Phrases or Terms? The Impact of Different Query Types

Daniela Becks, Thomas Mandl, Christa Womser-Hacker

University of Hildesheim, Information Science, Marienburger Platz 22,
31141 Hildesheim, Germany
{daniela.becks, mandl, womser}@uni-hildesheim.de

Abstract. At CLEF 2010, the University of Hildesheim took part in the Intellectual Property Track, which for the first time provided two separate tasks: the prior art candidate search and the classification task. We focused on the first one whose aim was to identify patent documents that state prior art of an invention. The University of Hildesheim submitted four monolingual English runs using term as well as phrase queries. Each of the experiments made use of the small topic set. With the help of the before mentioned and additional post runs, we tried to investigate the impact of phrase queries in contrast to simple terms. Compared to the results of last year, there seemed to be some improvements especially in case of the P@5 values which could be an effect of the implemented Okapi algorithm.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Performance, Experimentation

Keywords

Intellectual Property, Evaluation, Patent Retrieval System

1 Introduction

In 2009, for the first time the Intellectual Property Track took place within the context of the *Cross Language Evaluation Forum (CLEF)*. This year, the CLEF-IP Track 2010 has been organized again, but as a lab situated within the CLEF conference. In contrast to the previous IP Track, two different tasks have been provided by the organizers: the prior art candidate search and the classification task. [1] While the first one focused on identifying documents that state prior art of an invention, the second task aimed at classifying patents according to the IPC[1]. [1]

1.1 Test Collection and Topics

The test collection, a part of the MAREC collection, was provided by the IRF[2] and consisted of approximately 2.7 million documents from the *European Patent Office (EPO),* which were stored as XML files. [1, 2] Within the test collection, a document may be written in English, German or French. [1] Furthermore, the organizers provided a small (500 patents) as well as one large (2.000 documents) topic set, which contained application documents being assigned the code “A1” or “A2”. [1] Each of our runs made use of the small topic set.

1.2 Related Work

In the past years, many authors have argued that a query may be more precise if it contains phrases instead of *bag of words*. Therefore, a number of sophisticated approaches have been developed and adapted to the patent domain.

For example, Chu et al. 2008 explained a pattern-based method to identify *treatment relationships* of the form subject predicate object. Their approach is applied to US patent documents. More specifically, the authors concentrated on the medical domain. [3] A different method to extract phrases was provided by Koster and Beney 2009. They described a dependency parser which is able to extract so called *dependency triples* (word relation word). Within several experiments with different patent corpora, they figured out that dependency triples may be helpful in classifying documents. [4] The advantage of dependency relations has also been outlined by Ruge 1995, but in this case the author concentrated on the extraction of *head modifier pairs*. [5]

A look at these examples indicates that existing approaches mainly concentrate on the use of short phrases. At CLEF-IP 2010, the goal of the University of Hildesheim was to investigate the impact of the phrase length on the accuracy of the returned results. Furthermore, we aimed at finding out which phrases achieve the best retrieval results.

2 System Setup

To run our experiments, we set up a retrieval system utilizing Apache Lucene[3], which is mainly based on the traditional Vector Space Model. While in the standard implementation, the ranking is performed according to the well-known tf-idf

[1] *International Patent Classification*

[2] *Information Retrieval Facility*

[3] http://lucene.apache.org

approach, we integrated the Okapi algorithm[4] instead. We decided to do so, because a series of experiments with the CLEF-IP 2009 test collection revealed that the use of BM25 can particularly increase P@5 and mean average precision. In other words, more relevant documents are likely to appear at the beginning of the ranking list. Keeping in mind that patent searchers often spent a lot of time on analyzing the returned patents [6], this is especially advantageous in the intellectual property domain. More details on our approach are given in the following section.

2.1 Preprocessing and Indexing

At CLEF 2010, we only performed monolingual runs on an English index consisting of four different fields:

- Patent number (UCID)
- Title (INVENTION-TITLE)
- IPC codes (IPC)
- Abstract (ABSTRACT)

In a first step, abstract and title were extracted from the XML documents within the collection, whereas only the English content was taken into account. Furthermore, we decided to add the language-independent IPC codes, because in 2009, these particularly increased the recall of our retrieval system. [7] Finally, the patent number was stored into the index. In contrast to the above mentioned fields, the *UCID* simply served as an identifier and was not used within the search process.

