Patent Retrieval Experiments in the Context of the CLEF IP Track 2009

Daniela Becks, Christa Womser-Hacker, Thomas Mandl, Ralph Kölle

Information Science, University of Hildesheim,
Marienburger Platz 22
D-31141 Hildesheim, Germany

{becks, womser, mandl, koelle}@uni-hildesheim.de

Abstract

At CLEF 2009 the University of Hildesheim focused on the main task of the Intellectual Property Track which aims at finding prior art for a specified patent [cf. Information Retrieval Facility 2009]. The experiments of the University of Hildesheim concentrated on a baseline approach including stopword elimination, stemming and simple term queries. Furthermore only title and claim were included into the index as especially the second one is considered to be the most important patent part during a prior art search [cf. Graf/Azzopardi 2008: 64].

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

Since there is a growing number of patent applications the importance of the patent domain is increasing steadily [cf. Chu et al. 2008: 9]. This tendency can be seen not just in industry but also in sciences. In particular in information science patent documents play a vital role as corpus material because they bear a lot of differences especially at the terminology level [cf. Graf/Azzopardi 2008: 64; Kando 2000].
In 2009 the *Cross Language Evaluation Forum* offers a special track – the so called *Intellectual Property Track* - concentrating on this special domain. The test collection consisted of about one million patent documents of the European Patent Office [cf. Information Retrieval Facility 2009]. Furthermore different topic sets were provided by the organizers. The experiments of the University of Hildesheim focussed on the smallest topic set (500 patent documents).
In general a patent document consists of the following sections [cf. Graf/Azzopardi 2008: 63f]:

- Bibliographic data (e.g. inventor)
- Disclosure (e.g. title and detailed description)
- Claims

A look at the patent documents of the collection as well as the topic set reveals a similar structure. Furthermore one can see that the description is the longest part of the document. It is followed by the claim section which in general consists of a number of claims [cf. e.g. Patent number EP-1114924-B1].
As we wanted to investigate whether simple statistical methods work well in such a special domain we adopted a baseline approach including stopword elimination, stemming and simple term queries. In the future, our goal is to compare this approach with a more sophisticated linguistic indexing and to apply a bag of terminological resources.

2 Indexing approach for patent documents

As already said before, the University of Hildesheim adopted a baseline approach. A first analysis of the test collection as well as the topic set revealed that in each patent document one could find a German, English and French title. The same went for the claims. That´s why we decided to perform different monolingual runs – for English and German - including the above mentioned parts of the patent document. It should be said that we didn´t concentrate on French title and claims because German and English can be find more frequently in the patent domain.
All of the experiments were done using a simple retrieval system. More information about it will be given in section 2.2. In particular syntax and patent specific terms caused some difficulties during the system setup. Some of these problems will be discussed in more detail in the next section (2.1).

2.1 Patent Terminology

Patent documents are as special as the whole domain. Many scientists have already figured out the differences between patents and other kinds of documents. A very important fact is described in Schamlu 1985. Here the author points out that there are special rules which a patentee has to follow. Because of these rules some sentence constructions return in each patent text [cf. Schamlu 1985: 63 ff]. This fact is confirmed by Ahmad and Al-Thubaity [cf. 2003: 48].
In the test collection as well as in the topic set this is the case for constructions like “as set forth in claim” [cf. e.g. Patent number EP-1114924-B1]. The same goes for words like *comprises* or *claim*. As these words appear frequently we decided to add them to the English stopword list. In German titles and claims the words *umfasst* or *Anspruch* return in most patents. These ones were included into the German stopword list, too. Besides these text patterns patent documents contain a lot of technical or general terms [cf. Graf/Azzopardi 2008: 64] which strongly affected the retrieval experiments.
In particular the huge amount of technical terms made the parsing process more difficult. Parsing errors occurred for example because of German terms like *AGR-System* [cf. Patent number EP-1114924-B1]. As can be seen this word contains a hyphen which had to be removed before. The same went for numbers which frequently appeared in the claims. Besides parsing difficulties caused by the technical language the rather general vocabulary influenced the retrieval process. As many patentees use vague and general expressions [cf. Graf/Azzopardi 2008: 64] like “Verfahren und Vorrichtung zur” [cf. e.g. Patent number EP-1117189-B1] many relevant documents are returned if only simple term queries are conducted.
An overview over the retrieval system used for the experiments of the University of Hildesheim will be given in the next section (2.2).

