ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
**Indian Journal of Science and Technology,** *Vol 10(37), DOI: 10.17485/ijst/2017/v10i37/116435, October 2017*

# A Literature Review on Patent Information Retrieval Techniques

**Alok Khode[1]* and Sagar Jambhorkar[2]**

[1]Symbiosis Centre for Research and Innovation, Symbiosis International University, Lavale Campus, Pune – 412115, Maharashtra, India; alok@urdip.res.in
[2]Department of Computer Science, National Defense Academy, Khadakwasla, Pune – 412115, Maharashtra, India; sjambhorkar@yahoo.co.in

## Abstract

**Objective:** Patents are critical intellectual assets for any competitive business. They can prove to be a gold mine if retrieved, analyzed and utilized appropriately. Patentability search is an important step in the patent process and missing out any relevant patent may cause expensive legal consequences. As worldwide patent collection is growing rapidly, retrieval of this enormous knowledge source has become complex and exhaustive. This paper attempts to review the studies carried out in enhancing the relevance effectiveness of patent information retrieval. **Method/Analysis:** Literature review presents various research works that have been carried out to yield better results in patent retrieval task by refining existing information retrieval techniques or by using standard approaches at the various stages of the patent retrieval task. This work exclusively looks at literatures dealing with retrieval of patent text. **Findings:** Patent retrieval is not a completely solved research domain and general information retrieval approaches do not prove effective in this domain as patents are special documents posing various retrieval challenges. The review also highlights future research directions and will help researchers working in the domain of patent retrieval. **Application/Improvement:** Considering the various techniques and frameworks available and their limitations, there is a lot of scope in the field of patent retrieval techniques which makes room for further research to be taken up in this domain.

**Keywords:** Information Retrieval, Patent Retrieval, Patent Search, Query Expansion, Relevance Feedback, Semantic Analysis

## 1. Introduction

Patents have become critical asset for any innovative company, and with increased global competition, companies have started aligning their business strategies with their IP strategies. Patent informatics describes the science of searching, analyzing and presenting patent information to identify relationships and trends that would not be apparent while working with patent documents on one-on-one basis[1].

With the growth in the field of information and communication technology, patent searching has changed dramatically from manual catalogue base access to online access[2]. As worldwide patent collection is growing rapidly, retrieval and analysis of this enormous knowledge source have become complex, exhaustive, highly interactive and repetitive task, requiring lots of expertise with diverse search strategies[3].

*Author for correspondence

## 2. Patent Retrieval

The tasks performed by patent information users can be divided into patent retrieval, analysis and monitoring. The main goal in patent information retrieval is to find all the prior arts relevant to a given patent application. A patent examiner might be interested in getting at least one relevant patent document for a said patent application to avoid hampering the technology development of the application domain, whereas a corporate might require all possible patent documents in a said technology to avoid infringement as well as to develop new technology[4]. Patent retrieval task can be referred by a variety of different names depending on the end results such as novelty search, patentability search, infringement search, freedom-to-operate search, invalidity search, due diligence search[5,6]etc. The goals, relevance judgments and effectiveness requirements differ greatly depending upon the type of search tasks[7]. These searches are performed at various stages of patent documents life cycle[8] as illustrated in Figure 1.

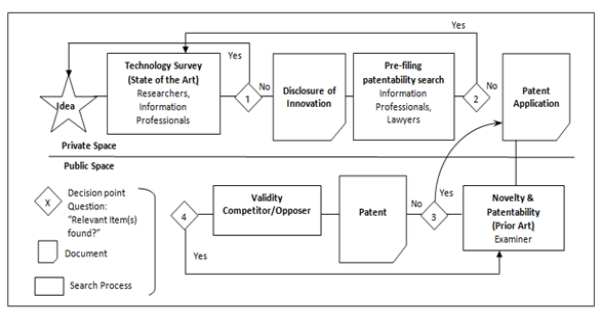


**Figure 1.** Life-cycle of a patented idea.

Increasing number of patent related information and ever growing need to access this information by various types of users motivates researchers to develop techniques and methodologies for efficient and effective patent retrieval. These users can be patent professionals, industrial and academic research communities, managers, venture capitalist, investors, patent attorneys etc.

There are various research areas in patent retrieval and mining such as evaluation of patent retrieval, automated patent classification, patent text retrieval, image-based patent retrieval and classification, multilingual patent retrieval etc. The objective of this literature review is to understand the existing research carried out in enhancing the effectiveness of the patent text information retrieval.

Despite the remarkable advancements in the area of information retrieval and search engine techniques, there still exists a gap between research in web search engine and techniques usually adapted in patent retrieval. Traditionally patent retrieval has been mostly investigated by database research communities; rather than the Information Retrieval (IR) research communities due to lack of test collections in this domain[6,9]. With recent information retrieval initiatives on patent information retrieval like CLEF-IP, PaI R, As PIRe, SIGIR, NTCIR, TREC, etc[10] by Information Retrieval Facility (IRF), Vienna, Association for Computing Machinery (ACM), USA, National institute of Standard and Technology (NIST), USA & National Institute of Informatics (NII), Japan and availability of domain specific test collection, patent retrieval has become an active area of research in information retrieval domain[9,11]. These initiatives have brought together leading researchers in information retrieval and those who practice and use patent search. This has led to an interdisciplinary dialogue between the information retrieval and the intellectual property (IP) communities thereby creating a discursive as well as empirical space for sustainable discussion and innovation[12].

Studies have also suggested that patent retrieval is not a completely solved research domain and general information retrieval approaches do not prove effective in the patent domain as patents are special documents posing various retrieval challenges[6,13–16].