Before being stored into the index, the text fields, including abstract and title, have been preprocessed. This preprocessing was divided into three common steps:

1. Stopword elimination
2. Tokenization
3. Stemming

Because we only performed monolingual English runs, we integrated a standard stopword list for English[5]. As described in [7], in patent documents some domain specific terms are likely to appear frequently and may therefore result in a comprehensive list of search results. To avoid this problem, the standard stopword list was enriched by this kind of words. After having removed the stopwords, the text of the abstract and the title was tokenized and finally stemmed with the Porter Stemmer[6].

2.2 Search Process

The experiments of the University of Hildesheim focused on the prior art candidate search task. Therefore, our aim has been to identify each existing document that states prior art of a given topic.[1] In the case of the Intellectual Property Track, a topic file is a patent document provided in XML format. A query is thus automatically

[4] http://nlp.uned.es/~jperezi/Lucene-BM25/

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

[6] http://snowball.tartarus.org/

constructed by extracting the text from the adequate fields of the topic file. After having extracted the text, we employed the preprocessing procedure described in Sect. 2.1.

In the context of CLEF -IP 2010, we experimented with the following query types:

1. Phrase queries consisting of just one term, term queries (run 1)
2. Phrase queries with a fixed length (runs 2-4)

3 Results and Analysis

Our experiments were separated into official and post runs. Further details on the experimental settings as well as the results can be found in next two sections.

3.1 Submitted Runs

We submitted four English runs within the prior art candidate search task.

As mentioned earlier, the experiments of 2009 made clear that the IPC codes are particularly useful to increase the recall of a retrieval system. [7] As a consequence, in each of our runs the classification information has been utilized. A relevant patent document has to share at least one IPC code with the topic file. Thus, the classification codes were connected by the Boolean operator *OR*.

Furthermore, the content of the title, the main claim and the introduction of the description has been extracted. The text of each field was either treated as single terms (run 1) or as one phrase (runs 2-4). Because this phrase, in general, might be quite long, we restricted the length to four. Phrases that exceeded this cutoff have been split into sub phrases consisting of four terms. As was the case for the classification codes, the single terms (run 1) were combined by the Boolean operator *OR*. An overview of the employed settings is given below:

1. **EN_BM25_Terms_allFields:** terms, IPC, title, main claim, description (background of the invention)
2. **EN_BM25_Phrases_title:** phrases, IPC, title
3. **EN_BM25_Phrases_des_cl:** phrases, IPC, description (background of the invention), main claim
4. **EN_BM25_Phrases_allFields:** phrases, IPC, title, main claim, description (background of the invention)

Some statistics according to the obtained results are provided in Table 1.

The results of our submitted runs (see Table 1) reveal that phrases extracted from the title seem to be most precise and effective, because this run achieved the highest mean average precision (0.0493) and the best recall (0.4816).

In contrast, the run utilizing terms (run 1) instead of phrases achieved the lowest MAP (0.041). With respect to our results, the hypothesis that phrase queries can better describe the information need than simple terms can be confirmed. Still, a mean average precision of about 0.049 and a recall of 48% are not satisfactory in the context of patent retrieval.

Table 1. Evaluation measures for the submitted runs

Run	**Recall**	**Precision**	**MAP**	**P@5**
EN_BM25_Terms_allFields	0.3298	**0.0125**	0.0414	0.0914
EN_BM25_Phrases_title	**0.4816**	0.0124	**0.0493**	0.0870
EN_BM25_Phrases_des_cl	0.3665	0.0109	0.0415	0.0922
EN_BM25_Phrases_allFields	0.3605	0.0116	0.0422	**0.0938**

3.2 Post Runs

To further investigate the effect of phrases in the context of patent retrieval, we performed some post runs. The main goal of these additional experiments has been to find out the optimal length of a phrase query. As a consequence, we decided to vary the cutoff of the phrases. While the official runs were based on phrases consisting of four terms, in the context of the additional runs a cutoff of three, five and six terms was tested out.

