On Term Selection Techniques for Patent Prior Art Search

Mona Golestan Far
NICTA & ANU
Canberra, Australia
mona.golestanfar@anu.edu.au

Scott Sanner
NICTA & ANU
Canberra, Australia
ssanner@gmail.com

Mohamed Reda Bouadjenek
INRIA & LIRMM, France
Montpellier, France
reda.bouadjenek@inria.fr

Gabriela Ferraro
NICTA & ANU
Canberra, Australia
gabriela.ferraro@nicta.com.au

David Hawking
Microsoft (Bing) & ANU
Canberra, Australia
david.hawking@acm.org

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 specifically applied to patent prior art search [11, 16]. 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 [14] 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 and effective term selection in patent prior art search.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
SIGIR'15, August 09 - 13, 2015, Santiago, Chile.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767801 .

2. BASELINE IR FRAMEWORK

We developed a baseline IR system for patent prior art search on the top of the Lucene search engine and LM (Dirichlet smoothing, and Jelinek-Mercer smoothing) scoring functions based on [2]. We used Lucene to index the English subset of the CLEF-IP 2010 dataset[1] that contains 2.6 million patent documents and a subset of 1281 topics (queries) in the English test set where we determined at least one valid, relevant English document was available. We used the default Lucene settings using the Porter stemming al-

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

gorithm and English stop-word removal. We also removed patent-specific stop-words as described in [8]. We indexed each section of a patent (title, abstract, claims, and description) in a separate field. However, when a query is processed, all indexed fields are targeted with an equal weight. 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 work [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. System performance 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 [16] and using either the LM or the BM25 scoring functions. We call this the *Patent Query* and use it as our 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 and exploits patent metadata and citation structures. All results in the paper use the 1348 English topic subset as reported in the PATATRAS evaluation [14]. Since the evaluation of our systems used a slightly smaller subset of 1281 topics as noted previously, we assume no relevant results were found by our systems for the 67 remaining topics of the 1348 topic subset in order to ensure a fair comparison to PATATRAS.

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 t in the top-100 retrieved documents for query Q as follows:

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

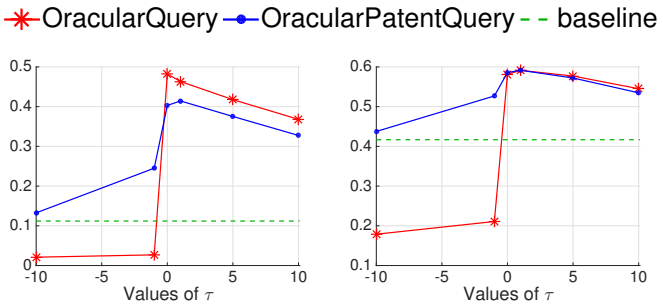
where $Rel(t,Q)$ is the average term frequency in retrieved relevant patents and $Irr(t,Q)$ is the average term frequency in retrieved irrelevant patents. We assume that words with a positive score are *useful words* since they are more frequent in relevant patents, while words with negative score are *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:

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

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

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



(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* (baseline), two variants of the *Oracular Query*, and *Top CLEF-IP 2010 Competitor (PATATRAS).*

		Baseline	PATATRAS	Oracular $\tau = 0$	Oracular (PQ), $\tau = 1$
LM	MAP	0.112	0.226	**0.482**	**0.414**
	Recall	0.416	0.467	0.582	0.591
BM25	MAP	0.123	0.226	**0.492**	**0.424**
	Recall	0.431	0.467	0.584	0.598

3.2 Baseline vs. Oracular Query

First, we investigate the ideal threshold setting τ for the oracular queries as shown in Figure 1. Notably, there is a rather unexpected steep drop-off in performance for both oracular queries when slightly noisy terms are included (i.e., τ just slightly less than 0). However, this dropoff is less pronounced for the *Oracular Patent Query* indicating that restriction to query terms in the reference patent may reduce the impact of the noisy terms that are present. While the *Oracular Query* and *Oracular Patent Query* peak at slightly different thresholds ($\tau = 0$ and $\tau = 1$, respectively), either value of τ yields good performance. However, values of $\tau > 1$ demonstrate a stronger relative decrease in performance due to the exclusion of a large number of useful terms.

