Improving Retrievability of Patents in Prior-Art Search

Shariq Bashir and Andreas Rauber
{*bashir,rauber*}*@ifs.tuwien.ac.at*

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Austria
http://www.ifs.tuwien.ac.at

Abstract. Prior-art search is an important task in patent retrieval. The success of this task relies upon the selection of relevant search queries. Typically terms for prior-art queries are extracted from the claim fields of query patents. However, due to the complex technical structure of patents, and presence of terms mismatch and vague terms, selecting relevant terms for queries is a difficult task. During evaluating the patents retrievability coverage of prior-art queries generated from query patents, a large bias toward a subset of the collection is experienced. A large number of patents either have a very low retrievability score or can not be discovered via any query. To increase the retrievability of patents, in this paper we expand prior-art queries generated from query patents using query expansion with pseudo relevance feedback. Missing terms from query patents are discovered from feedback patents, and better patents for relevance feedback are identified using a novel approach for checking their similarity with query patents. We specifically focus on how to automatically select better terms from query patents based on their proximity distribution with prior-art queries that are used as features for computing similarity. Our results show, that the coverage of prior-art queries can be increased significantly by incorporating relevant queries terms using query expansion.

1 Introduction

Patent retrieval falls into the recall-oriented application domain, where not missing a relevant patent is considered more important than retrieving only set of relevant patents at top rank results. This is particularly important in *prior-art search*, where missing one patent could result in a multimillion dollar lawsuit because of a patent infringement. The goal of searching a patent database for prior-art is to find all previously published patents on a given topic [9–11]. It is a common task for patent examiners and attorneys to decide, whether a new patent application is novel or contains technical conflicts with some already patented invention. Patent applications have complex structures and technical contents, which can create significant challenges for retrieval systems [7]. The vocabulary of patent applications is quite diverse, which leads to an extremely large dictionary. Writers are suspected to intentionally use many vague terms and expressions in order to avoid narrowing the scope of the invention. Combinations of general terms often have a special meaning that also has to be captured. Patent applications further contain many acronyms and new terminology. Furthermore, in order to pass the patent examination, writers tend to develop their own terminologies, which can cause serious terms mismatch problems [6]. The combination of these factors make prior-art search significantly different with other search tasks, such as web search.

Current prior-art retrieval systems use keyword search, where patent users (e.g. patent examiners or attorneys) extract relevant keywords from query patents, particularly from the claim field, to formulate their queries [9–11]. Here, *query patents* are new applications, which are examined for novelty. The success of the search highly depends upon the quality of queries terms selected from query patents. However, due to the above mentioned problems, selecting relevant keywords can be a difficult task. Some documents are retrieved by many queries, whereas others may never show up within the top-c documents retrieved for any reasonable query up to a certain length.

Retrievability measurement [1] is used to analyze the bias of retrieval systems and their capability of potentially retrieving each document in the corpus. Bias of retrieval systems denotes the characteristic of a system to give preference to certain features of documents, when it ranks results of any given query. For example, *PageRank* favors *"popular"* documents by evaluating the number of in-links of web pages in addition to pure content features. Similarly, *TFIDF* and *OKAPI-BM25* favor large terms frequencies and thus longer documents over shorter ones. Given the fact that only a limited number of top-ranked documents can be evaluated for any given query, we may arrive at the situation, that some set of documents cannot be retrieved for any plausible query by e.g., using all possible queries up to a certain length. To measure retrievability, a large set of potential queries (e.g. all combinations of all important keywords up to a pre-specified query length) are passed to a retrieval system, and the number of documents that can and that cannot be retrieved is evaluated. The resulting figure provides an estimate on the amount of bias introduced by a certain retrieval system, indicating its suitability for recall-oriented applications.

With prior-art queries generated only from query patents, our experiments using retrievability measurement indicate a large bias toward a subset of patents in state of the art retrieval systems [2]. A large subset of patents either has very low retrievability scores or could not be accessible via any query. In order to increase the coverage of prior-art queries, in this paper we reformulate (expand) queries using Query Expansion (QE) with Pseudo Relevance Feedback (PRF) [3]. Prior-art queries extracted from query patents may not contain all terms. Therefore, missing terms can be extracted from PRF documents. For selecting better patents for PRF, in this paper we propose a novel approach where relevant patents for PRF are identified based on their similarity with query patents via specific terms. The success of this approach, highly depends upon the selection of those terms from query patents, that produce the best PRF candidates. For example, those terms which appear closely with terms of prior-art queries in the same *claim*, *paragraph*, *sentence* or *phrase* can identify better patents for PRF as compared to using all terms of a query patent. This term selection problem can be considered a term classification problem. In this paper, we try to separate positive terms from others - according to their potential impact on the retrieval effectiveness. Finally, in order to evaluate how far this novel prior-art retrieval approach can increase the retrievability of patents in collection, we compare it with state of the art retrieval systems including different QE approaches. Our experiments indicate that patent retrievability can be improved significantly using QE with more sophisticated PRF patents selection.

