Exploring Keyphrase Extraction and IPC Classification Vectors for Prior Art Search

Manisha Verma and Vasudeva Varma

Search And Information Extraction Lab,
Language Technologies Research Centre,
International Institute of Information Technology,
Hyderabad, India
manisha.verma@research.iiit.ac.in, vv@iiit.ac.in

Abstract. In this paper we describe experiments conducted for CLEF-IP 2011 Prior Art Retrieval track. We examined the impact of 1) using key phrase extraction to generate queries from input patent and 2) the use of citation network and (International Patent Classification) IPC class vector in ranking patents. Variations of a popular key phrase extraction technique were explored for extracting and scoring terms of query patent. These terms are used as queries to retrieve similar patents. In the second approach, we use a two stage retrieval model to find similar patents. Each patent is represented as an IPC class vector. Citation network of patents is used to propagate these vectors from a node (patent) to its neighbors (cited patents). Similar patents are found by comparing query vector with vectors of patents in the corpus. Text based search is used to re-rank this solution set to improve precision. Two-stage system is used to retrieve and rank patents. Finally, we also extract and add citations present within the text of a query patent to the result set. Adding these citations (present in query patent text) to the results shows significant improvement in Mean Average Precision (MAP).

Keywords: Prior Art Retrieval, Patent Retrieval, Key phrase extraction, CLEF-IP track

1 Introduction

We participated in the CLEF-IP 2011 Prior Art Retrieval track to evaluate the performance of existing approaches and a new representation for patents on a large collection of documents and queries. Our goal was to use and evaluate key phrase extraction for constructing queries from input patents, impact of IPC class information and citation network of patents in the corpus on recall and contribution of citation mining in enriching initial set of search results.

IPC class information can be useful in filtering or re-ordering the search results. In CLEF-IP task, BiTeM group [1] have used IPC codes to filter patents which do not share at least one IPC code with the query patent. Key phrase extraction has been previously used to construct queries from input patents. In [2]

candidate n-grams are selected using a classifier. The authors manually annotate potential keywords to train the classifier. Extraction of citation information present in the the patent text has also improved (Mean Average Precision) MAP in previous CLEF-IP tracks [2, 6]. In CLEF 2011 we try and evaluate three ways to improve prior art retrieval. Firstly, we evaluate variations of a popular key phrase extraction technique (TextRank) for extracting and scoring terms of query patent. These terms are used as queries to retrieve similar patents. Secondly, we use a novel representation of patents and a two-stage retrieval approach to improve both precision and recall. Finally, we add citations extracted from the patent text to the search results, as it has improved MAP scores previously.

In Section 2, we explain briefly the approaches used for retrieving and reranking patents. The experiments, result and analysis are explained in Section 3 and Section 4 respectively. Conclusion and future work are discussed in Section 5.

2 Our Approach

2.1 Key Phrase Extraction

Reducing the input patent text to a query which can fetch related prior art was our first objective. In [5, 3] we explored several supervised and unsupervised key phrase extraction approaches to extract candidate terms to form a query. Since the training set of queries was small, we decided to use only unsupervised methods of term extraction from patents. In unsupervised approaches, TextRank outperformed *tf-idf* in selection of terms from a patent. TextRank uses the information around a word to calculate its importance whereas tf or $tf - idf$ scores do not reflect this information. Hence, we use TextRank to extract terms from a query patent and use following to assign weight to top 20 terms in the query.

TextRank: Since patents contain significant amount of text, co-occurrence information present in it can be safely used to determine key terms for a patent. Weight of each term w_i in the query is the score given by TextRank algorithm.

TextRank*idf: TextRank extracts words which are central/important for a patent. However, patents either use general or totally new terms to explain new concepts. TextRank extracts both these types of terms with efficiency (with the help of co-occurrence information) but a word's score does not reflect its rarity. Hence to capture rarer terms central to the query patent we use modified version of TextRank score to weigh each word in the query. Note that words are still selected on the basis of their TextRank score, only their weight in the query is determined by the following:
Weight of each term w_i in the query is the score $TextRank(w_i) * idf(w_i)$ where $TextRank(w_i)$ is the TextRank score of w_i,

$$idf(w_i) = \frac{log(\frac{N}{1.0+df(w_i)})}{log(2)} \tag{1}$$

N is the number of documents in the collection and $df(w_i)$ is the number of documents containing w_i.

