International Journal of Information Science and Management
Vol. 22, No. 2, 2024, 139-149
DOI: https://doi.org/10.22034/ijism.2024.1977932.0 / DOR: https://dorl.net/dor/20.1001.1.20088302.2024.22.2.9.8

Original Research

Using the Citation-Content-Based Approach to Patent Clustering

Narges Neshat
Professor, Knowledge and Information Science, National Library & Archive of Iran, Tehran, Iran.
Corresponding Author: n.neshat.sh@gmail.com
ORCID iD: https://orcid.org/0000-0002-8959-8007

Anahita Kermani
Ph.D. student in Knowledge and Information Science, Shahid Beheshti University, Tehran, Iran.
kermanianahita@gmail.com
ORCID iD: https://orcid.org/0009-0000-5543-1498

Received: 17 August 2022
Accepted: 15 January 2024

Abstract

patents are a significant competitive strategy to categorize commercial value based on the source information of technology; researchers use patent analysis as a practical tool to infer various types of information. This shows how important it is to retrieve and access them. Clustering is a method used in different fields to group similar natures. Citations are commonly used to cluster documents, and two methods are widely used for this purpose. The first method uses bibliographic coupling, and the second method identifies the words in the citation titles, also called co-citation. However, it is necessary to investigate which methods provide better patent clustering and retrieval results. This study examines citation contents instead of citations in building relevant groups of patents. Experimental research was done on a set of US patents. The analysis is divided into three phases. The first is appropriate databases to conduct patent searches according to the subject and objective of this study. The basic inventions and the experimental set were selected. Phase II, for developing a patent clustering system based on patent similarities and assisting the relationships among categories, we used fuzzy c-means (FCM) clustering because it can handle overlapping clusters similar to k-means. As fuzzy clustering is a kind of overlapping clustering, extended B Cubed precision and recall - measures for evaluating overlapping clustering - were used. Since patents can belong to multiple technology domains, in phase III, a Perl program was written to manage the matching process. The study involved creating two patent clusters using bibliographic coupling and citation title words, respectively. The results indicated that the bibliographic coupling method produced better clustering performance than the citation title words. Moreover, the cluster structure was more extensive in terms of exhaustivity than the citation title words. It's interesting to note that the use of cited patent title words resulted in a reduction of nearly 40% of the number of attributes. Additionally, when compared to the use of bibliographic coupling, the cited title words method had a nearly equal recall of clustering by cited patents in high exhaustivity. As a result, it appears that using cited title words may be preferable when the high exhaustivity approach is selected for patent clustering and retrieval.

Keywords: Clustering, US Patent Classification**,** Citation**,** Recall and Precision**,** Fuzzy C-Means Evaluation, Patent Citation, Bibliographic Coupling.

Introduction

Patents are documents containing both technical and legal information. In fact, according to patent law, a legal right is granted to inventors by governments in return for disclosing detailed technical information about their invention. An inventor should claim his/her invention through a patent application, and after a demonstration of eligibility, a patent is granted to the inventor. The patenting process can be carried out at international, regional, or national levels. Terms of eligibility are, to some extent, different in distinct patent offices, but novelty and non-obviousness are the general criteria a claimed invention should have to be patented (Habiba, 2004). To examine such criteria, examiners in patent offices should search through material, including patents, and compare them with the claimed invention in a patent application.

The use of patent and non-patent citations as indicators of innovation has increased dramatically in the last decade. As citations indicate the S&T precedents in inventions, they make it possible to track knowledge. The patent documentation has a higher technical value than the general scientific documents. Through effective retrieval of patent documents and awareness of their content, enterprises may find some effective solutions, gain access to new ideas about research and development (R&D), and avoid existing patent landmines at the same time. On the other hand, understanding the evolution of a technological field over time is a key task in patent analysis. Therefore, patents are a major competitive strategy to categorize commercial value based on the source information of technology; researchers use patent analysis as a practical tool to infer various types of information. This shows how important it is to retrieve and access them.

