Automatically Generating Queries for Prior Art Search

Erik Graf, Leif Azzopardi, Keith van Rijsbergen
University of Glasgow
{graf,leif,keith}@dcs.gla.ac.uk

Abstract

This report outlines our participation in CLEF-IP's 2009 prior art search task. In the task's initial year our focus lay on the automatic generation of effective queries. To this aim we conducted a preliminary analysis of the distribution of terms common to topics and their relevant documents, with respect to term frequency and document frequency. Based on the results of this analysis we applied two methods to extract queries. Finally we tested the effectiveness of the generated queries on two state of the art retrieval models.

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.3.7 Digital LibrariesH.3.7 [Information storage and retrieval]: Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Patent Retrieval, Prior Art Search, Automatic Query Formulation

1 Introduction

The formulation of queries forms a crucial step in the workflow of many patent related retrieval tasks. This is specifically true within the process of Prior Art search, which forms the main task of the CLEF-IP 09 track. Performed both, by applicants and the examiners at patent offices, it is one of the most common search types in the patent domain, and a fundamental element of the patent system. The goal of such a search lies in determining the patentability (See Section B IV 1/1.1 in [3] for a more detailed coverage of this criterion in the European patent system) of an application by uncovering relevant material published prior to the filing date of the application. Such material may then be used to limit the scope of patentability or completely deny the inherent claim of novelty of an invention. As a consequence of the judicial and economic consequences linked to the obtained results, and the complex technical nature of the content, the formulation of prior art search queries requires extensive effort. The state of the art approach consists of laborious manual construction of queries, and commonly requires several days of work dedicated to the manual identification of effective keywords. The great amount of manual effort, in conjunction with the importance of such a search, forms a strong motivation for the exploration of techniques aimed at the automatic extraction of viable query terms. Throughout the remainder of these working notes we will provide details of the approach we have taken to address this challenge. In the subsequent section we will provide an overview of prior research related to the task of prior art search. Section 3 covers the details of our experimental setup. In section 4 we report on the official results and perform an analysis. Finally in the last section we provide a conclusion and future outlook.

2 Prior Research

The remainder of this section aims at providing an overview of prior research concerning retrieval tasks related to the CLEF-IP 09 task. As the majority of relevant retrieval research in the patent domain has been pioneered by the NTCIR series of evaluation workshops [1], additionally a brief overview of relevant collections and the associated tasks is provided. Further we review a variety of successful techniques applied by participating groups of relevant NTCIR tasks.

First introduced in the third NTCIR workshop [9], the patent task has led to the release of several patent test collections. Details of these collections are provided in Table 1. From the listing in Table 1 we can see that the utilized collections are comparative in size to the CLEF-IP 09 collection, and that the main differences consist of a more limited time period and a much smaller amount of topics specifically for the earlier collections.

Workshop	Document Type	Time Period	# of Docs.	# of Topics
NTCIR-3	Patent JPO(J)	1998-1999	697,262	31
	Abstracts(E/J)	1995-1999	ca. 1,7 million	31
NTCIR-4	Patent JPO(J), Abstracts(E)	1993-1997	1,700,000	103
NTCIR-5	Patent JPO(J), Abstracts(E)	1993-2002	3,496,252	1223
NTCIR-6	Patent USPTO(E)	1993-2002	1,315,470	3221

Table 1: Overview of NTCIR patent test collections (E=English, J=Japanese)

Based on these collections the NTCIR patent track has covered a variety of different tasks, ranging from cross-language and cross-genre retrieval (NTCIR 3 [9]) to patent classification (NTCIR 5 [7] and 6 [8]). A task related to the Prior Art search task is presented by the invalidity search run at NTCIR 4 [4],5 [5], and 6 [6]). Invalidity searches are exercised in order to render specific claims of a patent, or the complete patent itself, invalid by identifying relevant prior art published before the filing date of the patent in question. As such, this kind of search, that can be utilized as a means of defense upon being charged with infringement, is related to prior art search. Likewise the starting point of the task is given by a patent document, and a viable corpus may consist of a collection of patent documents. In course of the NTCIR evaluations, for each search topic (i.e. a claim), participants were required to submit a list of retrieved patents and passages associated with the topic. Relevant matter was defined as patents that can invalidate a topic claim by themselves (1), or in combination with other patents (2). In light of these similarities, the following listing provides a brief overview of techniques applied by participating groups of the invalidity task at NTCIR 4-6:

- **Claim Structure Based Techniques**: Since the underlying topic consisted of the text of a claim, the analysis of its structure has been one of the commonly applied techniques. More precisely the differentiation between premise and invention parts of a claim and the application of term weighting methods with respect to these parts has been shown to yield successful results.

