A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications

Contribution ID: 5

ABSTRACT

Patents are used by legal entities to legally protect their inventions and represent a multi-billion dollar industry of licensing and litigation. In 2013, 302,948 patent applications were approved in the US alone – a number that has doubled in the past 15 years and which makes prior art search a daunting, but necessary task in the patent application process. In this work, we seek to investigate the efficacy of prior art search strategies from the perspective of the inventor who wishes to assess the patentability of their ideas prior to writing a full application. While much of the literature inspired by the evaluation framework of the CLEF-IP competition has aimed to assist patent examiners in assessing prior art for complete patent applications, less of this work has focused on patent search with queries representing partial applications. In the (partial) patent search setting, a query is often much longer than in other standard IR tasks, e.g., the description section may contain hundreds or even thousands of words. While the length of such queries may suggest query reduction strategies to remove irrelevant terms, intentional obfuscation and general language used in patents suggests that it may help to expand queries with additionally relevant terms. To assess the trade-offs among all of these pre-application prior art search strategies, we comparatively evaluate a variety of partial application search and query reformulation methods. Among numerous findings, querying with a full description, perhaps in conjunction with generic (non-patent specific) query reduction methods, is recommended for best performance. However, we also find that querying with an abstract represents the best trade-off in terms of writing effort vs. retrieval efficacy (i.e., querying with the description sections only lead to marginal improvements) and that for such relatively short queries, generic query expansion methods help.

Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Search and Retrieval

General Terms: Algorithms, Experimentation.

Keywords: Query Reformulation, Patent Search.

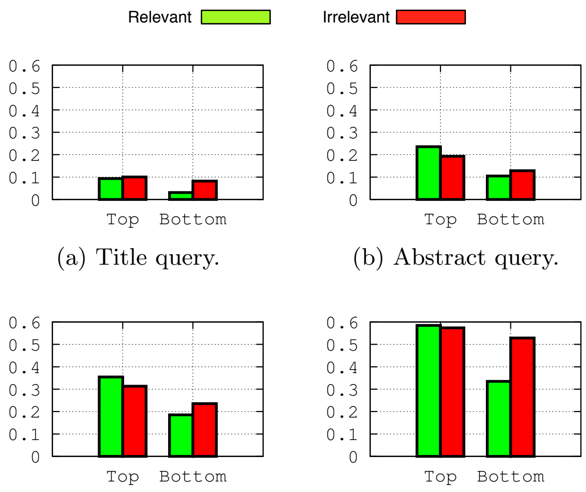
1. INTRODUCTION

Patents are used by legal entities to legally protect their inventions and represent a multi-billion dollar industry of licensing and litigation. In 2013, 302,948 patent applications were approved in the US alone[1], a number that has doubled in the past 15 years. Given that a single existing patent may invalidate a new patent application, helping inventors assess the patentability of an idea through a patent prior art search before writing a complete patent application is an important task.

Patent prior art search involves finding previously granted patents that may be relevant to a new patent application. The objective and challenges of standard formulations of patent prior art search are different from those of standard text and web search since [12]: (i) queries are (partial) patent applications, which consist of documents with hundreds or thousands of words organized into several sections, while typical queries in text and web search constitute only a few words; and (ii) patent prior art search is a recall-oriented task, where the primary focus is to retrieve all relevant documents at early ranks, in contrast to text and web search that are precision-oriented, where the primary goal is to retrieve a subset of documents that satisfy the query intent. Another important characteristic of patent prior art search is that, in contrast to scientific and technical writers, patent writers tend to generalize and maximize the scope of what is protected by a patent and potentially discourage further innovation by third parties, which further complicates the task of formulating effective queries. For instance, abstract and vague terms are sometimes pre-referred to concrete ones, e.g., recording means vs. recording apparatus; resources vs. battery life; machines located at point of sale locations vs. vending machines, etc.

While much of the literature inspired by the evaluation framework of the CLEF-IP competition has aimed to assist patent examiners in assessing prior art for complete patent applications, less work has focused on assessing the patentability of inventions before writing a full patent application. Furthermore, prior art search with queries that represent unfinished patent applications is generally desirable, since writing a full application is time-consuming and costly, especially if lawyers are hired to assist. Hence, in this paper we consider only sections which are more likely to be written by the inventor namely the title, the abstract, the description section, and an extended abstract (which we consider as the 5 first paragraphs of the description section),

[1] http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm



(a) Title query. (b) Abstract query.

(c) Extended Abstract (d) Description query.

