

Patent Query Reduction using Pseudo Relevance Feedback

Debasis Ganguly Johannes Leveling Walid Magdy Gareth J. F. Jones
School of Computing, Centre for Next Generation Localisation
Dublin City University, Dublin 9, Ireland
{dganguly, jleveling, wmagdy, gjones}@computing.dcu.ie

ABSTRACT

Queries in patent prior art search, being full patent applications, are very much longer than standard ad hoc search and web search topics. Standard information retrieval (IR) techniques are not entirely effective for patent prior art search because of the presence of ambiguous terms in these massive queries. Reducing patent queries by extracting small numbers of key terms has been shown to be ineffective mainly because it is not clear what the focus of the query is. An optimal query reduction algorithm must thus seek to retain the useful terms for retrieval favouring recall of relevant patents, but remove terms which impair retrieval effectiveness. We propose a new query reduction technique decomposing a patent application into constituent text segments and computing the Language Modeling (LM) similarities by calculating the probability of generating each segment from the top ranked documents. We reduce a patent query by removing the least similar segments from the query, hypothesizing that removal of segments most dissimilar to the pseudo-relevant documents can increase the precision of retrieval by removing non-useful context, while still retaining the useful context to achieve high recall as well. Experiments on the patent prior art search collection CLEF-IP 2010, show that the proposed method outperforms standard pseudo relevance feedback (PRF) and a naive method of query reduction based on removal of unit frequency terms (UFTs).

Categories and Subject Descriptors

H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval—*Query formulation, Relevance Feedback*; H.3.1 [INFORMATION STORAGE AND RETRIEVAL]: Content Analysis and Indexing—*Abstracting methods*

General Terms

Experimentation, Performance, Measurement

Keywords

Query Reduction, Patent Search, Pseudo-Relevance Feedback

1. INTRODUCTION

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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...\$10.00.

Patent prior art search involves ad hoc search for already filed patents which may invalidate or at least describe prior art work in a patent application, (henceforth referred to as query patent in this paper). The primary objective and challenges of patent prior art search are different from those of standard ad hoc text and web search. These differences can be summarised as follows: a) the queries are full patent applications comprising of hundreds of words on average in contrast to ad hoc search and web search where the queries typically constitute of two or three words; and b) 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 ad hoc and web search, which is precision oriented.

Another challenge in patent prior art search is the vocabulary mismatch between the existing filed patents and the query patent which arises often due to the obscure style of writing a patent (patentese). Long patent queries comprising of several hundreds of terms fail to represent a focused information need required for high precision retrieval. In addition to this, vocabulary mismatch can be aggravated if naive key word extraction methods are applied in an attempt to form a reduced query, thus in effect leading to a degradation of recall. A balance between precision and recall can be achieved by a careful trade-off between what to remove from the query patent and a reduced query applied to the patent retrieval system. This paper attempts to achieve an effective trade-off in reducing query patents for prior art patent search by utilizing pseudo-relevance information from top ranked retrieved documents.

When applied in standard ad hoc search tasks, pseudo-relevance feedback (PRF) expands the initial query by adding terms which occur most frequently in the assumed relevant top ranked documents from an initial retrieval step in attempt to retrieve relevant documents at higher ranks. However, the massive length of patent queries is not conducive to the query expansion step in PRF because precision at top ranks is low, focus is not clear and added terms are noisy, which in effect introduce further ambiguity to the query. Instead, a more intuitive process is to apply PRF in a reverse process of query reduction. In contrast to ad hoc IR, where query expansion is used to move the initial query representation closer to that of the pseudo-relevant documents, we use pseudo relevance information to make the feedback query more similar to the pseudo relevant documents by reducing the original query.

The remainder of this paper is organized as follows: Section 2 surveys related work on patent prior art search, Section 3 describes our proposed method of query reduction in detail, Section 4 describes the experiments and discusses the results, and finally Section 5 concludes the paper with directions for future work.