The remainder of the paper is organized as follows. Section 2 reviews related work in the field of prior-art patents retrieval. In Section 3, we introduce retrievability measurement for recall-oriented applications, which is used as a basis of our experiments. In Section 4, we explain the working of our prior-art retrieval approach, introducing the features used for classification (Section 4.1), learning accuracy (Section 4.2) and expanding queries using language modeling approach (Section 4.3). In Section 5, we evaluate the performance of our retrieval approach with other state of the art retrieval models using collection of patent documents under retrievability measurement.

2 Related Work

Osborn et al. [17] introduce a system that integrates a series of shallow natural language processing techniques into a vector-based document information retrieval system for searching relevant patents. Their methods are mainly based on the patterns of part-of-speech tags, where firstly phrases from the patents are extracted, and then these phrases are used as indexed features. Their methods use all contents of a patent for constructing the query vector, but ignore the structure information in the patent. Larkey [12] uses a probabilistic information retrieval system for searching and classifying US patents. In their approach, instead of searching in full patents, they select certain sections *(patent fields)* and portions of sections for reducing text and selecting dominant terms. Next, weights to different terms in these reduced patents are assigned based on their relative importance to different sections and term frequencies for efficient retrieval. Mase et al. [15] propose a two-stage retrieval strategy for patents search. In stage1, general text analysis and retrieval methods are applied to improve recall; while in stage2, top c patents retrieved from stage1 results are rearranged for improving precision by applying different text analysis and retrieval methods using the claim field.

Fujii [8] applies link analysis techniques to the citation structure for efficient patent retrieval. In their method, they first perform text based retrieval for obtaining *top* c patents, and then citation scores of these *top* c patents are computed based on PageRank and topic-sensitive citation-based methods. Finally, both the text-based and citation-based scores are combined for better ranking of these patents. Custis et al. [6], evaluate query expansion methods for legal domain applications. For this purpose, they systematically introduce query document terms mismatch into a corpus in a controlled manner and then measure the performance of retrieval systems as the degree of terms mismatch changes. A recent approach on patent retrieval considers a novel search scenario, in which users can pose full patents as a query instead of selecting relevant keywords from them in prior-art search [19]. They also explore the effect of different fields of patents as a search feature and further consider how these fields can be combined with learning techniques. A considerable improvement in relevance judgment results is reported using this approach.

3 Retrievability Measurement

"Retrievability" measures [1], how likely each and every document $d \in D$ can be retrieved within the top c ranked results for all queries in Q. More formally, retrievability $r(d)$ of $d \in D$ can be defined as follows.

$$r(d) = \sum_{q \in Q} f(k_{dq}, c) \tag{1}$$

Here, $f(k_{dq}, c)$ is a generalized utility/cost function, where k_{dq} is the rank of d in the result set of query $q \in Q$, c denotes the maximum rank that a user is willing to proceed down the ranked list. The function $f(k_{dq}, c)$ returns a value of 1 if $k_{dq} \leq c$, and 0 otherwise.

Retrievability inequality can be further analyzed using the **Lorenz Curve**. Documents are sorted according to their retrievability score in ascending order, plotting a cumulative score distribution. If the retrievability of documents is distributed equally, then the Lorenz Curve will be linear. The more skewed the curve, the greater the amount of inequality or bias within the retrieval system. The **Gini coefficient** G is used to summarize the amount of bias in the Lorenz Curve, and is computed as follows.

$$G = \frac{\sum_{i=1}^{n}(2 \cdot i - n - 1) \cdot r(d_i)}{(n-1)\sum_{j=1}^{n} r(d_j)} \tag{2}$$

where $n = |D|$ is the number of documents in the collection. If $G = 0$, then no bias is present because all documents are equally retrievable. If $G = 1$, then only one document is retrievable and all other documents have $r(d) = 0$. By comparing the Gini coefficients of different retrieval methods, we can analyze the retrievability bias imposed by the underlying retrieval system on the given document collection.

