Exploring Structured Documents and Query Formulation Techniques for Patent Retrieval

Walid Magdy, Johannes Leveling, Gareth J.F. Jones

Centre for Next Generation Localization
School of Computing
Dublin City University, Dublin 9, Ireland
{wmagdy, jleveling, gjones}@computing.dcu.ie

Abstract. This paper presents the experiments and results of DCU in CLEF-IP 2009. Our work applied standard information retrieval (IR) techniques to patent search. Different experiments tested various methods for the patent retrieval, including query formulation, structured index, weighted fields, document filtering, and blind relevance feedback. Some methods did not show expected good retrieval effectiveness such as blind relevance feedback, other experiments showed acceptable performance. Query formulation was the key to achieving better retrieval effectiveness, and this was performed through assigning higher weights to certain document fields. Further experiments showed that for longer queries, better results are achieved but at the expense of additional computations. For the best runs, the retrieval effectiveness is still lower than for IR applications for other domains, illustrating the difficulty of patent search. The official results have shown that among fifteen participants we achieved the seventh and the fourth ranks from the mean average precision (MAP) and recall point of view, respectively.

1 Introduction

This paper presents the experimental results of Dublin City University (DCU) in the CLEF-IP track 2009. We participated in the main task which is retrieving patents prior art. The aim of the task is to automatically retrieve all of citations for a given patent (which is considered as the topic) [5]. Only three runs were submitted, but additional unofficial experiments were performed for this task. Fifteen participants have submitted 48 runs; according to MAP scores, our best run achieved the seventh rank across participants and the 22nd across all runs. However, according to recall scores, our best run achieved the fourth rank across all participants and the fourth rank across all 48 submitted runs.

The paper is organized as follows: Section 2 describes the data for the task and an analysis of its nature; Section 3 presents all the experiments for this task; Section 4 shows the results; then Section 5 discusses these results; Finally, Section 6 concludes the paper and provides possible future directions.

2 Data Pre-Processing

More than 1.9M XML documents were provided representing different versions of 1M patents filed between 1985 and 2000. For our experiments, all different document versions for a single patent were merged into one document with fields updated from its latest versions. Patent structure is very rich, and some fields are present in three languages (English "EN", German "DE", and French "FR"), namely the title and claims. Only the patent 'title', 'abstract', 'description', 'claims', and 'classifications' fields are extracted from the patents. However, many patents lack some of these fields. The only fields that are present in all patents are the title and the classifications; the other fields are omitted in some patents. The "description" field is related to the "claims" field, and if the "claims" field is missing, then "description" is missing too. However, the opposite in not true, as some documents contain a "claims" field while the "description" field is missing. The "abstract" field is an optional part that is present in some patents. About 23% of the patents do not contain the claims and description fields, out of which 73% only have titles. 54% of the patents have claims in three languages (English, French, and German), and the remainder 23% of the patents have claims in the document language only (language of the 'description' field), these 23% are 68% English, 23% German, and 9% French.

In order to avoid language problems, the English fields only are selected. This step will lead to the loss of extra 7.4% of the patents which lack the claims and description fields (these are the German and French patents with claims only in one language). In addition, all non-English patents lack the abstract and description fields. The final outcome resulted in 30% of the collection suffering from missing most of the fields. This portion of the collection mostly comprises the titles only with a small portion of it containing abstracts too.

In order to maintain the full structure and overcome the lack of some fields in some patents, the abstract (if it exists) is copied to the description and claims fields; otherwise, the title is used instead.

3 Experimental Setup

In this section, different experiments for indexing and searching the data are discussed. After merging different versions of patents and extracting the relevant fields, some pre-processing is performed for the patent text in order to prepare it for indexing. Different methods were used for query formulation to search the collection.

Many experiments were performed on the training topics provided by the task organizers, however, a small number was submitted on the test data for the official runs. The training set contains 500 patent topics, which was sufficient to compare different methods and select the best for the official submissions. Official experiments were performed on the X-large topics set consisting of 10,000 patent topics. For each topic, the top 1,000 documents are retrieved.

3.1 Text Pre-Processing

Patent text contains many formulas, numeric references, chemical symbols, and patent-specific words (such as *method*, *system*, or *device*) that can cause a negative effect on the retrieval process. Some filtering of the text is done by removing predefined stop words[1], digits, and field-specific stop words.

To obtain the fields stop words, the field frequency for terms is calculated separately for each field. The field frequency for a term "T" in field "X" is the number of fields of type "X" across all documents containing the term "T". For each field, all terms with field frequency higher than 5% of the highest term field frequency for this field are considered as stop words. For example, for the "title" field, the following words have been identified as stop words: *method*, *device*, *apparatus*, *process*, etc; for another field such as "claims", the following words have been identified as stop words: *claim*, *according*, *wherein*, *said*, etc.

