IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. VIII (Nov – Dec. 2014), PP 48-52
www.iosrjournals.org

WordNet Sense Disambiguation Based Patent Search

S.R.Janani[1], C.Sathya[2], S.Poornima[3]
[1&2](Assistant Professor/Department of CSE, SNS College of Technology, Coimbatore, India)
[3](Assistant Professor/Dept of CSE& IT, Coimbatore Institute of Technology, Coimbatore, India)

Abstract: *Patent Search has attracted considerable attention recently for finding existing relevant patents and validating new patent application. Earlier try-and-see approach was used to find any relevant underlying patents. Recent development has involved partition methods to find relevant patents. However, that method mainly focuses on Effectiveness of the search but not the Efficiency. To address this problem, we propose a new user-friendly patent search method which focuses on both Effectiveness and Efficiency of the patent search method. Error correction, query expansion and query suggestion techniques are used to get the patent search effectively and to improve the efficiency of the patent search, The proposed system uses two tools called WordNet and POS Tagger. We have used PoS tagger and SynSets from WordNet for identifying the intention of the user for the given search keyword and to create relevant partitions based on the Synsets. Finally top-k answers are generated from those highly relevant partitions are grouped together and displayed to the user.*
Keywords: *Patent search, error correction, query suggestion, query expansion, PoS tagger, WordNet .*

I. Introduction

A patent is a set of exclusive right granted by a sovereign state to an inventor or their assignee for a limited period of time in exchange for detailed public disclosure of an invention. In this modern world, developments in technologies have led to a large number of inventions. At the same time, the person who invents or creates the idea deserves a copyright for their work, for this purpose Patents are evolved.

Now-a-days Patents play a vital role in protecting the intellectual property right. Patent search generally help the patent examiners to find the existence of any relevant patents which are previously published and to validate new patent applications.

It has become very popular and attracted much attention from industrial and academic communities. Some of the available patent search systems in online are Google patent search, Derwent Innovations Index (DII) and Unites States Patent and Trademark Office (USPTO).

Earlier methods of patent search involved try-and-see approach to check each and every underlying patent for existence of any relevant patents in order to validate or invalidate the patent application. That method was very tedious for finding relevant patents from a very large volume of data. So some many new methods are proposed for patent search.

The most popular method of searching patents from a database of previously issued patents involves the partitioning of the patents based upon the given search query. On entering the search query the system will generate partitions of the underlying patents which are related to the given query and those partitions are accessed in the order of most relevant partitions and it gives back the result in the basis of ranking of the results.

The major disadvantage of the above system of search is that, it mainly focuses only on the effectiveness of the search but not on the efficiency part of the search. It obviously increases the time of search, thereby reducing the efficiency of the search method. In the current world, time is basically a most important constraint. For increasing the efficiency of the search, a new method has been added with the existing method which focuses on the effectiveness of the search.

The proposed system uses two tools called WordNet and POS Tagger. The WordNet is the lexical database of English language and POS Tagger is used to identify the parts of speech tags in the document. In this proposed system WordNet and POS Tagger plays a significant role in identifying the meaning of the keyword so that the exact search intention of the user can be found out.

II. Related Work

Analysis the Patent is different from the other search tasks and it is quit challenging and difficult task, because it should concentrate more search task such as Novelty, Infringement search, etc. [1] Shows the survey of the following search tasks: State of the Art, Novelty, Patent-ability, Infringement, Opposition, Freedom to Operate, Due Diligence and also states that, a special Patent search system is needed to meet the searcher requirement.

Due to the complex structure of patents, retrievability of relevant information is difficult. The patent retrievability is increased in [2] by using the Prior – Art method. Patent queries are generated by using pseudo

relevance feedback with query expansion. Latent Dirichlet Allocation (LDA) method is implemented in [3] for finding the probability of keywords from many topics.

While letter by letter user types the query, the related terms are suggested based upon the topics in [4, 5] and LDA is suitable for finding the probability of keywords from many topics that is stated in [3]. This technique is used in the proposed system to improve the effectiveness of the search.

Support Vector machine is used in [6] to rank the results and accessing the large amount is tedious task and so new information access paradigms is introduced in [7] to improve the search quality. Leah S. Larkey proposes a system in [8] for searching and classifying the patent and it fully based on Inquery.