## 3. Information Retrieval

Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information. The main aim of information retrieval model is to find relevant knowledge-based information or a document that fulfill user needs[17]. Relevance is a central concept in Information Retrieval and it is used to work out effectiveness measures for information retrieval systems. Precision and recall are two traditional effectiveness measures: precision means the proportion of relevant documents out of those returned, whereas recall is the fraction of relevant documents that are retrieved[18]. The relevancy of retrieved information is not usually binary but is continuous, subjective and situational, depending upon user's judgment and needs which changes over a period of time.

## 4. Challenges Faced in Patent Information Retrieval

The following challenges make patent retrieval extremely complex and problematic for researchers as well as experienced patent professionals and requires robust and repeatable techniques/approaches[14–16].

a. Patent retrieval demands effective and efficient techniques as it is a multi-topical task and performed on diverse and large dataset[9,15,19].
b. In patent retrieval, a complete patent document is used as a query to represent the information need. It is a defining challenge to build queries from a whole patent document consisting of multiple fields[9,15,20–22].
c. Patent document makes use of vague, highly specialized, inventor specific non-standard acronyms/terms, involving of synonyms, homonyms and techno-legal words where keyword based search may not give desired results[23,24].
d. Patent retrieval is a recall oriented task as missing out even a single relevant document while performing patentability or freedom-to-operate search might lead to severe financial consequences due to lawsuit for patent infringement[14,25,26].
e. Many times retrieval techniques are only restricted to patent classification codes (IPC)[97]. This traditional approach is too general to meet the needs of users. The complexity of the classification system and sparse class assignments to the patent limits the patent search[27].

## 5. Research Work in the Field of Patent Retrieval

Initially research in patent search was largely undertaken by database community[6] but with the new initiatives[10] by the Information Retrieval (IR) community, various workshops and symposiums on patent retrieval have been started to promote the research and development. Hence patent retrieval has now become an active sub-domain of research in the area of information retrieval[3,9,11]. IR and Natural Language Processing (NLP) researchers have started investigating the patent informatics domain and are focusing research around patent retrieval with techniques and approaches[14,28–33] highlighted in Table 1.

**Table 1.** Various patent retrieval techniques and approaches

| | |
|---|---|
| Techniques: | query formulation, query expansion, summarization, relevance feedback etc. |
| Models: | Vector space model(VSM), semantic based processing, latent semantic analysis(LSA), language model, weighting techniques, probabilistic model etc. |
| Others: | Bibliomatric methodology, data mining, text mining, database management tools like OLAP, citation analysis. |

Patent search has changed little in the last 25 years and still the basic logic structure of Boolean is used which heavily relies on set theory and exclusion along with other types of patent retrieval techniques[34]. Study shows that Boolean search also known as "keyword" search remains the preferred method for locating relevant patents[35–37] since it is reproducible and helps patent professionals to defend their decisions as and when it is required[37].

Boolean search has certain limitations like many documents relevant to the query may be missed out, and many unrelated or irrelevant documents may be retrieved. It has no capability of determining which of those documents are of highest interest to the researcher[6,17,23,38,39]. Though using Boolean system, structured and precise queries can be created, it is an exhaustive task, requiring lots of experience and does not yield the desired result, especially in the patent domain[6,39]. As with infringement suits being filed at the rate of more than 10-per-day involving billions of dollar, missing even a single relevant patent documents may cost dearly[40]and hence the use of only Boolean/keyword search is not preferred for the patentability search[36].

To overcome the problem of Boolean model, researchers have proposed various alternative techniques by exploiting the patent structure. One of the proposed studies presented a Noise Elimination (NE) algorithm based on keyword weighted distribution to eliminate the noisy patent data from the search result[28]. However, the study has not evaluated the system using any patent benchmark dataset. In another study, patent retrieval system architecture is proposed in which patent metadata and the citation structures are exploited for creating restricted initial working sets of patents[41]. The method includes multiple indexing of the patents which are used for three languages. Additional indexes are also created for phrases and concepts extracted from external sources. Multiple retrieval models such as Language Model (LM) with Kullback-Leibler (KL) divergence and Okapi BM25

are used to generate ranked results from the indexes. Further, the study uses Support Vector Machines (SVM)[42] to merge the ranked results and again re-rank them using additional training sets created from the patent collection. Despite of complex design and high demand for resources, this system achieved the best result in the CLEF-IP 2010. This study is further enhanced by enriching the extracted citations using retrieved citations from non-European patent[43]. Patent family lookup is applied using OPS (Open Patent Service)[44] to identify a corresponding European patent. To search the index with less number of keywords, an additional keyword extraction module is introduced and search result is merged with other ranks lists using SVM. Citations from the patent document suggesting a scoring method based on PageRank algorithm[45] has also been proposed[46]. In this approach text based retrieval is performed for obtaining top patents, and then citation scores are calculated for these top patents. The study recommends that combining text-based and citation-based scores gives better results in patent retrieval.