Retrievability of patents is analyzed on a large collection of queries. Clearly, it is impractical to calculate the absolute $r(d)$ scores because the set of all query terms Q would be extremely large and require a significant amount of computation time as each query would have to be issued against the index for a given retrieval system. In order to perform measurements in a practical way, a subset of all possible queries is commonly used that is sufficiently large and contains relatively probable queries [1]. In query generation, we try to reflect the way patent examiners use for generating prior-art queries from the claim fields of query patents [9–11]. We first extract all frequent terms from the patents that have a term frequency greater than a minimum threshold ($\geq$3). For generating longer queries, single frequent terms are combined into *two*, *three* and *four* term combinations. For those patents that contain a large number of single frequent terms, the different term combinations become very large. To generate a similar number of queries for every patent, we put an upper bound of *90* queries generated from every patent.

4 Selecting Pseudo-Relevance Feedback Documents

In QE with PRF, it is normally assumed that the set of top documents retrieved by user queries is relevant for relevance feedback, and that learning expansion terms from these feedback documents can increase the effectiveness of search [14]. However, from the queries extracted from query patents, our retrievability results show a large bias toward some subset of patents [2, 3]. These high retrievable patents can skew the results, and due to this a large subset of patents either could become very low retrievable or could not be retrievable via any query. For selecting relevant patents for PRF, we consider a novel approach, where patents for PRF are identified based on their similarity with query patents over a subset of terms, rather than the overall document similarity. The success of this approach depends on two main factors. Firstly, appropriate terms need to be identified in the query patent via which to retrieve the best-matching documents for PRF that can help in improving retrievability during QE. As experiments below will show, these are terms that co-occur closely with the query terms stemming from the query patent. Secondly, we analyze which fields of a query patent (*title, abstract, description of patent, background summary*, and the *claim* field) should be considered for query expansion.

In a nutshell, the experimental set-up works as follows: Retrievability is analyzed for all patents in the corpus. 500 low-retrievable patents are identified to focus on improving their retrievability via PRF. For these we identify the 35 most similar documents to be used for PRF using the SMART similarity measure [16]. Using Language Modeling (LM) we then expand the queries based on the PRF documents and check whether retrievability of the original query patent has increased. The variable step is the identification of the PRF documents. Here, in each iteration we eliminate one term of the (low-retrievable) query patent, resulting in different documents being ranked high for PRF. Terms that, when removed, lead to a ranking of PRF documents that reduce retrievability by the resulting expanded queries, are obviously helpful in determining better documents for PRF. These terms serve as a training set to automatically learn which terms help in identifying better PRF documents. The characteristics of these positive terms are analyzed and described via a range of features that serve as a basis for automatically identifying them via a machine learning approach. Thus, during deployment, starting from a query patent only such terms are used to identify the best matching documents for PRF.

After applying QE using LM, we furthermore analyze, which sections of a patent these terms come from that are used for expanding queries that lead to higher retrievability. The individual steps are described in more detail below.

4.1 Query Patent Term Selection Features

Following the process outlined above, we obtain a set of terms that, when included in identifying PRF documents, lead to higher or lower retrievability. We now want to describe these terms via a set of features, focussing mostly on their positional relation to the original query terms. Our terms classification feature set thus consists of several measures that capture the proximity of terms in query patent with all terms in the original queries, based on [4, 5, 18, 21]. These features measure the closeness or compactness of the selected QP term with terms in the query. The underlying intuition is that, the more compact the terms are, the more likely it is that they are topically related, and thus higher the possibility that the terms help improve PRF. The feature set consists of 6 features, which are further computed seperately on all 6 individual fields of patents *(Title*, *Abstract*, *Claim*, *Background Summary*, *Description*, and *wholepatent)* capturing the positional relationship between the terms in the initial query and in the documents, plus a number of other features capturing the importance of candidate terms for query expansion. The total dimensionality of the feature set is 42.

The following sample query patent $\widehat{QP}$ will be used to explain how each of the features explained below can be calculated for a query $\widehat{q}$ containing query terms a and b with query patent term m.

Term Positions =	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Query Patent ($\widehat{QP}$) =	a	b	e	m	a	b	s	n	x	h	i	a	j	b	k	m
Query ($\widehat{q}$) =	a	b														

(f1): Average Minimum Distance to Single Query Term: This feature is defined as the average of the shortest distance between the occurrences of term t in QP with the terms of the query [21]. This feature rewards terms of a query patent that appear very close to query terms, e.g. in the same *phrase*, *sentence*, *paragraph* or *claim*.