The authors studied how to automatically transform a query patent into a search query and to find the answers by using the search query. [9] Mainly focuses on how the query words are extract from patents and weight them and whether to use noun phrases. Our problem is different from theirs as we focus on improving efficiency and quality to answer a keyword query.

III. Erection Of User-Friendly Patent Search Paradigm

The objective of the paper is to increase the effectiveness of the user search experience. The proposed method of search helps the user to find the relevant underlying patents effectively and to increase the user search experience by providing answers, which exactly matches the search intention of the user.

Keyword Identification from the Patent Documents

The main process is to work with the existing underlying documents. This work mainly deals with the collection of keywords from the patent documents. Usually, the patent documents are composed of a detailed explanation of the patent. Unique keywords are not used to describing the patent documents. The whole document processing at the time of search is a tedious process. So only the keywords used in the patent document is being taken for the search purpose.

The objective of considering only the keywords from the patent documents is that, whenever a user gives a search keyword, it only refers to the key which is being used in the document. So it is obvious that the explanation part in a document adds up as a junk which in turn slows down the searching process and also degrading the efficiency of the search. The partitioning concept is introduced for the large collection of documents. The documents are partitioned based on the concept or keyword under which the patents are dealing with.

The underlying patents are processed and the keywords used in a document is separated and formed as a dataset and it will be used for the searching process. Whenever a new application is approved for the patent right, this process will be repeated.

Loading patent database into POS tagger

Every individual processed patent document must be loaded into the POS tagger for identifying the tags used in the document. The data consists of the keywords, whose part of speech tags will be added with it. The tagged document collectively forms the database for the proposed search process. The documents in the patent database are loaded into the POS tagger and the tagged document is stored as a dataset.

Synset Creation of Queries

To load the data from dataset, select a file from the driver where a collection of patents are saved. Then the document is selected for which the phrases are identified word by word for the given patents. The keyword search helps to get the query keywords from the user and identifies the phrases for the given query keywords, so that the patents are initially identified and the relevant keywords are searched. These words are displayed as a result. For improving efficiency patent is partition into small partitions based on their topics and classes. Then given a query is used to find highly relevant partitions and answer the query for each relevant partition. Finally, the answers of each partition are combined and generate top answers of the patent-search query. In this module, the given keyword will be given to WordNet for Synset generation and the Synset will be used for patent matching. The retrieved patents will be ranked according to the degree of match with given keyword.