In various studies, patent structure has been analyzed and exploited for effective patent retrieval. Patent contains multiple topics and segmentation of the claim part of the patent document is used to find subtopics present in the target document. For each of the subtopic, query is created for the effective retrieval[31]. By calculating the entropy, which is the deviation degree of the appearance of the term in each subtopic, the importance of subtopics is calculated. For each subtopic in the query, relevant patents are retrieved with a relevant score based on the importance of the subtopic and then a final rank list is determined. The impact of various fields like 'title', 'claim', 'abstract,' 'background summary' on the effectiveness of the patent retrieval when used for query formulation has been studied[20]. However, extracting terms from a single field at a time has shown that 'background summary' gives better result as compared to other fields in a patent document. In another study, learning to rank approach is applied to combine language model based retrieval score (using term search and phrase search), IPC classification feature and low level feature such as Term Frequency (tf) and Inverse Document Frequency (idf) of search terms[21]. This has improved result as compared to the finding of previous research carried out by the same researchers[20]. In information retrieval, query model based approach has also been considered to improve its effectiveness. Patent retrieval architecture is proposed[14] in which query model is estimated based on weighted log-likelihood[47] and parsimonious language model[48]. It extracts terms from each field separately to build a query model. Result of this research suggests that description is the best field for extracting terms for query formulation whereas combination of different fields for building queries or merging the results is not effective. Queries can also be constructed by combining terms extracted from different fields rather than selecting terms from single field at a time, and weight them according to their log(tf)idf values may give better results[15]. However, study[49] highlights that Key Phrase Extraction (KPE) techniques work better than tf or tf-idf scores for selecting phrases, especially from "description" section, for invalidity search. The study explores various supervised and unsupervised KPE algorithms to construct an optimal query by extracting important phrases and keywords from a patent. It is empirically reported that more effective query can be formulated by utilizing 20-30 terms while giving higher importance to terms extracted from abstract, claims & description[15]. Approach based on multiple query representation for prior art search is also suggested[16] where set of similar queries are generated for a given patent application. Each query acts as an alternative representation of patent information and for each query; the retrieved results are treated as set of ratings. Using Collaborative Filtering (CF) algorithm, a final document ranking is achieved. A stage wise patent retrieval method considering the claim structure has shown improved effectiveness as far as precision is concerned[32]. In stage-1, query terms are extracted from claims to get top N patents to improve the recall. In stage-2, various text analysis and retrieval methods are used to re-rank the N patents.

Apart from complex patent retrieval systems, literature also suggests straight-forward and sophisticated search approaches in patent retrieval[50]. In one approach, researchers use simple and straightforward Information Retrieval (IR) technique using list of citations extracted from the patent numbers within the description field of some patent queries while in the another approach, much more sophisticated information retrieval techniques have been used[41,43]. Experiment of the study shows that the simple approach uses fewer resources, less time and efforts to achieve a statistically indistinguishable performance as compared to the advance techniques when patent applications contain citations. However, the advanced search technique is statistically better when no initial citations are provided.

Studies have also been carried out on query expansion to deal with term mismatch due to synonyms/polynyms.

In such techniques, initial query is formed by extracting the keywords from the patent application and then expansion terms are selected from a feedback process or from external knowledgebase such as WordNet[51] or Wikipedia[52], dictionaries, or query logs to enhance the initial query by mapping the concepts or synonyms to the initial information need. In one of the query expansion method, WordNet and Wikipedia for word-based and phrase-based query expansion is proposed[53]. From the query patent, keywords and key phrases are extracted using regular Expressions (RegEx). They are further splited into a Bag Of Words (BOW) and a bag of phrases (BOP). BOP and BOW are then expanded using Wikipedia and WordNet, respectively. Ranked list is generated using Okapi BM25 which is again re-ranked using three stage IPC based ranking algorithm. Analysis showed that it is not clear how phrases interact with the words which are expanded by WordNet in improving effectiveness. In the event when local collection fails due to the lack of relevant documents for query expansion, the use of external collections for query enrichment is justified[54]. However, it was observed that use of external knowledgebase slows down the retrieval process dramatically and WordNet is not an effective method for query expansion for patent search[55–57].

A proximity based framework is suggested[36] by utilizing independent claim (first claim) and International Patent Classification (IPC) classification as source for selection query expansion terms. In the study, a query specific patent lexicon is constructed from IPC definition. Further, the term proximity between query terms and expansion terms from patent lexicon is used to select most appropriate expansion terms. Weight for expansion terms is calculated by estimating query relatedness probability and concluded that proximity of expansion terms to query terms is a good indicator for the selecting terms.

Although there are several automatic query expansion approaches available, the Pseudo-Relevance Feedback (PRF) has shown its value in improving retrieval effectiveness[58]. Combination of the PRF and term-proximity distribution of initial set of relevant terms with respect to query terms improves the retrievability of patents when compared to standard PRF[23]. The patents for PRF are identified based on their similarities with query patents over a subset of terms, rather than the overall document similarity. Instead of using PRF for query expansion, study has also been carried out to evaluate the effect of PRF for patent queries reduction. In this technique, a patent application is decomposed into constituent text segments and the Language Modeling (LM) similarities is computed by calculating the probability of generating each segment from the top ranked documents[59]. This work achieves improvement over initial results using PRF which is satisfactory compared to all standard PRF methods which failed in the patent domain. A learning to rank framework has also been attempted to estimate the effectiveness of a patent document in terms of its performance in PRF by utilizing patent specific characteristics[60]. The method introduces a unigram query model by estimating the importance of each term according to a weighted log-likelihood based approach. A relevance model is then used to select the most appropriate terms (feedback terms) from the top retrieved documents by the initial search (feedback documents) to expand the original query. This prediction method obtained a statistically significant improvement over standard pseudo-relevance feedback. The influence of term selection on prior art is also investigated through the minimal interactive relevance feedback approach[61]. In this experiment, an oracular query based on judged relevant documents has been defined to demonstrate that Language Model (LM) and BM25 retrieval scoring models outperform the baseline when simple interactive methods for query reduction is employed.

To improve the retrieval performance, IPC codes have also been utilized in combination with the text contents. Many researchers have exploited the IPC codes for document ranking and filtering. Many a times, use of IPC codes can be helpful to identify the similarity between patent documents which even do not share similar terms[22]. To effectively narrow down the searching of patent document, a IPC enabled technique is proposed in combination with user selected key phrases and trained neural network classifier[62]. Since the IPC codes are semantic in nature and organized in taxonomy, it can be utilized effectively to group similar patents for retrieval, a cluster-based patent retrieval technique using IPC codes is proposed[13]. However, this study mainly analyzed result documents and might have missed few relevant patents[13]. The researcher also reports that it is difficult to discriminate among IPC-based clusters if searches are performed within themselves that share common terms. It is interesting to note that IPC-codes work well for state-of-art search[63] while they are not so effective for prior art search. Studies have also shown that variable length IPC codes also play an important role in retrieval. IPC codes with first 4 and 11 characters for re-ranking, along with weight

learned from data and inclusion of negative data instances improve the effectiveness[15].

