# 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, starting with the Description section of the reference patent and using LM and BM25 scoring functions. We find that an oracular relevance feedback system, which 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 improve on the baseline performance with any of them. However, we show that a simple, minimal feedback interactive 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, Data Analysis.


## 1. INTRODUCTION

Patent prior art search involves finding previously granted patents 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 [7]: (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 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] for patent prior art search. While prior work has largely focused on specific techniques for reformulation, we first build an oracular query formed from known relevance judgments for the CLEP-IP 2010 Prior Art test collection [12] in an attempt to derive upper bounds on performance of standard BM25 and LM retrieval algorithms for this task. Since the results of this evaluation suggest that query reduction methods can achieve state-of-the-art prior art search performance, we proceed to analyse four simple automated methods for identifying reference patent query terms for removal. Finding that none of these methods seems to independently yield promise for query reduction that strongly outperforms the baseline, 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 patent search.



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



## 2. BASELINE IR FRAMEWORK

We developed a baseline system for patent prior-art search using Lucene[1], which supports queries using the probabilistic model Okapi BM25 [14] as well as language models (LM: Dirichlet smoothing, and Jelinek-Mercer smoothing) [16]. We used this system to index the English subset of CLEF-IP 2010 dataset[2] with the default settings using the Porter stemming algorithm [13] and English stop-word removal. We also removed patent-specific stop-words as described in [7]. CLEF-IP 2010 contains 2.6 million patent documents, and the English test sets of CLEF-IP 2010 correspond to 1303 topics (queries). In our implementation, each section of a patent (title, abstract, claims, and description) is indexed

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

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

in a separate field. However, when a query is processed, all indexed fields are targeted, 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 [6]. Although this IPC codes filter may fail to retrieve relevant patents, we have chosen to keep it for the following reasons: (i) more than 80% of the reference patent queries share an IPC code with their associated relevant patents, and (ii) it makes the retrieval process much faster. We evaluate the results for the top 100 retrieved patents by Mean Average Precision (MAP) and Average Recall. We assume that users examine the top 100 patents [3].

We achieved the best performance while querying with the Description section as in previous work [15] and using either the LM or the BM25 scoring functions. We call this initial query: *Patent Query*, and we use it as our main baseline.

# 3. ORACULAR TERM SELECTION

In this section we seek to understand the adequacy of the baseline *Patent Query* as well as an upper bound on performance of the BM25 and LM models and the sufficiency of terms in the reference patent by developing an *Oracular Query*.

## 3.1 Oracular Query Formulation

We begin by defining an oracular relevance feedback system which extracts terms from the judged relevant documents. To do this, we calculate a relevance feedback (RF) score for each term in the top-100 retrieved documents as follows:

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

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 assumed 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 seek to evaluate the optimal threshold $\tau$ empirically.

We formulated two oracular queries. The first query was formulated by positive terms in top-100 documents as follows:

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

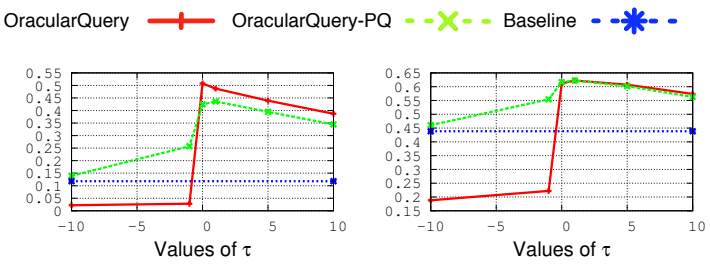
We formulated the second query by selecting only *useful terms* inside the reference patent:

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

## 3.2 Baseline vs. Oracular Query

In Table 1, we compare our two oracular relevance queries with both the baseline *Patent Query* and the *Top CLEF-IP 2010* system [5]. Here we see that the *Oracular Query* using $\tau = 0$ far outperforms the baseline and approximately performs twice as well on MAP as the best competitor in CLEF-IP 2010.

Next we investigate the ideal threshold setting $\tau$. Figure 1 illustrates that $\tau = 0$ is the best-performing value for *Oracular Query* while $\tau = 1$ is the best for *Oracular Patent Query*. The MAP for the *Oracular Patent Query* is lower than MAP




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

**Figure 1: System performance vs. the threshold $\tau$ for oracular query and oracular patent query.**

for the *Oracular Query*; nonetheless the terms selected from the reference patent itself are still sufficient to achieve performance significantly better than the Top CLEF-IP 2010 result. In addition, we remark on the rather unexpected steep drop-off in performance when the oracular query includes slightly noisy terms (i.e., *tau* just slightly less than 0) as defined previously.

Overall, our experiments related to oracular relevance feedback system suggest two important conclusions:

1. Query reduction should suffice for effective prior art patent retrieval.
2. Very precise methods for eliminating poor query terms in the reduction process.

# 4. QUERY REDUCTION: APPROXIMATING THE ORACULAR PATENT QUERY

The encouraging results of our Oracular Patent Query in the previous section motivate 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 used four approaches to refine the initial patent query:

1. removing document frequent terms (DF(t)>$\tau$),
2. keeping frequent terms in query (QTF(t)>$\tau$),
3. using pseudo relevance feedback to select query terms (PRF(t)>$\tau$),
4. removing general terms in IPC title.

In standard IR, removing terms, appearing a lot in the collection, helps the retrieval effectiveness. Inspired by this fact, we removed the words (in top-100 documents) with average term frequency higher than the threshold $\tau$ from the original query. As it can be seen in figure 2, unlike our assumption, removing frequent terms in top-100 documents ($DF(t) > \tau$) ruined the performance.