- **Document Section Analysis Based Techniques**: Further one of the effectively applied assumptions has been, that certain sections of a patent document are more likely to contain useful query terms. For example it has been shown that from the 'detailed descriptions corresponding to the input claims, effective and concrete query terms can be extracted' NTCIR 4 [4].

- **Merged Passage and Document Scoring Based Techniques**: Further grounded on the comparatively long length of patent documents, the assumption was formed that the occurrence of query terms in close vicinity can be interpreted as a stronger indicator of relevance. Based on this insight, a technique based on merging passage and document scores has been successfully introduced.
- **Bibliographical Data Based Techniques**: Finally the usage of bibliographical data associated with a patent document has been applied both for filtering and re-ranking of retrieved documents. Particularly the usage of the hierarchical structure of the IPC classes and applicant identities have been shown to be extremely helpful. The NTCIR 5 proceedings [5] cover the effect of applying this technique in great detail and note that, 'by comparing the MAP values of Same' (where Same denotes the same IPC class) 'and Diff in either of Applicant or IPC, one can see that for each run the MAP for Same is significantly greater than the MAP for Diff. This suggests that to evaluate contributions of methods which do not use applicant and IPC information, the cases of Diff need to be further investigated.' [5]. The great effectiveness is illustrated by the fact that for the mandatory runs of NTCIR the best reported MAP score for 'Same' was 0,3342 MAP whereas the best score for 'Diff' was 0,916 MAP.

As stated before our experiments focused on devising a methodology for the identification of effective query terms. Therefore in this initial participation, we did not integrate the above mentioned techniques in our approach. In the following section the experimental setup and details of the applied query extraction process will be supplied.

3 Experimental Setup

The corpus of the CLEF-IP track consists of 1,9 million patent documents published by the European Patent Office (EPO). This corresponds to approximately 1 million individual patents filed between 1985 and 2000. As a consequence of the statutes of the EPO, the documents of the collection are written in English, French and German. While most of the early published patent documents are mono-lingual, most documents published after 2000 feature title, claim, and abstract sections in each of these three languages. The underlying document format is based on an innovative XML schema [1] developed at Matrixware[2].
Indexing of the collection took place utilizing the Indri[3] and Lemur retrieval system[4]. To this purpose the collection was wrapped in TREC format. The table below provides details of the created indices:

Index Name	Retrieval System	Stemming	Stop-Worded	UTF-8
Lem-Stop	Lemur	none	Stop-worded	No
Indri-Stop	Indri	none	Stop-worded	Yes

Table 2: Clef-IP 09 collection based indices

As can be seen from the table we did not apply any form of stemming on both indices. This decision was based on the fact that the corpus contains a large amount of technical terms (e.g. chemical formulas) and tri-lingual documents. In order to increase indexing efficiency, stop-wording based

[1] http://www.ir-facility.org/pdf/clef/patent-document.dtd
[2] http://www.matrixware.com/
[3] http://www.lemurproject.org/indri/
[4] http://www.lemurproject.org/lemur/

on the English language was applied to all indices. A minimalistic stop-word list was applied in order to mitigate potential side effects. The challenges associated with stop-wording in the patent domain are described in more detail by Blanchard [2]. No stop-wording for French and German was performed. The creation of the Indri-Stop index was made necessary in order to allow for experiments based on the filtering terms by language. Lemur based indices do not support UTF-8 encoding and therefore did not allow for filtering of German or French terms by use of constructed dictionaries.

3.1 Effective Query Term Identification

As stated before the main aim of our approach lies in the extraction of effective query terms from a given patent document. The underlying assumption of our subsequently described method is, that such terms can be extracted based on an analysis of the distribution of terms common to a patent application and its referenced prior art.

