Does Patent IR profit from Linguistics or Maximum Query Length?

Daniela Becks[1], Maximilian Eibl[2], Julia Jürgens[1], Jens Kürsten[2], Thomas Wilhelm[2], Christa Womser-Hacker[1]

[1] University of Hildesheim, Information Science, Marienburger Platz 22, 31141 Hildesheim, Germany
{daniela.becks, juerge, womser}@uni-hildesheim.de
[2] Chemnitz University of Technology, Straße der Nationen 62 09111 Chemnitz, Germany
{eibl, jens.kuersten, thomas.wilhelm}@informatik.tu-chemnitz.de

Abstract. In 2011, the University of Hildesheim and Chemnitz University of Technology participated together in the CLEF Intellectual Property Track. We focused on the prior art candidate search, which was already provided for the third time. Our group submitted seven runs ranging from simple bag of words to linguistic phrases. The aim of our experiments was to examine the effectiveness of different query strategies. Especially, we wanted to evaluate the advantage of linguistic phrases in contrast to very long bag of words queries. Phrases were extracted using a special extraction component, which has been developed by the University of Hildesheim.

General Terms

Performance, Experimentation

Keywords

Intellectual Property, Evaluation, Patent Retrieval System, Natural Language Processing

1 Introduction

In 2011, the *Intellectual Property Track (CLEF-IP)* was organized for the second time as a lab within the context of the *Cross Language Evaluation Forum (CLEF)* conference. Since the beginning of the CLEF-IP track in 2009, different tasks have been proposed to the participants [1]:

- Prior art candidate search
- Classification task, two alternatives
- Image-based Classification
- Image-based Document Retrieval

Prior art search is a well known type of patent search that is performed to find out whether there exists prior art to a given patent application [1, 2]. In contrast, the goal of the classification tasks is to classify patents according to the International Patent Classification (IPC). Furthermore, two image-based tasks were introduced this year [1].

At CLEF-IP 2011, the University of Hildesheim and Chemnitz University of Technology did joint work. Each of our experiments concentrated on the prior art search, which has been organized for three years now.

The test collection, which was provided by the IRF[1], consisted of approximately 2.5 million documents stored as XML files. Each of the documents from the *European Patent Office (EPO)* was published before 2002 [1, 3]. The collection differed with respect to the last two years, because it contained about 400.000 extra documents from the *World Intellectual Property Office (WIPO)* [1, 3]. Beside the document collection, the organizers provided one topic set (about 4.000 patents assigned the code "A1" or "A2"), which equally consisted of German, English and French topic files [1].

2 System Setup

Our experiments were performed on the basis of the *Xtrieval* framework developed at Chemnitz University of Technology. This framework consists of four different components, three of which form the system core (1-3) [4]:

1. Indexing
2. Retrieval
3. Evaluation
4. User interface.

The Xtrieval framework was designed to make use of common retrieval API's, such as Lemur[2], Terrier[3] and Apache Lucene[4], for evaluation purposes. For the present experiments we used Apache Lucene in version 3.1 as retrieval core in Xtrieval [4]. Thus, the underlying retrieval model is the traditional Vector Space Model. More details on our approach are given in the following section.

[1] *Information Retrieval Facility*
[2] http://www.lemurproject.org/
[3] http://terrier.org/
[4] http://lucene.apache.org/

2.1 Preprocessing and Indexing

To index the document collection, the standard retrieval approach was followed. As a consequence, the text extracted from the XML documents was first preprocessed. This preprocessing included the following steps:

- stopword elimination
- tokenization
- stemming

In [5] it was mentioned that patent specific terms, which appear frequently are likely to result in comprehensive document lists. Following this, our group decided to use a customized stopword list, i.e. a standard stopword list[5] which was amended with a number of patent specific terms. This approach was already used in 2009 and 2010 [5, 6]. In a next step, the preprocessed text was stored into the index file. Because it proofed to be more effective during the experiments on the trainings set, we stored the text as a bag of words. In this context, each language was treated separately. Thus, the resulting index consisted of three fields, one single field per language.

Furthermore, the following patent parts were considered during the indexing process:

- Title
- Abstract
- Claim
- Description

Besides this textual information, the language-independent IPC codes were included into the index, because the experiments at CLEF 2009 showed that these are particularly advantageous to increase the recall of an information retrieval system [5].

2.2 Phrase Extraction

A lot of research has concentrated on the effectiveness of different kinds of phrases. At CLEF 2010, the University of Hildesheim investigated the effectiveness of terms and phrases in the context of patent information retrieval. The experiments using phrases significantly outperformed the term baseline, although a simple statistical approach was used [6]. Furthermore, in [7] the author focused on the advantage of statistical and syntactical noun phrases for interactive query expansion. In this case, linguistic phrases proofed to be effective for information retrieval [7]. Following this, we decided to investigate the influence of linguistic (dependency) phrases on the effectiveness of a patent retrieval system. In this context, different types of phrases were considered ranging from simple adjective noun to complex noun object relations.