Let $q = \{q_1, q_2, ..., q_m\}$ be the set of different query terms in a query q. $O_{q_i} = \{o_{i_1}, o_{i_2}, ..., o_{i_n}\}$ is the set of term occurrence positions of the query term q_i in QP p. $PD(q_i, t; p)$ denotes the distance between query term q_i and term t in p. Following [18], $f1$ is the distance between the closest occurring positions of term q_i and t, and can be measured through their occurring positions in p.

$$f_1(t) = \frac{\sum_{q_i \in q} PD(q_i; t|p)}{|q|} \tag{3}$$

$$PD(q_i; t|p) = min_{o_{i_k} \in O_{q_i}, o_{t_k} \in O_t} \left\{ \text{abs}(o_{i_k} - o_{t_k}) \right\} \tag{4}$$

where $|q|$ is the length of query q, and O_{q_i} and O_t are the sets of occurrences positions of terms q_i and t in p. In the example, the minimum value of $f1(m)$ using this feature with query terms a and b is $= ((5-4) + (6-4))/2 = 1.5$.

(f2): Pair-wise Terms Proximity Based on Minimum Distance: $f1$ captures the average minimum QP term distance with single terms of a query. However, a better feature could be with pairs of query terms [21]. $f2$ considers the minimum distance between the selected QP term and pairs of terms in the query. This feature is calculated as follows.

$$f_2(t) = min_{\hat{p}(q_i, q_j) \in q, t \neq q_i, t \neq q_j, q_i \neq q_j} \ \{PD2(q_i, q_j; t|p)\} \tag{5}$$

$$PD2(q_i, q_j; t|p) = min_{o_{i_k} \in O_{q_i}, o_{j_k} \in O_{q_j}, o_{t_k} \in O_t} \left\{ PD(q_i; t|p) + PD(q_j; t|p) \right\} \tag{6}$$

$\hat{p}(q_i, q_j)$ enumerates all possible terms pairs in q. $PD2(q_i, q_j; t|p)$ denotes the pair-wise distance between terms pair q_i and q_j in the query and term t in p. Similar to $f2$, it is the distance between the closest occurring positions of terms q_i, q_j and t. In the example, the minimum value of $f2(m)$ using this feature is $= (5-4) + (6-4) = 3$.

(f3): Pair-wise Terms Proximity Based on Average Distance: Instead of relying only on the minimum distance, this feature calculates the average distance between pairs of query terms q_i,q_j and term t in p [5]. This feature promotes those terms of a query patent that consistently occur closer to query term pairs in localized areas, e.g. in the same *claim*, *paragraph*, *sentence* or *phrase*. Given the set $\hat{p}(q_i, q_j)$ of all possible query terms pairs, the value of this feature can be calculated as follows. In the example, since there is only one query terms pair, therefore the pair-wise average proximity distance of $f3(m)$ using this feature is $= ((5-4) + (6-4))/1 = 3$.

$$f_3(t) = \frac{\sum_{\hat{p}(q_i, q_j) \in q, t \neq q_i, t \neq q_j, q_i \neq q_j} \ \{PD2(q_i, q_j; t|p)\}}{|\hat{p}(q_i, q_j)|} \tag{7}$$

(f4): Query Terms Difference Average Position: This feature considers the difference between the average positions of individual query terms with terms t in p. This feature first calculates the average positions of individual terms of query with t using the position vectors, and then these average positions are used for calculating average proximity positions [5]. This feature captures where terms of query and t are occurring together. In the example, the value of $f4(m)$ using this feature is $= abs((1+5+12)/3-(4+16)/2)+abs((2+6+14)/3+(4+16)/2)/2 = abs(6-10)+abs(7.33-10)/2 = 3.34$.

$$f_4(t) = \frac{\sum_{q_i \in q} abs\{\sum_{o_{i_k} \in O_{q_i}} \frac{o_{i_k}}{|O_{q_i}|} - \sum_{o_{t_k} \in O_t} \frac{o_{t_k}}{|O_t|}\}}{|q|} \tag{8}$$

(f5): Co-Occurrence with Single Query Term: In learning better expansion terms using classification approach all terms are considered useful for expansion, that co-occur frequently with query terms [4]. f(5) captures this co-occurrence.

$$f_5(t) = log \frac{1}{|q|} \sum_{q_i \in q} \frac{c(q_i, t|p)}{tf(q_i|p)} \tag{9}$$

where $c(q_i, t|p)$ is the frequency of co-occurrences of query term q_i and the term t within text windows of p. $tf(q_i|p)$ denotes the term frequency of q_i in p. The window size is empirically set to 20 terms. In the example, if window size is set to 3 then the number of co-occurrence of term m with query term a is 1 and with query term b is 3. Using equation 9, the value of $f5(m)$ is $= log(1/2 * ((1/3)+(3/3))) = -0.41$.