In Table 1, we compare our best oracular relevance queries with both the baseline *Patent Query* and the PATATRAS system. 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. More importantly, 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. The MAP and the recall for the best *Oracular Patent Query* are respectively lower than the MAP and the recall for the best *Oracular Query*. However, the query reduction approach inherent in the *Oracular Patent Query* is still sufficient to achieve MAP performance appreciably better than PATATRAS (for $\tau \geq 0$) with reduced sensitivity to the inclusion of noisy terms (when $\tau < 0$).

Hence, 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 Query:

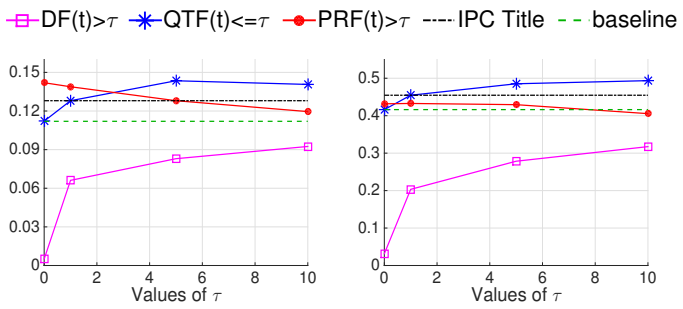
(i) In standard IR approaches, removing terms appearing highly frequently across documents in the collection can improve retrieval effectiveness. Inspired by this fact, after an initial run of the query, we removed terms with a high average document frequency (DF) over the top-100 documents ($DF(t) > \tau$). As seen in Figure 2 (magenta line), such pruning hurts performance. DF pruning continues increasing and converges to the baseline as $\tau \to \infty$ (i.e., no pruning).

(ii) Frequent terms inside long and verbose queries are considered important [13]. Thus, we hypothesize that removing terms appearing infrequently in the Patent Query may help and hence propose to remove terms with query term frequency (QTF) below a threshold τ ($QTF(t) \leq \tau$). Results in Figure 2 (blue line) indicate the performance is slightly better than the baseline when removing low QTF terms. The best MAP is achieved when $\tau = 5$ and it meets the baseline when $\tau = 0$ (i.e., all terms retained).

(iii) We use Pseudo Relevance Feedback (PRF) to select query terms [13] — the same as we did for the Oracular Relevance Feedback system (Section 3). We assume that the top 5 retrieved documents are relevant and the rest are irrelevant (this performed best), then we calculate PRF score based on this assumption. Terms that have PRF score higher than the threshold τ ($PRF(t) > \tau$), are selected from the Patent Query to reformulate a reduced query. Figure 2 (red line) shows that this approach is also unsuccessful at achieving a notable improvement over the baseline.

(iv) The titles of IPC codes indicate the intended content of patents classified under that code by using a single phrase or several related phrases linked together. We used words in IPC code titles for each patent query as stopwords to reduce the query, based on the assumption that these terms are common to all patents having the same IPC code label. As it can be seen in Figure 2 (black line), this approach slightly helps the performance.

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 the top 20 high-scoring terms (except for IPC Title Terms which are not scored, but simply displayed) and their associated RF scores for each approach. It can be seen that the four methods fail to clearly discriminate between useful and noisy terms. Important stemmed terms like "enzym" and "starch" would be pruned according to DF; in contrast, QTF and PRF both score "starch" highly and retain it, but also retain other noisy terms. Over half of the IPC Title Terms are noisy and appropriate to remove, but critical useful stemmed terms like "emulsifi" are also removed. Critically, all methods retain noisy terms (red/negative) and results from Section 3.2 showed that the inclusion of even slightly noisy terms can significantly hurt performance. Overall, all methods fail to retain only the oracular query terms (blue/positive) and do worse than PATATRAS.



