boldsymbol{S_i}$ and each cluster in $\boldsymbol{M}$ as $\boldsymbol{M_j}$. Any cluster $\boldsymbol{S_i}$ can be subdivided into smaller subclusters or fragments, where a fragment consists of those elements of $\boldsymbol{S_i}$ that have also been allocated to a single cluster $\boldsymbol{M_j}$. These common fragments are instances where both clusterings agree and they are the intersection between $\boldsymbol{S}$ and $\boldsymbol{M}$. Figure 1 shows an example of division of clusters into fragments. The numbers inside the box indicate the number of entities belonging to each fragment.

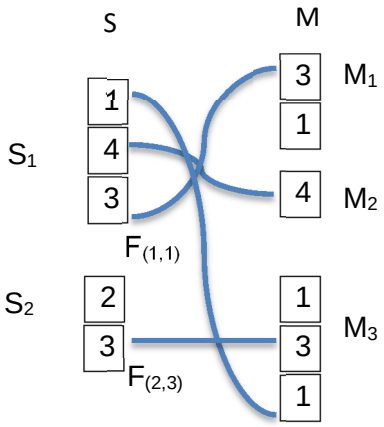

Figure 1. The division of clusters into fragments

3. Automatic Patent Clustering using Citation Information

Citation analysis is the most popular bibliometric approach and it 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 shown by Figure 2. For bibliographic coupling, citing documents are the subject for 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 C and D that are co-cited by specific documents [Lai and Wu 2005].

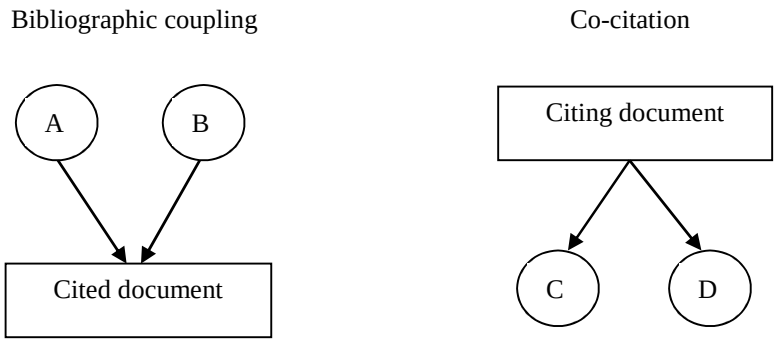

Figure 2. Examples of bibliographic coupling and co-citation

Adapted from Lai and Wu (2005)

Some papers have discussed the application of citation analysis to organize patent databases, highlighting how patents can be grouped in clusters' using patent's citation as connection between patents. Lai and Wu (2005) proposed an approach to create a patent classification system to replace International Patent Classification or United States Patent Classification system, to assist patent manager in understanding the basic patents for a specific industry and the evolution of the related technology field. Li et al. (2007) proposed to utilize patent citation information and considered the structure of patent citation networks for patent classification. They stated that '*a network of citations provides rich information about the relationships among patents as well as the relationship among their topics*'. They adopted a Kernel-based approach to capture content information and citation-related information in patents and the results showed that their proposal outperformed the kernels that did not use citation network structures. Liu and Shih (2011) combined content-based, citation-based and metadata-based classification methods to develop a hybrid-classification approach using a modified KNN algorithm. Some authors have used patent citation analysis for other purposes. Patent citations have been recognized as a source of data for the study of innovation and technical change [Trajtenberg 1990; Chakrabarti, Dror and Eakabuse 1993; Engelsman and Van Raan 1994; Hall, Jaffe and Trajtenberg 2002] and for measuring their economic value [Sapsalis, Van Pottelsberghe de la Potterie and Navon 2006]. Researchers as Morris and others (2001) and He and Hu (2001) used ANN and citations as attributes for clustering processes.

4. Methodology

In Meireles et al. (2014), the authors clustered documents by means of SOM, and using documents' citations as attributes for the clustering process. In this study, we found a specific field of knowledge to justify the use of citations, the patents databases; then, we adopted a similarity metric to compare different clustering algorithms and, finally, we added an auxiliary algorithm, K-Means, to evaluate the similarity between the resulting clusters of both methods. According Meireles et al. (2014), the general method here used is suitable for areas of restricted knowledge, where there is a significant number of common citations and where it becomes more difficult to find differences between words or expressions of semantically related documents to justify the creation of clusters. Our patent clustering method can be presented in three phases, which are shown by Figure 3.