Figure 1: Average Jaccard similarity between fields of topics and the corresponding (ir)relevant documents for different sets of top and bottom performing queries.

while we consider the claims section to be more likely to be written by a patent attorney.

To assess the difficulty of querying with partial patent applications, we refer to Figure 1. Here we show an analysis of the average Jaccard similarity[2] between different queries (representing the title, abstract, the extended abstract or descriptions of partial patent applications) and the labeled relevant (all) and irrelevant documents (top 10 irrelevant documents ranked by BM25 [23]). We show results for the top 100 and bottom 100 queries (100 queries that perform the best, and 100 queries that perform the worst) of CLEP-IP 2010 evaluated according to Mean Average Precision (MAP). Note that while the title section is usually composed of an average of six terms, the other sections are longer, ranging from ten to thousands of terms. There are three notable trends here: (i) term overlap increases from title to description since the query size grows accordingly; (ii) the bottom 100 performing queries tend to have much smaller term overlap with the relevant documents than the top 100 queries; and (iii) even in the best case of querying with very long description sections, the average term overlap indicates many terms of relevant documents are not found in the query.

Similar observations in the general patent prior art search literature [14] have led to a research focus on query reformulation. Therefore, we suggest an investigation of *query reformulation* [3] methods as a means for improving the term overlap between queries that represent partial patent applications and relevant documents, with the objective of assessing not only the performance of standard query reformulation methods, but also the effectiveness of query reformulation methods that exploit patent-specific characteristics.

In summary, to aid the patent inventor in developing an effective pre-application prior art search strategy, we seek to answer the following questions:

- What parts of a patent application should a patent inventor write first to achieve effective prior art search? What are the trade-offs in section writing effort vs. the retrieval performance of querying with that section? We assume the writing effort to be a function of word number.

- In query expansion, which patent section is the best source for term expansion?

- For query reformulation (both query expansion and reduction), which methods work best, and in which settings? Do patent-specific reformulation methods offer advantages over more generic IR reformulation methods?

To answer these questions, we perform a thorough comparative analysis of partial patent application query strategies and reformulation methods on the CLEF-IP patent prior art search datasets.

The rest of the paper is organized as follows: in Section 2, we present a variety of generic and patent-specific query reformulation methods; in Section 3, we present the evaluation results and analysis to answer the above questions; in Section 4 we discuss the related work on other patent-specific query reformulation methods, which are not considered in this paper; and in Section 5, we conclude with key observations from the evaluation that lead to concrete recommendations for patent prior art search with partial applications.

[2]The Jaccard similarity is used to measure the term overlap between two sets. Before applying the Jaccard similarity, patent-specific stopwords were removed, as suggested by [15].

2. QUERY REFORMULATION FOR PATENTS

Query Reformulation is the process of transforming an initial query Q to another query Q'. This transformation may be either an expansion or a reduction of the query. *Query Expansion* (QE) [6] enhances the query with additional terms likely to occur in relevant documents. Hence, given a query representation Q, QE aims to select an optimal subset T_k of k terms, which are relevant to Q, then build Q' such as $Q' = Q \cup T_k$. As for *Query Reduction* (QR) [11], it is the process that reduces the query such that superfluous information is removed. Hence, given a query representation Q, QR aims to select an optimal subset $T_k \subset Q$ of k terms, which are relevant to Q, then build Q' such as $Q' = T_k$.

The outline of the following subsections is as follows: Section 2.1 motivates query reduction for patent prior art search. Then, we describe the standard and patent-specific query reformulation methods that we evaluate in Section 3.

2.1 Utility of Query Reduction for Patents

While the title is usually composed by an average of six terms, the other sections are longer, ranging from ten to thousands of terms. Therefore, we investigate the impact of query reduction methods only when querying with long sections such as abstract, extended abstract or description.

Table 2.1 provides insight into the utility of query reduction for the abstract section of the Topic PAC-1019 from the CLEF-IP 2010 data collection. The baseline query, which is the original query (provided in the header row) after stemming and patent specific stopword removal, had an Average Precision (AP) of 0.280 and a Patent Retrieval Evaluation

Table 1: Sample of terms removed from the abstract section of CLEP-IP2010 Topic PAC-1019.

Topic: PAC-1019	
Abstract: A 5-aminolevulinic acid salt which is useful in fields of microorganisms, fermentation, animals, medicaments, plants and the like; a process for producing the same; a medical composition comprising the same; and a plant activator composition comprising the same.	

