Clustering algorithms performance analysis applied to patent database

Cinthia M. Souza
Institute of Mathematical Sciences and Informatics
Pontifical Catholic University of Minas Gerais
Belo Horizonte, Brazil
cinthia.mikaela@sga.pucminas.br

Magali R. G. Meireles
Institute of Mathematical Sciences and Informatics
Pontifical Catholic University of Minas Gerais
Belo Horizonte, Brazil
magali@pucminas.br

Paulo E. M. Almeida
Intelligent Systems Laboratory
Federal Center for Technological Education of Minas Gerais
Belo Horizonte, Brazil
pema@lsi.cefetmg.br

***Abstract*—The granularity of large patent classification systems hampers the reclassification process in which patent categories are broken down into smaller ones, suggesting new categories. As these groups belong to a constricted domain of knowledge, keywords and subject descriptors tend to be similar and therefore insufficient to differentiate documents. In this context, the identification of common cited references can be useful to define semantic relationship among patents. This work compares citation analysis based results obtained by three clustering algorithms, SOM networks, K-Means and Multi-SOM. An empirical experiment was conducted using a patent database from the United States Patent and Trademark Office with all patents of four subgroups classified by the Cooperative Patent Classification system. Practical results evaluated by statistical inference techniques showed that SOM performs better than the other algorithms to cluster that database. This study can contribute with the reclassification process for a subgroup level of current patent classification systems, demonstrating how citation analysis can be an alternative attribute to the automatic clustering process.**

***Index Terms*—Clustering algorithms, Computational intelligence, Knowledge representation, Patent database, Statistical inference**

I. Introduction

A patent is a public concession, whereby the government, in exchange for full disclosure of an invention, grants the inventor the right to exclude others for a limited time from making, using or selling this invention [1]. Patents are organized into classification systems according to their technical application and structural characteristics to aid the patenting and retrieval processes.

With the growth of digital patent collections, the number of patents at all levels of classification systems has been increasing and some groups need to be dismantled in order to generate new groups and facilitate access to information. Considering the patents subgroups, which are subsets in an equal knowledge area and have a lot of similar words in their abstracts, it is a challenge to identify common characteristics using words as attributes of the clustering process.

The main objective of this work is to evaluate the performance of some algorithms of patents clustering using citations as attributes. For this, three clustering algorithms will be used on a United States Patent and Trademark Office (USPTO) patent database. This work was divided into five sections. Section II presents a description of the implemented algorithms and some related works. Section III presents the database and proposed methodology, while Sections IV and V show the results and final considerations.

II. Theoretical Background

SOM networks are maps of artificial neurons developed by Teuvo Kohonen in the 1980s. These structures, based on topological maps present in the cerebral cortex, are responsible for the execution of the grouping process. Each input neuron is connected to an output neuron by its respective association weight. This network uses unsupervised learning. From the instant the network identifies the regularity between the input data, it generates internal representations to encode the input characteristics and automatically create new groups. These networks have the capacity for self-organization and are more similar to neurobiological structures than supervised networks. Many of the experiments reported in the literature describe the use of SOM in grouping documents so to organize them as an alternative format for information retrieval [2].

The Multi-SOM algorithm is an extension of the SOM algorithm. This algorithm uses simultaneously several maps of SOM to cluster input patterns. The amount of simultaneous maps is defined by the user. For the initial map, data training is performed using the SOM algorithm. For the generation of the next maps, the algorithm realizes the superposition and the communication between the previous and the current maps. To carry out the transition from one map to another, it is necessary to define the new nodes. These nodes are defined by using the

DOI reference number: 10.18293/SEKE2019-101

mean square composition of the four neighbors, directly from the lower level [3], [4].

Performing this procedure preserves the original neighborhood at the higher levels and also the topographical properties of the maps [3], [4]. The Multi-SOM algorithm used in this work was implemented using an R library called “multisom” [5] available for R language. This algorithm not only performs grouping of data but also estimates the optimum number of clusters. At the end, it returns the best result obtained for the input data.