Clustering is a method that is used in different fields to group similar natures. Among the applications of this method in the field of documents, we can mention its use in information retrieval and drawing scientific maps. Various factors are effective in document clustering, one of which is related to the type of feature that a document is represented by. Feature selection is a crucial step in creating a vector space, especially in document processing. When we need to group similar documents, the selection of features and good clustering algorithms play a significant role. Citations are commonly used to cluster documents, and two methods are widely used for this purpose. The first method uses bibliographic coupling, and the second identifies the words in the citation titles, which is also called co-citation. However, it is necessary to investigate which methods provide better patent clustering and retrieval results.

The application of cited references to achieve subject groups of documents was first proposed by Kessler (1963) in a method called bibliographic coupling. The method was based on the presumption that cited references are related to the subject matter of the citing article; therefore, assessing related articles would be possible by sharing cited references instead of subject-indexed terms.

Various motivations behind citation behavior (Meyer, 2000) in articles made this presumption somewhat wrong to the extent that it, on one hand, caused the development of other citation-based methods and analysis, like co-citation (Small, 1973) and, on the other hand, took attention into comparing subject with citation elements (Salton, 1963; Salton, 1971; Horri, 1981; Shaw, 1990; Shaw, 1991). Although in these researches, the preference for citation loneliness was not proved significantly (Salton, 1963; Horri, 1981), compounding citations with subject indexing showed improvement in final performance (Salton, 1963; Shaw, 1990; Shaw, 1991).

However, motivations behind patent citations are, to some extent, different from citations of scientific papers. Regarding the United States (US) patent system[1], they can be considered relevant materials that the patent applicants/examiners should cite during the patent application/examination process (USPTO, 2020). Citations in patents seem more indicative of technological relevance than impact, in contrast to scientific paper citations (Meyer, 2000). Thus, the presumption of citation analysis methods such as bibliographic coupling sounds more adjusted to the nature of patent citations than those of scientific papers.

Some studies introduced retrieval methods applying patent citations in the patent retrieval process (Fujii, 2007; Tiwana & Horowitz, 2009). However, historically, the usage of patent citations goes back to a suggestion from Seidel and Hartel, two patent attorneys, and the development of the Patent Citation Index by Garfield (Wouters, 1999). It is shown that integration of text-based and citation-based methods improves the effectiveness of invalidity search patent retrieval (Fujii, 2007).

To automate the patent classification process, patent citation analysis was also proposed, and the benefit of cited documents was proved. It was shown that the usage of a citation network performs better than direct citation in automatic patent classification (Li, Chen, Zhang & Li, 2007). Even so, IPC and UPC [USPC], two patent classification systems, need to be more general to meet the industry's needs. Instead, the three-phased method proposed within which the co-citation method determines the similarity of patents (Lai & Wu, 2005).

Although there are some comments on patent citation analysis (Kim, Suh & Park, 2008), the desirable categorization of patents has also resulted in using a bibliographic coupling method.

This study aims to determine whether citations in patents can be replaced with their word content to achieve relevant groups of patents. As cited matters are relevant to the patent invention, words contained in cited patent titles are more likely to be relevant to citing patents and cause desirable categorization.

Measurement of proximity

A. Clustering: Clustering is one of the methods used to develop categorization from a collection of documents. This method can be applied in the retrieval and development of scientific and technological maps (Leydesdorff, 1987; Kim, Suh & Park, 2008).

By clustering, documents are heuristically gathered to the same cluster based on similarity, while similarity is recognized based on common attributes. In this study, words of cited patent titles are evaluated for use instead of cited patents as attributes. Compared with clustering, classification is another method for categorization (Tan, Steinbach & Kumar, 2006). The significant difference between clustering and classification is that documents are placed in undetermined groups in clustering. Still, in classification, the groups are determined before putting documents on them, so clustering is recognized as unsupervised, and classification is known as a supervised organization (Ibid).

Three main steps for clustering are representing documents, defining proximity measures, and finally implementing clustering or groping (Jain, Murty & Flynn, 1999). Documents may be represented by their attributes determined for them as a vector. In this study, patents were represented once by their words of cited patents and once by cited patent number(s), as identification of cited patents. By proximity measure, the distance between documents is recognized. Then, documents are clustered based on the clustering algorithm.