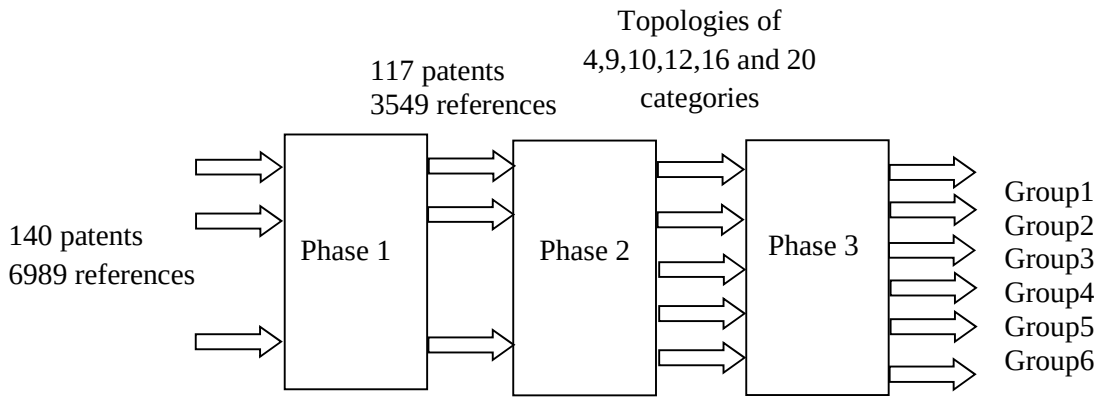

Figure 3. Representation of the methodological phases

In the first step, a group of patents from a restricted area of knowledge is selected and processed, so that data relating to patents and the patents cited as references in the document can be recorded in a database. The patents were choosen from four different subgroups from CPC classification system. These known subgroups were used to compare the groups created by the SOM network with this classification system. The database consisted of 140 patents from subgroups G06K 7/1443, G06K 7/1447, G06K 7/1452 and G06K 7/1456 of the subclass G06K from CPC classification system, called "Recognition of data, presentation of data, record carriers; handling record carriers". Some of these patents were classified in more than one subgroup. In these cases, only one subgroup for each patent was randomly assigned and so the number of patents was reduced to 117. A total of 6,989 references were registered for 117 patents. Of these, only 3,549 are not repeated. SOM network and K-Means algorithm input were then fed with 117 binary codes, each one with 3,549 binary digits representing absence or presence of a specific reference in a patent. Table 1 shows the available number of patents for each selected subgroup and the number of selected patents for the prototype database.

Table 1. Number of patents used in the experiment

Subgroups	Available patents	Selected patents
G06K 7/1443	343	28
G06K 7/1447	186	31
G06K 7/1452	65	29
G06K 7/1456	176	29
Total		117

The second phase of the experiment is the creation of the clusters (in the current case, by SOM and K-Means). In this work, five SOM network topologies were used to generate 4, 9, 12, 16 and 20 categories. The same number of clusters were created by K-Means algorithm, independently.

In the third phase, patent groups which were repeated in most of the topologies were identified. As SOM network and K-Means grouped these patents in a same cluster in different experiments, these groups should suggest a reclassification for these subgroups, from the patent database point of view.

5 Results

Two of the five SOM network outputs are analyzed in the next paragraphs. Figure 4 shows the nine clusters created by the first topology.

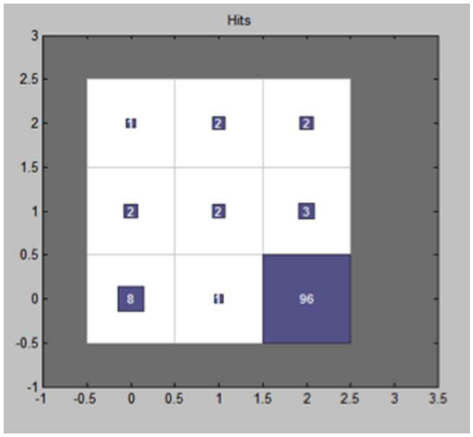

Figure 4. SOM network output for 3x3 topology (9 clusters)

For this topology, there were two clusters containing only one patent, four containing two patents, one containing 3 patents, one containing 8 and one containing 96 patents. The clusters presented in first column of Table 2 were numbered from 1 to 9 and identified with the final equivalent to the number of clusters generated by the topology. Patents grouped into some of these clusters, which are designated by letter P and by the reference number of the database, are identified in the third column. The fourth column shows in which subgroup CPC these patents are classified. The last column presents the number of common cited patents by the patents presented in the third column. The number between parentheses shows the number of citing patents in each cluster.