The task of query extraction therefore took place in two phases: In the first phase we contrasted the distribution of terms shared by source documents and referenced documents with the distribution of terms shared by randomly chosen patent document pairs. Based on these results the second phase consisted of the extraction of queries and their evaluation based on the available CLEF-IP 09 training data. In the following subsections both steps are discussed in more detail.

3.1.1 Analysing the common term distribution

The main aim of this phase lies in the identification of term related features whose distribution varies among source-reference pairs and randomly chosen pairs of patent documents. As stated before the underlying assumption is, that such variations can be utilized in order to extract query terms whose occurrences are characteristic for relevant document pairs. To this extent we evaluated the distribution of the following features:

1. The corpus wide term frequency (tf)
2. The corpus wide document frequency (df)

In order to uncover such variations the following procedure was applied: For a given number of n source-reference pairs an equal number of randomly chosen document pairs was generated. Secondly the terms common to document pairs in both groups were identified. Finally an analysis with respect to the above listed features was conducted.

As a result of this approach figure 1 depicts the number of common terms for source-reference pairs and randomly chosen pairs with respect to the corpus wide term frequency. In the graph, the x-axis denotes the collection wide term frequency, while on the y-axis the total number of occurrences of common terms with respect to this frequency is provided. Evident from the graph are several high-level distinctive variations: The first thing that can be observed is that the total number of shared terms of source-reference pairs is higher than for those of random pairs. Further the distribution of shared terms in random pairs, shown in blue, resembles a straight line on the log-log scale. Assuming that the distribution of terms in patent documents follows a Zipf like distribution this can be interpreted as an expected outcome. In contrast to this, the distribution of shared terms in source-reference pairs, depicted in red, varies significantly. This is most evident in the low frequency range of approximately 2-10000.

Given our initial goal of identifying characteristic differences in the distribution of terms shared within relevant pairs, this distinctive pattern can be utilized as a starting point of the query extraction process. Therefore, as will be evident in more detail in the subsequent section, we based our query extraction process on this observation.

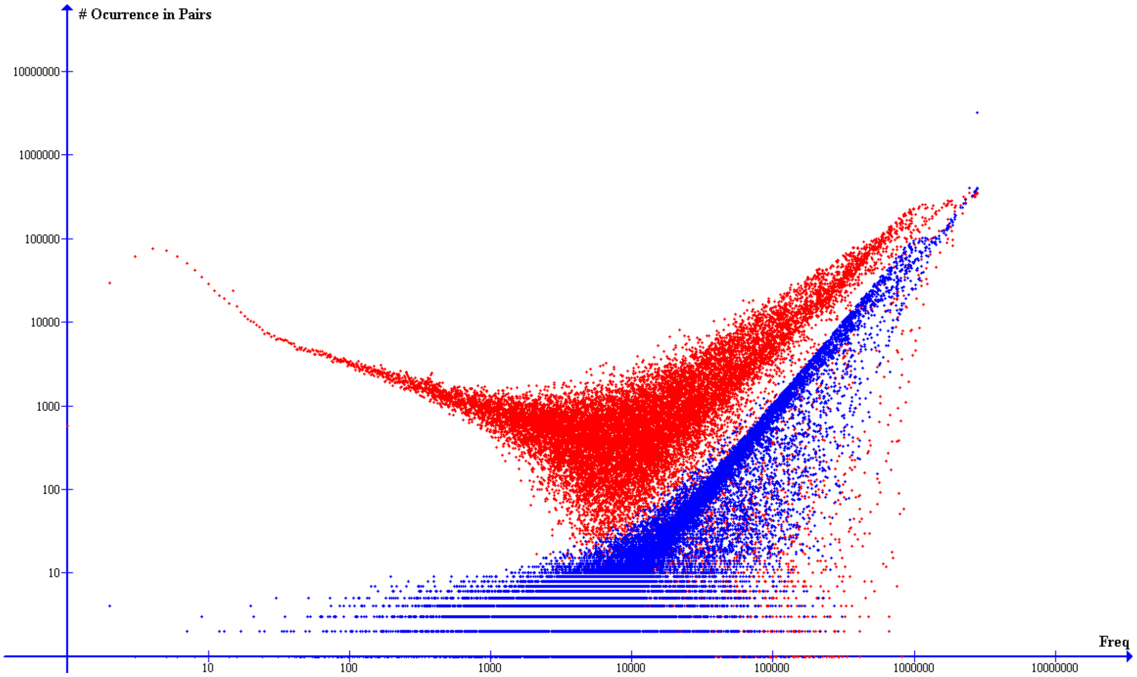


Figure 1: Distribution of shared terms: Source-Reference pairs versus Random Pairs