Term removed	P@5	P@10	R@10	AP	PRES
composit...	**0.600**	0.300	0.428	**0.360**	**0.829**
activ...	0.400	0.300	0.428	0.277	**0.809**
anim...	**0.600**	0.300	0.428	**0.345**	**0.798**
produc...	0.400	0.300	0.428	**0.286**	**0.797**
ferment...	0.200	0.300	0.428	**0.283**	**0.796**
microorgan...	**0.600**	0.300	0.428	**0.333**	**0.793**
compris...	0.400	0.300	0.428	0.271	**0.790**
medica...	0.400	0.300	0.428	**0.297**	**0.789**
medic...	0.400	0.300	0.428	**0.297**	**0.787**
field...	0.400	0.300	0.428	**0.282**	**0.782**
plant...	0.200	0.200	0.285	0.114	0.774
process...	0.400	0.300	0.428	0.279	0.764
acid...	0.400	0.300	0.428	0.252	0.693
salt...	0.200	0.200	0.285	0.216	0.663
aminolevulin...	0.000	0.100	0.142	0.026	0.352
Baseline	0.400	0.300	0.428	0.280	0.777

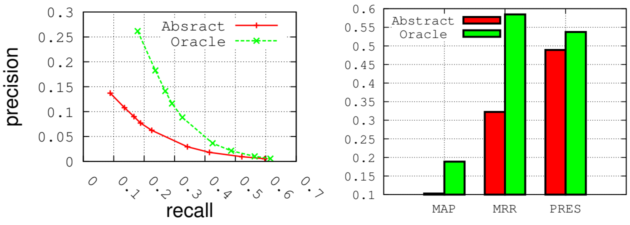


Figure 2: The utility of query reduction for 1304 abstract queries of the CLEF-IP 2010 dataset.

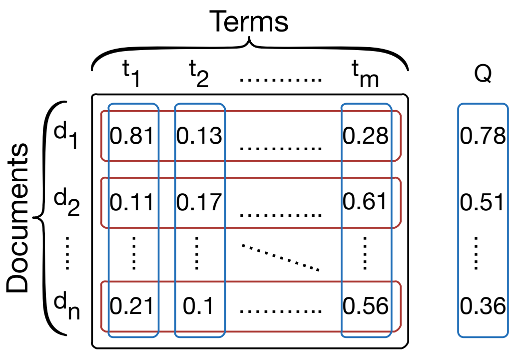


Figure 3: Notation used in MMR QE/QR.

Score (PRES)[3] [13] of 0.777 (its performance are provided in the footer row). We show the evaluation performance of the query after removing each term from the original query. The removed terms have been sorted in the order of decreasing PRES. We can observe that there are ten terms (highlighted in boldface) that if they are (individually) removed from the query, the PRES of the original long query increased.

Figure 2.1 shows the summary upper-bound performance for precision, recall, MAP, Mean Reciprocal Rank (MRR), and PRES that can be achieved for a set of 1304 abstract queries from the CLEF-IP 2010 data collection. "Baseline" refers to a probabilistic BM25 retrieval model [23] run using the Lucene search engine [21] and the original long query. "Oracle" refers to the situation where all terms with negative impact are removed from the original long query following the previous process. This gives us an upper bound on the performance that can be realized through query reduction for this set of queries. It is this statistically significant improvement in performance through query reduction that we can target for the abstract and the description sections.

2.2 Generic Query Reformulation Methods

The Rocchio Algorithm for Relevance Feedback: The Rocchio algorithm [25] is a classic algorithm of relevance feedback used mainly for query expansion. In brief, it provides a method of incorporating relevance feedback information into the vector space model representing a query [20]. The underlying theory behind Rocchio is to find a query vector $\overrightarrow{Q'}$, that maximizes similarity with relevant documents while minimizing similarity with irrelevant documents. Typically, a pseudo-relevance feedback (PRF) set of k top ranked documents obtained after an initial run of the query is considered as the set of relevant documents to build $\overrightarrow{Q'}$. We refer to this method as RocchioQE.[4]

Similarly, Rocchio can be used as a QR method. Basically, the idea is that once the Rocchio-modified query vector has been computed, it is possible to select only the terms that appear in the initial query Q and rank them using the Rocchio score and finally, select the top k terms with the highest score to build Q'. We refer to this approach as RocchioQR.

Maximal Marginal Relevance for Query Reformulation: As a general method for query reformulation, we also consider a method of "diverse" term selection — an adaptation of the *Maximal Marginal Relevance* (MMR) [5] algorithm for result set diversification. But, rather than use MMR for diverse document selection (as typically used), it is used here for diverse term selection — the hypothesis being that diverse term selection may improve coverage of relevant terms in the PRF set.