Table 2. Clusters obtained by topology 3x3

Clusters	Number of patents	Patents	Subgroup CPC G06K 7/	Number of common cited patents (citing patents)
C2_9	2	P3, P28	1443	320 (2)
C3_9	2	P41, P50	1447	143 (2)
C4_9	2	P97, P104	1456	149 (2)
C5_9	2	P51, P58	1447	154 (2)
C6_9	3	P99, P105, P117	1456	9 (2)
				16 (3)
C7_9	8	P45, P47, P48, P52, P54, P56, P57, P59	1447	146 (2)
				90 (3)
				62 (4)
				46 (5)
				21 (6)
				15 (7)
				6 (8)

OBS: C1_9 and C8_9 categories had only one patent and C9_9 grouped 96 patents.

For the second topology, there were four clusters containing only one patent, five containing two patents, one containing 3 patents, one containing 5 and one containing 95 patents. Figure 5 shows the twelve categories created by the second topology.

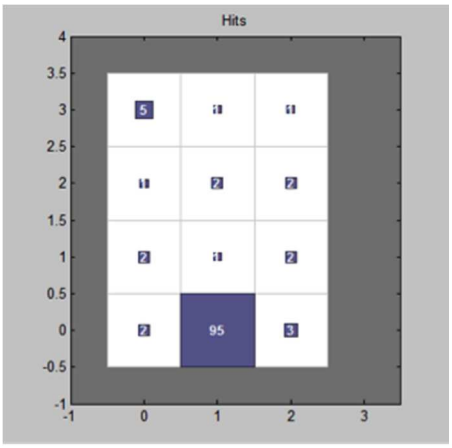

Figure 5. SOM network output for 4x3 topology (12 clusters)

Clusters were numbered from 1 to 12 and identified with the final equivalent to the number of clusters generated by this topology. Some of the characteristics of the generated clusters in this topology are presented in Table 3.

Table 3. Clusters obtained by topology 4x3

Categories	Number of patents	Patents	Subgroup CPC G06K 7/	Number of common cited patents (citing patents)
C1_12	5	P47, P48, P52, P54, P59	1447	53 (2)
				96 (3)
				77 (4)
				24 (5)
C5_12	2	P3, P28	1443	320 (2)
C6_12	2	P41, P50	1447	143 (2)
C7_12	2	P35, P44	1447	49 (2)
C9_12	2	P45, P57	1447	187 (2)
C10_12	2	P97, P104	1456	149 (2)
C12_12	3	P81, P86, P88	1452	12 (2)
				35 (3)

OBS: C2_12, C3_12, C4_12 and C8_12 had only one patent and C11_12 grouped 95 patents.

Among the experiments using topologies of 4, 9, 12, 16 and 20 clusters, six groups of patents, designated by S_i, where i varies from 1 to 6, stand out because they have been identified in the same cluster in at least three experiments. A summary of these results is shown in Table 4.

Table 4. Common groups between topologies

Groups	Patents (Number of cited patents)	Subgroup CPC G06K7/	Common cited patents (citing patents)	Titles	Topologies
S_1	P3 (388) P28 (320)	1443	320	P3: Image capture and processing system supporting a multi-tier modular software architecture; P28: Hand-supportable digital image capture and processing system supporting a multi-tier modular software architecture.	4,9,12 and 16
S_2	P41 (234) P50 (253)	1447	143	P41: Method for increasing the functionality of a media player/recorder device or an application program; P50: Identification documents and authentication of such documents.	4, 9, 12, 16 and 20
S_3	P45 (188) P57 (187)	1447	187	P45: Content containing a steganographically encoded process identifier; P57: Controlling a device based upon steganographically encoded data.	12,16 and 20
S_4	P47 (238) P48 (231) P52 (188) P54 (222) P59 (163)	1447	53 (2) 96 (3) 77 (4) 24 (5)	P47: Control signals in streaming audio or video indicating a watermark; P48: Connected audio content; P52: Methods and devices responsive to ambient audio; P54: Connected audio and other media objects; P59: Methods and devices responsive to ambient audio.	4, 9, 12 and 20
S_5	P81 (62) P86 (36) P88 (47)	1452	12 (2) 35 (3)	P81: Method of scanning indicia using selective sampling; P86: Optical scanners; P88: Method of scanning indicia using selective sampling.	12, 16 and 20
S_6	P97 (152) P104 (149)	1456	149	P97: Product provided with a coding pattern and apparatus and method for reading the pattern; P104: Product provided with a coding pattern and apparatus and method for reading the pattern.	9, 12, 16 and 20

