On Term Selection Techniques for Patent Prior-art Search

ABSTRACT

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-2010. We find a very clear term selection value threshold for use when choosing terms. A much more realistic approach in which feedback terms are extracted only from the first relevant document retrieved, still outperforms the best run. We 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 different approaches to identify negative terms but we were unable to improve on the baseline performance with any of them.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query Formulation

Keywords

Patent search, Query Reformulation, Data Analysis

1. INTRODUCTION

A patent is a set of exclusive rights granted to an inventor to protect their invention for a limited period of time. An important requirement for a patent to be granted is that the invention, it describes, is novel which means there is no earlier patent, publication or public communication of a similar idea. To ensure the novelty of an invention, patent offices as well as other Intellectual Property (IP) service providers mainly perform a search called 'prior art search'. The purpose of 'prior art search' is finding all relevant patents which may put the patent application at the risk of novelty invalidation or at least have common parts with patent application and should be cited [9] [15].

Patent retrieval has three main characteristics which makes it difficult compared to other IR applications: (1) the search starts with a query as long as a full patent application that helps users –usually patent examiners, inventors, or lawyers– avoid spending long hours to formulate a query; (2) it is recall-oriented, where not missing relevant documents is more important than appearing relevant documents at top of the list; (3) unlike the web application in which authors tend to highlight their work to be easily found through search engines, authors of the patents prefer to use a vague language to avoid the invalidation of their idea.

Many works has been conducted to improve the patent retrieval effectiveness so far. However, either the results showed quite small improvement or the proposed methods were complicated and computationally expensive. Overall, the works on patent search fall in five main categories [7] query reformulation(query expansion and query reduction), query term selection, query suggestions, using patent metadata and images for retrieval [8], and Cross-Language Information Retrieval [11].

In this work, we mainly emphasized on the problem from the term analysis perspective which ended in an effective minimal relevance feedback method. We investigated 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 found 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. A much more realistic approach in which feedback terms are extracted only from the first relevant document retrieved, still outperforms the winner. We 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 three different approaches to identifying negative terms but were unable to improve on the baseline performance with any of them.

2. BASELINE IR FRAMEWORK

We developed a Lucene-based[1] IR system with the possibility of using diverse generic IR models: TF-IDF, BM25,

SIGIR '15, August 9-13, 2015, Santiago, Chile
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

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

Language Models (Dirichlet smoothing, and Jelinek-Mercer smoothing) as our baseline system. We achieved the best baseline effectiveness querying with the Description section of the patent application as it is also mentioned in [16], and using LM and BM25 scoring functions. We call this initial query: *patent query*. We conducted our experiments on CLEF-IP[2]2010 data collection, with 2.6 million European patent documents and 1303 English test topics (queries). On the collection side, we only indexed English subset of each section of a patent (title, abstract, claims, and description), and IPC[3]code in a separate field. We also used the patent classification assigned to the query topics to filter search results to match at least one of the query IPC codes, as recommended in [6]. Our experiments showed that using IPC filter is itself a source of error because about 19% of relevant patents in CLEF-IP 2010 data collection do not share any classification code with their query. However, for our analysis, we kept the filter on since it made the matching process between the query and documents notably faster. We evaluate the results for top 100 retrieved patents by Mean Average Precision (MAP) and Average Recall. We assume that users examine the top 100 patents [2].

3. ORACULAR TERM SELECTION

The main complaint about patent search is an insufficient match between the content of patent queries and relevant patents[7][9]. However, considering the size of a patent query (usually thousand of words), the intuition is that there are enough terms to match the relevant patents.

3.1 Oracular Query Formulation

We started with *relevance feedback* where we have access to the judged relevant documents. We calculate a relevance feedback (RF) score for each term in top-100 retrieved documents as follows:

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

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 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 expected to see a higher performance for the queries which contain more *useful words*, but, surprisingly, we could not find any correlation between the performance and the presence of *useful words* in the query.

We hypothesized that a query, formulated by only the *useful terms*, is the best possible query we can make since they are all frequent in relevant patents but rare in irrelevant ones. We formulated two oracular queries. The first query was formulated by positive terms in top-100 documents as follows:

$$Oracular\ Query = \{t \in top - 100 | score_{RF}(t) > 0\} \qquad (2)$$

We formulated the second query by selecting only *useful terms* existing inside the patent query based on the hypothesis that a patent query contains sufficient words matched with the relevant patents:

$$Oracular\ Patent\ Query = \{t \in Q | score_{RF}(t) > 0\} \qquad (3)$$

The system performance to these two queries were encouraging. We discuss the detailed results in the next section.

3.2 Baseline vs. Oracular Query

The system performed very well for both *Oracular Query* and *Oracle Patent Query*. As it can be seen in Table 1, the *Oracular Query* far outperforms the baseline *Patent Query* and performs twice as well on MAP as the best competitor in CLEF-IP 2010 [5].

Table 1: System performance for the *Patent Query*, *Oracular Query*, and *Best Run Query*.

	Best Run	Pat.Query Weight:1	Oracular Query Weight:1
MAP	0.27	0.1181	0.5075*
A. Recall	-	0.4385	0.6118*

We used a threshold τ to formulate the *Oracular Query* and *Oracular Query* to include merely terms with the RF score higher than τ ($score_{RF}(t, Q) > 0$). Figure 1 illustrates that $\tau = 0$ is the best-performed 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 for *Oracular Query* which indicates that some positive terms from relevant patents are missed in the patent query. Further analysis showed an unexpected steep drop-off in performance when the oracular query is polluted with additional terms from the original patent query.