(f6): Co-occurrence with Pairs Query Terms: The previous feature considers only the co-occurrence of term t with individual terms of query q. This feature captures a stronger co-occurrence relation of term t with pairs of terms of the query [4]. Given the set $\hat{p}(q_i, q_j)$ of all possible pairs of query terms, the value of this feature can be calculated as follows.

$$f_6(t) = log \frac{1}{|\hat{p}(q_i, q_j)|} \sum_{\hat{p}(q_i, q_j) \in q, t \neq q_i, t \neq q_j, q_i \neq q_j} \frac{c(q_i, q_j, t|p)}{tf(q_i|p) + tf(q_j|p)} \tag{10}$$

$c(q_i, q_j, t|p)$ denotes the frequency of co-occurrences of term t with terms pair q_i and q_j of query q, within text windows of p. The window size is empirically set to 20 terms. In the example, if window size is set to 3 then the number of co-occurrence of term m with query term pairs a and b is 1. Using equation 10 the value of $f6(m)$ is $= log(1/2 * (1/6)) = -2.48$.

Other Features: Some other features that we consider for term classification purpose and use in our experiments are: (a) Sum of query terms and t in p, (b) Product of term frequencies of individual query terms and t in p. If the value of this feature is high, the probability of closer occurrence of term t with query terms will be high. (c) $fullcover(q, t, p)$ is the length of the patent segment that covers all occurrences of query terms with term t, (d) idf of term t, (e) $tfidf$ value of term t, and (f) length of patent.

4.2 Terms Classification and PRF Patents Selection

We use neural networks with radial basis function (RBF) to train a model for identifying terms that help in returning documents for PRF that improve the overall retrievability of a given patent. PRF documents are selected calculating the similarity between the query patent (with different terms removed) and the patent corpus using the SMART similarity measure [16], using the top 35 patents for PRF with subsequent query expansion via LM as described below.

The calculation of the majority of features describing the terms is based upon the proximity distribution of query terms and terms in patents. Thus 3000 queries that can retrieve the 500 low-retrievable patents are randomly selected for training the model, and further 800 queries are used for testing the accuracy of learning model. Using the RBF classifier, we obtain a classification accuracy of 72%.

4.3 Expanding Queries

We use *Language Modeling* (LM) [13] to select the most dominant terms from PRF patents for expanding the queries. Under LM each term w is ranked according to the sum of divergences between its prevalence in each relevance feedback patent it occurs and the importance of the term in the whole collection.

$$score(w) = \sum_{d \in K} P(d)P(w|d)P(q|d) \tag{11}$$

K is the set of PRF patents selected using the approach above. We assume that $P(d)$ is uniform over the set. After this estimation, the most $e = 35$ terms (words) from $P(w|K)$ are chosen for expanding the queries. The values $P(w|d)$ and $P(q|d)$ can be calculated following Equations 12 and 13.

$$P(q|d) = \prod_{i=1}^{m} P(q_i|d) \tag{12}$$

where q_i is the i^{th} query term, m is the number of terms in a query q, and d is a document model. Dirichlet smoothing [20] is used to estimate non-zero values for terms in the query which are not in a patent document. It is applied to the query likelihood language model as follows.

$$P(w|d) = \frac{|d|}{|d| + \lambda} P_{ML}(w|d) + \frac{\lambda}{|d| + \lambda} P_{ML}(w|D) \tag{13}$$

$$P_{ML}(w|d) = \frac{freq(w,d)}{|d|}, P_{ML}(w|D) = \frac{freq(w,D)}{|D|} \tag{14}$$

where $P_{ML}(w|d)$ is the maximum likehood estimate of a term w in document d, D is the entire collection, and λ is the smoothing parameter [20]. $|d|$ and $|D|$ are the lengths of a patent document d and collection D, respectively, $freq(w,d)$ and $freq(w,D)$ denote the frequency of a term w in d and D, respectively.

Queries Set	**Total Queries**	**Average Retrievability**	**Average Queries/Patent**
2 Terms Queries	$4,308,562$	512.83	78.27
3 Terms Queries	$2,908,972$	373.49	53.42
4 Terms Queries	$2,876,587$	282.24	51.78

Table 1. Queries sets properties used for Retrievability Measurement

Retr. Model/	**Two Terms Queries**					**Three Terms Queries**					**Four Terms Queries**				
Rank cut-off	**30**	**40**	**60**	**80**	**100**	**30**	**40**	**60**	**80**	**100**	**30**	**40**	**60**	**80**	**100**
TFIDF	0.48	0.51	0.50	0.49	0.49	0.50	0.51	0.51	0.48	0.48	0.63	0.62	0.62	0.62	0.62
BM25	0.58	0.54	0.51	0.50	0.50	0.56	0.53	0.52	0.50	0.50	0.67	0.65	0.64	0.63	0.63
Exact Match	0.79	0.77	0.75	0.71	0.67	0.90	0.87	0.83	0.76	0.70	0.91	0.88	0.84	0.78	0.72
LM	0.53	0.53	0.53	0.52	0.51	0.62	0.62	0.63	0.61	0.60	0.71	0.71	0.72	0.70	0.68
QP-with-TS	**0.39**	**0.39**	**0.38**	**0.38**	**0.37**	**0.38**	**0.37**	**0.37**	**0.36**	**0.36**	**0.54**	**0.53**	**0.53**	**0.52**	**0.51**
QP-without-TS	0.50	0.57	0.55	0.55	0.54	0.58	0.55	0.55	0.55	0.54	0.66	0.65	0.63	0.63	0.62