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

To run our experiments we extracted phrases using a special extraction component that has been developed at the University of Hildesheim. The underlying approach is called *rule based dependency parsing* and combines the rule based method described in [8] with dependency parsing. In case of rule based dependency parsing, dependency phrases are identified with the help of defined term pairs. The following example illustrates this approach:

A *METHOD FOR IMPLEMENTING ONLINE MAINTENANCE* ***IN*** *THE COMMUNICATION NETWORK* (Topic EP-1881641-A1)

In the example title the tool would extract the phrase "method for implementing online maintenance". The determiner "a" and the preposition "in" serve as patterns to identify the phrase. This approach was implemented using UIMA[6] and openNLP[7] as the basis of the extraction tool. A detailed description of the system developed by the University of Hildesheim can be found in [9]. By now, the extraction tool has been tested only on English documents. As a consequence, the phrase experiments concentrated on English patents only.

2.3 Search Process

Our group performed various experiments for the prior candidate search task. Given the topic file, the goal was to identify those patents that describe prior art. As already mentioned, a topic file is a patent which is assigned code "A1" or "A2" [1]. On the basis of these documents, the query was constructed automatically from the content of different patent parts. These included the following:

- Title
- Abstract
- Claim
- Description
- IPC

At CLEF-IP 2011, we experimented with the following query modifications:

1. Boolean queries consisting of terms (1-5)
2. Queries consisting of linguistic phrases (6)
3. Combined queries consisting of terms and phrases (7)

[6] http://uima.apache.org/
[7] http://incubator.apache.org/opennlp/

3 Results and Analysis

The University of Hildesheim and Chemnitz University of Technology submitted seven different runs. Each experiment made use of the same index file, but differed with respect to the query options. A detailed description of these runs as well as an overview of the results can be found in the next two sections.

3.1 Submitted Runs

Our group performed four multilingual (1-4) as well as three monolingual English runs (5-7) within the prior art candidate search task. An overview of the experimental settings is given below.

1. **CUT_UHI_CLEFIP_BOW:** query terms extracted from abstract, claims and title
2. **CUT_UHI_CLEFIP_BOW_DESC:** query terms extracted from abstract, claims, title and description
3. **CUT_UHI_CLEFIP_BOW_DESC_IPCR:** query terms extracted from abstract, claims, title and description, IPC
4. **CUT_UHI_CLEFIP_BOW_IPCR:** query terms extracted from abstract, title and claims, IPC
5. **CUT_UHI_CLEFIP_BOW_EN_ABSTRACT:** query terms extracted from abstract and title
6. **CUT_UHI_CLEFIP_BOW_EN_P:** linguistic phrases extracted from abstract and terms of title
7. **CUT_UHI_CLEFIP_BOW_EN_P_ABSTRACT:** linguistic phrases extracted from abstract and terms extracted from abstract and title

As can be seen, only the third experiment made use of all patent sections and the other runs were restricted to special fields. For example, the first run was restricted to abstract, claims and title. We did not use the whole IPC code, but included the first four digits only. Although the complete classification information proofed to be more accurate, we decided to use only the first four digits, because this significantly accelerated the search process.

Our runs were divided into two major categories. The first group of experiments (1-4) concentrated on English, French and German and was performed to investigate the effect of very long queries. In contrast, the second group of runs (5-7) made use of English terms and phrases only and aimed at investigating the effect of short, but precise queries. Phrases as well as terms were combined into Boolean queries using the operator *OR*. Independent of its type, the query was run against the language specific index field. Thus, an English query was searched within each index field that contained English content. These two approaches reflect two very distinctive perspectives on prior art search.

The results revealed that the first query strategy, which was based on using as many terms as possible, proofed to be more effective, because our best run achieved a map of 0.0914 (run 3). In this case, the query was formulated using terms extracted from all patent fields. Furthermore, the experiment that concentrated on claims, title and abstract (run 1) proofed to be effective (0.0824). Surprisingly, the recall of this run was similar to that of the fourth experiment (0.4318) using the IPC additionally. This indicates that the classification codes do not have any advantage over the title, abstract and claims. In contrast, the run that additionally concentrated on the description (run 2) achieved a lower recall (0.3993). Following this, we could summarize that the detailed description seems to be more advantageous with respect to the precision of a patent retrieval system while the remaining patent sections do have a positive effect on the completeness of the search results. Some statistics according to the obtained results are provided in Table 1.

Table 1. Evaluation measures for the submitted runs

Run	Recall	MAP	P@5	P@10
Run 1	**0.4318**	0.0824	0.1028	0.0751
Run 2	0.3993	**0.0914**	**0.1170**	0.0833
Run 3	0.3993	**0.0914**	**0.1170**	0.0833
Run 4	**0.4318**	0.0824	0.1028	0.0751
Run 5	0.4303	0.0580	0.0717	0.0541
Run 6	0.1899	0.0208	0.0282	0.0209
Run 7	0.3694	0.0446	0.0562	0.0428

The results in Table 1 further indicate that using linguistic phrases did not increase the effectiveness of the retrieval system. Instead, the results of this experiment (0.1899; 0.0208) did significantly fall below the recall and map values of the remaining runs. This aspect is quite surprising, because phrases, in general, are considered to be more effective than terms [7]. One reason for the negative results of this experiment might be the small number of phrases that were extracted from the abstract. For example, only three linguistic phrases were extracted from topic EP-1226990 with the help of the extraction tool. The enrichment of phrases with additional terms (run 6) led to higher recall and map values. This could be a hint that our phrase queries were too short.