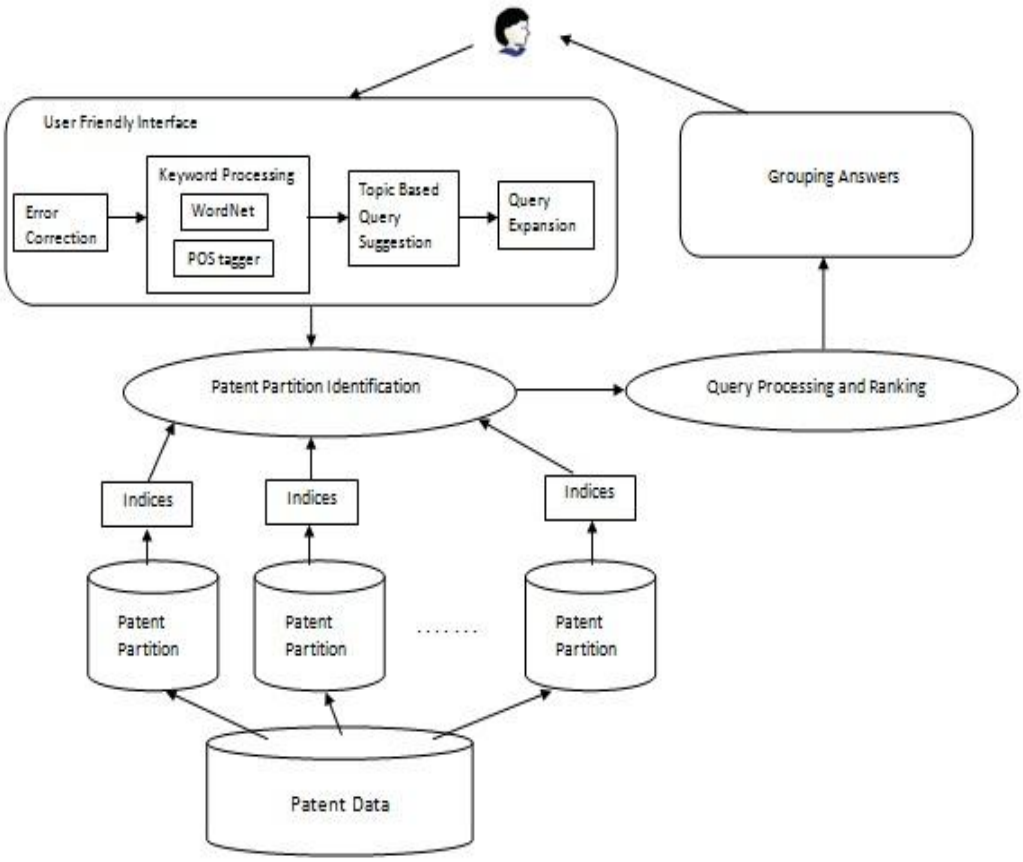

Fig.1 WordNet Sense Patent Search Architecture

Keyword processing by WordNet and POS tagger

During the search process as shown in the fig. 1, the keyword typed by the user is loaded into WordNet and POS tagger. The WordNet identifies the search intention of the user by matching the keyword with the Synset contained in it and the POS tagger identifies the part of speech tag for the given keyword. The WordNet selects the partitions which are to be searched for obtaining relevant partitions by relating the Synset with the indices of the patent partitions in the dataset. Hence the most relevant partitions for the given search keyword is selected.

Indexing of Patents

The patents are partitioned into different data partitions due to the following reasons. Patents inherently have different classes. The number of patents is very large and also it rapidly increases. For patent search query, only some classes or subclasses of patents could be relevant to the patent query.

Because of these reasons, patents are partitioned based on their classes and topics using the topic model as follows: First have to extract the topic of each patent and then partition the patents with the same topic into the same data partition. The patents in the same partition are highly relevant and those in different partitions are irrelevant.

For each partition, a well-known inverted index structure has to build and for each query keyword, the index structure has to use to find patents, that containing the keyword. Then, have to intersect the patents corresponding to different keywords to generate the most relevant patents. For each partition, any effective ranking function can be used to rank the patents in the partition. Patents in each partition are very relevant, so we can also do deeper ranking by considering the correlation between different patents. Trie structure is used to make the query suggestion easy. During the patent partition, keywords are constructed on the top, each keyword corresponds to a unique path from the root of the trie to a leaf node and each node on the path has a label of a character in the keyword. Inverted list of IDs of records that contain the corresponding keyword are stored in each leaf node.

IV. User-Friendly Interface

To identify users' query intention, several effective techniques are introduced to make the patent search user friendly and help the user to easily find relevant patents.

Error Correction

User typed query keywords may have typos; traditional methods cannot find the answers that contain the query keywords and they will not return any answer. So the traditional method is not user friendly. Alternatively, typos can be corrected using edit distance method and return answers for similar keywords. The edit distance between two keywords is the minimum number of edit operations (i.e., insertion, deletion, and substitution) of single characters needed to transform the first one to the second. For example, the edit distance value is 2 for "patent" and "paitant". If the edit distance is within a given threshold, then the two keywords are said to be similar.

To find similar keywords of a query keyword, filter-and-refine framework is used. In this method, first filter step is used to find a subset of keywords which may be potentially similar to the query keyword and then, it uses a verification step to remove those false positives and get the final similar keywords [8]. Although these methods can be used to efficiently suggest keywords for complete keywords, but they cannot support prefix keyword that the user is completing. To overcome this issue, the trie structure can be used to do efficient keyword correction and completion. Even if the users type a partial query keyword, by using trie structure relevant accurate keywords can be efficiently suggested. The vital idea is that if a prefix is not similar enough to a trie node, then no need to consider the keywords that are under the trie node. Because of this trie structure idea, similar keywords can be efficiently suggested.