Apart from patent retrieval techniques reviewed above, various other approaches based on syntactic and semantic processing, NLP, domain ontology, semantic web, vector space model, latent semantics analysis, probabilistic latent semantic analysis, and concept based retrieval etc[64]. have also been studied. A method is introduced to disambiguate query terms and predict whether expansion using noun phrases would improve the retrieval effectiveness[65]. Experiments at the University of Hildesheim[66] also confirm the results but still are outperformed by other systems which do not use phrases[67]. Researchers have proposed a system which identifies and classifies the biologically significant terms in the patents and integrates them with dictionaries based biomedical ontologies to create a biomedical semantic web[68]. Besides keyword search and queries linking the properties specified by one or more RDF triples, graph algorithms are utilized to determine the semantic associations between semantic web resources. Though system enable researchers to perform a single semantic search to retrieve all the relevant information about a biological concept, system is domain specific, not yet validated for effectiveness and scalability needs to be evaluated. Study shows that search tasks specific setting and tuning of retrieval system yield optimal effectiveness in patent retrieval in medicinal chemistry[69]. Furthermore, in another semantic web based approach, researchers suggest generating semantic annotations on patents by relying on the structure and semantic representation of patent documents[70]. A generated annotation comprises of a structure annotation, a metadata annotation and a domain based annotation, which are then merged into the Patent Semantic Annotation. The system is tested on biomedical patents with a very small dataset (~1000) and can be considered as a first step towards a semantic web-based patent retrieval/mining.

In paper[71] an automatic semantic annotation approach is proposed that integrates ontology-based techniques, structural template schemes, natural language processing and pattern learning to annotate patent documents from various aspects according to the structure and content characteristics of the patent document. This study shows semantic correlation between patent documents and generate abstract technical feature of a patent which helps in technology survey for new product design. This research may be further extended for the effective patent retrieval. A comparative analysis of the annotation of PubMed documents with Medical Subject Headings ontology (MeSH) terms and the assignment of the International Patent Classification (IPC) has been carried[27]. The study pointed out that complex class definitions rarely occur in patent text and the number of IPC class assignments to patents is low which limits the patent search severely. Authors propose that assignment of additional patent classes, combining them with search keywords and existing ontologies/taxonomies such as MeSH can improve the patent retrieval. In general, patent contains two core concepts i.e. "Problem" and "Solution", which constitute a particular technology. Study suggests that key-phrases can be annotated for these two semantic categories[72] to form two semantic clusters by grouping patent documents with the same "Problem" or "Solution" tag. Further, semantic cluster information is added to a conventional language model based retrieval method. However, in some cases, important documents cannot be added in the cluster as there are small numbers of "Problem" and "Solution" keywords in the collection.

From the literature, it is evident that Boolean search has certain limitations. To overcome the similarity issue between queries and patent documents caused by Boolean search, Vector Space Model (VSM) is also suggested where terms, documents and queries are represented as vectors[73]. This model was initially proposed by Gerard Salton[74]. The dimension of the vectors corresponds to the unique terms in the document collection and a value indicates the frequency of this term. A user query is also considered as a document in the search space. The query vector is compared with the document vector for finding similarity using cosine of the angle between the document and the query vector[75].

In most of the studies, complete patent application has been considered to retrieve the relevant patents for prior art. However, a novel approach is suggested which looks at the prior art from the inventor's perspective and consider ideas (partial application) to form a query rather than full application[76]. This study uses series of various query expansion and query reduction techniques and reports that Rocchio based relevance feedback for query expansion is more effective for short queries such as "title" while maximal marginal relevance based query expansion gives comparatively better result for medium length queries based on "abstract" or "description". The study also report that VSM model performs better for short queries while BM25 perform best when dealing with long queries.

Studies have shown that VSM and ontology based patent retrieval can be developed to improve precision of search results and rank them by similarities[77]. It is also reported that at preprocessing level, if additional morphological decompounding module is introduced, it positively influences the performance of VSM in discovering the similarities between patent claims[78]. Though VSM is a very effective method in information retrieval, it also has some limitations like large dimensionality due to long documents, search keywords must precisely match document terms and lack of semantic relationship between the data items[79]. In VSM model, indexing vocabulary changes as soon as changes occur in the document set, or in the indexing vocabulary selection algorithms, or in parameters of the algorithms, or if wording evolution occurs. To solve this issue, it is proposed, specifically for patent retrieval, to use IPC codes, to generate the indexing vocabulary for presenting all the patent documents[80].

To overcome the limitation of keyword searching and VSM, studies have suggested the use of Latent semantic analysis (LSA)[79,81]. In the seminal paper[82], Deerwester introduces a statistical technique "Latent Semantic Indexing" (LSI) which is based on a document-by-keyword matrix of large dimension where dimension is reduced using singular value decomposition (SVD). LSI also deals in the resolution to synonym and polysemy to some extent. In this statistical technique, latent concepts are derived by utilizing term co-occurrence. Words are considered semantically associated when they frequently occur together[83]. In literature, a patent document retrieval system is proposed which utilizes LSI to recognize synonymous expressions to address semantic and syntactic properties[84]. The system selects those patent documents whose abstract vectors lie in the neighborhood of the query vector to narrow down the search space. It then uses the form based template matching algorithm[85] to calculate the similarities of the document and the query. This approach uses only abstract field whereas many researchers have suggested that description/background section of the patent is more relevant. This system is yet to be verified.