In the case of QE, we call this diversified expansion method MMR Query Expansion (MMRQE). MMRQE takes as input a PRF set, which is used to build a document-term matrix of n documents and m terms as shown in Figure 3 (the TF-IDF is used to populate the matrix for each document vector). To represent the query Q in the documents' dimension as in Figure 3, we use the BM25 or TF-IDF score between each document d_i and the query. Hence, given a query representation Q, MMRQE aims to select an optimal subset of k terms $T_k^* \subset D$ (where $|T_k^*| = k$ and $k \ll |m|$, and D is the PRD set) relevant to Q but inherently differ-

[3]The PRES metric places more emphasis on high-recall retrieval by weighting relevant documents lower in the ranking more highly than MAP.

[4]We used the LucQE module, which provides an implementation of the Rocchio method for Lucene. `http://lucene-qe.sourceforge.net/`

ent from each other (i.e., diverse). This can be achieved by building T_k^* in a greedy manner by choosing the next optimal term t_k^* given the previous set of optimal term selections $T_{k-1}^* = \{t_1^*, \ldots, t_{k-1}^*\}$ (assuming $T_0^* = \emptyset$) using the MMR diverse selection criterion:

$$t_k^* = \underset{t_k \notin T_{k-1}^*}{\arg\max} [\lambda \cos(Q, t_k) - (1-\lambda) \max_{t_j \in T_{k-1}^*} \cos(t_j, t_k)] \quad (1)$$

Here, the first cosine similarity term measures relevance between the query Q and possible expansion term t_k while the second term penalizes the possible expansion term according to its cosine similarity with any currently selected term in T_{k-1}^*. The parameter $\lambda \in [0, 1]$ trades off relevance and diversity. For MMRQE, we found that $\lambda = 0.5$ generally provide the best results, according to our experiments on the CLEF-IP training dataset collection.

For QR, we can greedily rebuild the query from scratch, while choosing diversified terms from the query itself. Here, we call this approach MMR Query Reduction (MMRQR). Formally, given a query representation Q, MMRQR aims to select an optimal subset of k terms $T_k^* \subset Q$ (where $|T_k^*| = k$ and $k < |Q|$) relevant to Q but inherently different from each other (i.e., diverse). This can be achieved by building T_k^* in a greedy manner by choosing the next optimal term t_k^* given the previous set of optimal term selections $T_{k-1}^* = \{t_1^*, \ldots, t_{k-1}^*\}$ (assuming $T_0^* = \emptyset$) using an adaptation of the MMR diverse selection criterion. Note that we use all the sections of the patent documents in the PRF set to built the document-term matrix of n documents and m terms shown in Figure 3. For MMRQR, we found that $\lambda = 0.8$ generally provide the best results in our experiments on the CLEF-IP dataset collection.

The key insight we want to highlight is that MMRQE does not select expansion terms independently as in practical usage of Rocchio, but rather it selects terms that have uncorrelated usage patterns across documents, thus hopefully encouraging diverse term selection that covers more documents for a fixed expansion budget k and ideally, higher recall.

2.3 Patent-specific Query Reformulation Methods

Synonym Sets for Patent Query Expansion: Magdy et al. [14] proposed a patent query expansion method, which automatically generates candidate synonym sets (SynSet) for terms to use as a source of expansion terms. The idea for generating the SynSet comes from the characteristics of the CLEF-IP patent collection, where some of the sections in some patents are translated into three languages (English, French, and German). They used these parallel manual translations to create possible synonyms sets. Hence, for a word w in one language which has possible translations to a set of words in another language $w_1, w_2, \ldots, w_n$, this set of words can be considered as synonyms or at least related to each other. The generated SynSet is used for query expansion in two ways: (i) The first one used the probability associated with the SynSet entries as a weight for each expanded term in the query (denoted WSynSet). Therefore, each term was replaced with its SynSet entries with the probability of each item in the SynSet acting as a weight to the term within the query. (ii) The second one neglected this associated probability and used uniform weighting for all synonyms of a given term (denoted USynSet).