Topic-Based Query Suggestion

A novel model for effectively suggesting keywords as user's type in queries letter by letter. Topic model is used to estimate the probability of the next query keyword. If the query keyword in patents is more topically coherent with the previously typed query keywords, then high score would be obtained. Two important probabilities are focused to estimate the score of each keyword, 1.the probability of a keyword conditioned on topics, 2. The probability of a keyword sampling from a patent. Latent Dirichlet Allocation (LDA) model can be used to know about the keyword distribution over each topic from the underlying patents. This model can be classified as a soft-clustering technique, which allows a keyword to appear in multiple topics and takes into account the degree of a keyword belonging to each topic. A language model is used to learn about the distribution of keywords over a set of patents. The topic-based method and two probabilities are used to suggest the relevant keywords. This smoothing technique has been intensively examined and shown effective. It is the prior probability that the patent generates any term given the context s. A normalized ranking score is used to estimate it (the TF-IDF model).

Query Expansion

In most of the cases, users cannot understand the underlying data accurately. The user may type ambiguous or inaccurate keywords and in some cases the same keyword may have different representations. The WordNet can be used to expand these kinds of query keyword. If the particular query keyword is indexed by WordNet, then the relevant keywords can easily suggested. But if the keyword is not indexed in WordNet, then the relevant keywords cannot be recommended. To overcome this issue, two solutions are suggested that are Search engine and query logs to get the relevant keywords. If the user types the patent query keyword, the relevant keywords are suggested to expand the query by using the search engine and query logs [10].

Ranking and Query processing

After selecting the most relevant partitions, the query processing component relates the noun from the keywords, which are present in the patent database to the search keyword. This matching process keeps the number of occurrences of a particular keyword in a document as a record and these records are stored in the local database. Finally the searched relevant patents are ranked based upon the highest number of occurrences in a document. The document, which has the highest number of occurrences, is placed in the top and hence producing the top-k results for the search keyword.

V. Conclusion

In this paper, a new method is proposed to improve the efficiency of the search process and improve the user search experience by providing answers, which is exactly matches the search intention of the user, which include the advantages of the existing system. WordNet is used to find the exact search intention of the user by matching the meaning of the keyword used and the PoS Tagger identifies the parts of speech

components in the patent document. Hence the system easily matches the given keyword with the phrases generated by the PoS Tagger.

References

[1]. L.Azzopardi, W. Vanderbauwhede, and H. Joho, "Search System Requirements of Patent Analysts," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp.775-776, 2010.

[2]. S. Bashir and A. Rauber, "Improving Retrievability of Patents in Prior-Art Search," Proc. European Conf. Information Retrieval (ECIR), pp. 457-470, 2010.

[3]. D.M Blei, A.Y Ng, and M.I Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[4]. J. Fan, H. Wu, G. Li, and L. Zhou, "Suggesting Topic-Based Query Terms as You Type," Proc. Int'l Asia Pacific Web Conf. (APWEB), pp. 61-67, 2010.

[5]. G. Li, J. Feng, and C. Li, "Supporting Search-As-You-Type Using SQL in Databases," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 2, pp. 461-475, Feb. 2013.

[6]. Y. Guo and C.P. Gomes, "Ranking Structured Documents: A Large Margin Based Approach for Patent Prior Art Search," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), pp. 1058-1064, 2009.

[7]. S. Ji, G. Li, C. Li, and J. Feng, "Efficient Interactive Fuzzy Keyword Search,"Proc. Int'l Conf. World Wide Web (WWW), pp. 371-380, 2009.

[8]. L.S. Larkey, "A Patent Search and Classification System," Proc. Fourth ACM Conf. Digital Libraries, pp. 179-187, 1999.

[9]. C. Li, J. Lu, and Y. Lu, "Efficient Merging and Filtering Algorithms for Approximate String Searches," Proc. Int'l Conf. Data Eng. (ICDE), pp. 257-266, 2008.

[10]. Poorani.V, Iniyan.S, "An Efficient Approach for Patent Search Engine", International Journal of Emerging Technology & Research, pp 487-491, vol 1, Issue 1, Nov-Dec, 2013.