On Term Selection Techniques for Patent Prior Art Search

ABSTRACT

In this paper, we investigate the influence of term selection on retrieval performance on the CLEF-IP prior Art test collection, using the Description section of the patent query with Language Model (LM) and BM25 scoring functions. We find that an oracular relevance feedback system that extracts terms from the judged relevant documents far outperforms the baseline and performs twice as well on MAP as the best competitor in CLEF-IP 2010. We find a very clear term selection value threshold for use when choosing terms. We also noticed that most of the useful feedback terms are actually present in the original query and hypothesized that the baseline system could be substantially improved by removing negative query terms. We tried four simple automated approaches to identify negative terms for query reduction but we were unable to notably improve on the baseline performance with any of them. However, we show that a simple, minimal interactive relevance feedback approach where terms are selected from only the *first* retrieved relevant document outperforms the best result from CLEF-IP 2010 suggesting the promise of interactive methods for term selection in patent prior art search.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval Query Formulation

Keywords: Patent search, Query Reformulation.

1. INTRODUCTION

Patent prior art search involves finding previously granted patents, or any published work, such as scientific articles or product descriptions that may be relevant to a new patent application. The objective and challenges of standard formulations of patent prior art search are different from those of standard text and web search since [8]: (i) queries are reference patent applications, which consist of documents with hundreds or thousands of words organized into several sections, while typical queries in text and web search constitute only a few words; and (ii) patent prior art search is a recall-oriented task, where the primary focus is to retrieve all relevant documents at early ranks, in contrast to text and web search that are precision-oriented, where the primary goal is to retrieve a subset of documents that best satisfy the query intent. Another important characteristic of patent prior art search is that, in contrast to scientific and technical writers, patent writers tend to generalize and maximize the scope of what is protected by a patent and potentially discourage further innovation by third parties, which further complicates the task of formulating queries.

In this work, we focus on the task of query reformulation [1] specifically applied to patent prior art search [10, 13, 17]. While prior work has largely focused on specific techniques for query reformulation, in Section 3, we first build an oracular query formed from known relevance judgments for the CLEP-IP 2010 Prior Art test collection [13] in an attempt to derive an upper bound on performance of standard Okapi BM25 and Language Models (LM) retrieval algorithms for this task. Since the results of this evaluation suggest that query reduction methods can outperform state-of-the-art prior art search performance, in Section 4.1 we proceed to analyze four simple automated methods for identifying terms to remove from the original patent query. Finding that none of these methods seems to independently yield promise for query reduction that strongly outperforms the baseline, in Section 4.2 we evaluate an alternative interactive feedback approach where terms are selected from only the first retrieved relevant document. Observing that such simple interactive methods for query reduction with a standard LM retrieval model outperform highly engineered patent-specific search systems from CLEF-IP 2010, we conclude that interactive methods offer a promising avenue for simple but highly effective term selection in patent prior art search.

2. BASELINE IR FRAMEWORK

We developed a baseline IR system for patent prior art search on the top of the Lucene search engine[1], which processes queries using both BM25 [15] and LM (Dirichlet smoothing, and Jelinek-Mercer smoothing) [18] scoring functions. We used Lucene to index the English subset of CLEF-IP 2010 dataset[2] that contains 2.6 million patent documents and 1303 topics (queries) for the English test set. We used the default Lucene settings using the Porter stemming algorithm [14] and English stop-word removal. We also removed patent-specific stop-words as described in [8]. In our implementation, each section of a patent (title, abstract, claims,

SIGIR '15, August 9-13, 2015, Santiago, Chile

[1] `http://lucene.apache.org/`

[2] `http://www.ifs.tuwien.ac.at/~clef-ip/`

and description) is indexed in a separate field. However, when a query is processed, all indexed fields are targeted with an equal weight, since this generally offers best retrieval performance. We also used the International Patent Classification (IPC) codes assigned to the topics to filter the search results by constraining them to have common IPC codes with the patent topic as suggested in previous works [7]. Although this IPC code filter may prevent retrieval of relevant patents, we have chosen to keep it for the following reasons: (i) more than 80% of the patent queries share an IPC code with their associated relevant patents, and (ii) it makes the retrieval process much faster. The accuracy of the results is evaluated using two popular metrics — Mean Average Precision (MAP) and Average Recall — on the top-100 results for each query, assuming that patent examiners are willing to assess the top 100 patents [5]. We achieved the best performance while querying with the Description section as in previous work [17] and using either the LM or the BM25 scoring functions. We call this initial query the *Patent Query* and use it as our main baseline.

