CLEF-IP 2010: Prior Art Retrieval using the different sections in patent documents

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

Abstract

In this paper we describe our participation in the 2010 CLEF-IP Prior Art Retrieval task where we examined the impact of information in different sections of patent documents, namely the title, abstract, claims, description and IPC-R sections, on the retrieval and re-ranking of patent documents. Using a standard bag-of-words approach in Lemur we found that the IPC-R sections are the most informative for patent retrieval. We then performed a re-ranking of the retrieved documents using a Logistic Regression Model, trained on the retrieved documents in the training set. We found indications that the information contained in the text sections of the patent document can contribute to a better ranking of the retrieved documents. The official results have shown that among the nine groups that participated in the Prior Art Retrieval task we achieved the eigth rank in terms of both Mean Average Precision (MAP) and Recall.

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

Retrieval, Reranking

Keywords

Prior Art Search, Patent retrieval, CLEF-IP track

1 Introduction

In the literature on patent retrieval there is some disagreement on which part of the patent document would be the most informative for (text-based) document retrieval. Graf and Azzopardi conclude that the claims section is the most useful [2], while patent searchers themselves hold that the description is more useful [1]. Interestingly, the results of last year's CLEF-IP track (2009) showed that the use of the metadata such as IPC-R codes or name of inventor leads to substantial improvements in patent retrieval over approaches that focussed only on the text sections. [3].

For our participation to the CLEF-IP 2010 track[1], our goal was to compare the impact of the different patent sections on retrieval performance in a fixed benchmark data set. In this paper we describe our contribution to the track in which we examine the influence of both the IPC-R metadata and the information contained in the different text sections of the patent document on retrieval performance and re-ranking.

[1] http://www.ir-facility.org/research/evaluation/clef-ip-10/overview

2 Data Description

The CLEF-IP 2010 test collection provided by the organisation committee contains a corpus of 2.6 million patent documents pertaining to 1.3 million patents[2], a set of 300 patent documents that serve as training topics together with their relevance assessments and a set of 500 test topic patent documents for testing. The patents can contain text in three different languages: English, French and German. They are labelled with XML tags to help identify the different sections as well as the different metadata such as IPC-R code, the name of the inventor or the date of the application. The different patent documents correspond to the different stages in the evolution of a patent and will therefore contain different amounts of information, for example, a patent application (A1 document) will not contain as much information as a fully granted patent (B1 document). The information in the older version of the patent is often subsumed by the newer document, but older versions may contain unique information as well. This year we have decided to retrieve patent documents rather than whole patents[3].

3 Experimental Set-up

3.1 Patent Section Extraction

Using a perl script we extracted the English title, abstract, claims and description sections and the IPC-R codes[4] from the original XML files and saved them as plain text in respective text files. If a document did not contain a section or if the section was not in English, no corresponding text file was created. The most important characteristics of the five subcorpora that were created in this manner are shown in 1.

Table 1: number of English files generated from the corpus and both query sets

	title	abstract	claims	description	IPC-R	# of original patent documents
corpus	2,669,501	1,004,436	1,210,507	993,910	2,679,232	2,680,604
training	196	292	196	196	300	300
topic	339	490	339	339	490	500

3.2 Retrieval Step

In the retrieval step, we wanted to determine which section of the patent document is the most informative for patent retrieval. To this end, we performed six retrievals on the corpus using the training queries. The retrieved documents of the best-scoring system were used to train the re-ranking models as will be described in section 2.4.

For the retrieval step, all the text files in the subcorpora were saved in the Lemur format: Using a bash script, the text in the text files was lowercased, punctuation was removed and the appropriate XML tags for indexing by Lemur were added . Then the texts were indexed using the BuildIndex function of Lemur with the `indri` IndexType and a stop list for general English.

In total we built 6 indices: Titles only, Abstracts only, Claims only, Description only, IPC-R codes, and full-text. By full-text we mean that we concatenated title, abstract, claims and description; sections that were not available in the patent document were added as an empty string. If none of these sections were available in English, the patent document was not indexed.

[2] Please note the difference between a patent and a patent document: a patent is not a physical document itself but a name for a group of patent documents that have the same patent ID number.

[3] A whole patent can be constructed by concatenating different patent versions into one document or by constructing a document from the most recent version of every section in the patent documents

[4] We used the full IPC-R code up to the level of the subgroups, e.g. A01J 5/01.

The topics in the training set were preprocessed in the same manner in order to be used as queries in Lemur. If the original query XML document did not contain a section, it was not added to the lemur query file.