One of the common proximity measures is “Cosine similarity” (Huang, 2008). According

to this measure similarity of documents is calculated as follows:

$$\mathrm{sim}(d_i, d_j) = \cos\theta = \frac{\vec{d_i}.\vec{d_j}}{|\vec{d_i}||\vec{d_j}|}$$

Where "d" is the symbol of the document, $\vec{d_j}.\vec{d_k}$ is the dot product of the vector of documents i and j, and $|\vec{d_i}|$ is the length of the vector of document i and $|\vec{d_j}|$ is the length of the vector of document j.

So in this regard, the distance between documents is equal to:

$$Distance = 1 - sim$$

Interestingly, clustering algorithms can be classified into two main categories: hierarchical and "partitional"[2]. Hierarchical clustering involves organizing the members into a hierarchy, with the highest level containing all members and the lowest level consisting of clusters containing only one member each. On the other hand, partitional clustering involves determining a set number of centroids and assigning members to clusters based on their similarity to these centroids. Two famous examples of these clustering methods include agglomerative hierarchical clustering and k-means clustering, respectively. For large datasets, "partitional clustering" is more suitable than hierarchical clustering (Huang, 2008). Partitional clustering algorithms are more efficient and scalable for large datasets, while hierarchical clustering algorithms tend to be computationally expensive and may not perform well with large datasets. Additionally, Partitional clustering offers more flexibility in choosing the number of clusters, whereas hierarchical clustering generates a set hierarchy of clusters that may not be appropriate for all datasets.

Clustering can also be regarded as overlapping and non-overlapping. In non-overlapping ones, members can belong to one cluster, but members can be in more than one cluster in overlapping clustering (Jain et al., 1999). It is essential to consider the nature of the items being clustered when choosing a clustering algorithm.

According to Wedding (2009), the algorithm of fuzzy clustering is as follows:

1. For a data set consisting of *N* document, select the desired number of clusters, *k*, where $k < N$.

2. Generate a starting center point for each of the *k* clusters.

3. Calculate the distance from each of the *N* documents to each of the k clusters.

4. Assign a proportional or fuzzy membership of each of the *N* documents to each of the *k* clusters.

5. Find the new center point for each of the *k* clusters by finding the weighted average of the records.

6. Repeat steps 3, 4, and 5 until there are no changes in the cluster membership (or until some convergence criteria is met).

In this way, the membership degree (ibid) is calculated as follows:

$$u_k = \frac{1}{\sum_{i=1}^{j}\left(\frac{d_k}{d_i}\right)^p}$$

Where *k* is one of the *j* clusters, *d* values are the distance values and *p* is the power calculated (Wedding, 2009) by the following equation:

$$p = \frac{2}{(m-1)}$$

Where *m* is a fuzzy exponent, typically equals 2 (ibid).

B. Clustering Evaluation: Clustering evaluation measures can be divided into intrinsic and extrinsic measures (Amigo, Gonzalo, Artiles & Verdejo, 2008). Intrinsic measures determined the amount of similarity between members of the same clusters and dissimilarity between members in distinct clusters. Extrinsic measures are performed based on a pattern, and the resulting clustering is compared with it. This pattern, which is called "ground truth" (Tan et al., 2006), the golden or gold standard, is usually made by human experts (Amigo et al., 2008). Extrinsic measures are commonly used in the evaluation of text clustering (ibid).

Evaluation measures are not the same for both overlapping and non-overlapping clustering. Amigo et al. (2008) compared different extrinsic measures and found BCubed precision and recall the best among other measures. These measures are adjusted with non-overlapping clustering. They extended them to be adapted for overlapping clustering and developed EBP/R (Extended BCubed Precision/Recall).

To assess EBP/R, first multiplicity precision and recall should be calculated defined as follows (ibid):

$$Multiplicity\ Precision(e, e') = \frac{Min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|C(e) \cap C(e')|}$$

$$Multiplicity\ Recall(e, e') = \frac{Min(|C(e) \cap C(e')|, |L(e) \cap L(e')|)}{|L(e) \cap L(e')|}$$

where e and e' are two patents, L(e) the set of reference clusters and C(e) the set of clusters associated to e. Note that multiplicity precision is defined only when e and e' share some cluster, and Multiplicity Recall when e and e' share some reference clusters (Ibid).