K-Means is an unsupervised data clustering algorithm. The main idea of this algorithm is to define k centroids, one for each cluster. After defining the centroids, the algorithm associates the centroid with the closest data. Then the centroids are recalculated and the previous steps are repeated for the new centroids until the centroids no longer move or the stop criterion is met [6].

Meireles, Cendón and Almeida [7] presented a comparison of document clustering process using keywords and citations. The experiments were performed using a test database with 200 articles from a restricted knowledge domain. The first experiment used the keywords of the articles in the test database as attributes for the clustering process. The second experiment was carried out using the citations of the articles as attributes. Both tests were performed using a SOM network. The experimental results showed that, in a domain of restricted knowledge with a great similarity between documents, the use of keywords was not very efficient. On the other hand, the use of citations can be considered an important alternative.

Lai and Wu [8] proposed an approach to create a new patent classification system, to assist patent managers in evaluating the basic patents for a specific industry. Li, Chen, Zhang and Li [9] considered the structure of patent citation networks for patent classification. They adopted a Kernel-based approach to capture content information and citation-related information in patents, and their proposal outperformed Kernel’s which did not use citation network structures. Liu and Shih [10] combined content-based, citation-based and metadata-based classification methods to develop a hybrid-classification approach using a modified K-Nearest Neighbor (KNN) algorithm.

III. Proposed Approach

The database used in the experiment was extracted from USPTO. The Cooperative Patent Classification (CPC) system, which is used by USPTO, classifies patents into sections, classes, subclasses, groups, and subgroups. Figure 1 illustrates the organization of this patent database and it was highlighted the low level hierarchies subgroups used in this work.

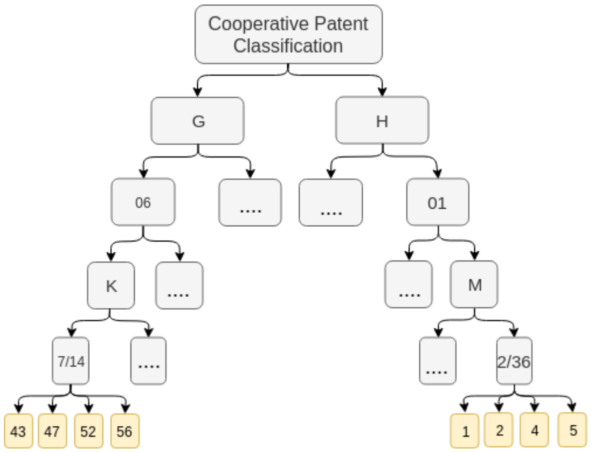

Fig. 1. Organization of the Database

In order to validate the proposed clustering process, two databases were created. Each database is composed by four subgroups randomly chosen in two distinct sections of the CPC system. In the hierarchy defined by CPC, the subgroups represent patent groups contextualized in the same area of knowledge. The first base is composed of subgroups G06K 7/1443, G06K 7/1447, G06K 7/1452 and G06K 7/1456 of the G06K subclass, named “recognition of data, presentation of data, record carriers, handling record carriers”, shown in Figure 1 only with their suffix 43, 47, 52 and 56. The second base is composed of subgroups H01M 2/361, H01M 2/362, H01M 2/364 and H01M 2/365 of the subclass H01M, named “processes or means, e.g. batteries, for the direct conversion of chemical energy, into electrical energy” represented in Figure 1 only with their suffix 1, 2, 4 and 5. Some patents of the subgroups are classified into more than one subgroup. In this work, only the first classification is considered and, therefore, the number of patents available is different from the number of patents selected for the database. Table I shows the name of the subgroups in the CPC system and the number of patents selected.

TABLE I
Database

Sections	Codes CPC	Number of patents available	Number of patents selected
G	G06K 7/1443	505	452
	G06K 7/1447	302	263
	G06K 7/1452	93	78
	G06K 7/1456	227	117
H	H01M 2/361	213	185
	H01M 2/362	126	101
	H01M 2/364	33	28
	H01M 2/365	139	59

The methodology used to cluster the documents was divided into three execution steps and one analysis phase. In the first step, the citations of the patents contained in the selected subgroups were extracted. These citations were taken from the section of documents referenced by the patent. From this process, two binary citation matrices per document were generated, one for each database, which inform the occurrence of a certain citation in each document. In these matrices, the digit 0 represents the absence of a citation in the document and the digit 1 represents the existence of a citation in the document. According to Borgman and Furner [11], the analysis of citations allows for the identification of relationships documents, regardless of the presence of equal terms. In this work, the occurrence of common quotes among patent documents is used as a mechanism to define the semantic relations between them.