In addition, we compare our results to *PATATRAS*, a highly engineered system developed by Lopez and Romary [7], which achieved the best performance in the CLEF-IP 2010 competition. This system uses multiple retrieval models (especially Kullback-Leibler divergence [1] and Okapi BM25) and exploits patent metadata and citation structures. While our evaluation excludes 22 of the 1303 topics for which no relevant English documents were available, the difference in MAP score between our evaluation and the full 1303 topic evaluation of PATATRAS is negligible.

3. ORACULAR TERM SELECTION

In this section we develop an *Oracular Query* to understand (a) the adequacy of the baseline *Patent Query*, (b) an upper bound on performance of the BM25 and LM models, and (c) the sufficiency of terms in the reference patent query.

3.1 Oracular Query Formulation

We begin by defining an oracular relevance feedback system, which extracts terms from the judged relevant documents. To this end, after an initial run of a given query, we calculate a Relevance Feedback (RF) score for each term in the top-100 retrieved documents as follows:

$$RF(t,Q) = Rel(t) - Irr(t) \qquad (1)$$
$$t \in \{\text{terms in top-100 retrieved documents}\}$$

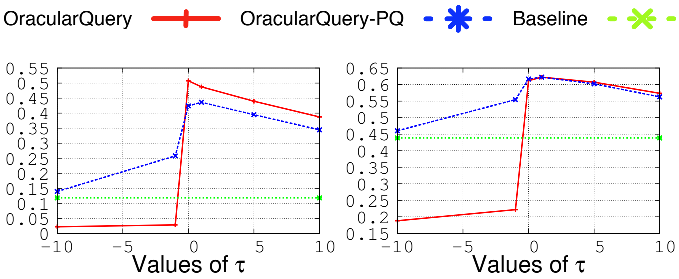
where $Rel(t)$ is the average term frequency in retrieved relevant patents and $Irr(t)$ is the average term frequency in retrieved irrelevant patents. We assume that words with a positive score are *useful words* since they are more frequent in relevant patents, while words with negative score are *noisy words* as they appear more frequently in irrelevant patents. We empirically seek to evaluate the threshold τ on $RF(t,Q)$ (defined below) yielding the best oracular query.

We formulate two oracular queries. The first query is formulated by selecting terms in the top-100 documents:

$$Oracular\ Query = \{t \in top - 100 | RF(t,Q) > \tau\} \qquad (2)$$

We formulate the second query by selecting terms that also occur in the reference patent query as follows:

$$Oracular\ Patent\ Query = \{t \in Q | RF(t,Q) > \tau\} \qquad (3)$$



(a) Mean Average Precision. (b) Recall.

Figure 1: System performance vs. the threshold τ for oracular query and oracular patent query.

Table 1: Performance for the *Patent Query*, two variants of the *Oracular Query*, and *Top CLEF-IP 2010 (PATATRAS)*.

		Baseline	PATATRAS	Oracular $\tau = 0$	Oracular(PQ) $\tau = 1$
LM	MAP	0.118	0.27	0.507	0.436
	Recall	0.438	N/A	0.612	0.622
BM25	MAP	0.129	0.27	0.518	0.446
	Recall	0.454	N/A	0.615	0.629

3.2 Baseline vs. Oracular Query

First, we investigate the ideal threshold setting τ. Figure 1 illustrates that $\tau = 0$ is the best-performing value for *Oracular Query* while $\tau = 1$ is the best for *Oracular Patent Query*. In general, the MAP and the recall for the *Oracular Patent Query* are lower than the MAP and the recall for the *Oracular Query* respectively; nonetheless the terms selected from the reference patent query itself are still sufficient to achieve MAP performance significantly better than the PATATRAS system. In addition, we remark on the rather unexpected steep drop-off in performance when the oracular query includes slightly noisy terms (i.e., τ just slightly less than 0) as defined previously.