Now, extended BCubed Precision and recall (Ibid) are defined as:

$$Precision\ BCubed = Avg_e\left[Avg_{e'.C(e)\cap C(e')\neq\emptyset}[Multiplicity\ precision(e, e')]\right]$$

$$Recall\ BCubed = Avg_e\left[Avg_{e'.L(e)\cap L(e')\neq\emptyset}[Multiplicity\ recall(e, e')]\right]$$

Material and Methods

An experimental study was done on a set of patents. The analysis is divided into three phases. In the first, appropriate databases to conduct patent searches according to the subject and objective of this study. The basic inventions and the experimental set were selected. In this regard, we chose a set of 717 US patents cited in 75 US patents belonging to the 977/774 class of US Patent Classification (USPC). A collection of 75 patents was selected as reference clustering and cited US patents were considered as test collection. Notably, references to a patent can be considered as relevant matters to that patent (Graf & Azzopardi, 2008). As citations in US patents are prepared or approved by experts (patent examiners) (USPTO, 2020) it is possible to imagine a patent as a reference cluster that references are cluster members. So, a collection of patents can be considered as reference clustering. In this regard, a collection of patent references can be considered as a test collection that resulted from clustering on them

and can be evaluated according to citing patents.

Attributes (cited patents and words of cited patent titles) were semi-manually extracted and stored in Excel Software and weighted according to the tf-idf method using Rapid Miner3, an open-source software. Notably, non-US patents (foreign patents) were not considered attributes; only US-cited patents/patent applications and US-cited patent titles/patent application titles were considered attributes.

The cited patent titles were separated into words, and stopwords were removed before weighting the words. The rest were stemmed using Porter stemmer, which is available in RapidMiner.

In phase II, we used fuzzy c-means (FCM) clustering to develop a patent clustering system based on patent similarities and assist the relationships among categories. FCM can handle overlapping clusters in a way similar to k-means. As fuzzy clustering is a kind of overlapping clustering, extended BCubed precision and recall—measures for evaluating overlapping clustering—were used. Since patents can belong to multiple technology domains, in phase III, a Perl program was written to manage the matching process.

Why Used USPTO? United States Patent and Trademark Office (USPTO) is recognized as one of the credible intellectual property offices acting at a national level and its published patents are called "US patents". According to United States patent laws, "Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor" (35 USC 101). A utility patent, plant patent, and design patent are three particular types of patent in USPTO.

Components of US patents usually comprise of title, abstract, background of the invention, summary of the invention, brief description of drawings, detailed description of the invention, claims, cited references, and drawings:

"The title should be brief but technically accurate and descriptive" (USPTO, 2012) and "must be as short and specific as possible" (37 CFR 1.72 quoted in USPTO (2012). The abstract contains "the nature and gist of the technical disclosure" (37 CFR 1.72. *Title and abstract*, quoted in USPTO (2012), can be understood by rapid examination. In the background of the invention, the "field of the invention," "description of related art," and, if applicable, the problem solved by the invention are explained (USPTO, 2012). The summary points to the specific subject matter of the patent "in one or more clear, concise sentences or paragraphs", ignoring generalities that may be the same for previous patents. A brief description of drawings appears when there are some drawings. This part describes "the several views of the drawings" (USPTO, 2012). In the detailed description explanation should be "be in such particularity as to enable any person skilled in the pertinent art or science to make and use the invention without involving extensive experimentation" (USPTO, 2020). Claims are the legal part of the patents that determine the limitation of the rights. Cited references are prior arts "consisting of patents or printed publications" which may be provided by the patent applicant or examiner _ but approved by an examiner that is relevant to the subject matter of the patent. If needed, some drawings may be provided as patents by the inventor (USPTO, 2020).

Results

The two types of clustering were evaluated based on 6 membership thresholds identified experimentally according to variation in cluster size (the number of patents in a cluster) while changing the membership degree. Figures 1 and 2 show variations in cluster size in different

thresholds when attributes are cited patents and words of cited patent titles.