2.2 System Setup

The experiments of the University of Hildesheim were done using a simple retrieval system based on Apache Lucene[1]. This framework provides classes for stemming, indexing and searching. We followed the traditional retrieval process which is listed below.
As the collection consisted of patents in XML format a parser was necessary to first read out the content of the documents. We finally decided to integrate the SAX Parser[2] because using the DOM Parser seemed to take a lot of time. The major reason for this might be that patents are much longer than other document types [cf. Graf/Azzzopardi 2008: 64]. Following Iwayama et al. patent documents are even 24 times longer than newspaper articles [cf. 2003: 254].

[1] http://lucene.apache.org/java/
[2] http://www.saxproject.org/

- **Stopword removal and stemming**

Because we performed monolingual runs with English and German terms we integrated one specific stopword list[3] for each language. As described in section 2.1 we added some patent specific terms to these lists.
After having removed all the stopwords we employed a stemmer to the text. While the Lucene Standard Analyzer was chosen for English the German text was stemmed using the German Analyzer provided with Lucene.

- **Indexing**

To avoid having one huge file a separate index per language was created. Finally, a German as well as an English index existed and could be used as the basis of the actual search process. The following fields were included into the index file (cf. Table 1).

Index field	Part of patent
UCID	Patent number
CLAIM-TEXT	Claims (including all claims available in a patent)
INVENTION-TITLE	Title of the invention

Table 1: Index structure

- **Search process**

Prior art search is performed to determine whether an invention or part of it already existed [cf. Graf/Azzopardi 2008: 64]. Any document that states prior art is relevant to the query. In the context of the Intellectual Property Track a query is said to be a given patent.
After having parsed the topic files which existed in XML format stopwords were removed. Again only title and claims were considered for query formulation. Finally, we employed the same stemmer as during the indexing process. The remaining terms were used as simple term queries.

3 Results and Analysis

The experiments of the University of Hildesheim concentrated on the main task of the Intellectual Property Track 2009. Because of the above mentioned parsing problems as well as the time that was needed to adapt the system to the specialties of the domain we actually submitted only one run. Furthermore to investigate the influence of the single parameters some post runs were performed. A detailed description of the post runs can be found in section 3.2.

3.1 Submitted runs

The University of Hildesheim submitted one run within the main task. For this run we concentrated on German titles and claims only. The search process was based on the German index file. Equivalent to the described German approach a second run considering only English titles and claims was performed. We didn´t submit this run, but analyzed the results based on the provided relevance assessments.
Unfortunately the results of our Hildesheim_MethodeA_Main_S run as well as the equivalent run with English terms did not satisfy our expectations.
One problem might be the huge amount of results returned by the system. This caused that among the submitted result list a lot of relevant patents were missing. As we examined our results in more detail it was shown that some relevant documents have been actually found by the system, but didn't appear among the first 1000 patents of the ranking list. Furthermore the results revealed that the retrieval results seemed to be better if the topic was

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

of type “B1” meaning that the patent has already been granted [cf. Graf/Azzopardi 2008: 67]. For example in case of topic number EP1169314 nine patent documents were said to be relevant. The system of the University of Hildesheim returned seven of them.
Unfortunately the performance as a whole remained relatively bad. To find out if the results of the experiments can be improved in some way further runs were performed.

3.2 Post runs

As described before the results of the submitted run did not satisfy our expectations. During an intensive analysis we tried to figure out which approach could lead to the improvement of the system´s performance. Finally, the following post runs were performed. It should be said that these experiments had been run using a smaller topic set of the first 50 XML documents only.

