Prior Art Retrieval Using Various Patent Document Fields Contents

Metti Zakaria Wanagiri and Mirna Adriani

Fakultas Ilmu Komputer, Universitas Indonesia
Depok 16424, Indonesia
metti.zakaria@ui.edu, mirna@cs.ui.ac.id

Abstract. In this paper, we report our approach to retrieve patent documents based on the prior art. We use the standard Information Retrieval (IR) techniques which contain indexing and retrieval processes. We use various combinations of document fields for the query formulation. Based on the evaluation summary, we achieve the best result for the combinations of invention-title, description and claims fields in terms of precision and recall.

Keywords: patent retrieval

1 Introduction

There are a lot of inventions that have been invented in the industry and sciences. The number of inventions is growing from time to time as there is a high demand and need from human to have better and easier life, such as the living environment, working environment and so on. For example, around April 2010, Apple Inc. developed a new portable tablet computer called iPad which is one of the latest patented inventions. One of the functions is it can enable humans to read any e-book documents anytime and anywhere.

An invention can be granted an exclusive right called patent by the national government for a limited period of time in exchange for public disclosure of those inventions. This exclusive right granted to an inventor is the right to prevent others from making, using, selling or distributing the patented invention without permission. So, with this exclusive right, an inventor can fully protect its patented invention from any misuses in the given period of time.

According to World Intellectual Property Organization (WIPO) of United Nations, patent inventions/applications consist of patent specifications, official forms and correspondence relating to the applications. A patent specification is a document that describes the invention which generally contains the invention/application title, section detailing the background and overview of the invention, a description of the invention and embodiments of the invention and claims, which set out the scope of the protection. It also includes an abstract which provides a summary of the invention. The claims of a patent specification define the scope of protection of a patent granted by the patent and describe the invention in a specific legal style.

As the number of patent applications increases, the patent domain is considered quite important. Since there are many new inventions that are being set out for patent granting, then it should be a justification on those new inventions. A new invention should be checked whether there are any existing patents which may invalidate them. So a patent specification or patent document plays a vital role in differentiating any inventions.

In 2009, the Cross Language Evaluation Forum (CLEF) launches a track called CLEF-IP which focuses on Intellectual Property domain. It investigates the use of Information Retrieval techniques for patent document retrieval. The main task in this track is to find any existing patent documents that may invalidate a new invention who apply for its patent. Jarvelin and Preben [1] use an automatic query generation algorithm. They compare queries generated by human experts to those generated by system and the automatic generated queries achieve the better performance. Lopez and Romary [2] use multiple retrieval models for producing several sets of ranked results. Then they apply Multiple SVM regression models to merge the results. Toucedo and Losada [9] build queries by extracting terms from some textual patent documents fields using *inverse document frequency* (idf) and give preference to the title terms. BM25 retrieval model is used and the best result is achieved when the title terms and the standard parameters of BM25 retrieval model are used.

Mukherjea and Bamba [5] also develop a retrieval system for biomedical patents called BioPatentMiner. It integrates information from the patents with knowledge from biomedical ontologies to create a Semantic Web. Takaki et al. [8] propose the invalidity patent search by applying an associative document retrieval method, in which a document is used as a query to search for other similar documents. They use subtopics or compositional elements extraction to extract subtopics which correspond to an element constituting the claim section. Then content words which mainly nouns are extracted from each compositional element as query terms. Evaluation results show that the method used was effective in the patent search. Mase et al. [3] use two retrieval stages which consists of query term extraction from claim text, query term weighting without term frequency (tf) and *using measurement terms* (terms that accompanied by numerical values) and text retrieval using claims as targets. Evaluation results show that the effectiveness of the method varies depending on the test sets used.

In this paper, we report our participation in the 2nd CLEF-IP. We focused on the Prior Art Candidate Search (PAC) task to find patent documents that are likely to constitute prior art to a given patent application (patent topic).The remaining of this paper is organized as follows: section 2 discusses our retrieval system for patent documents, section 3 describes the experiments, section 4 describes the evaluation summary and section 5 is the conclusion.