3.2 Structured Indexing

Indri [6] was used to create a structured index for patents. A structured index keeps the field structure in the index (Figure 1). This structured index allows searching specific fields instead of searching in the full document. It also allows giving different weights for each field while searching. As shown in Figure 1, "DESC1" and "CLAIM1" are sub-fields for the description "DESC" and claims "CLAIMS" fields respectively. "DESC1" is the first paragraph in the description field; typically it carries useful information about the field of the invention and what the invention is about. "CLAIM1" is the first claim in the claims sections, and it describes the main idea of the invention in the patent. The field "CLASS" carries the IPC classification [7] information of the patent of which the three top classification levels are used, the deeper levels are discarded (example: B01J, C01G, C22B).

As mentioned earlier, for patents that lack some fields, the empty fields are filled with the abstract if it exists or with the title otherwise. Pre-processing includes stemming using the Porter stemmer [4].

```
<DOC>
      <DOCNO>patent number</DOCNO>
      <TEXT>
            <TITLE>title</TITLE>
            <CLASS>3rd level classification</CLASS>
            <ABSTRACT>abstract</ABSTRACT>
            <DESC>
            <DESC1>1st sentence in description</DESC1>
            Rest of patent description
            </DESC>
            <CLAIMS>
            <CLAIM1>1st claim</CLAIM1>
            Rest of patent claims
            <CLAIMS>
      </TEXT>
</DOC>
```

Fig. 1. Structured text for a patent in TREC format

[1] http://members.unine.ch/jacques.savoy/clef/index.html

3.3 Query Formulation

Query formulation can be seen as one major task in patent retrieval ([1], [5]). As a full patent is considered to be the topic, extracting the best representative text with the proper weights is the key enabling for good retrieval results.

Using the full patent as a query is not practical due to the huge amount of text in one patent. Hence, text from certain fields was extracted and tested to search the structured index with different weights to different fields. Various combinations of fields were employed, using different weights, enabling/disabling filtering using third level classification, and enabling/disabling blind relevance feedback [6].

The patent topic text was pre-processed in the same way as in the indexing phase by removing stop words and digits, in addition to removing special characters, symbols and all words of small length (one or two letters).

Similar to the indexed documents, only English parts are used, which means all non-English patent topics will miss the abstract and description fields to be used in the search. However, the amount of text present in claims and titles should be sufficient to create a representative query. In patent topics, claims and titles are always present in all three languages. Two types of experiments for the query formulation were conducted, the first type focused on using the short text fields to create the query from the patent topic. The short text fields that were used for constructing the queries are "title", "abstract", "desc1" (first line in description), "claim_main" (first sentence in first claim), "claim1" (first claim), and "claims". The second type of experiments tested using the full patent description as the query, which does not exist for non-English patent topics; hence, the already existing translated parts of the non-English patents are used instead.

The aims behind both types of experiments are to check the most valuable parts that better represent the patent, and to check the possibility of reducing the amount of query text which leads to less processing time without reducing the quality of results.

3.4 Citations Extraction

One of the strange things about patents, and that is thought to be neglected or forgotten by the track organizers, is the presence of some of the cited patents numbers within the text of the description of the patents. These patent numbers have not been filtered out of the text of the patent topics, which can be considered as the presence of part of the answer within the question. Despite of this fact, we have not focused on building extra experiments based on this information as it can be considered as a hack for finding the cited patents. In addition, in real life this information is not always presented in the patent application, and hence, creating results on it can be considered as a misleading conclusion in the area of patent retrieval.

However, in the results, adding this information to the tested methods is reported to demonstrate the impact of using this kind of information. Results shows that a misleading high MAP can be achieved but with a very low recall, and recall is usually the main objective for the patent retrieval task.

For the X-large topics collection which contains 10,000 patent topics, 36,742 patent citations were extracted from the patent topics, but only 11,834 patents

citations were found to be in the patent collection. The 11,834 patent citations are extracted from 5,873 patent topics, leaving 4,127 topics with nothing extracted from them. Only 6,301 citations were found to be relevant leading to a MAP of 0.182 and a recall of 0.2 of the cited patents to these topics. The format of the cited patent number within the description text varies a lot; hence, we think that more cited patents could be extracted from the text if more patterns for the citations were known to us.