- De_ohne_Snowball

  In this case we concentrated on German terms and removed stopwords. In contrast to the submitted run the Snowball Stemmer[4] was included instead of the German Analyzer.

- En_ohne_Snowball

  This post run is the English equivalent to the first one. In this case the Standard Analyzer has been replaced by the Snowball Stemmer.

- De_mit_Snowball

  In case of this run again German terms were used, but we didn´t remove stopwords.

With the help of the first two post runs we wanted to investigate whether the retrieval results can be influenced by the choice of analyzer. The experiments made clear that the performance of the retrieval system strongly depends on the included analyzer. Figure 1 shows the results for the German runs.

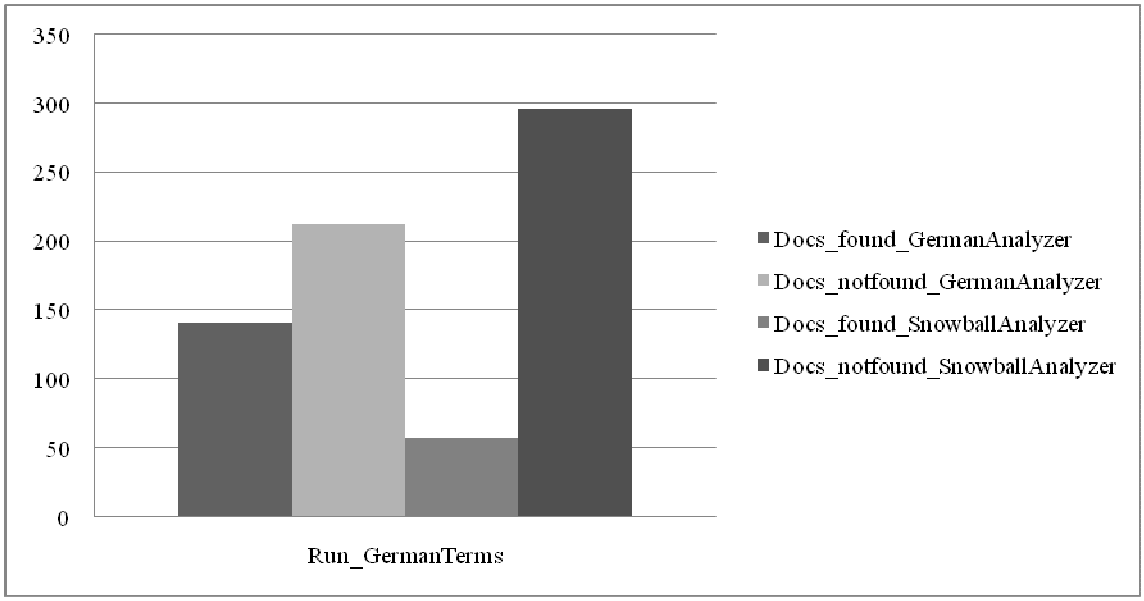

Figure 1: Relevant documents retrieved/ not retrieved in case of German runs using different analyzers (only 50 topics)

Figure 1 illustrates the number of relevant documents that were retrieved / not retrieved by our system. It should be said before that only the results for the first 50 topics of our submitted run were taken into account. This has been necessary because the post runs were performed with a smaller topic set of 50 documents. Otherwise we had not been able to compare the results of our submitted and our post runs.
As can be seen clearly in figure 1 the number of documents found decreased as the German Analyzer was replaced by the Snowball Analyzer. In case of our submitted run (left side) nearly 140 of 354 relevant patents

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

were found. Instead during the post run (right side) the system only returned about 50 relevant documents. To summarize the observations we have to state that in case of our German runs the Snowball Analyzer didn´t work well.
In contrast the results of the English run improved as the Snowball Analyzer had been included. This can be clearly seen in figure 2. Although the number of documents which were not found by the system is relatively high the number of relevant patents found obviously increased.
It should be said that these results only refer to a relatively small topic set of 50 XML documents. Maybe this would change if one would take into account the whole topic set. As a single run took us a lot of time we decided in a first attempt to use only part of the given topic set.