3.2 Query extraction

Based on the characteristic variations of the distribution of terms common to source-reference pairs our query term extraction process uses the document frequency as selection criterion. The applied process hereby consisted of two main steps. Based on the identification of the very low document frequency range as most characteristic for source-reference pairs, we created sets of queries with respect to the df of terms (1) for each training topic. These queries were then submitted to a retrieval model, and their performance was evaluated by use of the available training relevance assessments (2).

Following this approach two series of potential queries were created via the introduction of two thresholds.

- **Document Frequency (df) Based Percentage Threshold**: Based on this threshold, queries are generated by including only terms whose df lies below an incrementally increased limit. To allow for easier interpretation the incrementally increased limit is expressed as $\frac{df}{N} * 100$, where N denotes the total number of documents in the collection. A percentage threshold of 0.5% therefore denotes, that we include only terms in the query that appear in less than 0.5% of the documents in the collection.

- **Query Length Threshold**: A second set of queries was created by utilization of an incrementally increased query length as underlying threshold. In this case for a given maximum query length n, a query was generated by including the n terms with the lowest document frequency present in the topic document. The introduction of this threshold was triggered by the observation that the amount of term occurrences with very low df varies significantly for the topic documents. As a consequence of this a low df threshold of 1000 can yield a lot of query terms for some topics, and in the extreme case no query terms for other topics.

We generated queries based on a percentage threshold ranging from 0.25%-3% with an increment of 0.25, and with respect to query lengths ranging from 10-300 with an increment of 10. The performance of both query sets was then evaluated by utilization of the large training set of the main task with the BM25 and Cosine retrieval models.

Figure 2 depicts the MAP and Recall scores based on a series of df-threshold based queries using the BM25 retrieval model. Scores for the Cosine model based on varying query length are shown in Figure 3.

The first thing we observed from these training topic runs, is that the applied methods of query formulation return promising results for both retrieval models, and that this is the case for both query extraction strategies. Secondly, BM25 always exhibited a higher performance with respect to both MAP and the number of retrieved documents. The higher slope of the graph showing the performance of the cosine retrieval model is not mainly induced by the properties of the model itself, but rather through the length of the applied queries. The average query length for a percentage threshold of 0.25 (the first data point) for example was 198.276. By applying lower df thresholds, which would result in considerably shorter queries, a similar graph can be witnessed for the performance of BM25. During our training phase the percentage-threshold method showed slightly better results. We believe that a possible explanation may consist of an increased potential for topic drift that can be introduced by allowing for the inclusion of terms with higher df for large query length thresholds.

4 Results and Analysis

In the following a description of the submitted runs and an analysis of their performance will be conducted. In total our group submitted five runs. While the performance of one of our runs is in line the with the observations based on the training set, the performance of the other four runs resulted in a completely different and order of magnitudes lower results. Unfortunately this was induced by a bug occurring in the particular retrieval setup utilized for their creation. The amount of analysis that can be drawn from the official results is therefore very limited. After obtaining the official qrels we re-evaluated the baseline run of these four runs in order to verify the observed tendencies of the training phase.