The same experiment was also performed with the use of K-Means. Patents groups repeated in the majority of the five runs, which were found with variation of k parameter, were identified and are presented in Table 5 as M_j, where j varies from 1 to 7.

Table 5. Common groups among K-Means

Groups	Patents	Subgroup CPC G06K7//	K parameter values
M_1	P3, P28	1443	4,9,12,16 and 20
M_2	P41, P50	1447	9,12,16 and 20
M_3	P46, P51, P58	1447	9,12,16 and 20
M_4	P45, P57	1447	9,12,16 and 20
M_5	P47, P48, P52, P54, P56, P59	1447	9 and 16
M_6	P97, P104	1456	12,16 and 20
M_7	P6, P23	1443	12,16 and 20

Considering only the clusters identified by this method, 6 by SOM and 7 by K-Means, and taking into account two facts: (1) Four of these clusters are exactly the same; (2) One of them has 5 of 6 entities in common. Then, MoC index for both approaches can be calculated with Equation 1, yielding a final value of *0.699*. This calculation will be used afterwards, in the discussion section, and presented as an objective measure of similarity between SOM and K-Means methods, while clustering the tested database.

6. Discussion

All the patents of these groups, identified in most of the SOM topologies and by K-Means, are related to a same subgroup of CPC system, as shown by Table 4 and Table 5. For example, in S4, the five patents are from CPC subgroup G06K7/1447. However, three groups of patents, belonging to the same CPC subgroup, were associated by SOM networks to different clusters. This clustering conducted by SOM suggests that the CPC subgroup should be reformulated. In some of these groups, the patents are closely related to a content, such as those identified in S3, which have been filed on the same date, have the same inventor and the same assignee, but have different publication dates. These patents have the same number of drawings, but different number of claims. These patents of S3 should be member of a new subgroup. The same analysis could be applied to the patents belonging to groups S5 and S6.

There are some specific issues related to a patent database. Some patents are classified in more than one subgroup, which contradicts the theory of classification, in which an entity must be associated with only one class within a set of mutually exclusive classes that do not overlap each other [Jacob 2004]. This method could help to choose only one of the subgroups, that one more related to the patent.

Given that the range of MoC index should be between 0 and 1, the result obtained can be interpreted as a similarity of almost 70% between the clusters obtained by the two methods implemented. This fact can confirm that, for the database used, citations can be used as a relevant attribute for the patent clustering process. After all, this can be understood as an objective indication of the relevance of citations as attributes to the general process of patents clustering and classification.

7. Final remarks

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 from their common properties, and thus better understanding the surrounding context [Meireles et al. 2014]. According to Hjorland [2002], classification systems organize the logical structures of categories and concepts in a domain, as well as the semantic relationship between these concepts. With the increasing number of patents and the development of new technologies, these classification systems should be constantly reviewed to avoid accumulation of patents on certain subgroups. To use cited patents in common, as clustering attributes, can be an alternative process to create new groups in subgroups level of classification systems, where patent offices organize patent applications into very large topic taxonomies. In a restricted domain of knowledge as these subgroups, it is difficult to use words as units of knowledge representation, since the subject and, consequently, the words are similar. To break down a subgroup into other ones, this work explored the relationship between citing and cited documents.

The objective of this work was to identify, among four selected subgroups of a specific classification system, other groups that could generate a new cluster, and to suggest a new distribution of patents. An empirical experiment with five different SOM topologies and five runs of K-Means with different k parameters were used to identify groups of patents, which were clustered together in most of those topologies and runs.