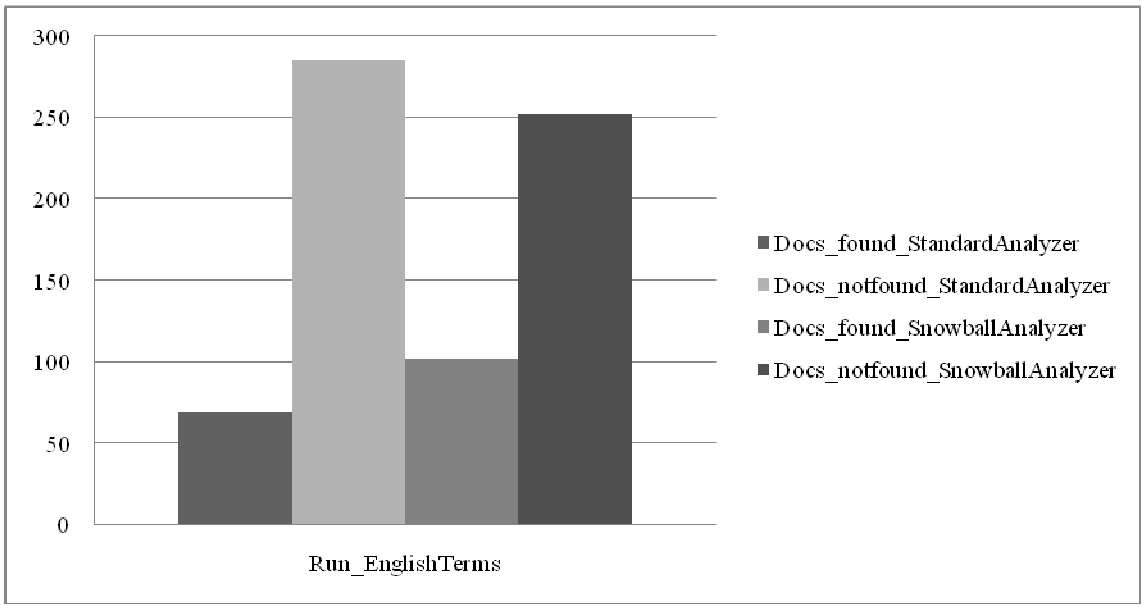

Figure 2: Relevant documents retrieved/ not retrieved in case of English runs using different analyzers (only 50 topics)

The first two post runs already revealed a relation between the implemented analyzer and the number of relevant documents returned by the retrieval system. Furthermore the influence of stopwords was investigated.
Our submitted run focussed on the basic retrieval approach which included stopword removal, but as the patent domain differs in many ways this might not be the best approach. That´s why a further post run without removing stopwords was performed. The results are illustrated in figure 3.