In the second step, those matrices were used as inputs for each one of the three algorithms discussed in Section II. For each algorithm, the experiments were repeated for 30 times, to account for statistical validation. For the SOM network and K-Means, the number of clusters k was defined as 4, which was the number of subgroups of the database used for validation. The Multi-SOM does not need to receive as input the number of clusters. However, it is necessary to define the dimensions of the first map. Thus, the first map dimension was defined as 6 x 6.

In the third step, an algorithm was implemented to evaluate the correspondence between the groups generated by SOM, Multi-SOM and K-Means algorithms and the original CPC subgroups presented in Table I. Finally, in the analysis phase, an objective comparison was performed using statistical inference, by hypothesis tests. In this test, the hypothesis tested (H_0) was the equality of average between the number of patents identified in step 3 and the number of patents selected in each subgroup. In this test, a categorization algorithm will be considered more efficient when the average is closer to zero. At the end, it was possible to infer if some of the algorithms were better or worse than the others, to solve that clustering problem. The alternative hypothesis (H_1) used for the Kruskal-Wallis H test and boxplot was the average difference in the number of patents clustered (ADPC) by each algorithm in the clusters, i.e. the difference between the number of patents identified, for example in group G1, and the number of patents selected in the corresponding original CPC subgroup. A low ADPC informs that the associated clustering algorithm groups together a number of patents similar to those of the original CPC subgroups.

IV. Experiments

The experiments were divided into two phases. In the first phase, the tests were carried out with the database composed by the patents of section G. A total of 10,148 citations were extracted from the 910 patent documents. The matrices generated in this process were used for the clustering process. Fig. 2 shows the distribution of patents in the four groups generated by SOM algorithm, refered here as G_G1_S, G_G2_S, G_G3_S and G_G4_S. Kruskal-Wallis H test was performed to compare those samples from each of the groups generated, the test suggested that there were statistical differences between the three samples (with p_0 = 5.31 $\times 10^{-18}$ for the group G_G1 and with p_0 = 3.08 $\times 10^{-18}$ for the group G_G2). To find the difference of samples, a comparison was performed by means of boxplot representations. This resulted in the plots of Fig. 3 and Fig. 4. It is possible to infer, with 95% of confidence, that SOM algorithm performs better than K-Means and Multi-SOM to cluster patents of this database by means of citations, as SOM averaged differences between number of patents identified on step 3 and patents selected in each subgroup (H_1) could not be rejected. Therefore, it is clear from this analysis that SOM performed much better than K-Means and Multi-SOM, as their ADPC is smaller than those by its conterparts.

To the groups identified by K-Means algorithm, it was added the termination KM (G_G1_KM, G_G2_KM) and to those identified by Multi-SOM, it was added the termination MS (G_G1_MS, G_G2_MS).

