

# Prior art retrieval using the claims section as a bag of words

Suzan Verberne and Eva D'hondt  
Information Foraging Lab, Radboud University Nijmegen  
(s.verberne|e.dhondt)@let.ru.nl

## Abstract

We describe our participation in the 2009 CLEF-IP task, which was targeted at prior-art search for topic patent documents. Our system retrieved patent documents based on a standard bag-of-words approach for both the Main Task and the English Task. In both runs, we extracted the claim sections from all English patents in the corpus and saved them in the Lemur index format with the patent IDs as DOCIDs. These claims were then indexed using Lemur's BuildIndex function. In the topic documents we also focussed exclusively on the claims sections. These were extracted and converted to queries by removing stopwords and punctuation. We did not perform any term selection. We retrieved 100 patents per topic using Lemur's RetEval function, retrieval model TF-IDF. Compared to the other runs submitted for the track, we obtained good results in terms of nDCG (0.46) and moderate results in terms of MAP (0.054).

## Categories and Subject Descriptors

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

## General Terms

Indexing, Bag-of-Words, Queries

## Keywords

Prior Art Retrieval, CLEF-IP

## 1 Introduction

The CLEF-IP track was launched by the Information Retrieval Facility (IRF) in 2009 to investigate IR techniques for patent retrieval<sup>1</sup>. It is part of the CLEF 2009 evaluation campaign<sup>2</sup>.

The task of the track is: “to find patent documents that constitute prior art to a given patent”. The given patent in this task description serves as a topic for the retrieval task. The Main Task is to retrieve prior art for topic patents in any of the three following languages: English, French and German. Three facultative subtasks use parallel monolingual topics in one of these three languages.

---

<sup>1</sup>See [http://www.ir-facility.org/the\\_irf/clef-ip09-track](http://www.ir-facility.org/the_irf/clef-ip09-track)

<sup>2</sup>See <http://www.clef-campaign.org/>

## 2 Our methodology

### 2.1 Data selection

The CLEF-IP corpus consists of EPO documents with publication date between 1985 and 2000, covering English, French, and German patents (1,958,955 patent-documents pertaining to 1,022,388 patents, 75GB) [3]. The XML documents in the corpus do not correspond to one complete patent each but one patent can consist of multiple XML files (representing documents that were produced at different stages of a patent realization).

We decided to focus on the claims sections of the patents, because we found that many of the English patent documents did not contain abstracts. Moreover, we expected the claims section to be the most informative part of a patent.

In the CLEF-IP 2009 track the participating teams were provided with 4 different sets of topics (S,M,L,XL). We opted to do runs on the smallest set (the S data set) for both the Main and the English task. This set contained 500 topics. Because the information in these topics was different for both tasks (the topics for the Main Task contained the abstract content as well as the full information of the granted patent except for citation information, while the topic patents for the English Task only contained the title and claims elements of the granted patent [3]), we focussed only on the (English) claims sections from all topic patents.

### 2.2 Query formulation

There has been much research on the topic of query term extraction/query formulation [1]. However, we chose not to distil any query terms from the extracted claims section but took all words in the claims section as one long query (weighted in retrieval with TF-IDF). The reason for this was twofold. First, adding a term selection step in the retrieval process makes the retrieval process more prone to errors because it requires the development of a smart selection process. Second, by weighting the query and document terms using TF-IDF, a form of term selection is carried out in the retrieval and ranking process.

### 2.3 Indexing using Lemur

We extracted the claims sections from all English patents in the corpus. after we had removed all XML markup from the texts in a preprocessing script. Since a patent may consist of multiple XML documents, which correspond to the different stages of the patent realization process, one patent can contain more than one claims section. In the index file, we concatenated the claims sections pertaining to one patent ID into one document. We saved all patent claims in the Lemur index format with the patent IDs as DOCIDs. They were then indexed using the BuildIndex function of Lemur with the indri IndexType and a stop word list for general English<sup>3</sup>.

## 3 Results

We performed runs for the Main and English Task with the methodology described above. Since we used the same set-up for both runs, we obtained the same results. These results are in Table 1. The first row shows the results that are obtained if all relevant assignments are taken into consideration; the second row contains the results for the highly-relevant citations only [2].

## 4 Discussion

Although the results that we obtained with our ClaimsBOW approach may seem poor on first sight, they are not bad compared to the results that were obtained in runs by other participants. In terms of nDCG, our run performs well (ranked 6th of 70 runs); in terms of MAP our results

---

<sup>3</sup>This stop word list can be provided by the authors upon request.

Table 1: Results for the clefip-run ‘ClaimsBOW’ on the small topic set using English claims sections for both the Main Task and the English Task.

	P	P5	P10	P100	R	R5	R10	R100	MAP	nDCG
All	0.0129	0.0668	0.0494	0.0129	0.2201	0.0566	0.0815	0.2201	0.0540	0.4567
Highly-relevant	0.0080	0.0428	0.0314	0.0080	0.2479	0.0777	0.1074	0.2479	0.0646	0.4567

are moderate (ranked around 35th of 70 runs). The low performance achieved by almost all runs (except for the one submitted by Humboldt University) shows that the task at hand is a difficult one.

## References

- [1] Kazuya Konishi. Query Terms Extraction from Patent Document for Invalidity Search. In *Proceedings of NTCIR-5 Workshop Meeting*, pages 312–317, 2005.
- [2] Florina Piroi, Giovanna Roda, and Veronika Zenz. CLEF-IP 2009 Evaluation Summary. Technical report, Information Retrieval Facility, 2009.
- [3] Florina Piroi, Giovanna Roda, and Veronika Zenz. CLEF-IP 2009 Track Guidelines. Technical report, Information Retrieval Facility, 2009.