2.2 IPC-Vector based Retrieval

Patents contain meta-data other than text which can be leveraged to improve retrieval accuracy. A patent has manually assigned classification code, defining broad area of the invention. It also cites other patents to discuss similar inventions in the past. Our approach is to combine both the classification and citation information to represent a patent. Each patent is manually assigned one or more International Patent Classification (IPC) codes. We use this information to represent each patent as an IPC class vector. Citation network of patents is used to propagate these vectors from a node (patent) to its neighbors (cited patents). Thus, each patent vector is a weighted combination of its neighbor IPC information and its own. Vector formation and propagation are explained in [4]. Two stages of the system are :

1. **Stage 1 :** Converting the query and corpus patents into vectors using IPC codes and citation network. For the runs submitted in CLEF-IP, we use only cosine similarity to retrieve similar vectors.
2. **Stage 2 :** Re-Ranking top K documents using text of the query patent. We use *tf-idf*, TextRank and $TextRank * idf$ score to select and weigh top 20 words in the query.

2.3 Citation extraction from queries

Some query patents contain cited patent numbers within the text of their description. These patent numbers were not filtered out of the text of the query patent, which can be added to the search results. Adding this information to the experimental results is reported to demonstrate the impact of using this kind of information.

For the large topic collection containing 3973 query patents, citations were extracted from 1419 patent topics and found to be IDs of patents in the indexed collection. Other extracted citations that do not exist in the collection were discarded.

3 Evaluation

3.1 Data

The data collections are extracts of the MAREC[1] dataset, containing over 2.6 million patent documents pertaining to 1.3 million patents from the European

[1] http://www.ir-facility.org/prototypes/marec

Patent Office with content in English, German and French, and extended by documents from the WIPO. The queries have been translated to English from German and French with the help of Google Translator[2]. Only English translations of original patents are used for making queries. The data has been indexed using Lemur[3] toolkit. All the fields of a patent (title, abstract, description, claims, citations and IPC class information) have been indexed. Of 1.3 million patents, 0.8 million patents cite at least one patent in the corpus and 0.64 million patents are cited by at least one patent. Dimension of concatenated IPC class vector for this dataset is 79963, of which level 1 has 875, level 2 has 8631 and level 3 has 70457 dimensions respectively. 3973 query patents were provided including 1351 English and 2622 German and French patents.

3.2 Evaluation Method

In the CLEF-IP Workshop, we use the mean average precision (MAP), Recall at 100 (R@100), R@200 and R@1000 as evaluation measures. For CLEF-IP Prior Art Search task we compare the following methods:

Base: Simple Text Retrieval, 20 words, from the query patent, with high *tf-idf* values are used to form a weighted query. The weight of each word is its *tf-idf* score.

TextRank: 20 words, from the query patent, with high TextRank values are used to form a weighted query. The weight of each word is its TextRank score.

TextRank∗idf: 20 words, from the query patent, with high TextRank values are used to form a weighted query. The weight of each word w_i in the query is $TextRank(w_i) * idf(w_i)$.

Since limited number of runs could be submitted for the task, it was found on the training data that $TextRank * idf$ performed the best. Hence, we did not submit the results of Base and TextRank.

COS: Cosine similarity (COS) has been used to measure similarity between a patent and query IPC vectors. The process for generating IPC vectors for patents in the corpus is explained in 2.2.

COS, tf-idf: IPC information present in the patent is used to make the vector. Cosine is used to calculate similarity between a patent and query. For a query patent top 1000 similar patents are retrieved. These patents are re-ranked using query generated by TextRank method. It does not contain citations extracted from the patents.

COS, TR: IPC information present in the patent is used to make the vector. Cosine Similarity is used to calculate similarity between a patent and query. For a query patent top 1000 similar patents are retrieved and re-ranked using queries generated by TextRank method mentioned above.