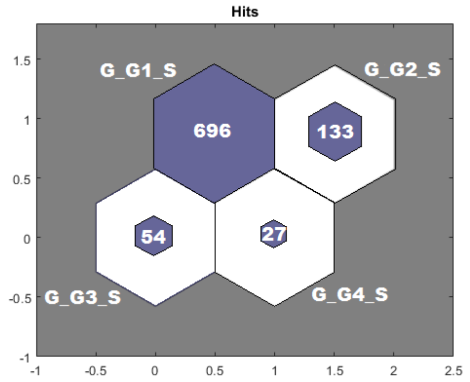

Fig. 2. Typical result of the clustering process using a SOM network

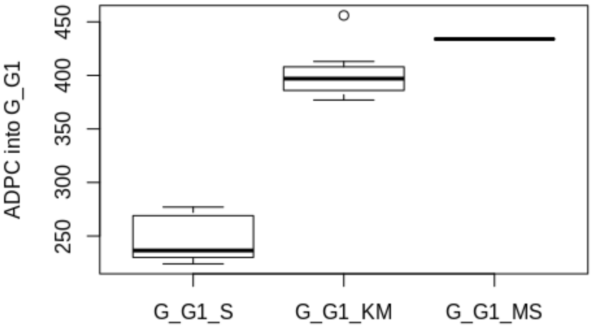

Fig. 3. Boxplot representation of results for G_G1

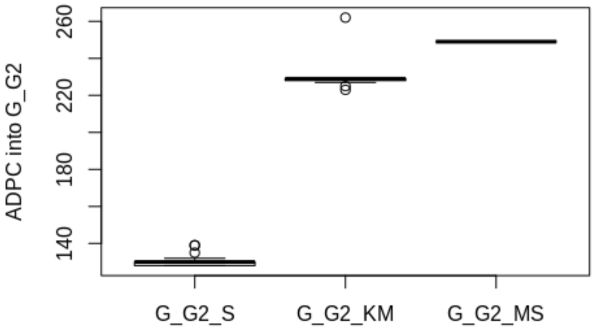

Fig. 4. Boxplot representation of results for G_G2

From the four groups created in the experiments, it was possible to identify two of them, which had a majority of patents, one from G06K 7/1443 and the other from the G06K 7/1447 subgroup. The other two groups had patents from the four used subgroups. Analyzing the first one created by SOM, named G_G1_S, with 696 averaged patents, 62.75% belonged to G06K 7/1443 subgroup of the documents database. K-Means and Multi-SOM created a big group with averages of 844.6 and 886 patents, respectively. From these, 58.60% and 54.85% in average were from the G06K 7/1443 subgroup. This indicates that K-Means and Multi-SOM algorithms failed to identify differences between documents from their citations, keeping the vast majority of patents in a single group.

The second group analyzed, created by SOM, named G_G2_S, had 133 patents in average. A total of 100% belonged to the G06K 7/1447 subgroup of the documents database. K-Means and Multi-SOM algorithms kept, in average, 100% and 28.57% of the patents from the G06K 7/1447 subgroup in G_G2_KM and G_G2_MS. But, while SOM was able to keep an average of 133 patents on G_G2_S, K-Means and Multi-SOM clustered only 35 patents in G_G2_KM and 14 patents in G_G2_MS respectively, in average. Table II presents the results obtained.

TABLE II
RESULT OF THE FIRST PHASE OF THE CLUSTERING PROCESS

Groups	Average cluster size	Average hit percentage
G_G1_S	696	62.75%
G_G2_S	133	100%
G_G1_KM	845	58.6%
G_G2_KM	35	100%
G_G1_MS	886	54.85%
G_G2_MS	14	28.57%

In the second phase, the tests were performed with the database composed of patents of section H. A total of 2,755 citations were extracted from the 373 patent documents. Fig. 5 shows the distribution of patents in the four groups generated by SOM algorithm, refered here as H_G1_S, H_G2_S, H_G3_S and H_G4_S. The Kruskal-Wallis H test suggested that there were statistical differences between the three samples of the three groups generated (with p_0 = 1.26 $\times 10^{-18}$ for the group H_G1, p_0 = 8.99 $\times 10^{-19}$ for the group H_G2 and p_0 = 6.05 $\times 10^{-17}$ for the group H_G3). To find the difference of samples, a comparison was performed by means of boxplot representations. The Fig. 6, 7 and 8 present a boxplot representation of the results in terms of ADPC. It is possible to infer, with 95% of confidence, that SOM algorithm performs better than K-Means and Multi-SOM to cluster patents of this database by means of citations, as SOM averaged differences between number of patents identified on step 3 and patents selected in each subgroup (H_1) could not be rejected. Again, it is clear that SOM's ADPC is significantly smaller than ADPC obtained by K-Means and Multi-SOM.