As mentioned in [11] terms inside verbose queries are also important. Hence, we kept frequent words inside the query while removing document frequent words. The results in Figure 2 indicate that it is better to keep all query terms ($QTF(t) > \tau$).

Table 1: System performance for the *Patent Query*, two variants of the *Oracular Query*, and *Top CLEF-IP 2010*.

| | Baseline (BM25) Weight:TF | Baseline (LM) Weight:TF | Baseline (LM) Weight:1 | Best Run | Oracular Weight:1 | Oracular(PQ) Weight:1 |
|---|---|---|---|---|---|---|
| MAP | 0.159 | 0.162 | 0.118 | 0.27 | 0.507 | 0.436 |
| A. Recall | 0.545 | 0.549 | 0.438 | - | 0.612 | 0.622 |

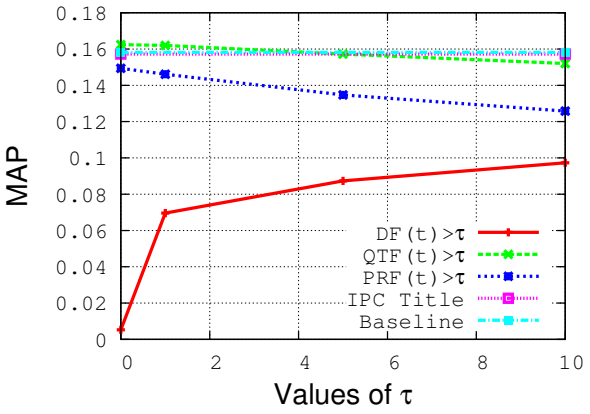



**Figure 2: System performance vs. the threshold $\tau$ for four query reduction approaches.**

The third approach we used to reduce the query was Pseudo Relevance Feedback (PRF). PRF is an automated process without user interaction which assumes the top k ranked documents are relevant and the others are irrelevant. Again, it can be seen in Figure 2 that the results for query reduction using PRF were below the baseline. In fact, we could not find any heuristic correlates between $PRF(t)$ and $RF(t)$.

Finally, we used words in IPC code title to reduce the query assuming they are general words in all patents, which belong to the same category and may consider as stop-words. However we hurt the effectiveness by pruning them out.

Figure 3 is an anecdotal example for a sample query, which can explain why these four approaches did not work. It shows the abstract and a pair of terms, and RF score of each term, where $t \in \{t|DF(t)/QTF(t)/PRF(t) > 10\}$. It can be seen that high scored terms are polluted with the sufficient amount of noise to ruin the retrieval effectiveness.

Unfortunately, none of the proposed query reduction approaches for query reduction worked better than the baseline, which leads us to investigate interactive methods for reduction in the next section.

## 4.2 Semi-automated Interactive Reduction

Our anecdotal 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 results. Using this approach, Table 2

```
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 def 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: Anecdotal example: it shows the abstract, and $t : RF(t)$ pair of a sample query, $\{t|DF(t)/QTF(t)/PRF(t) > 10\}$. Useful terms are highlighted in blue and the noisy ones in red.**

shows that we can double the MAP in comparison to our baseline and also outperfor the best system from CLEF-IP 2010.

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

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

| | $k=1$ $\tau=0$ | $k=1$ $\tau=1$ | $k=3$ $\tau=0$ | $k=3$ $\tau=1$ |
|---|---|---|---|---|
| MAP | 0.3028 | 0.3040 | 0.3879 | 0.3872 |
| A. Recall | 0.5040 | 0.5090 | 0.5757 | 0.5787 |

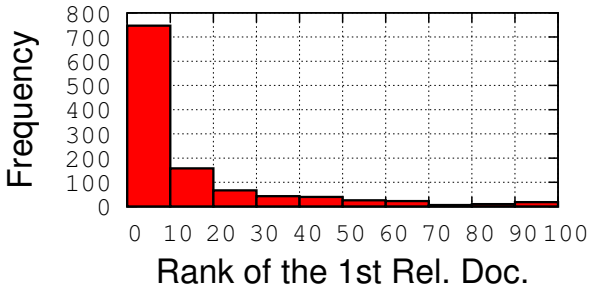



**Figure 4: The distribution of the first relevant document rank over test queries.**

low user effort while achieving state-of-the-art performance.

## 5. RELATED WORK

Our work is different from pioneer studies on patent retrieval, as we closely looked into the problem rather than solutions to figure out the causes that generic IR models which are based on term matching process, do not work efficiently in patent domain. Magdy et al. [8] studied works on query expansion in patent retrieval and discussed that standard query expansion techniques are less effective, where the initial query is the full texts of query patents. Mahdabi et al. [10] used term proximity information to identify expansion terms. Ganguly et al. [2] 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. [4] provided diverse query suggestion using aspect identification from a patent query to increase the chance of retrieving relevant documents. Mahdabi et al. [9] used linked-based structure of the citation graph together with IPC classification to improve the initial patent query.

## 6. CONCLUSIONS

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 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 conclude that interactive methods offer a promising avenue for simple but highly effective term selection in patent prior art patent 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] D. Ganguly, J. Leveling, W. Magdy, and G. J. Jones. Patent query reduction using pseudo relevance feedback. In *CIKM*, 2011.

[3] H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proceedings of the third symposium on Information interaction in context*, pages 13–24. ACM, 2010.

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

[5] P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. In *CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation*, 2010.

[6] 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.

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

[8] 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*, pages 19–24. ACM, 2011.

[9] P. Mahdabi and F. Crestani. Patent query formulation by synthesizing multiple sources of relevance evidence. *ACM Transactions on Information Systems (TOIS)*, 32(4):16, 2014.

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

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

[12] 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.

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

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

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

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