Furthermore, the abstract might not be the adequate patent section to construct phrase queries because of the existence of noisy terms [10]. As can be seen in the above table, integrating the terms from the detailed description (run 2 and 3) achieved the best results. Therefore, this patent section might be a better basis for generating phrase queries.

3.2 Influence of query length

Besides the retrieval effectiveness, the duration of the experiment and the query length were measured. Some statistics of these aspects are displayed in the following table.

Table 2. Run time and query length

Run	Run time in s	Min query length	Max query length
Run 1	6164	147	3896
Run 2	**38535**	531	21255
Run 3	13436	533	**21259**
Run 4	**1003**	149	3898
Run 5	2815	**1**	409
Run 6	3591	**1**	292
Run 7	3621	**1**	606

Table 2 illustrates that especially the query length differs significantly across the experiments. Having a maximum query length of about 21.250 terms, the longest queries were constructed in case of the second and third run. This result does not surprise much, because both experiments made use of the description, which normally is the longest section of a patent. In contrast, the shortest queries were generated using phrases extracted from the abstract and title terms (292.0). Although these queries contained very precise knowledge, the experiments utilizing the longer queries (run 2 and 3) achieved better retrieval results. In particular, the mean average precision of the patent retrieval system is influenced positively by very long queries. This fact is indicated by the results of the second experiment (map: 0.0914).

Although it seems to be quite effective, the use of very long queries has one disadvantage, because the search process is slowed down. As can be seen, the experiment which concentrated on the description (run 2) ran about eleven hours. In case of a realistic prior art search this might be problematic.

4 Outlook

At CLEF 2011, the University of Hildesheim and Chemnitz University of Technology did joint work. Our experiments concentrated on the following two different aspects:

1. The effect of linguistic phrases extracted from patent documents
2. The effect of very long queries with maximum number of terms

The results reveal that very long queries seem to be more effective in the context of patent information retrieval, but they significantly slow down the search process. In contrast, short queries which were constructed of phrases extracted from the abstract did not show any positive effect on neither recall nor mean average precision, but they

are advantageous with respect to the run time. This raises the question of how long a patent searcher is willing to wait for accurate search results.

In the future, we will have to think about a combined search strategy that takes into account terms as well as phrases, because both query types seem to have some advantages. Furthermore, we will have to improve the generation of phrase queries, because phrases extracted from the abstract did not improve the retrieval effectiveness. Using a different part of the patent, e.g. the detailed description, might show some improvements with respect to recall or map.

References

1. Piroi, F.: CLEF- IP 2011. Track Guidelines, 2011.
2. Graf,E; Azzopardi, L. (2008): A methodology for building a patent test collection for prior art search. In: Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA), S.60-71.
3. Information Retrieval Facility (2011): Documents in the CLEF-IP Corpus. http://www.ir-facility.org/collection (verified: 05.08.2011)
4. Kürsten, Jens; Wilhelm, Thomas; Eibl, Maximilian (2008): Extensible Retrieval and Evaluation Framework: Xtrieval. In: Baumeister, J.; Atzemüller, M. (2008): Proceedings of the LWA – Workshop FGIR, Würzburg, S.107-110.
5. Becks, D.; Womser-Hacker, C.; Mandl, T.; Kölle, R (2010): 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, Berlin et. al: Springer, S.491-496.
6. Becks, D.; Mandl, Th.; Womser-Hacker, Ch. (2010): Phrases or Terms? The Impact of different Query Types. In: Working Notes of 11th Workshop of the Cross-Language Evaluation Forum, CLEF 2010, Padua, Italy.
7. Vechtomova, O. (2006): Noun Phrases in Interactive Query Expansion and Document Ranking. In: Information Retrieval, 9(4), S.399-420.
8. Jaene, H.; Seelbach, D. (1975): Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Berlin, Köln, Frankfurt(Main): Beuth.
9. Becks, D.; Schulz, J. (2011): Domänenübergreifende Phrasenextraktion mithilfe einer lexikonunabhängigen Analysekomponente. In: Griesbaum, J.; Mandl, Th,; Womser-Hacker, Ch.(Hrsg.), Information und Wissen: global, sozial und frei?, Schriften zur Informationswissenschaft, Band 58, Boizenburg: Werner Hülsbusch, S.388-392.
10. Jochim, C.; Lioma, C.; Schütze, H. (2010): Preliminary Study into Query Translation for Patent Retrieval. In: Proceedings of the PaIR'10 Workshop, Toronto, Ontario, Canada, S.57-66.