4 Submitted Runs and Results

Some of the tested methods seemed to be ineffective for our IR experiments. Blind relevance feedback (FB) and structured search have negative impact on the results (best FB run achieved 0.05 MAP) . All experiments with blind relevance feedback led to a degradation in the MAP to around 60% of the original runs without feedback, and this can stem from the low quality of the highly ranked results. Structured retrieval was tested by searching each field in the patent topic to its corresponding field in the index. Different weights for fields were tested; however, all experiments led to lower MAP and recall than searching in the full index as a whole without directing each field to its correspondent. Since patent documents were treated as full documents neglecting their structure, patent topics which were used for formulating the queries were tested by giving different weights to the text in each short field and compared to using the full description for formulating the query. Assigning higher weight to text in "title", "desc1", and "claim_main" has been proven to produce the best results across all runs for using the short fields.

Three runs were submitted to CLEF-IP 2009 on the official topics with the same setup which returned the best results in training. The three runs tested how to better use the short fields to generate the query. The common setup for the three runs was as follows:

1. The patent document is treated as a full document, neglecting its structure.
2. English text only is indexed with stemming (Porter stemmer).
3. Stop words are removed, in addition to digits and words consisting of less than two letters.
4. A query is formulated from the following fields with the following weights: 5×title + 1×abstract (English topics only) + 3×desc1 (English topics only) + 2×claim_main + 1×claims.
5. Additional bi-grams with a frequency in the text higher than one were used in query. The text of the fields: "title", "abstract", "desc1", and "claim_main" was used for extracting the bi-grams words.

The difference between the three runs is as follows:

- Run 1: No filtering to the results is performed.
- Run 2: Filtering is performed for all results that do not match up to the third level classification code of the patent topic (at least one common classification should be present).
- Run 3: The same as 2nd run, but removing query words consisting of less than three letters.

Runs were submitted on the X-large topic collection that contains 10,000 patent topics. The average time for running this amount of topics was around 30 hours (about ten seconds on average for retrieving results of one topic on a standard 2GB RAM, Core2Duo 1.8GHz PC).

Later experiments tested the use of the full description text of a patent topic to generate the query after removing all terms appeared only once. The average amount of time taken to search one topic was found to be slightly higher than 1 minute, which is more than 6 times the average time taken for searching using the short fields.

Table 1 shows the results of the 3 submitted runs [5]. In Table 1, it is shown that the 3rd run got the best results from the precision and recall perspective. The 1st run yields the lowest performance, which shows that applying the filtering over the results based on the patent classification codes is useful. For all runs (official and training ones), the retrieval effectiveness is relatively low when compared to other IR tasks; this can stem from the nature of patent document itself in addition to the task of finding cited patents which are relevant to the patent topic from the conceptual point of view, not from the word matching. This is discussed in the next section in detail.

In Table 2, the additional experiments when using the description of the patent topics to search the collection are compared to the best run in Table 1 (Run 3). In addition, adding the extracted citations from the description to both results is reported. From Table 2, it can be seen that using the description text for searching is on average 11% better than using the best combination of the short fields from the precision perspective, and this was statistically better when tested using Wilcoxon statistical significance test with confidence level of 95% [3]. Furthermore, combining the results with the extracted citations from the text leads to a huge improvement in the MAP, where these citations are considered as the top ranked results in the final list then added the results from the searching. When combining both results, if more than one citation is extracted from the text on one topic, they are ordered according to their position in the search result list, otherwise, the extracted citations are ordered randomly in the top of the list. Although the impact of adding the extracted citations to the results list is high, it can not be considered as an information retrieval result, as no search effort is done for retrieving these documents, and building a conclusive method for searching patents can not be generalized based on these results as it is not the common case for most of the cited patents.

Table 1. Recall (R) and MAP for the 3 submitted runs in CLEF-IP 2009.

Run #	R	MAP
Run 1	0.544	0.097
Run 2	0.624	0.107
Run 3	0.627	0.107

Table 2. Recall (R) and MAP for the best submitted run compared to using patent topic description for search with and without adding extracted citations.

	Official Results		With Extracted Citations	
	R	**MAP**	**R**	**MAP**
Run 3	0.627	0.107	0.660	0.200
Description	0.627	0.119	0.668	0.209

5 Discussion

In this section, results are analyzed to identify the reasons behind the low retrieval effectiveness for the patent retrieval task. In order to analyze this problem, the overlap between short fields of each topic in the training data and its relevant cited patents is computed; in addition, the overlap between short fields of the topics and the top five ranked non-relevant documents is calculated. The reason behind selecting the number "five" is that the average number of relevant documents for all topics is between five and six. The overlap is measured using two measures: 1) cosine measure between each two corresponding fields of the two compared patents; 2) percentage of zero overlap (no shared words) between two corresponding fields of the two compared patents. The same pre-processing is done for all patents and topics, where stop words are removed (including digits), and the comparison is based on the stemmed version of words. From Figure 2 and 3, it seems that relying on common words between topics and relevant documents for patent retrieval is not the best approach. Figure 3 shows that the cosine measure between the top ranked non-relevant documents to the topic is nearly twice as high as for the relevant documents for all fields. The same is shown in Figure 4, where surprisingly, 12% of the relevant documents for topics have no shared words in any field with the topics. This outcome has proven the importance of introducing different approaches for query formulation instead of relying on word matching in the patent topics only.