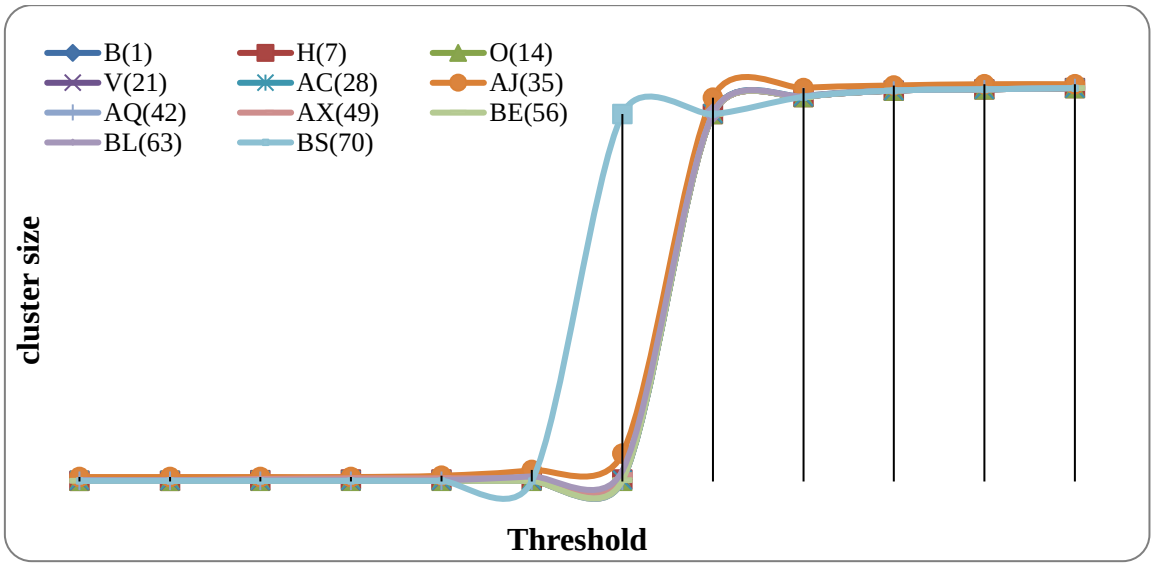

Figure 1: Cluster size variation (cited patents as an attribute)

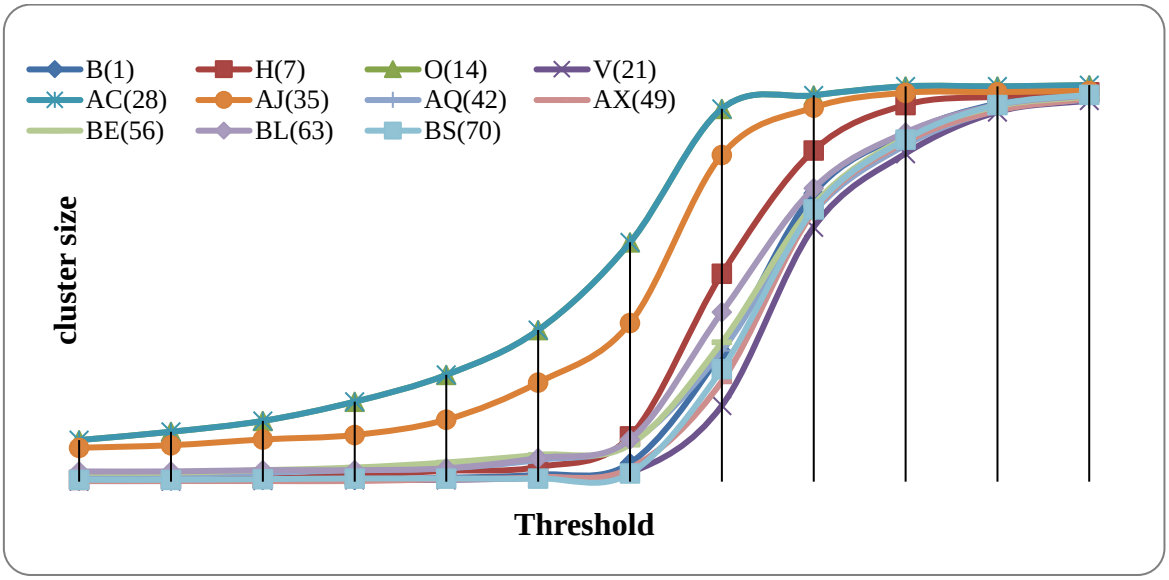

Figure 2: Cluster size variation (words of cited patent titles as an attribute)