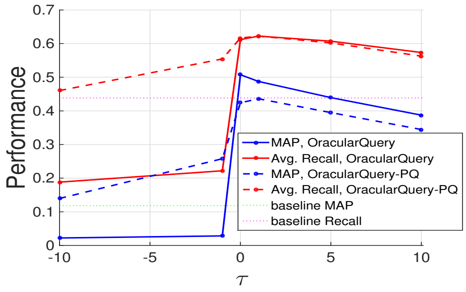

Figure 1: System performance vs. the threshold τ for oracular query and oracular patent query.

Overall, our experiments related to oracular relevance feedback system suggest two main solutions:

1. Query reduction should suffice for effective prior art patent retrieval.

2. A very precise methods for eliminating poor query terms are needed in the reduction process.

4. QUERY REDUCTION: APPROXIMATING THE ORACULAR QUERY

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

[3] International Patent Classification

4.1 Automated Reduction

We 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 used four approaches to refine the initial patent query:

1. removing document frequent terms,
2. keeping frequent terms in query [14],
3. using pseudo relevance feedback to select query terms,
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 with average term frequency (in top-100 documents) higher than the threshold τ 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 [14] terms inside verbose queries are also important. So, we kept frequent words inside the query while removing document frequent words. It can be seen in Figure 2 that keeping terms with term frequency higher than a threshold τ helped and we got the performance when keeping all query terms but it is close to the baseline.

We used pseudo relevance feedback (PRF) as the third feature to reduce the query. 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 $score_{RF}(t)$ and $score_{PRF}(t)$. Figure 3 is an anecdotal example for a sample query which can explain the reason that PRF did not work. It shows the query abstract and a pair of PRF terms, with $score_{PRF}(t) > 10$, and RF score of each term. It can be seen that terms with high PRF score have a negative RF score which means words from PRF contaminated with sufficient amount of noise to ruin the retrieval effectiveness. We used words in IPC code title to reduce the query because as it can be seen in Figure 2 the majority of them are negative terms as they are general words in all patents belonging to the same category. However we hurt the effectiveness by pruning them out.

Unlike our initial assumption, non of the standard proposed query reduction approaches for query reduction worked better than the baseline.

4.2 Reduction by Relevance Feedback

All our attempts to improve the system effectiveness without accessing the relevance feedback were quite in vein because the features we recognized were tightly the combination of the useful words and noisy words and the system performance is too sensitive to the existence of a noisy word or the absence of the useful terms. So, we decided to apply much more realistic approach in which feedback terms are extracted only from the first ranked relevant document retrieved. Table 2 shows that we can double the MAP by only the first-ranked relevant document. Fig. 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, such an interactive approach

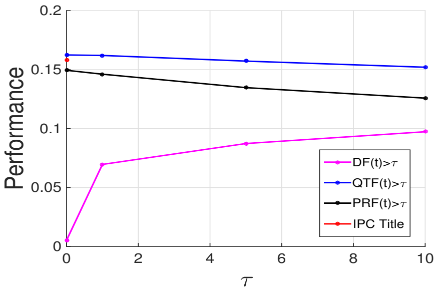

Figure 2: System performance vs. the threshold τ for four query reduction approaches.

```
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 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 $term$: $score_{RF}(term)$ pair of a sample query. Useful terms are highlighted in blue and the noisy ones in red.

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	0.3028	0.3040*	0.3879	0.3872
A. Recall	0.5040	0.5090	0.5757	0.5787

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

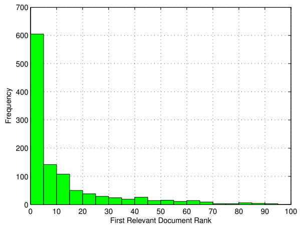

Figure 4: The distribution of the first relevant document rank over test queries which have TPs

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. [10] 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. [13] used term proximity information to identify expansion terms. Ganguly et al. [1] 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. [3] provided diverse query suggestion using aspect identification from a patent query to increase the chance of retrieving relevant documents. Mahdabi et al. [12] 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 different 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 finding a golden standard from relevance feedback, we examined the most obvious features such as: document frequent words, query frequent words, IPC definition words, and pseudo relevance feedback that might correlate RF score for terms in top retrieved documents. We showed that these feature helps very little because they are a complicated mixture of useful terms and noisy words that can not be separated easily. Finally, we showed that we can double the MAP with minimum user interaction. For future works, we plan to analyse more features which are independent from the relevance feedback but correlate with RF score. Inspired by some excellent works proposing query reduction and term selection techniques for the long non-patent queries[14][4], we are also going to apply them for patent retrieval.

7. REFERENCES

[1] D. Ganguly, J. Leveling, W. Magdy, and G. J. Jones. Patent query reduction using pseudo relevance feedback. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1953–1956. ACM, 2011.

[2] 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.

[3] Y. Kim and W. B. Croft. Diversifying query suggestions based on query documents. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 891–894. ACM, 2014.

[4] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 564–571. ACM, 2009.

[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] M. Lupu, A. Hanbury, et al. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7(1):1–97, 2013.

[8] M. Lupu, F. Piroi, and A. Hanbury. Evaluating flowchart recognition for patent retrieval. In *The Fifth International Workshop on Evaluating Information Access (EVIA)*, pages 37–44, 2013.

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

[10] 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.

[11] W. Magdy and G. J. Jones. Studying machine translation technologies for large-data clir tasks: a patent prior-art search case study. *Information Retrieval*, 17(5-6):492–519, 2014.

[12] 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.

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

[14] K. T. Maxwell and W. B. Croft. Compact query term selection using topically related text. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 583–592. ACM, 2013.

[15] F. Piroi, M. Lupu, and A. Hanbury. Overview of clef-ip 2013 lab. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 232–249. Springer, 2013.

[16] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 808–809. ACM, 2009.