On the Construction of Ideal Query for Patent Prior-art Search

M O
NICTA & ANU
Canberra, Australia
name.surname@nicta.com.au

W Z
NICTA & ANU
Canberra, Australia
name.surname@nicta.com.au

X Y
NICTA & ANU
Canberra, Australia
name.surname@nicta.com.au

ABSTRACT

Patent prior-art search aims to find all relevant patents which may invalidate the novelty of a patent application or at least have common parts with patent application and should be cited. Patent search has been the centre of attention in IR communities for years, however it has lower retrieval effectiveness compared to other IR applications. In this work, we focused on the causes of failure rather than solutions. We started with relevance feedback to get a golden standard, then we concentrated on heuristics correlate with our RF standard. Finally, we showed that features other than relevance feedback can not be helpful because they are a complex mixture of useful words and noisy words. Finally, we got a considerable improvement by user feedback with a minimum effort.

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 [8] [14].

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

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.

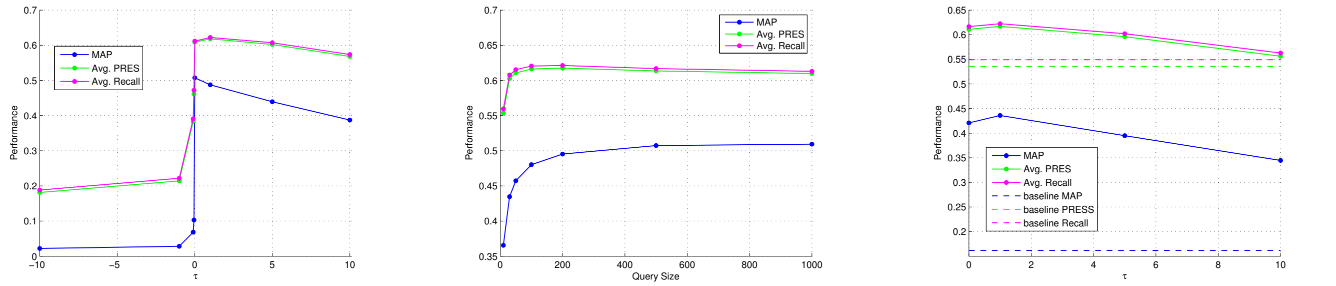
Many works has been conducted to improve the patent retrieval effectiveness so far. However, either the results showed quite small improvement or the proposed methods were complicated and computationally expensive. Overall, the works on patent search fall in five main categories: query reformulation(query expansion and query reduction), query term selection, query suggestions, using patent meta-data and images for retrieval [7], and Cross-Language Information Retrieval [10].

In this work, we mainly emphasized on the problem from the term analysis perspective which ended in an effective minimal relevance feedback method. 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-2014. 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 last yearâĂŹs 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, Language Models (Dirichlet smoothing, and Jelinek-Mercer smoothing) as our baseline system. We achieved the best

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



(a) (b) (c)

Figure 1: How score threshold(τ) and query size controls the performance. (a) Performance versus the score threshold. (b) Performance versus the query size. (c) System performance when we reduced the query by RF: $query = Q \cap (useful\ terms)$**, where** Q **is the patent query and** $useful\ terms = \{t|score_{RF}(t) > \tau\}$**.**

baseline effectiveness querying with the Description section of the patent application as it is also mentioned in [15], and using LM and BM25 scoring functions. We conducted our experiments on CLEF-IP[2]2010 data collection, with 2.6 million European patent documents and 1303 English 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 [5]. 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 makes the matching process between the query and documents notably faster.

3. IDEAL QUERY

The main complain about patent search is insufficient match between the content of patent queries and relevant patents[6][8]. However, we have the intuition that there are sufficient terms in a patent query containing thousands words to be matched with the relevant patents. So, in this section, we focused on term analysis to figure out the main causes that the system fails in retrieving relevant documents at top of the result list.

We started our analysis using *relevance feedback*, in which the user gives feedback on the relevance of documents in an initial set of results to improve the final result set. 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 percentage of *useful words* in the query.

3.1 Ideal Query Formulation

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 the ideal query as follows:

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

Table 1 compares the baseline performance, where the query is the full patent application, with the performance of the ideal query. It can be seen that MAP jumps from 0.1618 to *0.5075*, which means the ideal query considerably performs better than the baseline.

Table 1: System performance for the baseline and ideal query.

	Pat.Query Weight:TF	Pat.Query Weight:1	Ideal Query Weight:Score(t)	Ideal Query Weight:1
PRES	0.5355	0.4268	0.6086	0.6087
MAP	0.1618	0.1181	0.4617	0.5075*
A. Recall	0.5491	0.4385	0.6129	0.6118

3.2 Patent Query and Useful Terms

Our previous experiments led us to the hypothesis that a patent query contains sufficient words matched with the relevant patents. To prove our idea, we formulated a query by selecting only RF *useful terms* existing inside patent query as follows:

$$query = \{t|t \in \{Q \cap (useful\ terms)\}\} \qquad (3)$$

The results were encouraging, as MAP was improved from 0.1618 to *0.44*.

3.3 Analyse the Results

The main results related to ideal query formulation has been summarized in Figure 1. Figure 1-a shows how the RF score threshold τ controls the performance, It can be seen that it is better to include all terms with positive RF score. On top of that, we can see that the system is *over-sensitive* to the *noisy words* ($\tau < 0$). Adding words with negative RF score can sharply hurt the performance. Figure 1-b indicates

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

that formulating a query with up to 200 *useful words* helps the performance whereas there is no significant improvement when we add more than 200 right words. Finally, Figure 1-c explicitly shows that a patent query contains sufficient words to perform well.