In clustering produced by cited patents, changing the threshold from 0.014 to 0.013 causes tangible changes in cluster size for most clusters. In cited patent titles words-based clustering, changes in cluster size start at a threshold of 0.014 and end at a threshold of 0.009. In this regard, the broader range (0.014 to 0.009) is considered for evaluation.

In the evaluation process, thresholds of 0.009 and 0.014 are respectively selected as high and low levels of exhaustiveness. Thresholds of 0.01, 0.011, 0.012, and 0.013 between them were also examined. Table 1 shows the EBP and EBR of clustering in selected thresholds.

Table 1
EBP/EBR in cited patents/words of cited patent titles

	Cited Patents		Words of Cited Patent Titles	
Threshold	EBP	EBR	EBP	EBR
0.014	1	0	_	_
0.013	1	0.7	1	0.26
0.012	1	0.73	1	0.52
0.011	1	0.75	1	0.63
0.010	1	0.75	1	0.70
0.009	1	0.76	1	0.73
Mean	1	0.615	1	0.568
Median	1	0.74	1	0.63

Figure 3 shows a better comparison between the EBP/R of clustering by citation and cited titles words.

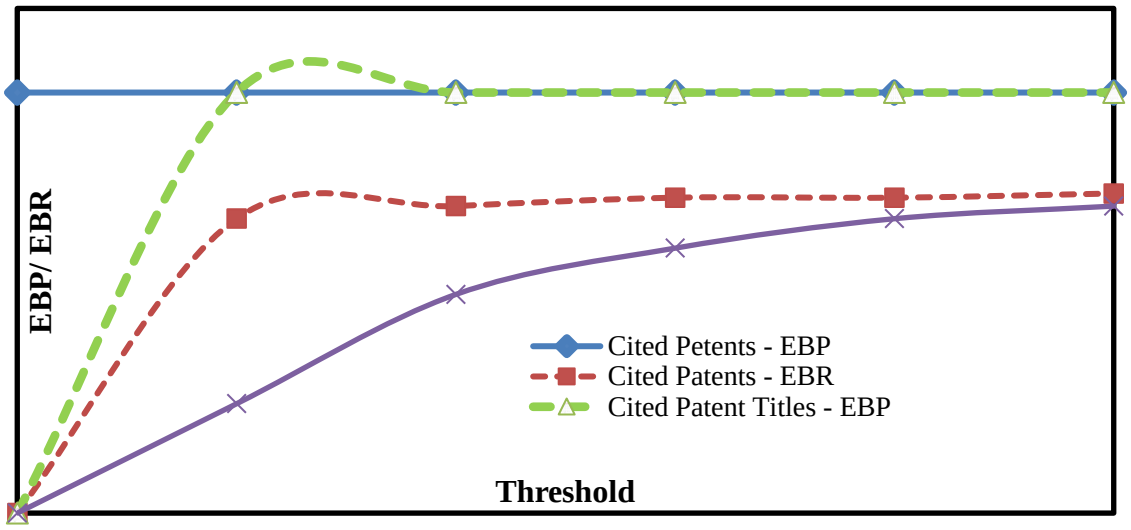

Figure 3: EBP and EBR in cited patents versus words of cited patent titles

As Figure 3 shows, for both clustering in different thresholds, EBP is higher than EBR. For both clustering, EBP equals the maximum value (1) in all thresholds, but EBR is less than 1 and is higher when using cited patents as attributes. Moving from low to high exhaustivity in evaluation, values of EBP have no changes, but EBR varied in different thresholds in both clusters. The rate of changes in EBR is very low when cited patents are used, while the changes are more frequent when words of cited patent titles are used; this can show more constancy of cluster structure power in clustering by cited patents. Also, EBP and EBR are definable in a wider range when using cited patents. This can show the existence of cluster structures in a wider range for cited patents.