4.1 Description of Submitted Runs and Results

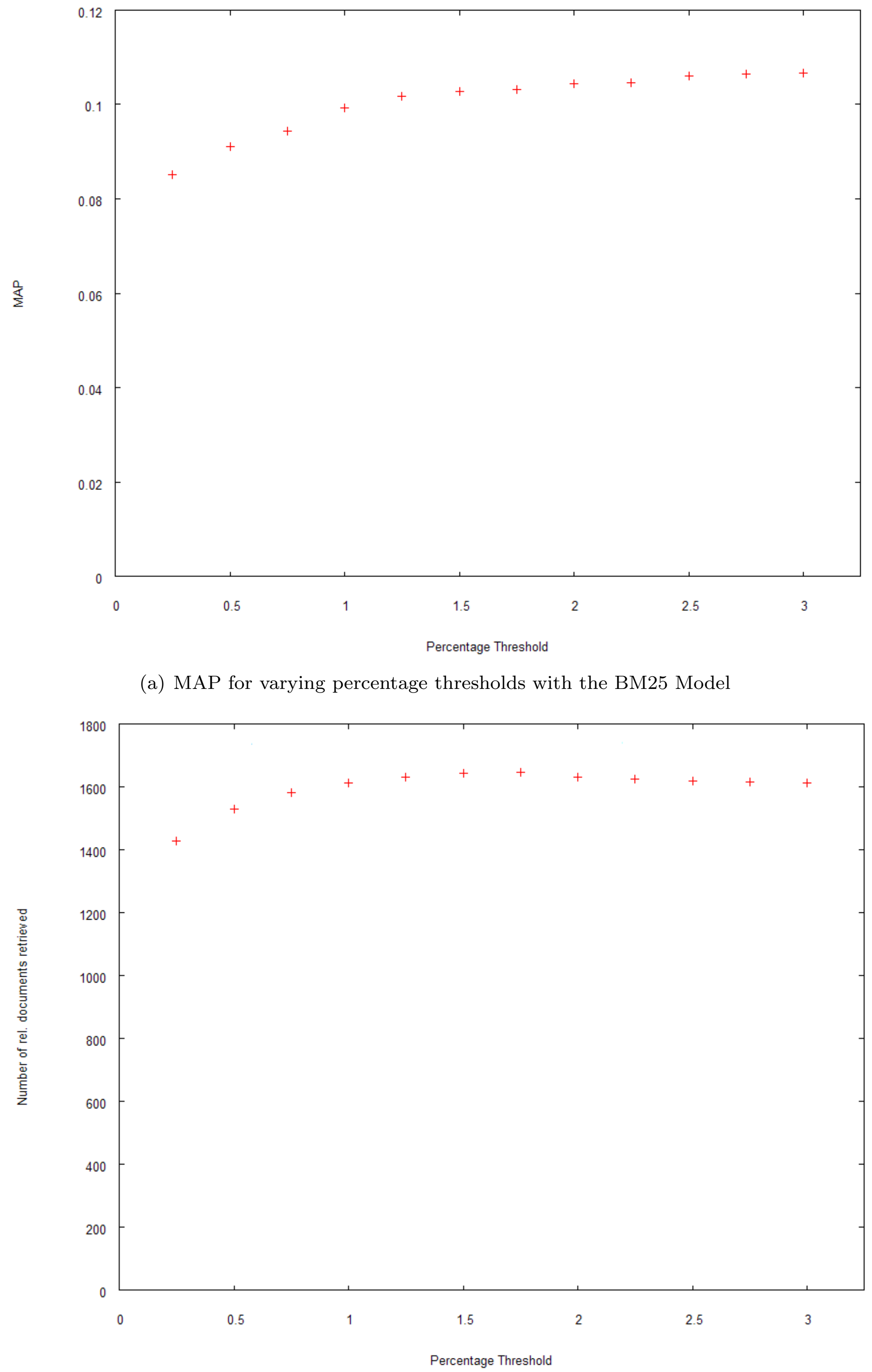
We participated in the Main task of this track with four runs for the Medium set that contains 1,000 topics in different languages. All four runs where based on the BM25 retrieval model using standard parameter values (b = 0.75, k1 = 1.2, k3 =1000), and utilized a percentage threshold of 3.0. These runs are listed below:

- BM25medStandard: No filtering of query terms by language was applied. Query terms where selected solely considering their df.
- BM25EnglishTerms: German and French terms were filtered out.
- BM25FrenchTerms: English and German terms were filtered out.
- BM25GermanTerms: English and French terms were filtered out.

Additionally we submitted a run for the XL set consisting of 10000 topics. This run also utilized a threshold of 3.0, used the Cosine retrieval model, and filtered out French and German terms via the utilization of dictionaries that were constructed based on the documents in the Clef-IP 09 corpus. Table 4 lists the official results of the above described runs.