[2] http://translate.google.com
[3] http://www.lemurproject.org/

For the runs (COS,tf-idf), (COS,TR) and (TextRank∗idf) we also add the extracted citations from query patents .

4 Results And Discussion

The results for the submitted runs are shown in Table 1. In the runs without citations, methods using vector representation of patents perform well in terms of Recall. Importance of re-ranking documents is evident from COS method (using only vector representation to find similar patents) results, as the MAP is still low. However, re-ranking the top 1000 documents does not result in significant change in MAP value either. This is primarily due to the approach used for re-ranking top documents. After the submission of the results, it was found linear combination of the COS and Text (Re-rank using queries generated by $tf - idf$, TextRank etc) score resulted in higher MAP values [4]. The high recall is due to the representation of a patent as vectors and propagation of vectors in the citation graph.

The TextRank∗idf method has low recall as compared to vector based methods. This is primarily due to limited coverage of queries. The queries created by using only the patent text cannot be used for retrieving documents which share meta information such as IPC Class information and citations. Such queries may not ensure very high recall while still managing high precision.

Citation addition to the initial set of results improves performance of TextRank∗idf significantly but pushes down the MAP for COS methods. The lowering of MAP may be due to the relevance judgments given for the queries. The relevance judgements contain more than one level of relevance. It may be the case that documents with higher priority in relevance judgments which were ranked higher by COS methods had been pushed to lower ranks due to citation addition which inturn resulted in low MAP values.

Table 1. Comparison of methods for Prior Art Search

Method	MAP	R@100	R@200	R@1000
TextRank*idf	0.055	0.200	0.26	0.399
COS	0.049	0.255	0.354	0.600
COS,tf-idf	**0.061**	**0.288**	**0.385**	**0.601**
COS,TR	0.057	0.281	0.378	0.595
TextRank*idf + Citation	**0.097**	**0.244**	**0.297**	**0.423**
COS,tf-idf + Citation	0.055	0.269	0.362	0.599
COS,TR + Citation	0.052	0.262	0.355	0.594

5 Conclusion and Future Work

In CLEF-IP 2011 we experimented with some approaches for query formation from an input patent. We also explored a two-stage approach to find related

prior art. First, a vector based representation which uses IPC information of patent and its neighbors to retrieve similar patents. This representation proves effective in increasing the recall. Then, re-ranking top 1000 documents in second stage is used to improve precision. We also used extracted citations from the query patent text to improve the results. Vector based representation proved to be effective in increasing recall, however improvement in precision was not achieved with simple re-ranking of documents. Approaches like TextRank which use co-occurence information in the text to find out key terms were better than frequency based measures of selecting words from the text. Citation extraction and addition certainly proved instrumental in increasing mean average precision. However, ways of citation addition to the result set needs further investigation. An extension to this work for future participation would be to use a learning-to-rank approach to re-rank top documents. It would be interesting to observe effects of combining both vector representation with patent text to avoid re-ranking.

References

1. J. Gobeill, E. Pasche, D. Teodoro, and P. Ruch. Simple pre and post processing strategies for patent searching in clef intellectual property track 2009. In *Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation: text retrieval experiments*, CLEF'09, pages 444–451, Berlin, Heidelberg, 2009. Springer-Verlag.
2. P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for Prior Art Search. In *CLEF 2010 - Conference on Multilingual and Multimodal Information Access Evaluation*, Padua Italie, 2010.
3. M. Verma and V. Varma. Applying key phrase extraction to aid invalidity search. In *Proceedings of the 13th International Conference on Artificial Intelligence and Law*, ICAIL '11, pages 249–255, New York, NY, USA, 2011. ACM.
4. M. Verma and V. Varma. Patent search using IPC Class vectors. To Appear In *Proceedings of the 4th international workshop on Patent information retrieval*, PaIR '11.
5. R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proc. of EMNLP*, 2004.
6. W. Magdy, J. Leveling, and G. J. F. Jones. DCU @ CLEF-IP 2009: Exploring standard IR techniques on patent retrieval. In *10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30-October 2, Revised Selected Papers*, Lecture Notes in Computer Science (LNCS). Springer, 2010.