2. RELATED WORK

The real life working principle undertaken by patent examiners to manually formulate queries for invalidating claims, involves selecting high frequency terms from the text of query-patent claim. Some early work on keyword extraction to form a reduced query, modelled on this real-life methodology of patent examiners includes that of [11, 4]. More recent work by Xue and Croft [17] advocates the use of full patent text as the query to reduce the burden on patent examiners and concludes with the observation that usage of the whole patent text with raw term frequencies gives the best mean average precision (MAP). Recent work in the CLEF-IP¹ task has shown that best retrieval results are obtained when terms are used from all the fields of the query patents [14]. Fujii [2] showed that retrieval effectiveness can be improved by merging IR methods with citation extraction. Magdy et.al. [9] show that the best performing run of CLEF-IP 2010 uses citations extracted by training a Conditional Random Field (CRF), whereas the second best run uses a list of citations extracted from the patent numbers within the description field of some patent queries. They also show that a simple IR approach of using terms from all fields of a patent-query with a frequency of at least two, merged with extracted citations achieves a statistically indistinguishable performance compared to the best run which employs sophisticated methods of retrieval using two complementary indices, one constructed by extracting terms from the patent collection and the other built from terminological resources such as the Wikipedia. As the baseline run for this paper we follow the simpler approach of the second ranked participating group and show that our method of query reduction produces better results without using citations.

Magdy and Jones [7] report that MAP can be a misleading metric for patent prior art search because of its inherent characteristic of favouring precision more than recall, and proposed the PRES (Patent Retrieval Evaluation Score) metric which measures the system recall and quality of ranking in one score. Our experiments report an improvement in both MAP and PRES over the baseline.

Ad hoc IR on news and web data has been shown to improve both in MAP and average recall measures by the use of PRF, due to the fact that additional terms from pseudo-relevant documents bridge the vocabulary gap between the query and the documents relevant to it in the collection [16, 15]. However, query expansion is associated with the risk of additional terms contributing to a drift in the original information need followed by a degradation of retrieval effectiveness in the feedback step [13]. Unfortunately all existing work on PRF coupled with query expansion for patent prior art search tasks report a degradation in MAP [5, 12, 8].

3. QUERY REDUCTION

3.1 Motivation

The important observations to be made from existing work as discussed in Section 2 are that patent prior art search achieves best retrieval performance when: i) information from all the fields of the query patents are used; ii) unit frequency terms (UFTs), i.e. terms which occur only once in the patent query are eliminated; and iii) no PRF is applied. This provides the motivation to devise a more intelligent technique of reducing queries in comparison to a naive frequency based cut-off and to develop techniques of exploiting the potential benefits of PRF. The work presented in this paper hence tries to reduce queries intelligently by utilizing pseudo-relevance. The main idea behind the approach is that since patent queries are long, retrieval performance can suffer because of a strong likelihood of presence of ambiguous terms which tend to

boost the scores of non-relevant documents. However it is not always reasonable to assume that low frequency terms are the harmful ones. Instead, retrieved pseudo-relevant documents can give us an estimation of the distribution of useful terms.

3.2 Term Context

Term proximity plays a key-role in IR. Local Context Analysis (LCA) hypothesizes that good expansion concepts tend to co-occur with query terms in the top ranked documents [16]. LCA decomposes the document text into fixed length passages and ranks candidate expansion terms by computing co-occurrences weighted by the *idf* (inverse document frequency) of query terms, the assumption being that co-occurrence of a term with a rare query term carries more weight. Relevance Based Language Model (RLM) theoretically establishes the LCA principle, where it is assumed that the expansion terms are generated from an underlying relevance model, which generates both the relevant documents and the query terms [6]. The probability of generating an expansion term from the relevance model is estimated by the co-occurrence of that term with the query terms.

In our proposed method, we utilize the context of terms in a slightly different way. In one of the variants, we investigate the constituent sentences and in the other, we decompose the query into fixed length non overlapping word windows. Our query reduction method works by removing from the query a subset of segments (*sentences* or *word windows*), least similar to the top ranked documents.

3.3 Query Reduction Algorithm

Algorithm 1 outlines the working steps of our proposed query reduction method. Line 6 calls the procedure `DecomposeQuery` which either breaks up the query text into component sentences or fixed length non-overlapping windows of *usize* words. The parameter *usize* thus controls the amount of contextual information that is used for computing the Language Model (LM) similarities. Line 10 aggregates the LM similarities of a query segment q_s with each $|R|$ top ranked document. The LM similarity equation that we use involves Jelineck-Mercer smoothing as described in [3] and is shown in Equation 1.