4.2 Analysis

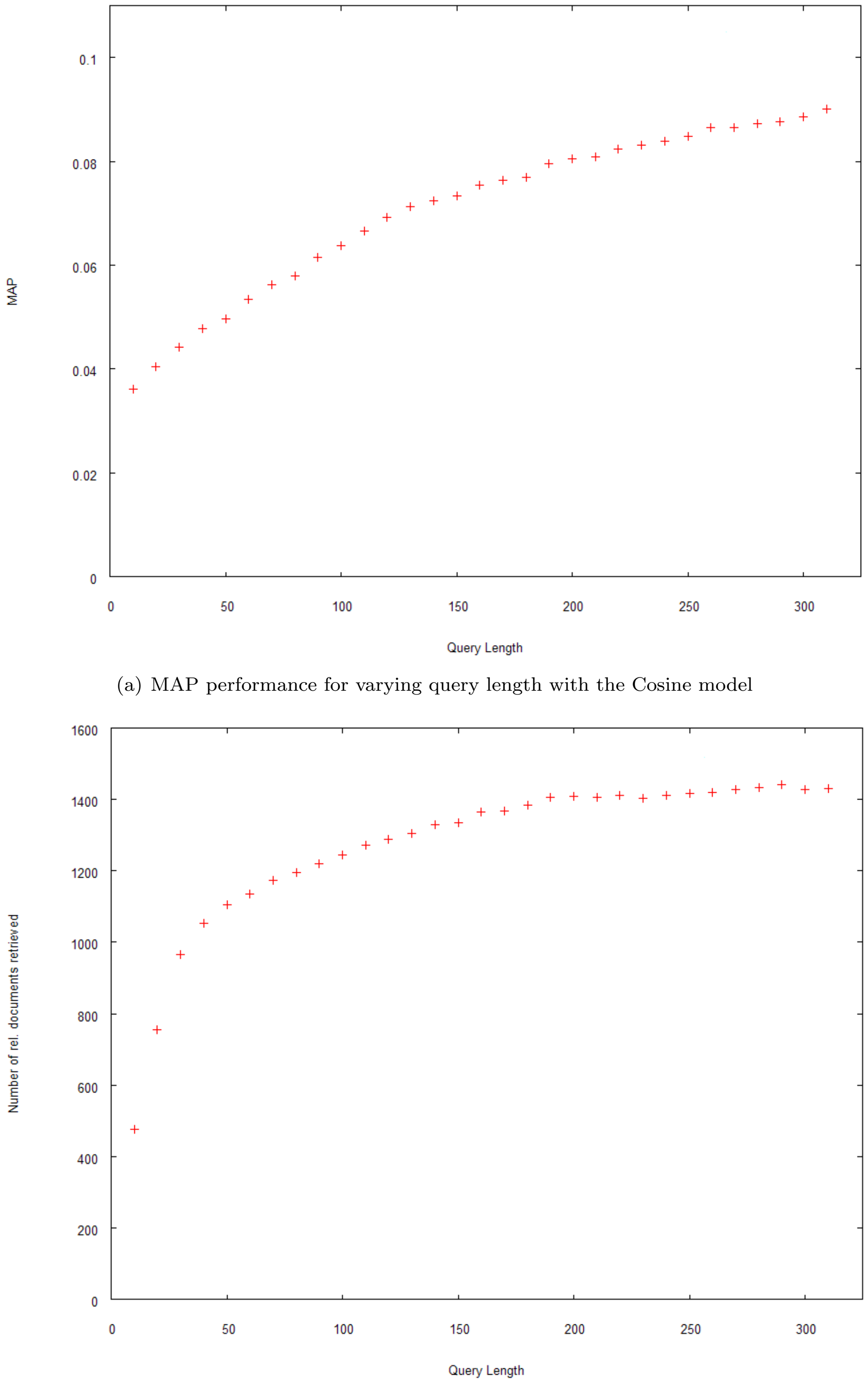
While the CosEnglishTerms run showed comparable performance to the observations during the training phase outlined in Figure 3, it can be seen from the results that the performance of the BM25 based runs was significantly lower than the observed results in Figure 2 . Therefore first