For each query in the query file we retrieved 100 documents and ranked them according to the TF-IDF ranking model as implemented in Lemur. Table 2 shows the results of the retrievals on the 6 indices with their respective training queries. The results are given for Precision (P) and Recall (R) respectively at position 5, 10, 50 and 100 in the result list as well as the Mean Average Precision (MAP) score.

Table 2: BOW retrieval on training set.

	MAP	P	P5	P10	P50	R	R5	R10	R50
title	0.0995	0.0102	0.0602	0.0378	0.0143	0.0931	0.0319	0.0408	0.0674
abstract	**0.1139**	0.0129	0.0555	0.0432	0.0204	0.1262	0.0277	0.0419	0.0995
claims	0.0866	0.0088	0.0515	0.0375	0.0156	0.0938	0.0273	0.0373	0.0754
description	0.0646	0.0140	0.0428	0.0314	0.0296	0.1479	0.0777	0.1074	0.1479
full-text	0.0615	0.0138	0.0232	0.0149	0.0138	**0.1577**	0.0148	0.0149	0.0667
IPC-R	0.0677	**0.0156**	0.0373	0.0260	0.0187	0.1546	0.0196	0.0269	0.0887

The index with the IPC-R codes proved to be the most informative for patent retrieval in terms of Recall and Precision, although results are quite low for all six retrievals. Based on these results, we decided to proceed with only the retrieval results from the IPC-R subcorpus to the second step of our approach.

3.3 Re-ranking Step

It seems that in a retrieval task, conceptual information (as encoded in the IPC codes) works better than ‘surface’ textual information. However, we wanted to examine the influence of the different text sections on the positions of the retrieved results in the set.

We aimed to improve the ranking of the retrieved documents on the basis of the textual information present in the different sections of the patent document. As a predictor of relevance for the sections, we used the cosine similarity between corresponding sections of the topic and each of the retrieved documents.

We extracted this information as follows: For each topic–document pair from the training result set, we extracted the title, abstract, claims and description sections (if present) from both the topic and the retrieved document. We then calculated the cosine similarity between the sections of the respective documents using a python script which was based on the script by Dennis Muhlestein[5].

For each query–document pair we obtained a vector with 4 features: cosine similarity titles, cosine similarity abstracts, cosine similarity claims, and cosine similarity descriptions. In order to determine the importance of each of these features (and thereby each of the sections), we trained a Logistic Regression Model (LRM). The criterium variable was the relevance score of the retrieved document in the training relevance assessments.[6]

We used the lrm function from the Design package in R to train this model. We then used the LRM (trained on the training data) to predict an alternative ranking for the retrieved documents. We created two variants of the model: one with only these four features, and one in which the TF-IDF score for the retrieval with IPC-R codes was added as a fifth feature. We did not perform any step-wise model selection but rather combined all predictors at once.

[5] http://allmybrain.com/2007/10/19/similarity-of-texts-the-vector-space-model-with-python

[6] We only considered documents to be either ‘relevant’ or ‘non-relevant’ and did not adhere to the subdivision (‘relevant’ or ‘highly relevant’) made by the CLEF-IP organisers.

4 Results

In this section we present the results of our models in terms of MAP, Precision and Recall for both the training data (table 3) and the test data (table 4). The evaluation of the two re-ranking models on the training data was performed using 5-fold cross-validation. The P, R and MAP results are the averages over the five folds. Between the brackets is the standard deviation.

Table 3: Results on training set

	MAP	P	P5	P10	P50	R	R5	R10	R50
Baseline using IPC-R	0.0677	0.0156	0.0373	0.0260	0.0187	0.1546	0.0196	0.0269	0.0887
Re-ranking no TF-IDF	0.0858 (0.973)	0.0156 (0.0216)	0.0589 (0.1379)	0.0431 (0.0920)	0.0215 (0.0328)	0.1546 (0.2204)	0.0334 (0.0859)	0.0471 (0.1051)	0.111 (0.1782)
Re-ranking with TF-IDF	0.0870 (0.519)	0.0156 (0.0216)	0.0595 (0.1361)	0.0455 (0.0943)	0.0219 (0.0331)	0.1546 (0.2204)	0.0326 (0.0788)	0.0494 (0.1058)	0.114 (0.1814)

Table 4: Results on topic set

	MAP	P	P5	P10	P50	R	R5	R10	R50
run-1-small (no TF-IDF)	0.0274	0.0255	0.0786	0.0593	0.0314	0.1253	0.0228	0.0344	0.0857
run-2-small (with TF-IDF)	0.0291	0.0255	0.0826	0.0643	0.0331	0.1253	0.0240	0.0376	0.0902

The re-ranking model that incorporates the TF-IDF score of the retrieval set performs slightly better than the other model in both the training and the test results. In terms of Recall and Precision we performed slightly better than during our participation in the CLEF-IP 2009 track but compared to the other teams in this year's track we achieved low scores.