$$\log P(q_s|d) = \sum_{t \in q_s} n(t) \log \left(1 + \frac{\lambda_t P(t|d)}{(1 - \lambda_t) P(t)} \right) \quad (1)$$

$$\log P(q_s|R) = \sum_{d \in R} \log P(q_s|d) \quad (2)$$

Algorithm 1 QueryReduction($q, R, \tau, usize, window$)

```

1:  $q$  : The original patent query.
2:  $R$  : Set of pseudo-relevant documents retrieved after an initial run.
3:  $\tau$  : Fraction of segments to retain in the query. The remaining (i.e.
   (1 -  $\tau$ )) fraction of segments are removed from the query.
4:  $usize$  : Number of words in each window.
5:  $window$  : A boolean flag set to 0 for decomposing into sentences.
6:  $S \leftarrow DecomposeQuery(q, usize, window)$ 
7: for  $i = 1$  to  $|R|$  do
8:   {Compute and aggregate the LM similarities}
9:   for  $j = 1$  to  $length[S]$  do
10:     $S[j].sim \leftarrow S[j].sim + LMSim(S[j], R_i)$ 
11:   end for
12: end for
13: Sort the set  $S$  such that  $S_\alpha.sim \geq S_\beta.sim \quad \forall \alpha < \beta$ 
14: for  $i = 0$  to  $\tau \cdot length[S]$  do
15:    $RQ \leftarrow RQ \cup S_i$ 
16: end for
17: return  $RQ$ 

```

¹<http://www.ir-facility.org/clef-ip>

We show the log-transformed equation of the multinomial term generation model, $n(t)$ being the term frequency of t in query segment q_s . The accumulated score as shown in Equation 2 computes the combined probability of generating a query segment q_s from each pseudo-relevant document. The idea is that a query segment q_s with a low probability of generation from the pseudo-relevant document language model is likely to be *out-of-context* with respect to the actual information need of the query patent.

In Line 13 we sort the query segments by the accumulated LM scores and retain the top τ fraction of the segments in the reduced query. The parameter τ can thus be used to control the amount of reduction. A higher value of τ would remove less segments and a lower value would remove more. Another parameter that is used to control the amount of proximity information is the window size *wsiz*e. When *wsiz*e is set to 1, each window is comprised of an individual term and the method reduces to pure term deletion without any contextual evidence. A value of *wsiz*e which is too small may miss the desired proximity information, whereas a value which is too high may run the risk of removing useful information from the query. Therefore this parameter has to be tuned carefully. A simple way to avoid tuning the parameter *wsiz*e is to rely on the implicit semantic context represented by natural sentence boundaries instead of working with word windows.

4. EXPERIMENTS

4.1 Description and Parameter Settings

To evaluate our approach, we use the patent collection CLEF-IP 2010, which is comprised of 2.68 million patents from the European Patent Office (EPO). We restrict our retrieval experiments only to the English subset of the collection which constitutes 68% of the collection. Relevance assessments are provided for 1348 topics which are patent applications having *title*, *abstract*, *claims* and *description* fields. In addition to using a standard list of stopwords we also removed formulas, numeric references, chemical symbols and patent jargons such as *method*, *system*, *device*. A Porter stemmer [10] was used to stem the words. SMART² with LM (Equation 1), was used for retrieval, with λ set to 0.4.

4.2 Choosing the Baseline

In order to choose the strongest possible baseline to compare our method against, we attempted naive query reduction and PRF using RLM [6]. We start with a retrieval run by using all terms from all fields of a patent query and then generate another retrieval run with a shorter query obtained by removing the UFTs. We see that this naive query reduction method results in 62.8% improvement of MAP as compared to the run which uses all terms. PRF was attempted on this improved run with different settings for the number of pseudo-relevant documents to use (R), and the number of query expansion terms (T). All the PRF runs show a degradation in MAP. Table 1 summarises the results and shows the best feedback run that we obtained. We choose the best configuration, i.e. the second run of Table 1 as the baseline for the query reduction experiments.