In Table 1, we compare our two oracular relevance queries with both the baseline *Patent Query* and the PATATRAS system. Here we see that the *Oracular Query* using $\tau = 0$ far outperforms the baseline and approximately performs twice as well on MAP as the PATATRAS system, the best competitor in CLEF-IP 2010. In general we found BM25 and LM to offer very similar performance. Our subsequent results use only LM due to space limitations although results for BM25 are very similar.

Overall, our experiments related to oracular relevance feedback system suggest two important conclusions: (1) query reduction should suffice for effective prior art patent retrieval; and (2) very precise methods for eliminating poor query terms in the reduction process are needed.

4. QUERY REDUCTION: APPROXIMATING THE ORACULAR QUERY

The gain achieved using the Oracular Patent Query method motivates us to explore various methods to approximate the terms selected by this query without "peeking at the answers" provided by the actual relevance judgements. We first attempt this via fully automated methods and then proceed to evaluate semi-automated methods based on interactive relevance feedback methods.

4.1 Automated Reduction

We use the following four simple approaches to reduce the initial patent queries:

(1) In standard IR approaches, removing terms appearing highly frequently across documents in the collection can improve retrieval effectiveness. Inspired by this fact, after an initial run of the query, we removed terms with a high average document frequency (DF) over the top-100 documents ($DF(t) > \tau$). As illustrated in Figure 2, such pruning significantly hurts performance.

(2) Frequent terms inside long and verbose queries have been shown to be important [12]. Hence, we only keep high query TF (QTF) terms ($QTF(t) > \tau$) but remove from this set any document frequent terms ($DF(t) > 0.01$). Figure 2 indicates that this approach performs slightly better than the baseline.

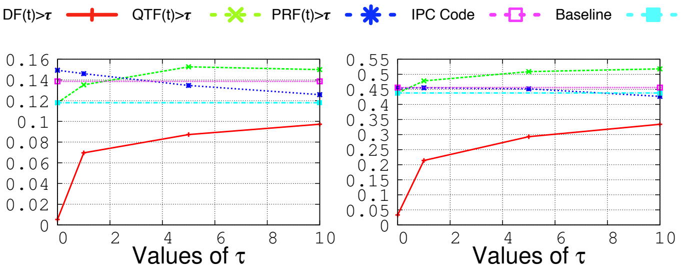
(3) The third query reduction approach is to select query terms using pseudo-relevance feedback (PRF) [1, 12]. We calculated a PRF score similar to RF score assuming that the top-k ranked documents are relevant. We selected the query terms which have high PRF score ($PRF(t) > \tau$). As Figure 2 illustrates, this approach does not notably outperform the baseline.

(4) Finally, we used words in IPC code title of each patent query to reduce the query, based on the assumption they are common to all patents, which belong to the same category and may be considered as stop-words. In our experiment, we removed the IPC title terms from a selection of frequent query terms ($QTF(t) > 5$). We can see in Figure 2 that the results drop slightly compared to approach (2), where $\tau = 5$.

Figure 3 shows an anecdotal example for a sample query about an invention related to "emulsifier" to help explain why these four approaches fail. It shows the raw abstract of the invention, and terms and their associated RF scores for each approach. In Figure 3 terms are chosen for each approach as follows: $\{t|DF(t)/QTF(t)/PRF(t) > 10\}$. IPC title terms are the words appearing in the IPC title and do not have any score. It can be seen that the four methods fail clearly to discriminate between useful and noisy terms. As one example, important stemmed terms like "enzym" and "starch" have been removed by DF pruning in (1), which hurts query quality. As another example, retaining IPC code title terms yields more noisy terms than useful terms (19 out of 32, and few of them with a very negative score like "amylos" or "saccharid"). Overall, while some methods like PRF work better than others for query reduction, all methods may retain highly negative terms and results from Section 3.2 showed that the inclusion of even slightly negative terms can significantly hurt performance.

4.2 Semi-automated Interactive Reduction

Our sample analysis of specific queries and terms selected via our oracular approach suggests that automated methods fall far short of optimal term selection. This leads us to explore another approach of approximating the oracular query derived from relevance judgements by using a subset of relevance judgements through interactive methods. Specifically, to minimize the need for user interaction, in this section we analyse the performance of an oracular query derived from only the first relevant document identified in the search re-



(a) Mean Average Precision.

(b) Recall.