Studies have also shown that results can be enhanced by using text clustering along with tailored Singular Value Decomposition (SVD) parameters to the specific patent corpus. Such techniques can be used to address ambiguities in language[81]. An appropriate selection of a number (k) to truncate SVD is very important as low k results in missing out some important factors while a high k generates noise resulting in an equivalent VSM[86]. In patent searching, accuracy is very important and the choice of k during SVD has shown to have substantial effect on accuracy[83,87,88], and in theory there is no method which exists to determine the optimum value of k and hence empirical testing is a must[86]. In another study, it is reported that the value of k=80 is sufficiently enough in the truncation of SVD for obtaining satisfying results in the patent retrieval process[89]. The study also highlights that Latent Semantic Indexing (LSI) slightly improve results compared to the standard VSM. However, dimensionality reduction technique in SVD and LSA do not yield the best results when working with smaller patent dataset[89]. Although LSI works well on large data set, it comes at a cost since it requires huge storage size and more computation time[90]. To overcome such limitation, a divide-and-conquer approach is proposed for retrieving similar patents from a large-scale patent collection. The approach first divides patents into 200 categories based on IPC and for each category, LSI is applied repeatedly for reducing dimension and extracting features.

Though latent semantic indexing is known to improve retrieval effectiveness, developing an accurate latent semantic based search engine in patent domain is still an active research area[86]. Latent Semantic Analysis may find its way as an assisting technology, rather than relying fully on its ability (or inability) to detect document similarity[36]. Its theoretical foundation also remains unsatisfactory and incomplete to a larger extent. A more principled novel method for unsupervised learning, called Probabilistic Latent Semantic Analysis (PLSA), is proposed by Hofmann which is based on a statistical latent class model and possesses a sound statistical foundation[91,92]. In case of LSA, empirical testing needs to be carried out to determine the optimal value of k[86] while in PLSA a probabilistic approach is used to determine the value of k[90]. With experimentally verified substantial performance gains, Probabilistic Latent Semantic has a wide range of applications in text mining and information retrieval. However, this technique is not much explored in the patent domain. In paper[93], researchers propose the use of PLSA on large corpora of the patent documents for the classification of newly drafted patent documents. The study claims that the results are superior to all methods reported in scientific and technical prior art so far, however the evaluation detail is not reported in the study. Study shows that more resources to process large corpus of patent documents and efficient and effective models of parallel and grid computing may be of help[94]. Google's patented Map Reduce programming

model[95] provides an efficient framework for processing large data sets with a parallel, distributed algorithm on a cluster and provides efficient and reliable distributed data storage required for applications involving large data sets. Various benefits of Map Reduce implementation have been reported in the literature[96]; however, its application in the patent retrieval is yet to be explored.

It is apparent from the literature review that various research works have been carried out to yield better results in patent retrieval task by refining existing IR techniques or by using standard approaches at the various stages of the patent retrieval task or by combining multiple techniques. A summary for various research studies reviewed in this papers is given in Table 2.

**Table 2.** Summary of research studies carried out in patent information retrieval

| Techniques/ Approaches | Publications | Methodology | Remarks/limitations | |
|---|---|---|---|---|
| IPC Based | 34 | IPC based Knowledge representations of data, BM25, | | Broad categorization, Complex class definitions, Sparse class assignments Class definition revised every year |
| | 13 | IPC-based clusters | Difficult to discriminate documents in same cluster | |
| | 20, 31, 62 | Exploited the IPC codes for document ranking and filtering | | |
| Patent features and Query Formulation | 45 | Claim and Citation linking, PageRank algorithm, BM25 | | Citation may improve the accuracy Not all patents are cited, Have different degree of relevancy, Biased retrieval as TFIDF and OKAPI-BM25 favor large terms frequencies , Due to unusual vocabulary usage, bigram, unigram do not work |
| | 41,43 | Citation from descriptions, multiple indexing for languages, concept from Wikipedia, LM,KL divergence, BM25, OPS, Key-phrase extraction, SVM | Best run in CLEF-IP 2009, 2010. Complex architecture, manually annotation to train the Classifier | |
| | 31 | Claim segmentation, multiple queries, Entropy calculation, BM25 | | |
| | 20,21 | Utilize all fields, learning to rank, unigram, LM, tf-idf, unigrams | One field a time, "background summary" is important section | |
| | 14 | Utilized all fields, LLQM, IPC filtering, BM25, | "Description" is the best field, combination of fields & merging ranked list is not effective. | |
| | 76 | Query formulation from partial patent application | prior art from the inventor's perspective | |
| | 61 | oracular query, LM and BM25 | minimal interactive method perform better | |

## 6. Research Directions

The literature review points out to multiple research directions which may be used for the effective patent retrieval. This is depicted in the Figure 2. Further, it is also evident that research need to consider the following basic questions before deriving any new patent retrieval technique:

- Are conventional retrieval techniques effective in patent information retrieval?
- How to exploit structured and unstructured information in the patent documents for better retrieval?
- What are the effects of technological domains on patent retrieval?
- Does the length of patent document affect the effectiveness?
- What effect does the size of patent dataset have on efficiency and effectiveness of patent retrieval?
- How different sections of patent documents influence the effectiveness?