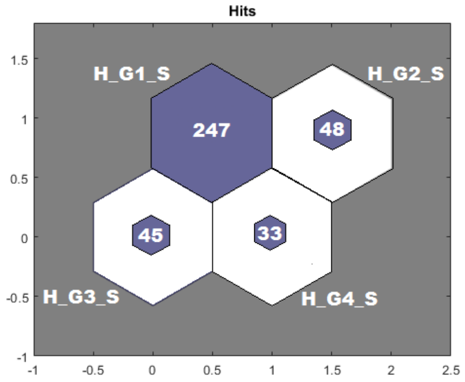

Fig. 5. Typical result of the clustering process using a SOM network

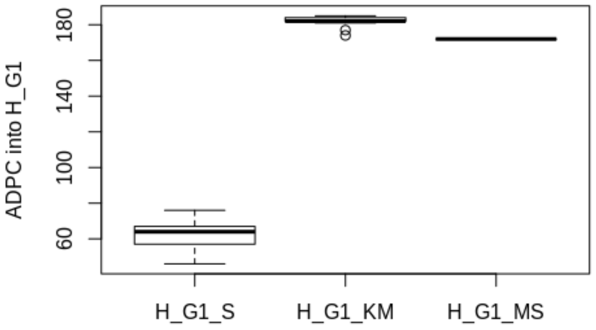

Fig. 6. Boxplot representation of results for H_G1

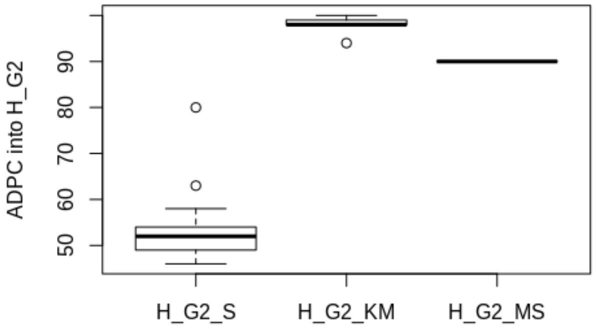

Fig. 7. Boxplot representation of results for H_G2

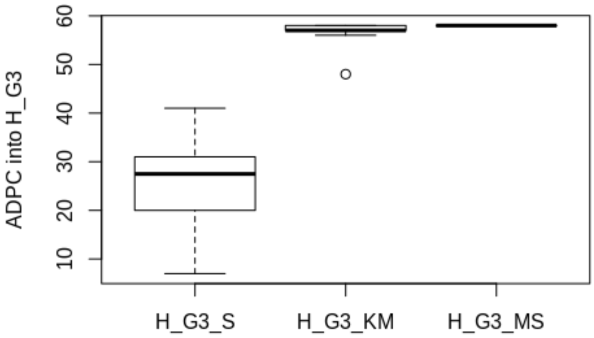

Fig. 8. Boxplot representation of results for H_G3

From the four groups created by the algorithms, it was possible to identify three with more similarity to the subgroups created by the specialists. These are equivalents to subgroups H01M 2/361, H01M 2/362 and H01M 2/365. These groups were named as H_G1, H_G2 and H_G3, respectively. The H_G1_S group created by SOM contains 247 patents in average, of which 57.82% were correctly grouped. K-Means and Multi-SOM have created the group H_G1_KM with 367 and H_G1_MS with 357 patents in average. From these, 57.17% and 56.58% were correctly grouped.

Analyzing the groups H_G2_S and H_G3_S it was identified that only SOM managed to create groups of a relevant size. The H_G2_S group created by SOM contains 48 patents in average, of which 78.30% were correctly grouped. The H_G3_S group created by SOM contains 45 patents in average, of which 57.41% were correctly grouped. Therefore, it is possible to state that the SOM has a more satisfactory result than K-Means and Multi-SOM, since it can better identify the differences between patent documents. K-Means and Multi-SOM clustered the vast majority of patents into a single group. These algorithms were not able, in these experiments, to identify differences between the documents that allowed them to be clustered in different groups. We believe that SOM could perform better than Multi-SOM to solve a given problem, even being nothing but a special case of Multi-SOM, because SOM can be more specific and specialized than Multi-SOM, thus being more precise. On the other hand, Multi-SOM is more general, and perhaps capable of dealing better with different instances of the problem. The percentage of patents correctly grouped by algorithms is very close, in some cases, this is due to the fact that the number of patents in the generated groups is very small. Table III shows the groups created, the average size of each cluster and the average percentage of patents correctly classified.