Because the second official run (phrases taken from the title) achieved the best retrieval results, this served as the baseline for our post runs.

Similar to the **EN_BM25_Phrases_title** run, only the content of the English title field and the IPC codes have been taken into account. As described in Sect. 3.1, a relevant patent is supposed to share at least one classification code with the topic file. The text extracted from the title field was considered to be one phrase whose length was restricted to three/ five/ six terms. Titles that exceeded this cutoff have been split into sub phrases (see Sect. 3.1). An overview of the employed settings is given below:

1. **EN_BM25_Phrases(3)_title:** phrases, IPC, title, cutoff = 3 terms
2. **EN_BM25_Phrases(5)_title:** phrases, IPC, title, cutoff = 5 terms
3. **EN_BM25_Phrases(6)_title:** phrases, IPC, title, cutoff = 6 terms

In Table 2, the results of the post runs are provided.

Table 2. Evaluation measures for the post runs

Run	**Recall**	**Precision**	**MAP**	**P@5**
EN_BM25_Phrases(3)_title	0.4703	**0.0122**	0.0479	**0.0860**
EN_BM25_Phrases(5)_title	0.4912	0.0119	0.0495	0.0828
EN_BM25_Phrases(6)_title	**0.4954**	0.0118	**0.0500**	0.0844

As can be seen in table 2, the results of the post runs look quite similar to those of our baseline (**EN_BM25_Phrases_title**). In spite of this, the third post run, which was performed with title phrases restricted to a length of six terms, achieved a higher mean average precision (0.05) and a higher recall (0.0495).

Therefore, we might summarize that longer phrases tend to increase the map and the recall of a retrieval system. Still, this hypothesis longs for further investigation.

4 Outlook

At CLEF-IP 2010, the University of Hildesheim conducted experiments, which aimed at investigating the impact of phrase queries. Particularly, we draw attention to the length and kind of phrases integrated into the query.

Our results reveal that using phrases, especially phrases extracted from the title, instead of terms seems to be advantageous wrt mean average precision. Therefore, the hypothesis that queries should contain phrases instead of terms can be confirmed.

Additionally, with the help of some post runs, we were able to figure out that queries consisting of longer phrases tend to increase the map of a retrieval system, but this aspect longs for further investigation.

In the future, we will have to think about a more sophisticated method to extract phrases from the topic files, because by now, we have not concentrated on the semantic of the phrases. The implementation of a semantic approach might further improve the results.

References

1. Piroi, F.: CLEF- IP 2010. Track Guidelines, 2010.
2. Information Retrieval Facility (2010): Data Collection. http://www.ir-facility.org/research/evaluation/clef-ip-10/test-collection (verified: 5.08.2010)
3. Chu, A.; Sakurai, S.; Cardenas, A.F.: Automatic Detection of Treatment Relationships for Patent Retrieval. In Proceedings of the PaIR'08 workshop, Napa Valley, California, USA, Pages: 9-14, 2008.
4. Koster, C.H.A., Beney, J.G.: Phrase-based Document Categorization revisited. In Proceedings of the PaIR'09 workshop, Hong Kong, China, 2009.
5. Ruge, G.: Wortbedeutung und Termassoziation: Methoden zur automatischen semantischen Klassifikation. Olms: Hildesheim, 1995.
6. Azzopardi, L.; Joho, H.;Vanderbauwhede, W.: A Survey on Patent Users Search Behavior, Search Functionality and System Requirements. Report: Available online: www.ir-facility.org/research/technical-reports/files/irf_tr_2010_00001.pdf
7. Becks, D.; Womser-Hacker, C.; Mandl, T.; Kölle, R.: Patent Retrieval Experiments in the Context of the CLEF IP Track 2009. In Peters, C.; Di Nunzio, G.M.; Kurimo, M.; Mandl, T.; Mostefa, D.; Penas, A.; Roda, G. (Eds.): Multilingual Information Access Evaluation I – Text Retrieval Experiments, Proceedings of CLEF 2009, Corfu, Greece, 2010 (to appear).