## 7. Conclusion

With the rapid increase in the worldwide patent data and its use, effective retrieval of patent information is important for businesses and innovations. The literature review highlights the fact that patent is a special document and its retrieval is a challenging task. Various information retrieval models, algorithms and techniques have been suggested by researchers, however no single technique is proved to be effective for patent retrieval. The knowledge based retrieval frameworks proposed are in their infancy stage. They are not fully validated and tested on actual patent data sets. Studies on patent query formulations using query expansion and query reduction techniques have seldom shown enhancement of effective retrieval. The use of IPC at post-processing may yield better results, if combined with patent text, for ranking and filtering. Considering the various techniques and frameworks available and their limitations, there is a lot of scope in the field of patent retrieval techniques which makes a room for further research to be taken up in this domain.

## 8. References

1. Trippe AJ. Patinformatics: Tasks to tools. World Patent Information. 2003; 25(3):211–21. Crossref
2. Adams S. The text, the full text and nothing but the text: Part 2–the main specification, searching challenges and survey of availability. World Patent Information. 2010; 32(2):120–8. Crossref
3. Joho H, Azzopardi LA, Vanderbauwhede W. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. Proceedings of the third symposium on Information interaction in context. USA; 2010. p. 13–24. Crossref
4. Trippe A, Ruthven I. Evaluating real patent retrieval effectiveness. Current Challenges in Patent Information Retrieval. Springer Science & Business Media: Berlin Heidelberg; 2011. p. 125–43. Crossref
5. Hunt D, Nguyen L, Rodgers M. Patent searching: Tools & techniques. John Wiley & Sons: New Jersey, USA; 2012. p. 208. PMid:22948932 PMCid:PMC3614147
6. Bonino D, Ciaramella A, Corno F. Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Information. 2010; 32(1):30–8. Crossref
7. Hansen P, Järvelin A, Järvelin A. Exploring manual and automatic query formulation in patent IR: Initial query construction and query generation process. Journal of Documentation. 2013; 69(6):873–98. Crossref
8. Lupu M, Hanbury A. Patent Retrieval. Foundations and Trends in Information Retrieval. 2013; 7(1):1–97. Crossref
9. Azzopardi L, Vanderbauwhede W, Joho H. Search system requirements of patent analysts. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. Switzerland; 2010. p. 775–6. Crossref
10. New initiatives on patent retrieval [Internet]. [cited 2013 Oct 9]. Available from: Crossref
11. Tait J, Lupu M, Berger H, Roda G, Dittenbach M, Pesenhofer A, Graf E, Van Rijsbergen CJ. Patent search: An important new test bed for ir. Proceedings of 9th Dutch–Belgian Information Retrieval Workshop (DIR). Netherlands; 2009. p. 56–63.
12. Lupu M, Mayer K, Tait J, Trippe AJ. Current challenges in patent information retrieval. Springer Science & Business Media: Berlin Heidelberg. 2011; 29(1):418. Crossref
13. Kang IS, Na SH, Kim J, Lee JH. Cluster-based patent retrieval. Information processing & management. 2007; 43(5):1173–82. Crossref

14. Mahdabi P, Keikha M, Gerani S, Landoni M, Crestani F. Building queries for prior-art search. Information Retrieval Facility Conference. Austria; 2011. p. 3–15. Crossref
15. Cetintas S, Si L. Effective query generation and postprocessing strategies for prior art patent search. Journal of the American Society for Information Science and Technology. 2012; 63(3):512–27. Crossref
16. Zhou D, Truran M, Liu J, Zhang S. Using multiple query representations in patent prior-art search. Information Retrieval. 2014; 17(5–6):471–91. Crossref
17. Salton G, McGill, MJ. Introduction to modern information retrieval. McGraw-Hill: USA; 1983.
18. Dominich S. Relevance effectiveness in information retrieval. Mathematical Modelling: Theory and Applications. Springer International Publishing. Switzerland; 2001. p. 215–32.
19. Agatonovic M, Aswani N, Bontcheva K, Cunningham H, Heitz T, Li Y, Roberts I, Tablan V. Large-scale, parallel automatic patent annotation. Proceedings of the 1st ACM workshop on Patent information retrieval, USA; 2008. p. 1–8. Crossref
20. Xue X, Croft WB. Automatic query generation for patent search. Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong; 2009. p. 2037–40. Crossref
21. Xue X, Croft WB. Transforming patents into prior-art queries. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, USA; 2009. p. 808–9. Crossref
22. Graf E, Frommholz I, Lalmas M, Van Rijsbergen K. Knowledge modeling in prior art search. Information Retrieval Facility Conference, Austria; 2010. p. 31–46. Crossref
23. Bashir S, Rauber A. Improving retrievability of patents in prior-art search. European Conference on Information Retrieval, Ireland; 2010. p. 457–470. Crossref
24. Atkinson KH. Toward a more rational patent search paradigm. Proceedings of the 1st ACM workshop on Patent information retrieval, USA; 2008. p. 37–40. Crossref
25. Magdy W, Jones GJ. Examining the robustness of evaluation metrics for patent retrieval with incomplete relevance judgments. International Conference of the Cross-Language Evaluation Forum for European Languages, Italy; 2010. p. 82–93.
26. Mahdabi P, Gerani S, Huang JX, Crestani F. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, Ireland; 2013. p. 113–22. Crossref
27. Eisinger D, Tsatsaronis G, Bundschus M, Wieneke U, Schroeder M. Automated patent categorization and guided patent search using IPC as Inspired by MeSH and PubMed. Journal of biomedical semantics. 2013; 4(1):S3. Crossref PMid:23734562 PMCid:PMC3632996
28. Lee W, Leung CK, Song JJ. Reducing noises for recall-oriented patent retrieval. IEEE Fourth International Conference on Big Data and Cloud Computing(BdCloud), Australia; 2014. p. 579–86. Crossref
29. Markellos K, Perdikuri K, Markellou P, Sirmakessis S, Mayritsakis G, Tsakalidis A. Knowledge discovery in patent databases. Proceedings of the eleventh international conference on Information and knowledge management, USA; 2002. p. 672–74. Crossref
30. Fall CJ, Törcsvári A, Benzineb K, Karetka G. Automated categorization in the international patent classification. ACM SIGIR Forum, Canada. 2003; 37(1):10–25. Crossref
31. Takaki T, Fujii A, Ishikawa T. Associative document retrieval by query subtopic analysis and its application to invalidity patent search. Proceedings of the thirteenth ACM international conference on Information and knowledge management, USA; 2004. p. 399–405. Crossref
32. Mase H, Matsubayashi T, Ogawa Y, Iwayama M, Oshio T. Proposal of two-stage patent retrieval method considering the claim structure. ACM Transactions on Asian Language Information Processing (TALIP). 2005; 4(2):190–206. Crossref
33. Lai KK, Wu SJ. Using the patent co-citation approach to establish a new patent classification system. Information Processing & Management. 2005; 41(2):313–30. Crossref
34. Benson CL, Magee CL. A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field. Scientometrics. 2013; 96(1):69–82. Crossref Crossref
35. Connett-Porceddu M, Ashton DE, Bacon N, dos Remedios N, Nottenburg C, Okada S, Quinn G, Wei Y, Jefferson RA. Analysis of trends in search and retrieval of intellectual property-related information. Black Mountain: Cambia; 2005.
36. Gibbs A. Boolean patent search: comparative patent search quality/cost evaluation super Boolean vs. legacy Boolean search engines. Patent café: USA; 2006.
37. Adams S. A practitioner's view on PaIR. Proceedings of the 4th workshop on Patent information retrieval, UK; 2011. p. 37–8. Crossref
38. Blair DC, Maron ME. An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM. 1985; 28(3):289–99. Crossref
39. Cao Y, Fan J, Li G. A user-friendly patent search paradigm. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(6):1439–43. Crossref
40. Jochim C, Lioma C, Schütze H. Expanding queries with term and phrase translations in patent retrieval.