2 Patent Documents Retrieval System

In this section, we describe our retrieval system using standard Information Retrieval (IR) techniques for indexing and retrieving patent documents.

2.1 Extracting Patent Fields

Before indexing process is carried out, we need to extract the patent fields in the multilingual document collection. There are about 60 different fields in a patent document, however their contents are not always informative and important. So we need to figure out which fields that are considered important to a corresponding patent application. First, we randomly take some patent documents from the CLEF-IP 2010 corpus. Then we extract all of the document fields by recognizing the associated tags. Then we create a list of unique patent fields

2.2 Indexing Documents

We choose a number of informative fields from the list of unique patent fields in which the contents considered to be valuable and represent all the information about the corresponding patent application. There are 30 chosen patent fields (see Table 1) that are used in the indexing process.

Table 1. Patent Fields for Indexing Process

abstract	address	agents
applicant	application-reference	claim
claim-text	claims	classification-ecla
classification-ipc	classification-ipcr	classification-symbol
classifications-ipcr	colspec	copyright
country	date	dates-of-public-availability
description	designated-states	doc-number
doc-page	document-id	invention-title
inventor	inventors	patent-citations
patent-document	priority-claim	priority-claims

2.3 Query Formulation

The CLEF-IP 2010 topic documents are categorized into two sets: the large topic set and small topic set. Each topic document is a patent document in XML format which has the same structured data as the patent documents in CLEF-IP 2010 corpus. Both sets come in three different languages: English, French and German (see Table 2).

Table 2. Two Sets of Topic Documents

Topic Sets	**Number of Docs**
Large Topic Set	2005
Small Topic Set	500
Total	**2505**

Our task in this track is to find all relevant patent documents in the collection that invalidate a given topic documents. In this case, we build some appropriate and effective queries from the topic documents.

In this query formulation process, we use the standard term weighting algorithm of TF-IDF [6[. Essentially, TF-IDF works by determining the relative proportion of words in a specific document compared to the inverse proportion of that word over the entire document corpus. This calculation determines how relevant a given word is in a particular document.

So, given a document collection *D* and a document *d* ϵ *D*, the calculation of TF-IDF for a word *w* is

$$w_d = f_{w,d} * log\ (N/f_{w,D}) \qquad (1)$$

where $f_{w,d}$ is the number of times *w* appears in *d*, *N* is the number of documents in *D* and $f_{w,D}$ is the number of documents in *D* in which *w* appears [6].

For each of the topic documents from both sets, we apply these steps of query formulation:

1. Extracting the contents from three patent fields: `invention-title`, `description` and `claims`,
2. Extracting words from the extracted contents by applying the standard weighting algorithm of TF-IDF,
3. Retrieving top 10 words with high TF-IDF, and
4. Forming the 10 words as one query.

As there are three patent fields that we used for query formulation, we define three possible combinations that will be used for our experiments. The combinations are:

1. `claims`
2. `invention-title` + `description`
3. `invention-title` + `description` + `claims`

So following the steps above, after we extract the contents from each patent field, we combine the contents based on the combinations above and then extract the top 10 words as the query. Finally, there are three sets of query that we will use in the retrieving process. Table 3 shows the details of the query sets.

Table 3. Details of Query Sets

Query Set	Patent Fields	Number of Queries
QS-1	invention-title + description	2505
QS-2	claims	2505
QS-3	invention-title + description + claims	2505

3 Experiments

For the experiments, we use CLEF-IP 2010 corpus which contains around 2 million patent documents from European Patent Office (EPO). Each patent document is an XML file containing structured data with different fields delimited by specified tags.

We index the documents using Indri[1] which is part of the Lemur[2] Toolkit. Indri retrieval model is based on a combination of language model and inference network frameworks [7]. We remove stopwords from the corpus but we don't stem the words. We don't use any cross language technique in those runs therefore no language specific methods are used.