5 Discussion

In this section we will discuss (a) the retrieval results on the training set and (b) analyse the re-ranking models used.

One of our goals was to determine which section of the patent document is the most informative for patent retrieval in terms of recall and precision. The results in table 2 showed that for a bag-of-words approach the IPC-R codes in the patents were the most informative of all the patent sections. During our post-evaluation analysis we discovered that the low scores for the individual text sections are more likely an artefact of our data selection process rather than an adequate reflection of their performance in a retrieval task. Table 1 showed that there are considerable differences in size between the different text section corpora and thus in the number of patent documents that could be retrieved for a specific query. Moreover, we found evidence that some relevant patent documents were impossible to retrieve for certain queries. For example, if a relevant document for a query consisting of a claims section did not have a claims section itself, it did not feature in the claims subcorpus and could therefore not be retrieved. Consequently, we cannot draw a definite conclusion about the relative importance of the separate text sections for patent retrieval. The full-text corpus and the IPC-R corpus, however, did not suffer from these drawbacks. We found it interesting that the IPC-R outperformed the full-text retrieval, though the difference between the results is small. The major advantage of the IPC-R section is -predictably- the fact that it is language-independent, conceptual and has a limited 'vocabulary' of terms that can be used. For future work it would be interesting to examine the differences in retrieval results by using more general and more specific IPC codes as retrieval terms.

Our second goal was to examine the impact of the text sections on the re-ranking of retrieved documents: When we look at the results in table 3 and 4, it seems that the use of the information in the respective text sections of the query and retrieved document can lead to an improvement in the ranking of the relevant results. However, the high standard deviation values for the five folds show that our training set of 300 queries is too small to make any definite conclusions about the improvements made by the models. This may be a consequence of the fact that the models were not trained on optimal data but on rather poor retrieval results. Though they seem to boost the ranking of the retrieved documents, they contain enough noise to diminish the accuracy.

In order to evaluate the importance of the different text sections in the re-ranking of the retrieval results, we rank them in table 5 according to the coefficient that was assigned to them in the Logistic Regression Model. We find that all texts sections except for the description have a significant influence on the re-ranking of the retrieval results. The correlation analysis reported in table 6 shows a high correlation between the cosine similarity of the claims and description sections. Consequently, the coefficient for the claims section should be interpreted as being caused by the combination of the cosine similarities for the claims and description sections. Of all the text sections the abstracts have the most impact in the re-ranking process. This was to be expected as the abstracts are most likely to contain the keywords that are specific to the field of the invention.

Table 5: Impact of different text sections on re-ranking

Feature	Coefficient	p-value
Cosine similarity between abstracts	3.43541	0.00
Cosine similarity between claims	2.09992	0.001
Cosine similarity between titles	1.22437	0.00
TF-IDF value from retrieval data	0.01143	0.00
Cosine similarity between descriptions	-1.18692	0.022

Table 6: Correlation analysis of predictors

Pearon's r	Cos.sim. titles	Cos.sim. abstracts	Cos.sim. claims	Cos.sim. descriptions	TF-IDF value
Cos.sim. titles	1	0.151	0.368	0.336	0.066
Cos.sim. abstracts	0.151	1	0.202	0.250	-0.001
Cos.sim. claims	0.368	0.202	1	0.905	0.046
Cos.sim. descriptions	0.336	0.25	0.905	1	0.029
TF-IDF value	0.066	-0.002	0.046	0.029	1

6 Conclusion

In our contribution to the CLEF-IP 2010 Prior Art Retrieval task we examined the impact of different sections of patent documents on the retrieval and re-ranking of patent documents. Using a standard bag-of-words approach in Lemur we found that the IPC-R sections are more informative for patent retrieval than a full-text representation of the patent document. We then performed a re-ranking of the retrieved documents using a Logistic Regression Model, trained on the retrieved documents in the training set. Looking at the improved MAP scores, we found indications that the information contained in the separate text sections of the patent document can contribute to a better ranking of the retrieved documents.

References

[1] Eva D'hondt. Lexical issues of a syntactic approach to interactive patent retrieval. In *Proceedings of the 3rd BCSIRSG Symposium on Future Directions in Information Access*, 2009.

[2] Erik Graf and Leif Azzopardi. A methodology for building a patent test collection for prior art search. In *Proceedings of EVIA2008*, 2008.

[3] Patrice Lopez and Laurent Romary. Multiple retrieval models and regression models for prior art search. In *Proceedings of CLEF 2009*, 2009.