Information Retrieval Facility Conference, Austria; 2011. p. 16–29. Crossref
41. Lopez P, Romary L. Multiple retrieval models and regression models for prior art search. In: CLEF Workshop, Technical Notes, Corfu, Greece; 2009.
42. Joachims T. Learning to classify text using support vector machines: Methods, theory and algorithms. Kluwer Academic Publishers: Norwell; 2002. Crossref
43. Lopez P, Romary L. Experiments with citation mining and key-term extraction for prior art search. CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation, Italy; 2010.
44. Open Patent Service [Internet]. [cited 2017 Mar 23]. Available from: Crossref.
45. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the web. Stanford University: USA; 1999.
46. Fujii A. Enhancing patent retrieval by citation analysis. Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Netherlands; 2007. p. 793–94. Crossref
47. Meij E, Weerkamp W, de Rijke M. A query model based on normalized log-likelihood. Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong; 2009. p. 1903–6. Crossref
48. Hiemstra D, Robertson S, Zaragoza H. Parsimonious language models for information retrieval. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, UK; 2004. p. 178–85. Crossref
49. Verma M, Varma V. Applying key phrase extraction to aid invalidity search. Proceedings of the 13th International Conference on Artificial Intelligence and Law, USA; 2011. p. 249–55. Crossref
50. Magdy W, Lopez P, Jones GJ. Simple vs. sophisticated approaches for patent prior-art search. European Conference on Information Retrieval, Ireland; 2011. p. 725–8. Crossref
51. WordNet: A large lexical database of English [Internet]. [cited 2015 Mar 07]. Available from: Crossref
52. Wikipedia: A multilingual, web-based, free-content encyclopedia [Internet]. [cited 2017 Jul 17]. Available from: Crossref.
53. Al-Shboul B, Myaeng SH. Query phrase expansion using wikipedia in patent class search. Asia Information Retrieval Symposium, UAE; 2011. p. 115–26. Crossref
54. Kwok KL, Chan M. Improving two-stage ad-hoc retrieval for short queries. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Australia; 1998. p. 250–6. Crossref
55. Voorhees EM. Using WordNet for text retrieval. In WordNet, An Electronic Lexical Database. The MIT Press; 1998. p. 285–303.
56. Liu S, Liu F, Yu C, Meng W. An effective approach to document retrieval via utilizing WordNet and recognizing phrases. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, UK; 2004. p. 266–72. Crossref
57. Magdy W, Jones GJ. A study on query expansion methods for patent retrieval. Proceedings of the 4th workshop on Patent information retrieval, UK; 2011. p. 19–24. Crossref
58. Xu J, Croft WB. Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Switzerland; 1996. p. 4–11. Crossref
59. Ganguly D, Leveling J, Magdy W, Jones GJ. Patent query reduction using pseudo relevance feedback. Proceedings of the 20th ACM international conference on Information and knowledge management, UK; 2011. p. 1953–6. Crossref
60. Mahdabi P, Crestani F. Learning-based pseudo-relevance feedback for patent retrieval. Information Retrieval Facility Conference, Austria; 2012. p. 1–11. Crossref
61. Golestan Far M, Sanner S, Bouadjenek MR, Ferraro G, Hawking D. On Term Selection Techniques for Patent Prior Art Search. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Chile; 2015. p. 803–6. Crossref
62. Trappey AJ, Hsu FC, Trappey CV, Lin CI. Development of a patent document classification and search platform using a back-propagation network. Expert Systems with Applications. 2006; 31(4):755–65. Crossref
63. Criscuolo P, Verspagen B. Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. Research Policy. 2008; 37(10):1892–908.
64. Jimeno-Yepes A, Berlanga-Llavori R, Rebholz-Schuhmann D. Ontology refinement for improved information retrieval. Information Processing & Management. 2010; 46(4):426–35. Crossref
65. Mahdabi P, Andersson L, Keikha M, Crestani F. Automatic refinement of patent queries using concept importance predictors. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, USA; 2012. p. 505–14. Crossref
66. Becks D, Mandl T, Womser-Hacker C. Phrases or Terms? The Impact of Different Query Types. In: Working Notes of 11th Workshop of the Cross-Language Evaluation Forum (notebook Papers); 2010. PMCid:PMC2964634