(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:14.6,enzym:29.5,amylos:-20.1,oil:8.6, dispers:-8.7,ph:-4.6,dry:-6.2,heat:-2.3,product:-5.5, slurri:-11,viscos:8,composit:-4,reaction:-2,food:-12, agent:5,debranch:-11,reduc:-6,fat:-13,prepar:-0.8, hour:-5

QTF Terms:starch:14.6,emulsifi:6.7,succin:-3.5, enzym:29.5,emuls:12.7,hydrophob:5.4,anhydrid:-5.5, reaction:-2,octenyl:-0.7,stabil:3.6,alkenyl:0.06, reagent:1.2,carbon:0.1,potato:3.7,alkyl:-0.3,wt:-4.6, ether:2,enzymat:-3.4,convers:10.4,chain:-5.5

PRF Terms:starch:14.6,encapsul:17.5,chees:-4,oil:8.6, hydrophob:5.4,agent:5,casein:-2.2,degrad:17,deriv:12, tablet:5.3,debranch:-11,imit:-1,viscos:7.8,oxid:6, activ:6,osa:9.3,funnel:2.7,amylas:26,amylopectin:-7, maiz:20.6

IPC Title Terms:cosmet:3.8,toilet:0.2,prepar:-0.8, case:0.5,accessori:-0.01,store:-0.4,handl:0.07, pasti:-0.2,amylos:-20,fibrou:-0.01,pulp:-1.3, constitut:-0.06,paper:1.3,impregn:-0.1,emulsifi:6.7, wet:-0.3,dispers:-9,saccharid:-12,produc:-0.6,agent:5

Figure 3: The top 20 terms scored by each of four methods on a sample query (except for IPC Title Terms which are not scored); whether the term is pruned or retained depends on the approach, cf. (i)–(iv). Numerical oracular scores $RF(t, Q)$ are provided indicating whether the term was actually useful (blue/positive) or noisy (red/negative).

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 evaluate the impact of minimal user interaction, we next analyze the performance of an Oracular Patent Query

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

	$k=1$ $\tau=0$	$k=1$ $\tau=1$	$k=3$ $\tau=0$	$k=3$ $\tau=1$
MAP	0.288	**0.289**	**0.369**	0.368
Recall	0.479	0.484	0.547	0.550

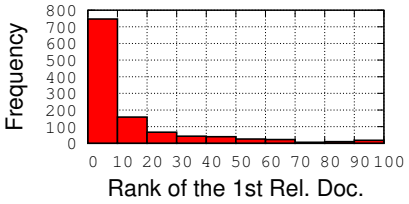


Figure 4: The distribution of the first relevant document rank over test queries.

(Equation 3) derived from *only* the top-k ranked relevant documents identified in the search results (for small k) — we assume that the remaining documents in the top-100 are irrelevant. Using this approach, Table 2 shows that we can double the MAP in comparison to our baseline and also outperform the PATATRAS system by identifying only the *first* relevant document.

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. Bashir et al. [1] proposed query expansion with pseudo-relevance feedback that used machine learning for term selection. Verma and Varma [15] used IPC codes instead of using the patent text to query, 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. [12] 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; the least similar segments to the pseudo-relevant documents are removed from the query. 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] and Bouadjenek et al. [2] studied different query expansion and reduction techniques for patent search on CLEF-IP 2010, and reported little improvement with automatic methods. Magdy et al. [10] further compare the best two systems in CLEF-IP 2010.

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.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

7. REFERENCES

[1] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, 2010.
[2] M. R. Bouadjenek, S. Sanner, and G. Ferraro. A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications. In *ICAIL*, 2015.
[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 *ACL workshop on Patent corpus processing*, 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. Experiments with citation mining and key-term extraction for prior art search. In *CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation*, 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] W. Magdy, P. Lopez, and G. J. Jones. Simple vs. sophisticated approaches for patent prior-art search. In *Advances in Information Retrieval*, pages 725–728. Springer, 2011.
[11] P. Mahdabi and F. Crestani. Patent query formulation by synthesizing multiple sources of relevance evidence. *ACM Transactions on Information Systems*, 2014.
[12] P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *SIGIR*, 2013.
[13] K. T. Maxwell and W. B. Croft. Compact query term selection using topically related text. In *SIGIR*, 2013.
[14] F. Piroi. Clef-ip 2010: Prior art candidates search evaluation summary. Technical report, IRF TR, Vienna, 2010.
[15] M. Verma and V. Varma. Patent search using IPC classification vectors. In *PaIR*, 2011.
[16] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *SIGIR*, 2009.