Table 2. Gini coefficient *(*G) values of different retrieval systems, for different *rank cut-off factors (c)*. As c increases, G steadily decreases indicating that lower bias is experienced when considering longer ranked lists.

5 Experiments

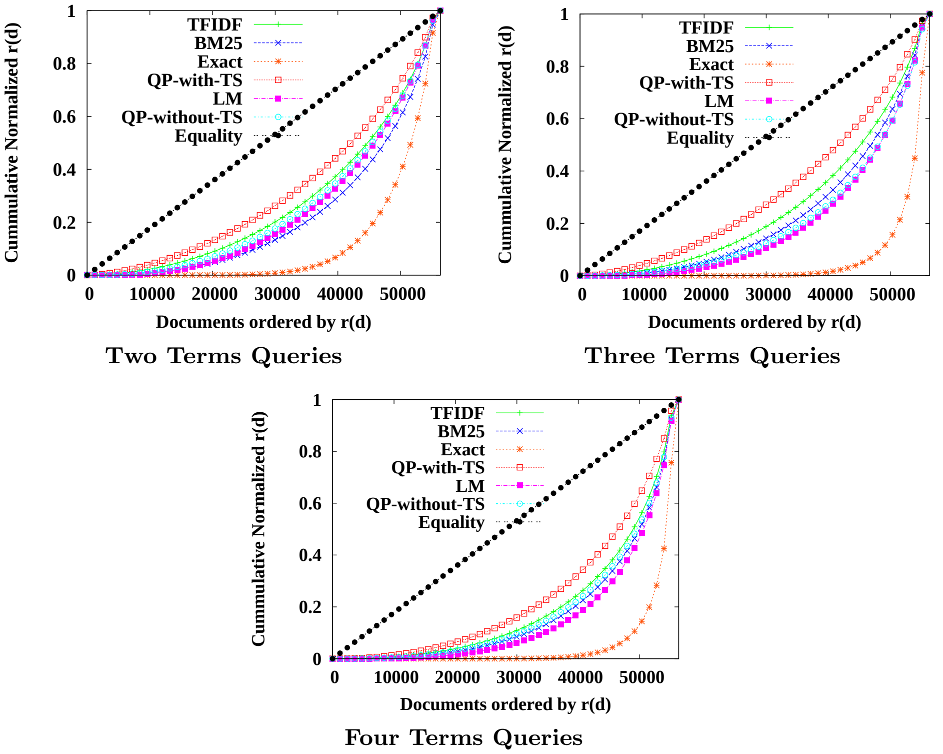
For experiments, we use a collection of freely available patents from the US patent and trademark office, downloaded from *(http://www.uspto.gov/)*. We collect all patents that are listed under *United State Patent Classification* (USPC) classes *422 (Chemical apparatus and process disinfecting, deodorizing, preserving, or sterilizing)*, and *423 (Chemistry of inorganic compounds)*. There are a total of $54,353$ patents in our collection, with an average patent size of $3,317.41$ terms (without stop words removing).

In query generation, we only consider the *claim field* of every patent, as this is the section that most professional patent searchers use as their basis for query formulation [9, 11]. For retrieval we index the full text of all patents *(Title, Abstract, Background Summary, Claim, Description)*. This reflects the default setting in a standard full text retrieval engine. Before indexing, we remove stop words and stem the words. For indexing and querying we use the *Apache LUCENE*[1] IR toolkit. Four state-of-the art retrieval systems along with our proposed prior-art retrieval approach are used for evaluating retrievability inequality and prior-art queries coverage. The retrieval systems that we evaluate are; **TFIDF**, OKAPI retrieval function **(BM25)**, **Exact Match** model, Language Modeling with term smoothing **(LM[2])** [20], our PRF patents selection approach based on query patents similarity with terms selection **(QP-with-TS)**, and PRF patents selection based on query patents similarity using all terms of query patents **(QP-without-TS)**. In QP-with-TS we select terms from all fields of query patents. We create a set of queries from each query patent using two terms as well as three and four terms combinations (Section 3). Table 1 shows the properties of different query sets.