We can conclude two important ideas: (1) a patent query contains sufficient useful terms to achieve an acceptable performance. (2) Noisy terms can highly ruin the IR effectiveness. Therefore, to improve patent prior-art search, we need to reformulate the initial patent query using term selection, and query reduction rather than query expansion. In addition, it is very important to identify and prune all the noisy words out because they are highly harmful.

4. QUERY REFINEMENT

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 (3) using pseudo relevance feedback to select query terms, and (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.

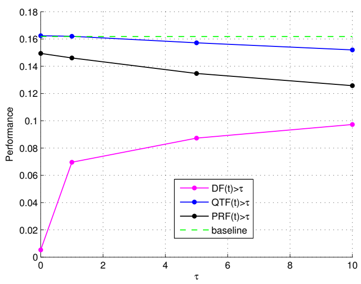


Figure 2: System performance vs. the threshold τ for three query reduction approaches.

As mentioned in [13] 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.

```
PAC-1612
Abstract: A wireless communication method for transmitting data from
        at least one master to one or more slaves positioned at various
         spatial locations and configured for generally simultaneous
        reception of the data. The method includes dividing the data
        into a number of portions, transmitting at least some of the
        portions using different transmission configurations for the
        different portions, having one or more of the slaves measure
        the quality of transmission associated with the group of
        different transmission configurations, and processing the
        quality measurements to determine new transmission
        configurations for use in transmitting the data.

PRF Terms: commun:-69.43159, transmiss:-58.168427, wireless:-7.68421,
        telephon:-25.17895, recept:-37.810528, slave:-31.0421, deleg:-22.368422,
        turn:-18.536846, master:-35.778954, origin:-4.7473674, schedul:12.852628,
        control:-14.842104, frequenc:60.34737, station:-76.26316, electron:-
        8.442106, perform:-9.71579, band:16.789476, termin:-40.04211,
        indic:3.6210496, reason:-6.642107, apparatus:-6.2421055, determin:-
        8.97895, complet:-8.842103, prohibit:-3.8947372, state:-9.557897,
        link:1.1157892, hop:24.378946, lan:-8.3368435, assign:-9.68421,
        f1:14.926317, short:-3.6000004, pattern:17.58947, paramet:1.5473672,
        serv:-4.5684214, permit:0.62105286

IPC def Terms: network:  -28.557888, traffic:  -3.2526314, resourc:
        -1.2947367, manag:  -9.652633, local:  -6.1368427, wireless:
        -7.68421, schedul:  12.852628, select:  13.473684, alloc:
        -10.042107, switch:  -14.463159, interconnect:  -0.5578947, transfer:
        -4.5052633, inform:  -43.400013, signal:  -30.71579, memori:
        -9.094736, input:  -9.736838, plan:  -0.22105263, coverag:
        -0.4526316, tool:  -0.15789473, deploy:  -0.07368421, partit:
        0.0526316, cell:  -7.6842103, structur:  -0.94736844, orthogon:
        -4.1789474, multiplex:  -11.621052, walsh:  -0.29473686, code:
        -27.842102, supervisori:  -0.16842104, monitor:  -6.7157907, test:
        -1.2210523, arrang:  -0.5578947, spread:  -2.7263162, spectrum:
        3.6000001, direct:  -1.810526, sequenc:  -3.9052625, modul:
        -16.0, frequenc:  60.34737, hop:  24.378946, radio:
        -15.242106, transmiss:  -58.168427, radiat:  -0.3157895,  mobil:
        -6.831579, mainten:  -0.75789464, administr:  -0.07368421, path:
        -1.6631576, configur:  -9.094739, lan:  -8.3368435, wide:
        0.021052642, wan:  -0.021052632
```

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.

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

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

6. RELATED WORK

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

	$k=1$ $\tau=0$	$k=1$ $\tau=1$	$k=3$ $\tau=0$	$k=3$ $\tau=1$
PRES	0.4965	0.5016	0.5699	0.5727
MAP	0.3028	0.3040^*	0.3879	0.3872
A. Recall	0.5040	0.5090	0.5757	0.5787

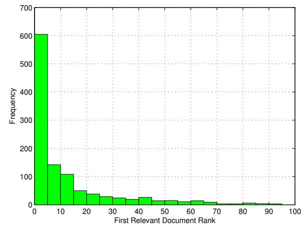


Figure 4: The distribution of the first relevant document rank over test queries which have TPs

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. [9] 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. [12] 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. [3] provided diverse query suggestion using aspect identification from a patent query to increase the chance of retrieving relevant documents. Mahdabi et al. [11] used linked-based structure of the citation graph together with IPC classification to improve the initial patent query.

7. 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[13][4], we are also going to apply them for patent retrieval.

8. 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.

9. REFERENCES

[1] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *Advances in Information Retrieval*, pages 457–470. Springer, 2010.

[2] 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.

[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. 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.

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

[7] 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.

[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*, pages 19–24. ACM, 2011.

[10] 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.

[11] 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.

[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 *Proceedings of*

the 36th international ACM SIGIR conference on Research and development in information retrieval, pages 583–592. ACM, 2013.

[14] 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.

[15] 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.