4.3 Sentence Removal

The following parameters were varied in our experiments: i) the number of pseudo-relevant documents to use in computing the probabilities of generating every query sentence; and ii) the fraction of sentences to retain in the query. To find the optimal range of the parameters, we varied τ in [0.1, 0.9] in steps of 2 for 5, 10, 20 and 50 pseudo-relevant documents. Figure 1 shows that the best results

²ftp://ftp.cs.cornell.edu/pub/smart

Table 1: Choosing the strongest baseline

| Run description | Parameters | | Avg. query length | Metrics | |
|--------------------------|------------|----|-------------------|---------------|---------------|
| | R | T | | MAP | PRES |
| BL_1 : Using all terms | - | - | 628.21 | 0.0785 | 0.3903 |
| BL_2 : Removing UFTs | - | - | 381.90 | 0.1278 | 0.4604 |
| BL_3 : BL_2 and PRF | 5 | 50 | 678.21 | 0.0719 | 0.2649 |

are obtained when τ is set to 0.9. Retrieval results with too few or too much pseudo-relevance are unstable, as seen from the sharp drop of the line $R = 10$ from 0.7 to 0.9, sharp peak of $R = 5$ line from 0.7 to 0.9, and low IR effectiveness for the line $R = 50$ in [0.1, 0.5]. An important observation to be made here is that using a very small number of pseudo-relevant documents may result in an unpredictable retrieval performance because the estimated probabilities of generating query segments from too few a documents may not be accurate.

4.4 Word Window Removal

In this section, we investigate the effect of varying the window size *wsiz*e on retrieval. We take the best settings as found in Section 4.3, i.e. we use $R = 20$ and $\tau = 0.9$ and use discrete valued window sizes of 5, 20, 50 and 100. The right graph of Figure 1 shows that optimal results are obtained by employing windows of 20 words. The figure also portrays a degradation of retrieval performance for windows which are too short, as seen from the left-most point of the graph. This can be attributed to the fact that very short windows imply a higher chance of a match of all the constituent words within a pseudo-relevant document. For example, the probability of generating a one-term window from a document d is $P(t_1|d)$, whereas for a two-term window, the probability is $P(t_1|d)P(t_2|d) \leq P(t_1|d)$. Thus, the accumulated LM scores being comparatively higher are not very reliable for very short windows. On the other hand, windows which are too long may result in too much information being removed from the queries resulting in a degradation of MAP. The window version of the query reduction mechanism outperforms the sentence based one. This again can be attributed to the length variation of sentences, some being too short and the others too long. Too short ones suffer from unreliable estimates of generation probabilities from the pseudo-relevant documents, whereas too long ones contribute to removal of too much of information. Word windows of fixed lengths can overcome this length variation problem.

4.5 Discussion

Table 2 shows the results of segment based query reduction. An interesting observation is that the average number of unique query terms for the best performing segment removal runs, both for sentence and window, are higher than the baseline, which shows that even UFTs can play an important role in retrieval. We find that the average number of query terms for the window based removal is higher, thus contributing to a higher recall as evident from the higher value of PRES without compromising average precision, as evident from the higher value of MAP. The new method is able to outperform the strongest baseline BL_2 by 7.28%. Although the improvement is not statistically significant, as evaluated by Wilcoxon test on per-topic average precisions, the percentage gain is non-trivial, keeping in mind the fact that all standard PRF methods fail on the patent prior art search leading to a degradation in IR effectiveness. In fact we see that window based query reduction significantly outperforms the best PRF by a 90.68% increase in

Figure 1: Effect of varying the number of pseudo-relevant documents (R) and the fraction of sentences to retain (τ) on MAP (left). Effect of window size ($wsiz$ e) variation with the best settings of sentence reduction, i.e. $R = 20$ and $\tau = 0.9$ (right).

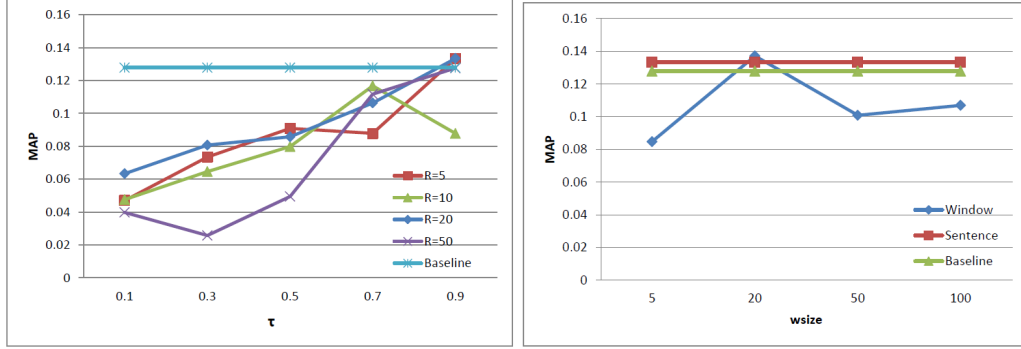


Table 2: Query Reduction performance on CLEF-IP 2010 topic set.