Figure 1 shows the retrievability inequality of different retrieval systems using Lorenz Curves with *rank cut-off* factor of 30. For other *rank cut-off* parameters configurations, we show the retrievability inequality of different retrieval approaches using

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

[2] In LM, top 40 patents are used for PRF, and top 35 terms from PRF patents are used for expansion.



Two Terms Queries

Three Terms Queries

Four Terms Queries

Fig. 1. Lorenz Curves visualizing the retrievability inequality of different retrieval systems, with *rank cut-off factor (c)* = 30. *Equality* refers to an optimal system which has no bias.

Gini coefficient values in Table 2. We can see that as the *rank cut-off* factor increases, the Gini coefficient tends to decrease slowly on all different queries sets. This indicates that the retrievability inequality within the collection is mitigated by the willingness of the user to search deeper down into the ranking, as expected. If users are willing to examine only the top documents, then they will face a greater degree of retrieval bias.

On almost all rank cut-off factors with different lengths of query sets, the Lorenz curve and Gini coefficient values of our prior-art retrieval approach *(QP-with-TS)* are less skewed and have lower Gini coefficient values than other retrieval approaches. This indicates that our prior-art retrieval approach makes individual patents more easily retrievable. The *Exact Match* method, which is widely used in professional patent retrieval systems, consistently shows the worst performance. In *LM* approach, considering top documents in queries relevant for PRF also does not perform too well with respect to providing potential access to all patents. The performance results of *QP-without-TS* are worse than *QP-with-TS*. This happens because in patent retrieval domain a large diversity exists in patent lengths. Therefore, when calculating PRF patents similarity using all terms of query patent *(whole patent)*, longer patents have more chance that they can increase their similarity values with query patents. This results in a higher bias in retrievability scores. The retrievability results of *TFIDF* are better than other retrieval approaches.

Patent Field/	**Two Terms Queries**					**Three Terms Queries**					**Four Terms Queries**				
Rank cut-off	**30**	**40**	**60**	**80**	**100**	**30**	**40**	**60**	**80**	**100**	**30**	**40**	**60**	**80**	**100**
Whole Patent	**0.39**	**0.39**	**0.38**	**0.38**	**0.37**	**0.38**	**0.37**	**0.37**	**0.36**	**0.36**	**0.54**	**0.53**	**0.53**	**0.52**	**0.51**
Description	0.44	0.44	0.44	0.43	0.43	0.40	0.40	0.39	0.39	0.39	0.56	0.55	0.55	0.54	0.53
Claim	0.51	0.51	0.50	0.50	0.49	0.46	0.46	0.45	0.44	0.44	0.60	0.59	0.59	0.58	0.58
Abstract	0.47	0.47	0.46	0.46	0.46	0.44	0.44	0.44	0.44	0.43	0.58	0.58	0.57	0.57	0.57
Summary	0.47	0.47	0.46	0.46	0.44	0.43	0.43	0.43	0.42	0.42	0.59	0.58	0.57	0.56	0.55

Table 3. Effect of different fields of patents on Gini coefficient for query patent terms selection, with different *rank cut-off factors (c)*.

The Gini coefficient values of Table 3 summarize our analysis which field of query patent is better for terms extraction in *QP-with-TS* approach. Given the results shown in Table 3, it is clear that query patent terms extracted from the description field can better decrease the retrievability inequality as compared to other fields. However, the performance results of selecting terms from the whole patent are much better than individual fields' results. From the results, it is interesting to notice, that the performance results of summary and abstract fields are much better than claim field results, which is mostly used in prior-art retrieval. This is because writers in claim fields, for protecting the invention, may tend to use language that extends the scope of patents, which may create serious term mismatch problems. Other fields like description, abstract and background summary are mainly written for the technical use, where authors briefly try to describe their description of invention relative to the field.

6 Conclusions

This paper evaluates the coverage of prior-art queries extracted from query patents using retrievability measurement. In experiments, retrievability shows large bias toward subset of patents using state of the art retrieval systems. Due to bias, a large number of patents either have very lower retrievability scores or could not be retrievable via any query. The main reason behind low retrievability is the presence of large terms mismatch in patents. For increasing the retrievability of patents, we consider query expansion approach with pseudo relevance feedback (PRF). In this way, missing terms of query patents are retrieved from related PRF patents. For better identifications of PRF patents, a novel approach is presented where patents for PRF are identified based on their similarity with query patents via selected terms. We identify relevant terms from query patents based on their proximity distribution with prior-art queries. Using this approach, an increase in the retrievability of individual patents is obtained, which indicates that this prior-art retrieval approach provides better opportunity for retrieving individual patents in search space.