Figure 2: System performance vs. the threshold τ for four query reduction approaches.

```
PAC-1293
Abstract: The invention relates to an emulsifier,
a method for preparing said emulsifier, and to
its use in various applications, primarily food
and cosmetic applications. The invention also
relates to the use of said emulsifier for the
creation of an elastic, gelled foam. An
emulsifier according to the invention is based on
a starch which is enzymatically converted, using
a specific type of enzyme, and modified in a
specific esterification reaction.

DF Terms: starch:14.64, enzym:29.49, amylos:-20.15,
oil:8.63, dispers:-8.66, ph:-4.55, dry:-6.21, heat:-2.26,
product:-5.48, slurri:-11.48, viscos:7.77, composit:-4.49,
reaction:-1.97, food:-11.94, agent:5.19, debranch:-10.58,
reduc:-6.37, fat:-12.83, prepar:-0.82, hour:-5.42,
waxi:19.41, deriv:11.97, content:-3.38, aqueou:0.38,
saccharid:-11.95, ml:-0.79, cook:-10.04, modifi:5.65,
solid:5.50, sampl:6.27, mix:2.48, minut:-1.68, dri:-0.91,
gel:-9.85, activ:5.98, corn:-5.27, alpha:12, sprai:-2.74

QTF Terms: starch:14.64, emulsifi:6.72, succin:-3.46,
enzym:29.49, emuls:12.66, hydrophob:5.45, anhydrid:-5.47,
reaction:-1.97, octenyl:-0.66, stabil:3.64, alkenyl:0.06,
reagent:1.17, carbon:0.12, potato:3.74, alkyl:-0.33,
wt:-4.57, ether:1.96, enzymat:-3.45, convers:10.44,
chain:-5.53, atom:0.03, ph:-4.55, treat:-0.89,
ammonium:-1.96, food:-11.94, amylos:-20.15,
glucanotransferas:-0.86, glycidyl:-0.40, glycosyl:-0.02,
dry:-6.21, deriv:11.97, transferas:0.89, foam:-0.49,

PRF Terms: starch:14.64, encapsul:17.50, chees:-4.22,
oil:8.63, hydrophob:5.45, agent:5.19, casein:-2.19,
degrad:17.13, deriv:11.97, tablet:5.30, debranch:-10.58,
imit:-1.13, viscos:7.77, oxid:5.97, activ:5.98, osa:9.32,
funnel:2.68, amylas:26.06, amylopectin:-7.14, maiz:20.61,
blend:-3.17, waxi:19.41, convert:31.81,

IPC title Terms:cosmet:3.77, toilet:0.18, prepar:-0.82,
case:0.47, accessori:-0.01, store:-0.37, handl:0.07,
pasti:-0.17, substanc:-1.21, fibrou:-0.01, pulp:-1.28,
constitut:-0.06, paper:1.26, impregn:-0.11, emulsifi:6.72,
wet:-0.28, dispers:-8.66, foam:-0.49, produc:-0.57,
agent:5.19, relev:0.18, class:0.053, lubric:-0.38,
emuls:12.66, fuel:-0.011, deriv:11.97, starch:14.64,
amylos:-20.15, compound:-0.63, saccharid:-11.95,
radic:1.03, acid:-3.19
```

Figure 3: Four query reduction approaches on a sample query. Top terms retained by each method are shown. Numerical oracular scores $RF(t, Q)$ are provided indicating whether the term was useful (blue/positive) or noisy (red/negative).

Table 2: System performance using minimal relevance feedback. τ is RF score threshold, and k indicates the number of first relevant retrieved patents.

	$k=1$	$k=1$	$k=3$	$k=3$
	$\tau=0$	$\tau=1$	$\tau=0$	$\tau=1$
MAP	0.303	0.304	0.388	0.387
Recall	0.504	0.509	0.576	0.579

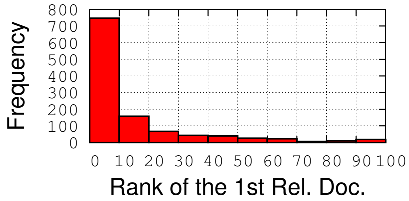


Figure 4: The distribution of the first relevant document rank over test queries.

sults. Using this approach, Table 2 shows that we can double the MAP in comparison to our baseline and also outperform the PATATRAS system.