Discussion

Patent and non-patent citations are the references provided in the search report that are used to assess an invention's patentability and help define the legitimacy of the claims of the new patent application. As they refer to the prior art, they indicate the knowledge that preceded the invention and may also be cited to show the lack of novelty of the citing invention. However, citations also indicate the legal boundaries of the claims of the patent application in question. They, therefore, serve an important legal function since they delimit the scope of the property rights awarded by the patent. If patent B cites patent A, it means that patent A represents a piece of previously existing knowledge upon which patent B builds or to which patent B relates and over which B cannot have a claim. Hence, citations may be used to preclude the issuance of a patent or limit the scope of the protection to what was specifically known at the time of filing the patent application.

Before patent clustering, attribute selection is a crucial step. One such attribute is citation. In patents, citations have legal implications, and citation lists are related matters that report on the patentability of claimed inventions. The legal aspects of citation in patent documents make patent citation an essential element in patent clustering and retrieval. Clustering can be used for both citing and cited patents. In the clustering of citing patents, citation may be used in two ways: 1) clustering using citation title words and 2) clustering using bibliographic coupling. However, the question remains as to which method has a more effective role in patent clustering. Therefore, this research was conducted to evaluate and compare these two clusters.

An experimental study was conducted to determine the better method for patent clustering and retrieval. The study involved creating two patent clusters using bibliographic coupling and citation title words, respectively. The results indicated that the bibliographic coupling method produced better clustering performance than the citation title words. Moreover, the cluster structure was more extensive in terms of exhaustivity than the citation title words.

Clustering is a technique that groups similar entities and can be used in information retrieval and scientific mapping. This involves grouping documents into clusters so that similar documents are placed together. In document clustering, it's essential to select the right algorithm and feature type. However, citation-based features have also been found to be effective, as both types of features yield clustering evaluations of approximately above 50% of the measured values, indicating that citation-based features are not only suitable for patent clustering but have promising results in the retrieval and classification of patents. Lai & Wu (2005) so applied the co-citation analysis to propose a methodology for establishing a patent classification system. Their approach was composed of three parts: selecting basic patents, assessing the similarities of the basic patents, and establishing a patent classification system. Research by Fujii (2007) also showed a combination of text and citation information improved retrieval accuracy for USPTO patents. Also, Li et al. (2007) in their research, suggested that patent citation information, especially the citation network structure information, can address the patent classification problem. Of course, they adopted a kernel-based approach and designed kernel functions to capture content information and various citation-related information in patents. Their evaluation results showed that a labelled citation graph kernel that uses citation network structures is significantly better than kernels that only use citation information. However, it seems more research is necessary in other technological fields and the use of different clustering algorithms.

Conclusion

The study involved creating two patent clusters using bibliographic coupling and citation title words, respectively. The results indicated that the bibliographic coupling method produced better clustering performance than the citation title words. Moreover, the cluster structure was more extensive in terms of exhaustivity than the citation title words. It is important to note that the volume of citations will increase over time, requiring more storage space for attributes. However, storage space for words extracted from cited titles can be reduced by various methods like removing stop words and stemming. Also, the use of cited patent title words resulted in a reduction of nearly 40% in the number of attributes. Additionally, when compared to the use of bibliographic coupling, the cited title words method had a nearly equal recall of clustering by cited patents in high exhaustivity. As a result, it appears that using cited title words may be preferable when the high exhaustivity approach is selected for patent clustering and retrieval.