| Run Description | Parameters | | | Avg. # of query terms | MAP Relative gains over baseline | | | PRES Relative gains over baseline | | | | |
|---------------------|------------|--------|----------|-----------------------|----------------------------------|--------|--------|-----------------------------------|---------------|--------|--------|--------|
| | R | τ | $wsiz$ e | | Absolute | BL_1 | BL_2 | BL_3 | Absolute | BL_1 | BL_2 | BL_3 |
| Sentence removal | 20 | 0.9 | - | 474.99 | 0.1333 | 69.80% | 4.30% | 85.39% | 0.4615 | 18.24% | 0.23% | 74.21% |
| Word window removal | 20 | 0.9 | 20 | 547.59 | 0.1371 | 74.65% | 7.28% | 90.68% | 0.4674 | 19.75% | 1.52% | 76.44% |

MAP. The improvements in PRES are very small for both variants of the reduction algorithm.

5. CONCLUSION

The main contribution of the paper is the adaptation of pseudo-relevance for query reduction, instead of using it for query expansion. This is particularly useful in the domain of patent search because queries in patent prior art search are full patents and very much longer than ad hoc search queries. The method has been demonstrated to work well on the CLEF-IP 2010 patent prior art search task. Our work is a step ahead towards using PRF to improve MAP in patent prior art search, where all the previous PRF approaches have been reported to fail to improve retrieval effectiveness over initial retrieval.

There are several possible avenues along which our work can be extended in future, including a) devising a method for predicting whether query reduction would improve retrieval effectiveness for a query analogous to selective query expansion [1]; and b) exploring pseudo-relevance based summarisation techniques to reduce the length of a query analogous to the method of query based multi-document summarisation.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project.

6. REFERENCES

- [1] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A framework for selective query expansion. In *CIKM 2004*, pages 236–237. ACM, 2004.
- [2] A. Fujii. Enhancing patent retrieval by citation analysis. In *SIGIR*, pages 793–794. ACM, 2007.
- [3] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center of Telematics and Information Technology, AE Enschede, 2000.
- [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*, pages 41–45, Stroudsburg, PA, USA, 2003.
- [5] K. Kishida. Experiment on pseudo relevance feedback method using taylor formula at NTCIR-3 patent retrieval task. In *NTCIR-3*, 2003.
- [6] V. Lavrenko and B. W. Croft. Relevance based language models. In *SIGIR 2001*, pages 120–127. ACM, 2001.
- [7] W. Magdy and G. J. Jones. PRES: a score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010*, pages 611–618, 2010.
- [8] W. Magdy, J. Leveling, and G. J. F. Jones. Exploring structured documents and query formulation techniques for patent retrieval. In *CLEF-2009*, pages 410–417, 2010.
- [9] W. Magdy, P. Lopez, and G. J. F. Jones. Simple vs. sophisticated approaches for patent prior-art search. In *ECIR*, pages 725–728, 2011.
- [10] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [11] T. Takaki. Query terms extraction from patent document for invalidity search. In *NTCIR-5*, 2005.
- [12] H. Takuechi, N. Uramoto, and K. Takeda. Experiments on patent retrieval at NTCIR-5 workshop. In *NTCIR-5*, 2005.
- [13] E. L. Terra and R. Warren. Poison pills: harmful relevant documents in feedback. In *CIKM 2005*, pages 319–320. ACM, 2005.
- [14] M. Z. Wanagiri and M. Adriani. Prior art retrieval using various patent document fields contents. In *CLEF-2010 (Notebook Papers/LABs/Workshops)*, 2010.
- [15] R. H. Warren and T. Liu. A review of relevance feedback experiments at the 2003 Reliable Information Access (RIA) workshop. In *SIGIR 2004*, pages 570–571. ACM, 2004.
- [16] J. Xu and W. B. Croft. Improving the effectiveness of informational retrieval with Local Context Analysis. *ACM Transactions on information systems*, 18:79–112, 2000.
- [17] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In *SIGIR*, pages 808–809, 2009.