TABLE III
RESULT OF THE SECOND PHASE OF THE CLUSTERING PROCESS

Groups	Average cluster size	Average hit percentage
H_G1_S	247	57.82%
H_G2_S	48	78.30%
H_G3_S	45	57.41%
H_G1_KM	367	57.17%
H_G2_KM	3	58.70%
H_G3_KM	1	36.06%
H_G1_MS	357	56.58%
H_G2_MS	3	68.67%
H_G3_MS	2	50%

V. Conclusion

With the increasing number of patents and the development of new technologies, the classification systems employed by patent offices should be constantly reviewed to avoid accumulation of documents on certain subgroups. In a restricted domain of knowledge such as the subgroups of CPC system, it is difficult to use words as units of knowledge representation in an automatic clustering process because the subject descriptors and the words tend to be similar.

The main contribution of this work is to evaluate the performance of three clustering algorithms on a restricted knowledge domain, based on CPC sub-groups. The experiments brought the theory of citation analysis to a practical application of interest to the academic and industry communities. For the given scenarios, SOM networks showed superior performances compared with K-Means algorithm and Multi-SOM networks. Most of patent offices professionals and researchers in the domain of information retrieval and applied machine learning deal with the upper levels of classification hierarchies (class and subclass levels) and only some have tracked the problem on a more fine-grained classification (group and subgroup levels), as done in this work. For future work, it is expected to perform the comparison of the clustering process in a larger scale, using the upper hierarchy of the CPC system.

Acknowledgements

The authors would like to thank the financial support of the Pontifical Catholic University of Minas Gerais (PUC Minas), the Federal Center for Technological Education of Minas Gerais (CEFET-MG), the National Council for Scientific and Technological Development (CNPq, grant 429144/2016-4) and the Foundation for Research Support of the State of Minas Gerais (FAPEMIG, grant APQ 01454-17).

References

[1] T. Hufker and F. AlpertL, “Patents: a Managerial Perspective”, Journal of Product and Brand Management, vol. 3, pp. 44-54, 1994.

[2] X. Luo, A. N. Zincir-Heywood, “A comparison of som based document categorization systems”, in: Proceedings of the International Joint Conference on Neural Networks, vol. 3, pp. 1786-1791, 2003.

[3] X. Polanco, C. François, and J-C. Lamirel, “Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach”, Scientometrics, vol. 51, pp. 267-292, 2001.

[4] I. Khanchouch, M. Charrad, and M. Limam, “A Comparative Study of Multi-SOM Algorithms for Determining the Optimal Number of Clusters”, International Journal of Future Computer and Communication, vol. 4, pp. 198-202, 2015.

[5] S. Chair, M. Charrad, and N. Ghazzali, “A new r package for multi-som clustering”, in Conférences Conjointes Francophones sur la Sciences des Données AAFD & SFC, 2016.

[6] D. Xu, Y. Tian, “A comprehensive survey of clustering algorithms”, Annals of Data Science, vol. 2, pp. 165-193, 2015.

[7] M. R. G. Meireles, B. V. Cendón and P. E. M. Almeida, “Comparação do processo de categorização de documentos utilizando palavras-chave e citações em um domínio de conhecimento restrito”, Transinformação, Campinas, vol. 28, pp. 87-96, 2016 (in portuguese).

[8] K-K. Lai and S-J. WU, “Using the patent co-citation approach to establish a new patent classification system”. Information processing & management, Elsevier, vol. 41, pp. 313-330, 2005.

[9] X. Li, H. Chen and Z. Zhang and J. Li, “Automatic patent classification using citation network information: an experimental study in nanotechnology”, in Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL ’07), pp. 419-427, 2007.

[10] D-R. Liu and M-J. Shis, “Hybrid-patent classification based on patent-network analysis”, Journal of the Associationfor Information Science and Technology, Wiley OnlineLibrary, vol. 62, pp. 246-256, 2011.

[11] C. L. Borgman and J. Furner, “ Scholarly communication and bibliometrics”, Annual review of information science and technology, v. 36, n. 1, p. 2-72, 2002.