The main contribution of this research was to show that SOM networks and K-Means algorithm could identify clusters successfully using bibliographic coupling. It is known that citation analysis is limited by several practical constraints. Often, the authors of documents are not aware of potentially relevant coupling and may even deliberately omit bibliographic coupling. Furthermore, citations appear chronologically and older patents cannot possibly contain citations of newer patents. Nevertheless, the citations may become an alternative to be considered for the creation of new groups, where documents are semantically related and other layers of knowledge can be used to establish relationships between them.

Acknowledgement

This research was supported by grants from "*Fundo de Incentivo à Pesquisa*" (FIP / PUC Minas).

References

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). Modern information retrieval. (2nd.ed.). England: Pearson.

Borgman, C. L. and Furner, J. (2002). Scholarly communication and bibliometrics, Annual Review of Information Science and Technology, *36* (1*)*, 2-72.

Chakrabarti, A. K; Dror, I. and Eakabuse, N. (1993). Interorganizational transfer of knowledge: an analysis of patent citations of a defense firm, IEEE Transactions on Engineering Management, 40 (1), 91-94.

Croft, W. B., Metzler, D. and Strohman, T. (2010). Search Engines: Information Retrieval in Practice. Boston: Addison Wesley.

Engelsman, E. C. and Van Raan, A. F. J. (1994). A patent-based cartography of technology, Research Policy, 23(1), 1-26.

Hall, B. H., Jaffe, A. B. and Trajtenberg, M. (2002). The NBER patent citations data file: lessons, insights and methodological tools. In A. B. Jaffe and M. Trajtenberg (Eds.), Patents, citations & innovations (pp. 403-459). Cambridge, MA, London: MIT Presss.

Haykin, S. (1994). Neural Networks: a comprehensive foundation. New Jersey: Prentice Hall.

He, Y. and Hui, S. C. (2001). PubSearch: a web citation-based retrieval system. Library hi tech, 19, 274-285.

Hjorland, B. (2002). Domain analysis in information science: eleven approaches – traditional as well as innovative, Journal of Documentation, 58, 422-462.

Jacob, E. (2004). Classification and categorization: a difference that makes a difference, Library Trends, 52(3), 515-540.

Kukolj, D. et al. (2012). Comparison of Algorithms for Patent Documents Clusterization. In: MIPRO Proceedings of the 35th International Convention, Opatija, Croatia, 995-997.

Lai, K-K. and Wu, S-J. (2005). Using the patent co-citation approach to establish a new patent classification system, Information Processing & Management: an International Journal, 41(2), 313-330.

Li, X., Chen, H., Zhang, Z. and Li, J. (2007). Automatic patent classification using citation network information: an experimental study in nanotechnology, In: Proceedings of the seventh ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'07). Vancouver, Canada.

Liu, D-R. and Shih, M-J. (2011). Hybrid-patent classification based on patent-network analysis, Journal of the American Society for Information Science and Technology, 62(2), 246-256.

Meireles, M. R. G., Cendón, B. V. and Almeida, P. E. M. (2014). Bibliometric Knowledge Organization: A Domain Analytic Method Using Artificial Neural Networks, Knowledge Organization, 41(2), 145-159.

Meireles, M. R. G., Ferraro, G and Shlomo, G. (2016). Classification and information management for patent collections: a literature review and some research questions, Information Research, 21(1).

Morris, S. A., Wu, Z. and Yen, G. (2001). A SOM mapping technique for visualizing documents in a database. In: Proceedings of the International Joint Conference on Neural Network, Washington, D. C., 1914-1919.

Pfitzner D., Leibbrandt R. and Powers D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings, Knowledge and Information Systems, 19, 361-394.

Sapsalis, E., Van Pottelsberghe de la Potterie, B. and Navon, R. (2006). Academic versus industry patenting: an in-depth analysis of what determines patent value, Research Policy, 35 (10), 1631-1645.

Smith, H. (2002). Automation of patent classification, World Patent Information, 24(4), 269-271.

Tikk. D., Biró, G. and Törcsvári, A. (2008). A hierarchical Online Classifier for Patent Categorization, 244-267.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations, The Rand Journal of Economics, 21(1), 172-187.

Widodo, A. and Budi I. (2011). Clustering Patent Document in the Field of ICT (Information & Communication Technology). In: International Conference on Semantic Technology and Information Retrieval, Putrajaya, Malaysia, 203-208.