Endnotes

1. Note that different patent offices may have distinct laws and rules for patents. For example, citing all relevant matter to the claimed invention by applicant in EPO is not as rigorous as it is in USPTO.
2. In the field of data analysis, "partitional clustering" refers to a method of clustering data points into distinct groups or clusters. This method offers more flexibility in choosing the number of clusters, making it suitable for a wider range of datasets. In contrast, "hierarchical clustering" produces a fixed hierarchy of clusters that may not always be the best fit for certain types of datasets.
3. http://rapid-i.com/content/view/181/190/

References

Amigo, E., Gonzalo, J., Artiles, J. & Verdejo, F. (2008). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4), 461-486. https://doi.org/10.1007/s10791-008-9066-8

Fujii, A. (2007). Integrating content and citation information for the ntcir-6 patent retrieval task. Paper presented at the *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access* (pp. 377-380). Tokyo, Japan. Retrieved from https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/76.pdf

Graf, E. & Azzopardi, L. (2008). A methodology for building a patent test collection for prior art research. In *Proceeding of the 2nd International Workshop on Evaluating Information* Access (EVIA), December 16, 2008, Tokyo, Japan (pp. 60-71). Retrieved from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings7/pdf/EVIA2008/11-EVIA2008-GrafE.pdf

Habiba, S. (2004). Iran patent system after acceptance of Trade-Related Aspects of Intellectual Property Rights (TRIPS). *Law & Political Science*, 66, 145-181. Retrieved from https://journals.ut.ac.ir/article_11229_adfb2c1c942393c295f8f640669c5e08.pdf [in Persian]

Horri, A. (1981). *A Comparison of Citation Similarities and Index Term Similarities for Linking Subject Related Documents*. Ph.D. Theses, Case Western Reserve University, Cleveland.

Huang, A. (2008, April). Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand (Vol. 4, pp. 9-56).

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3), 264-323. https://doi.org/10.1145/331499.331504

Kessler, M. M. (1963). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, 1(4):169-187. https://doi.org/10.1016/0020-0271(63)90016-0

Kim, Y. G., Suh, J. H. & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, 34(3), 1804-1812. https://doi.org/10.1016/j.eswa.2007.01.033

Lai, K.-K. & Wu, S.-J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing and Management*, 41(2), 313-330. https://doi.org/10.1016/j.ipm.2003.11.004

Leydesdorff, L. (1987). Various methods for the mapping of science. S*cientometrics*, 11(5), 295-324. https://doi.org/10.1007/BF02279351

Li, X., Chen, H., Zhang, Z. & Li, J. (2007, June). 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* (pp. 419-427). Vancouver, BC, Canada.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. *Scientometrics*, 49(1), 93-123. https://doi.org/10.1023/A:1005613325648

Salton, G. (1963). Associative document retrieval techniques using bibliographic information. *Journal of ACM*, 10(4), 440-457. https://doi.org/10.1145/321186.321188

Salton, G. (1971). Automatic indexing using bibliographic citations. *Journal of Documentation*, 27 (2), 98 - 110. https://doi.org/10.1108/eb026511

Shaw Jr, W. M. (1990). Subject indexing and citation indexing- Part II: An evaluation and comparison. *Information Processing & Management*, 26(6),705-718. https://doi.org/10.1016/0306-4573(90)90047-6

Shaw Jr, W. M. (1990). Subject indexing and citation indexing- Part I: Clustering structure in the cystic fibrosis document collection. *Information Processing & Management*, 26(6), 693-703. https://doi.org/10.1002/(SICI)1097-4571(199110)42:9%3C669::AID-ASI5%3E3.0.CO;2-Y

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265-269. https://doi.org/10.1002/asi.4630240406

Tan, P. N., Steinbach, M. & Kumar, V. (2006). *Cluster analysis: Basic concept and algorithm introduction to data mining*. Boston, Massachusetts: Pearson Addison-Wesley.

Tiwana, S. & Horowitz, E. (2009, November). Find cite: Automatically finding prior art patents. In *Proceedings of the 2nd international workshop on Patent information retrieval* (pp. 37-40). Hong Kong, China.

USPTO. (2012). *Manual of Patent Examining Procedure (MPEP)*. Final Revision

USPTO. (2020). *Manual of Patent Examining Procedure (MPEP)*. E9R-10.2019

Wedding, D. K. (2009). *Extending the data mining software packages sas enterprise miner and spss clementine to handle fuzzy cluster membership: implementation with examples*. Master of Science Central Connecticut State University, Connecticut.

Wouters, P. (1999). *The Citation Culture*. Ph.D. Theses. University of Amsterdam, Amsterdam. Retrieved from http://garfield.library.upenn.edu/wouters/wouters.pdf