(a) MAP for varying percentage thresholds with the BM25 Model

(b) Number of retrieved rel. documents for varying percentage thresholds with the BM25 Model

Figure 2: Results for varying percentage thresholds for the BM25 model



(a) MAP performance for varying query length with the Cosine model

(b) Number of retrieved rel. documents for varying query length with the Cosine model

Figure 3: Results for varying query length for the cosine model

run id	P	P5	P10	P100	R	R5	R10	R100	MAP	nDCG
BM25medstandard	0.0002	0.0000	0.0001	0.0002	0.0238	0.0000	0.0001	0.0033	0.0002	0.0389
BM25EnglishTerms	0.0001	0.0002	0.0001	0.0001	0.0159	0.0002	0.0002	0.0017	0.0002	0.0318
BM25FrenchTerms	0.0001	0.0004	0.0002	0.0002	0.0123	0.0004	0.0004	0.0027	0.0003	0.0270
BM25GermanTerms	0.0001	0.0002	0.0001	0.0001	0.0159	0.0002	0.0002	0.0017	0.0002	0.0318
CosEnglishTerms	0.0036	0.0854	0.0600	0.0155	0.4667	0.0808	0.1100	0.2599	0.0767	0.4150

Table 3: Official Run Results

of all, it is not possible for us to draw any conclusions towards the effect of the applied filtering by language from these results. In retrospective analysis we identified that this almost complete failure in terms of performance was linked to applying the BM25 model to the Indri indices created to allow for language filtering. While this problem has not yet been resolved and we were therefore not able to re-evaluate the language filtering based runs, we re-evaluated the BM25medstandard run using a Lemur based index and the released official qrels. This resulted in the below listed performance, that is in the same range of what we witnessed during the BM25 training phase. It confirms our observation that BM25 seems to be more effective than the Cosine model.

run id	P5	P10	P100	R	MAP
BM25medstandard	0.1248	0.0836	0.0188	0.511	0.1064

Table 4: Re-evaluated run result

5 Conclusion and Future Outlook

Based on one of the submitted runs and our training results this preliminary set of experiments has shown that our proposed method of automatic query formulation may be interpreted as a promising start towards effective automatic query formulation. As such a technique may significantly facilitate the process of prior art search through the automatic suggestion of efficient keywords, it is planned to extend our experimentation in several directions. These extensions include the consideration of a patent document's structure (i.e. title, description, claims) in the selection process, and the introduction of a mechanism that will allow the weighted inclusion of term related features in addition to the document frequency.

6 Acknowledgements

We would like to express our gratitude for being allowed to use the Large Data Collider (LDC) computing resource provided and supported by the Information Retrieval Facility[5]. We specifically also want to thank Matrixware[6] for having co-funded this research.

References

[1] National institue of informatics test collection for ir systems (ntcir). http://research.nii.ac.jp/ntcir/.

[2] Antoine Blanchard. Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information*, 29(4):308 – 316, 2007.

[3] European Patent Office (EPO). *Guidelines for Examination in the European Patent Office*, December 2007.

[5] http://www.ir-facility.org/

[6] http://www.matrixware.com/

[4] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of patent retrieval task at ntcir-4. In *Proceedings of NTCIR-4 Workshop Meeting*, 2004.

[5] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of patent retrieval task at ntcir-5. In *Proceedings of NTCIR-5 Workshop Meeting*, 2005.

[6] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of the patent retrieval task at the ntcir-6 workshop. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 359–365, 2007.

[7] Makoto Iwayama, Atsushi Fujii, and Noriko Kando. Overview of classification subtask at ntcir-5 patent retrieval task. In *Proceedings of NTCIR-5 Workshop Meeting*, 2005.

[8] Makoto Iwayama, Atsushi Fujii, and Noriko Kando. Overview of classification subtask at ntcir-6 patent retrieval task. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 366–372, 2007.

[9] Makoto Iwayama, Atsushi Fujii, Noriko Kando, and Akihiko Takano. Overview of patent retrieval task at ntcir-3. In *Proceedings of NTCIR-3 Workshop Meeting*, 2002.