We run three experiments based on three sets of query and retrieve top 1000 patent documents which are relevant to each query from sets. In the experiments, we combined the title, description, and claims that occurred o the documents. These three experiments or runs are the submitted runs for CLEF-IP 2010. For all of the runs, we use both large and small topic sets. Table 4 shows the details of the submitted runs.

Table 4. Details of Experiments

Run ID	Run Name	Query Set	Topic Set
ui-1	ui_title&desc_Run1_PAC_all	QS-1	Large + Small
ui-2	ui_claims_Run2_PAC_all	QS-2	Large + Small
ui-3	ui_title-desc-claims_Run3_PAC_all	QS-3	Large + Small

4 Evaluation

The results of our submitted runs using large and small topic set are shown on Table 5.

Table 5. The Performance of the Submitted Runs

Topic Set	Run ID	P	R	MAP	NDCG
Large	ui-1	0.0064	0.2937	0.052	0.1705
	ui-2	0.0059	0.2827	0.0457	0.1592
	ui-3	**0.007**	**0.3301**	**0.0581**	**0.1898**
Small	ui-1	0.0062	0.2859	0.0425	0.1551
	ui-2	0.0059	0.2756	0.0441	0.1559
	ui-3	**0.007**	**0.3332**	**0.0537**	**0.1846**

We present our evaluation summary in Table 5 in four measures: precision (P), recall (R), Mean Average Precision (MAP) and NDCG. The retrieval performance of

[1] http://www.lemurproject.org/indri/

[2] http://www.lemurproject.org/

all the topics sets show that the recall is much higher that the precision. The MAP of the large topic set (0.0581) is higher than the small topic set (0.0537).

Our results have motivated us to explore more on the patent fields' contents that are valuable for retrieval process. Furthermore the query formulation process needs to be improved using different approach.

5 Conclusion

This year we participate in the Patent Retrieval track in CLEF-IP 2010. We use standard IR techniques for retrieving patent documents. We identify several fields that are used in the indexing process. For the retrieval process, we combine several fields such as title, description, and claims. The evaluation shows that the precision is much lower than the recall.

There are still rooms for improvement such as adding more context to the query using query expansion or relevance feedback and also using different term weighting algorithm..

References

1. Jarvelin, A., Preben, H.: UTA and SICS at CLEF-IP. In: 1st CLEF-IP, Corfu, Greece (2009)
2. Lopez, P., Romary, L.: Multiple Retrieval Models and Regression Models for Prior Art Search. In: 1st CLEF-IP, Corfu, Greece (2009)
3. Mase, H., Matsubayashi, T., Ogawa, Y.: Proposal of Two Stage Patent Retrieval Method Considering the Claim Structure. ACM Transactions on Asian Language Information Processing 4(2) (2005)
4. Michel, J.: Considerations, challenges and methodologies for implementing best practices in patent office and like patent information departments. World Patent Information 28:132-135 (2006)
5. Mukherjea, S., Bamba, B.: BioPatentMiner: An Information Retrieval System for Biomedical Patents. In: VLDB '04: Proceedings of the Thirtieth International Conference on Very Large Data Bases, pp. 1066-1077 (2004)
6. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. In: Information Processing & Management, 24(5), pp. 513-523 (1988)
7. Strohman, T., Metzler, D., Turtle, H., Croft, B.: Indri: A language-model based search engine for complex queries (extended version). Technical Report IR-407, Department of Computer Science, University of Massachusetts (2005)
8. Takaki, T., Fujii, A., Ishikawa, T.: Associative Document Retrieval by Query Subtopic Analysis and its Application to Invalidity Patent Search. In: Proceedings of CIKM (2004)
9. Toucedo, J.C., Losada, D.E.: University of Santiago de Compostela at CLEF-IP09. In: 1st CLEF-IP, Corfu, Greece (2009)