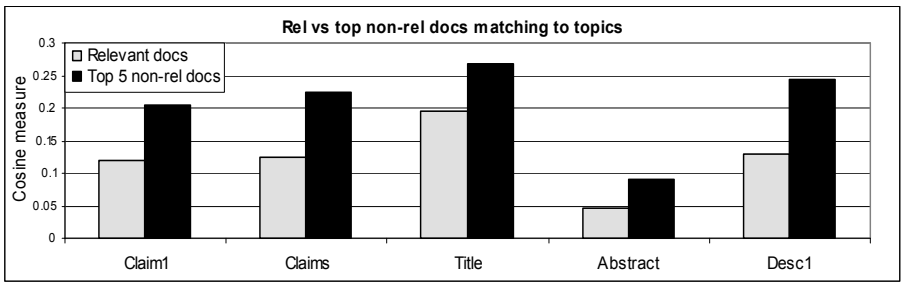

Fig. 2. Cosine measure between fields of topics and the corresponding ones in relevant and top retrieved documents

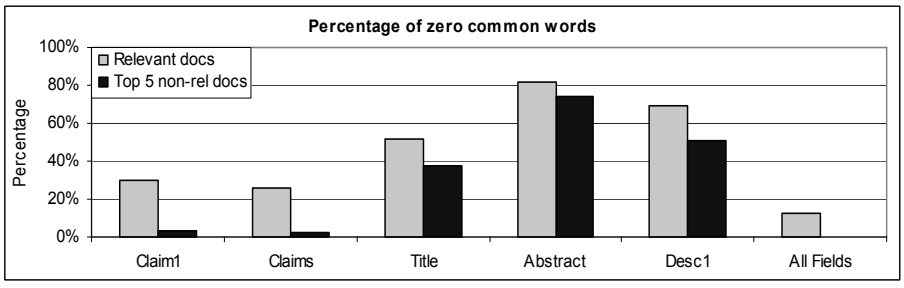

Fig. 3. Percentage of fields with zero common (shared) words between that of topics and the corresponding ones in relevant and top retrieved documents

6 Conclusion and Future Work

In this paper, we described our participation in the CLEF-IP track 2009. Standard IR techniques were tested focusing mainly on query formulation. Our experiments illustrated the challenge of the patent search task, where an additional analysis showed that depending on word matching is not the best solution as in other IR applications. Our best result was obtained by treating patents as a full document with some pre-processing by removing standard stop words in addition to patent-specific stop words. In the query phase, it was shown that the more text is present in the query the better the results are. However, the computational cost is much higher. For using the short fields for query formulation, text is extracted from these fields and higher weights are assigned to some fields. When using the full patent description text, 11% improvement in the retrieval is achieved, but 6 times the processing time is required. Some additional experiments showed the poor effectiveness of using blind relevance feedback or using the patent structure in index.

For future work, more investigation is required for checking the best use of patent structure in both index and query phases. Machine learning can be a useful approach for identifying the best weights for different fields. Furthermore, query expansion through the conceptual meaning of words is a potential approach to be tested. Finally, machine translation can be a good solution to overcome the problem of multi-lingual documents and queries.

7 Acknowledgment

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project

8 References

[1] Fujii A., M. Iwayama, and N. Kando. Overview of patent retrieval task at NTCIR-4. In *Proceedings of the fourth NTCIR workshop on evaluation of information retrieval, automatic text summarization and question answering, June 2–4, Tokyo, Japan*, (2004)

[2] Graf E. and L. Azzopardi. A methodology for building a patent test collection for prior art search. EVIA-2008 Workshop, NTCIR-7, (2008)

[3] Hull D. Using statistical testing in the evaluation of retrieval experiments. In SIGIR '93,pp 329–338, New York, NY, USA, (1993)

[4] Porter M.F. An Algorithm for Suffix Stripping, *Program* 14 (3) (1980), pp. 130–137

[5] Roda G., J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. *CLEF working notes 2009, Corfu, Greece*, (2009)

[6] Strohman T., D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis,* (2004)

[7] IPC (International Patent Classification): http://www.epo.org/patents/patent-information/ipc-reform.html