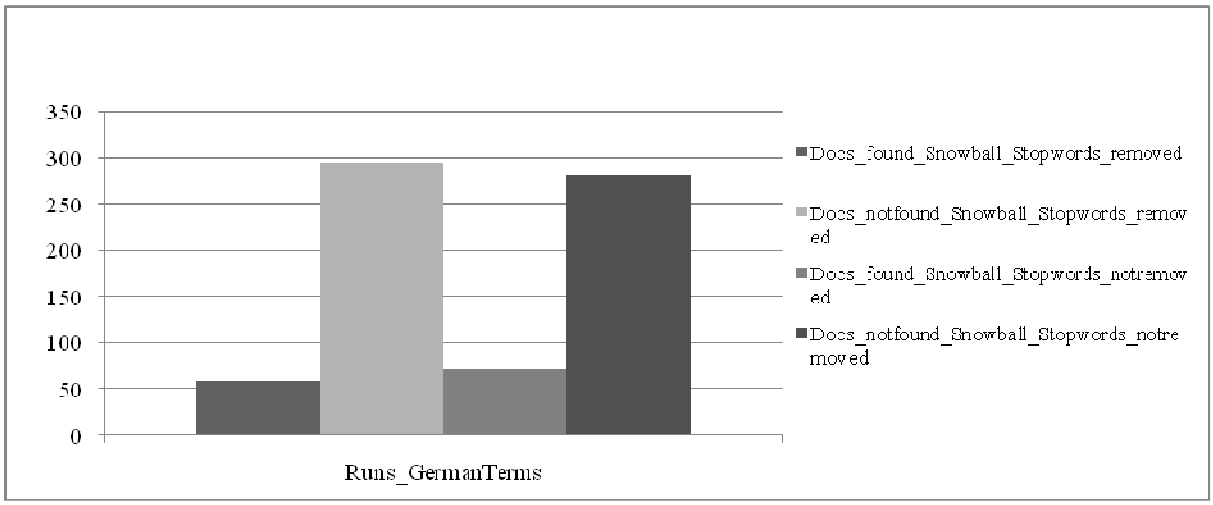

Figure 3: Relevant documents retrieved/ not retrieved in case of German runs with and without stopword removal (50 topics)

Figure 3 illustrates that there are little differences between the results of the run with and the one without stopwords. In both cases the Snowball Analyzer was used. If we have a look at the results in more detail we will realize that there is a little increase of the number of relevant documents found in case the stopwords haven´t been removed before. Even it is quite surprising there seems to be a positive relation between the existence of stopwords and the documents returned. This might be another hint that the patent domain needs to be treated in a special way.

4 Outlook

The patent domain is quite different from other domains. Especially the linguistic features like terminology and text structure bear a lot of difficulties. Because of these problems the University of Hildesheim only submitted one run with German terms. With the help of further post runs we were able to figure out that the analyzer used during the stemming process strongly influences the retrieval results. Furthermore there seems to be little evidence that in the patent domain stopwords should not be removed before stemming, but this needs further investigation.
If we have a look at the results as a whole they are not quite good. In the future we will have to adopt our retrieval system to the specialties of the patent domain. This includes the implementation of a better analyzer as well as a robust parser. We will also have to investigate whether simple term queries are sufficient in the area of patent retrieval.
Overall we need to implement a more sophisticated search strategy. The experiments of this year provide a good baseline for our further experiments.

References

[1]Ahmad, Khurshid; Al-Thubaity, AbdulMosen (2003): *Can Text Analysis Tell us Something about Technology Progress?*

[2]Chu, Aaron; Sakurai, Shigeyuki; Cardenas, F. Alfonso (2008): *Automatic Detection of Treatment Relationships for Patent Retrieval.* In: Proceedings of the PaIR 2008, Octobre 30, 2008, Napa Valley, USA, pp. 9-14.

[3]Graf, Erik; Azzopardi, Leif (2008): *A methodology for building a test collection for prior art search.* In: Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA), December 16, 2008, Tokyo, Japan, pp. 60-71.

[4]Information Retrieval Facility (2009): *CLEF–IP09 Track.* <http://www.ir-facility.org/the_irf/clef-ip09-track> (19.08.2009, 17:20)

[5]Iwayama, Makoto; Fujii, Atsushi; Kando, Noriko; Marukawa, Yuzo (2003): *An Empirical Study on Retrieval Models for Different Document Genres: Patents and Newspaper Articles.*In: Proceedings of the ACM SIGIR 2003, July 28 – August 1, 2003, Toronto, Canada, pp. 251-258.

[6]Kando, Noriko (2000): *What Shall We Evaluate? - Preliminary Discussion for the NTCIR Patent IR Challenge (PIC) Based on the Brainstorming with the Specialized Intermediaries in Patent Searching and Patent Attorneys*. In: ACM-SIGIR Workshop on Patent Retrieval, July 28, 2000, Athens, Greece, pp. 37-42.

[7]Schamlu, Mariam (1985): *Patentschriften – Patentwesen. Eine argumentationstheoretische Analyse der Textsorte Patent am Beispiel der Patentschriften zu Lehrmitteln.* Indicium-Verlag, München.