67. Piroi F, Lupu M, Hanbury A, Sexton AP, Magdy W, Filippov IV. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. Cross-Language Evaluation Forum CLEF (notebook Papers). Padua, Italy; 2010.
68. Mukherjea S, Bamba B. BioPatentMiner: an information retrieval system for biomedical patents. Proceedings of the Thirtieth international conference on Very large data bases, Canada; 2004. p. 1066–77.
69. Pasche E, Gobeill J, Kreim O, Oezdemir-Zaech F, Vachon T, Lovis C, Ruch P. Development and tuning of an original search engine for patent libraries in medicinal chemistry. BMC bioinformatics. 2014; 15(1):1–9. Crossref
70. Ghoula N, Khelif K, Dieng-Kuntz R. Supporting patent mining by using ontology-based semantic annotations. Proceedings of the IEEE/WIC/ACM international Conference on Web intelligence, USA; 2007. p. 435–8. Crossref
71. Wang F, Lin LF, Yang Z. An Ontology-Based Automatic Semantic Annotation Approach for Patent Document Retrieval in Product Innovation Design. Applied Mechanics and Materials. 2014; 446. p. 1581–90. Crossref
72. Kim Y, Ryu J, Myaeng SH. A Patent Retrieval Method using Semantic Annotations. International Conference on. Knowledge Discovery and Information Retrieval (KDIR 2009), Portugal; 2009. p. 211–18.
73. Silva IR, Souza JN, Santos KS. Dependence among terms in vector space model. Database Engineering and Applications Symposium IDEAS'2004, Portugal; 2004. p. 97–102. Crossref
74. Salton G. The SMART retrieval system—experiments in automatic document processing. Prentice Hall: USA; 1971. p. 556.
75. Aswani Kumar C, Radvansky M, Annapurna J. Analysis of a vector space model, latent semantic indexing and formal concept analysis for information retrieval. Cybernetics and Information Technologies. 2012; 12(1):34–48. Crossref
76. Bouadjenek MR, Sanner S, Ferraro G. A study of query reformulation for patent prior art search with partial patent applications. Proceedings of the 15th International Conference on Artificial Intelligence and Law, USA; 2015. p. 23–32. Crossref
77. Lim SS, Jung SW, Kwon HC. Improving patent retrieval system using ontology. 30th Annual Conference of IEEE on Industrial Electronics Society IECON' Korea. 2004; 3. p. 2646–9.
78. Andersson L. A vector space analysis of Swedish patent claims with different linguistic indices. Proceedings of the 3rd international workshop on Patent information retrieval, Canada; 2010. p. 47–56.
79. Zhang C, Song W, Li C, Yu W. Self-adaptive GA, quantitative semantic similarity measures and ontology-based text clustering. IEEE International Conference on Natural Language Processing and Knowledge Engineering NLP-KE' China; 2008. p. 1–8. Crossref
80. Chen YL, Chiu YT. An IPC-based vector space model for patent retrieval. Information Processing & Management. 2011; 47(3):309–22. Crossref
81. Ryley JF, Saffer J, Gibbs A. Advanced document retrieval techniques for patent research. World Patent Information. 2008; 30(3):238–43. Crossref
82. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American society for information science. 1990; 41(6):391–407. Crossref
83. Kontostathis A, Pottenger WM. A framework for understanding Latent Semantic Indexing(LSI) performance. Information Processing & Management. 2006; 42(1):56–73. Crossref
84. Chen L, Tokuda N, Adachi H. A patent document retrieval system addressing both semantic and syntactic properties. Proceedings of the ACL-2003 workshop on Patent corpus processing, Japan. 2003; 20:1–6. Crossref
85. Chen L, Tokuda N. Robustness of regional matching scheme over global matching scheme. Artificial Intelligence. 2003; 144(1):213–32. Crossref
86. Ryley J. Latent semantic indexing for patent information [Internet]. [cited 2007 Sep 12]. Available from: Crossref
87. Ding CH. A similarity–based probability model for latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, USA; 1999. p. 58–65. Crossref
88. Moldovan A, Bot RI, Wanka G. Latent semantic indexing for patent documents. International Journal of Applied Mathematics and Computer Science. 2005; 15(4):551–60.
89. Magerman T, Van Looy B, Song X. Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. Scientometrics. 2010; 82(2):289–306. Crossref
90. Aono M. Leveraging Category-based LSI for Patent Retrieval. Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, Japan; 2007. p. 373–6.
91. Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, USA; 1999. p. 50–7. Crossref
92. Hofmann T. Probabilistic latent semantic analysis. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Sweden; 1999. p. 289–96.

93. Kumar R, Math S, Tripathi RC, Tiwari MD. Patent classification of the new invention using PLSA. Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia, India; 2010. p. 222–5. Crossref
94. Venner J. Pro Hadoop. Apress: New York, USA; 2009. Crossref
95. Dean J, Ghemawat S. System and method for efficient large scale data processing. United States patent US 7,650,331; 2010.
96. Lämmel R. Google's MapReduce programming model-Revisited. Science of computer programming. 2008; 70(1):1–30. Crossref
97. WIPO- International Patent Classification (IPC) provides for a hierarchical system of language independent symbols for the classification of patents and utility models according to the different areas of technology to which they pertain [Internet]. [cited 2017 Jun 15]. Available from: Crossref