7 Acknowledgments

We would like to thank anonymous reviewers for providing helpful comments on improving the quality of paper. We would also like to acknowledge Higher Education Commission of Pakistan (HEC) and Austria Exchange Service (OAD) for providing scholarship and support to Shariq Bashir for conducting his PhD.

References

1. L. Azzopardi, V. Vinay. Retrievability: an evaluation measure for higher order information access tasks. *In Proc. of CIKM '08*, pages 561–570, October 26-30, 2008, Napa Valley, California, USA.
2. S. Bashir, A. Rauber Analyzing Document Retrievability in Patent Retrieval Settings. *In: Proc. 20th Intl Conf on Database and Expert Systems Applications (DEXA 2009)*, pp. 753-760, Aug 31 - Sep 4 2009, Linz, Austria.
3. S. Bashir, A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. *In Proc. of CIKM '09*, pages 1863–1866, November 2-6, 2009, Hong Kong, China.
4. G. Cao, J-Y. Nie, J. Gao, S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. *In Proc. of SIGIR'08*, pages 243-250, 2008, Singapore, Singapore.
5. R. Cummins, C. O'Riordan. Learning in a pairwise term-term proximity framework for information retrieval. *In Proc. of SIGIR'09*, 2009, Pages 251–258, Boston, MA, USA.
6. T. Custis, K. Al-Kofahi. A new approach for evaluating query expansion: query-document term mismatch. *In Proc. of SIGIR '07*, pages 575–582, July 23-27, 2007, Amsterdam, The Netherlands.
7. C. J. Fall, A. Torcsvari, K. Benzineb, G. Karetka. Automated categorization in the international patent classification. *ACM SIGIR Forum*, Volume 37, Issue 1 (Spring 2003), Pages 10–25.
8. A. Fujii. Enhancing patent retrieval by citation analysis. *In Proc. of SIGIR'07*, 2007, Pages 793–794, Amsterdam, The Netherlands.
9. H. Itoh, H. Mano, Y, Ogawa. Term distillation in patent retrieval. *ACL '03: Proceedings of the ACL-2003 workshop on Patent corpus processing*, 2003, pp. 41–45, Sapporo, Japan.
10. K. Konishi. Query terms extraction from patent document for invalidity search. *In Proc. of NTCIR '05: NTCIR-5 Workshop Meeting*, 2005, Tokyo, Japan.
11. K. Konishi, A. Kitauchi, T. Takaki. Invalidity patent search system at NTT data. *In Proc. of NTCIR-4 Workshop Meeting*, 2004, Tokyo, Japan.
12. L. S. Larkey. A Patent Search and Classification System. *In Proc. of 4th ACM Conference on Digital Libraries*, 1999, Pages 179–187, Berkeley, CA, USA.
13. V. Lavrenko, W. B. Croft. Relevance based language models. *In Proc. of SIGIR '01*, 2001, Pages 120–127, New Orleans, Louisiana, USA.
14. K. S. Lee, W. B. Croft, J. Allan. A cluster-based resampling method for pseudo-relevance feedback. *In Proc. of SIGIR'08*, pages 235–242, 2008, Singapore.
15. H. Mase, T. Matsubayashi, Y. Ogawa, M. Iwayama, T. Oshio. Proposal of two-stage patent retrieval method considering the claim structure. *ACM Transactions on Asian Language Information Processing*, 2005, Pages 190–206, Volume 4, Issue 2 (June 2005).
16. M. Murata, T. Kanamaru, T. Shirado, H. Isahara. Using the k-nearest neighbor method and SMART weighting in the patent document categorization subtask at NTCIR-6. *In Proc. NTCIR-6 Workshop Meeting*, 2007, Tokyo, Japan.
17. M. Osborn, T. Strzalkowski, M. Marinescu. Evaluating Document Retrieval in Patent Database: A Preliminary Report. *In Proc. of CIKM '97*, 1997, Pages 216–221, Las Vegas, Nevada, USA.
18. T. Tao, C. Zhai. An exploration of proximity measures in information retrieval. *In Proc. of SIGIR'07*, 2007, Pages 295–302, Amsterdam, The Netherlands.
19. X. Xue, W. B. Croft. Transforming patents into prior-art queries. *In Proc. of SIGIR'09*, pages 808–809, 2009, Boston, MA, USA.
20. C. Zhai, J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, Volume 22(2), pages 179–214, 2004.
21. J. Zhao, Y. Yun. A proximity language model for information retrieval. *In Proc. of SIGIR'09*, 2009, Pages 291–298, Boston, MA, USA.