Furthermore, to establish the minimal interaction required by this approach, Figure 4 indicates that the baseline methods return a relevant patent approximately 80% of the time in the first 10 results and 90% of the time in the first 20 results. Hence, such an interactive approach requires relatively low user effort while achieving state-of-the-art performance.

5. RELATED WORK

In this work, we focused on the development of an oracular query in order to address a number of fundamental questions regarding query reformulation and their efficacy in terms of approximating the oracular query. Previous works have not formulated such an oracular query, but nonetheless have inspired our investigation of query reformulation techniques as we briefly discuss below.

Bashir et al. [2] proposed query expansion with pseudo-relevance feedback that used machine learning for term selection. Verma and Varma [16] proposed a different approach, which instead of using the patent text to query, use its IPC codes, which are expanded using the citation network. Itoh et al. [4] proposed a new term selection method using different term frequencies depending on the genre in the NTCIR-3 Patent Retrieval Task. Mahdabi et al. [11] used term proximity information to identify expansion terms. Ganguly et al. [3] adapted pseudo-relevance feedback for query reduction by decomposing a patent application into constituent text segments and computing the Language Modelling (LM) similarities of each segment from the top ranked documents. The least similar segments to the pseudo-relevant documents removed from the query, hypothesizing it can increase the precision of retrieval. Kim et al. [6] provided diverse query suggestion using aspect identification from a patent query to increase the chance of retrieving relevant documents. Magdy et al. [9] discussed that standard query expansion techniques are less effective in patent retrieval, where the initial query is the full texts of query patents.

6. CONCLUSION

In this paper, we looked at the patent prior art search from a term selection perspective. While previous works proposed different solutions to improve retrieval effectiveness, we focused on term analysis of the patent query and top-100 retrieved patents. After defining an oracular query based on relevance judgements, we established both the sufficiency of the standard LM retrieval scoring models and query reduction methods to achieve state-of-the-art patent prior art search performance. After finding that automated methods for query reduction approaches fail to offer significant performance improvements, we showed that we can double the MAP with minimum user interaction by approximating the oracular query through a relevance feedback approach with a single relevant document. Given that such simple interactive methods for query reduction with a standard LM retrieval model outperform highly engineered patent-specific search systems from CLEF-IP 2010, we concluded that interactive methods offer a promising avenue for simple but highly effective term selection in patent prior art search.

7. REFERENCES

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition.* Pearson Education Ltd., Harlow, England, 2011.

[2] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, 2010.

[3] D. Ganguly, J. Leveling, W. Magdy, and G. J. Jones. Patent query reduction using pseudo relevance feedback. In *CIKM*, 2011.

[4] H. Itoh, H. Mano, and Y. Ogawa. Term distillation in patent retrieval. In *Proceedings of the ACL-2003 workshop on Patent corpus processing - Volume 20*, 2003.

[5] H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *IIiX*, 2010.

[6] Y. Kim and W. B. Croft. Diversifying query suggestions based on query documents. In *SIGIR*, 2014.

[7] P. Lopez and L. Romary. Patatras: Retrieval model combination and regression models for prior art search. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, pages 430–437. Springer, 2010.

[8] W. Magdy. *Toward higher effectiveness for recall-oriented information retrieval: A patent retrieval case study.* PhD thesis, Dublin City University, 2012.

[9] W. Magdy and G. J. Jones. A study on query expansion methods for patent retrieval. In *Proceedings of the 4th workshop on Patent information retrieval*, 2011.

[10] P. Mahdabi and F. Crestani. Patent query formulation by synthesizing multiple sources of relevance evidence. *ACM Transactions on Information Systems*, 2014.

[11] P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *SIGIR*, 2013.

[12] K. T. Maxwell and W. B. Croft. Compact query term selection using topically related text. In *SIGIR*, 2013.

[13] F. Piroi. CLEF-IP 2010: Retrieval experiments in the intellectual property domain. In *CLEF 2010 LABs and Workshops, Notebook Papers, 22-23 September 2010, Padua, Italy*, 2010.

[14] M. F. Porter. An algorithm for suffix stripping. In *Program*, volume 14, pages 130–137. 1980.

[15] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-2. In *TREC*, pages 21–34, 1993.

[16] M. Verma and V. Varma. Patent search using IPC classification vectors. In *PaIR*, 2011.

[17] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *SIGIR*, 2009.

[18] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, 2001.