Patent Lexicon for Query Expansion: Mahdabi et al. [19] proposed to build a query-specific patent lexicon based on definitions of the International Patent Classification (IPC). The lexicon is simply built by removing general and patent-specific stop-words from the text of IPC definition pages. Each entry in the lexicon is composed of a key and a value. The key is an IPC class and the value is a set of terms representing the mentioned class. Then, the lexicon is used to extract expansion concepts related to the context of the information need of a given query patent. To this end, the IPC class of the query patent is searched in the lexicon and the terms matching this class are considered as candidate expansion terms. The proposed approach tries to combine these two complementary vocabularies (i.e. terms of the query and the IPC codes). Note that all the levels of the IPC codes are used to build the lexicon. In this paper we refer to this patent query expansion method as IPC Codes.

Language Model for Query Reduction: In [8], the authors proposed a query reduction technique, which decomposes a query (a patent section) into constituent text segments and computes Language Model (LM) similarities by calculating the probability of generating each segment from the top ranked documents (PRF set). Then, the query is reduced by removing the least similar segments from the query. We refer to this method as LMQR.

IPC Codes for Query Reduction: Based on the intuition that, terms in the IPC code definition may represent ”stop-words”, especially if they are rare (infrequent in the patent application), a query can be reduced as follows: each patent application, take the definitions of the IPC codes which are associated to it. Then, (ii) rank the terms of the query according to the difference in their frequency in the query and their frequency in the class code definition. Finally, (iii) remove bottom terms of this ranking from the query (i.e. good terms are terms that occur a lot in the query, and few in the class code definition, whereas bad terms are those that occur few in the query, and a lot in the class code definition). In the evaluation section we denote this approach IPC-StopWords.

3. EXPERIMENTAL EVALUATION

In this section we first explain the experimental setup for evaluating the effectiveness of patent prior art search with partial applications. Then, we discuss the results of QE and QR methods in Sections 3.2 and 3.3 respectively.

3.1 Experimental Setup

For our experiments, we used used the Lucene IR System[5] to index the English subset of CLEF-IP 2010 and CLEF-IP 2011 datasets[6] [22, 24] with the default settings for stemming and stop-word removal. We also removed patent-specific stop-words as described in [12]. CLEF-IP 2010 contains 2.6 million patent documents, and the English test sets of CLEF-IP 2010 correspond to 1303 topic sources of partial patent application queries. We also experimented with the CLEF-IP 2011 dataset.

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

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

In our implementation, each section of a patent (title, abstract, claims, and description) is indexed in a separate field so that different sections can be used, for example, as source of expansion terms. However, when a query is processed, all indexed fields are targeted, since this generally offers best retrieval performance. We report both MAP and PRES on the top 1000 results.

3.2 Query Expansion Results

In this section, we discuss the results of partial patent queries with the QE methods described in Section 2. The configuration options and associated questions that were considered are the followings:

- **Partial patent query type:** We consider a query of a partial patent application to consist of either the title, the abstract, the extended abstract or the full description section. Recall that we do not consider the Claim section, since it is more likely to be written by a lawyer than the inventor. Hence, critical questions are: what part of a partial application an inventor should write to obtain the best search results? And what QE methods work best for each type of query?
- **Query expansion source:** We consider the abstract, claims, and description sections as different term sources to determine which section offers the best source of expansion terms, e.g., are words in the claims of particularly high value as expansion terms? We omit the use of the title as a source of expansion terms noting that this configuration performed poorly due to the relative sparsity of useful expansion terms in the titles of the PRF set.
- **Relevance model:** For initial retrieval of documents in the *pseudo-relevant* feedback set (PRF) and subsequent re-retrieval, there are various options for the relevance ranking model. In this work, we explore a probabilistic approach represented by the popular BM25 [23] algorithm, as well as a vector space model (VSM) approach, TF-IDF [26]. A natural question is which relevance model works best for query expansion for patent prior art search?
- **Term selection method:** We consider the different query expansion methods described above, i.e. RocchioQE, MMRQE, IPC Codes, WSynSet, USynSet and ask what is the best QE method for patent search?

To summarize all the results obtained over all the above configurations, Figures 4, 5, 6 and 7 show the MAP and PRES obtained for all the QE methods (on CLEF-IP 2010 and CLEF-IP 2011), while selecting the optimal number of terms used for the expansion (the number of terms that maximizes the performance for each method). From these results, we make the following observations:

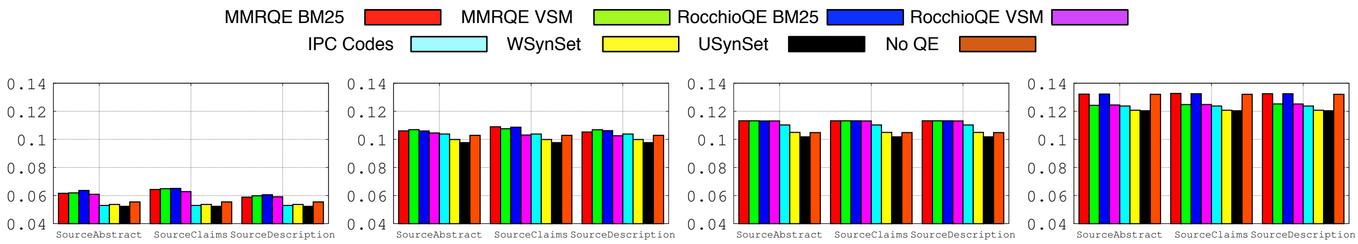
1. The best partial application section to use for querying is the description section. attribute this to the fact that the description section has more content along with relevant terms that define the invention since a detailed summary of the invention is described therein.
2. However, perhaps a better trade-off in terms of writing effort vs. retrieval performance is to query with the abstract or the extended abstract. Relative to the description, they take much less effort to write the abstract and the extended abstract. Further, querying with the abstract or the extended abstract provide a substantial boost in retrieval performance compared to the title (about 165% for MAP). In contrast, querying with the description offer only marginal performance gains (about 10% to 30% for MAP) compared to using the abstract or the extended abstract.
3. Query expansion is not useful for very long queries (i.e. description) since no method outperforms the baseline. This indicates that in advanced writing stages of the patent preparation process, QE is not useful.
4. As for query expansion, MMRQE is less effective than Rocchio for short queries such as title or abstract, whereas it appears to provide slightly better comparative results for the medium length queries (i.e., abstract) and long query (i.e., description). This suggests diverse term selection may be helpful for long queries.
5. The description section does not appear to be a good source for expansion, likely since its content is too broad and it contains many irrelevant terms.
6. When dealing with short and medium-length queries(i.e., title, abstract, and extended abstract), VSM performs better than BM25, while for very long queries (i.e., description), BM25 performs the best.
7. In general, generic QE methods like Rocchio tend to outperform patent-specific QE methods, although among patent-specific methods, the IPC Codes approach seemed to work best.

To give an insight of the effect of MMRQE and Rocchio over the performance, Table 2 shows some queries where QE methods improved the performance. Terms in bold are terms chosen by MMRQE, whereas terms underlined are terms chosen by Rocchio. Terms added by the two methods are both in bold and underlined. First of all, it is interesting to notice that even if there are common terms selected to expand the queries by both MMRQE and Rocchio, the lists of MMRQE contain more diversified terms (at least in the two first examples). For the two first examples, relevant patents talk about a similar idea than the applications, but using different examples and applications (the writers of a patent use complex and ambiguous terms to generalize the coverage of the invention). Hence, for the first query, key terms like: *rotor, blend, and suction*, were able to capture the scope of the relevant patents to allow either retrieving them (improving PRES), or pushing them to the top of the ranking (improving MAP). As for the third query, MMRQE expand the query with general terms, e.g. *result, includ, extend, plural*, which probably encourage retrieving irrelevant patents.

3.3 Query Reduction Results

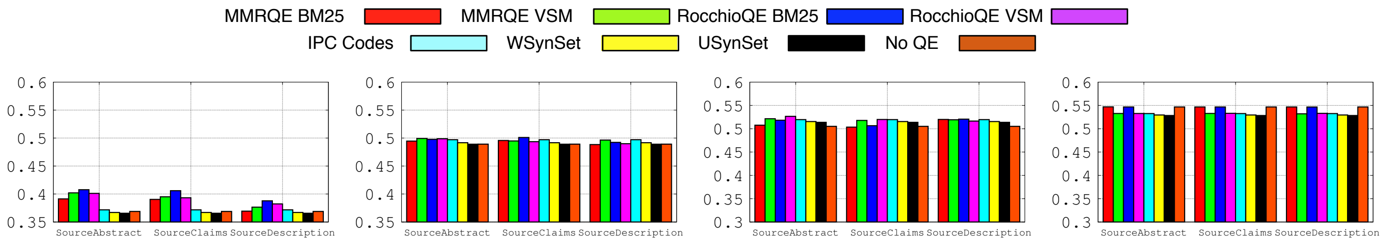
Next we discuss the results of the evaluation performed on the QR methods described in Section 2. As for QE, we carry out comprehensive experiments with the following configuration options and associated questions to consider:

- **Partial patent query type:** We apply QR methods to a query of a partial patent application, consisting of



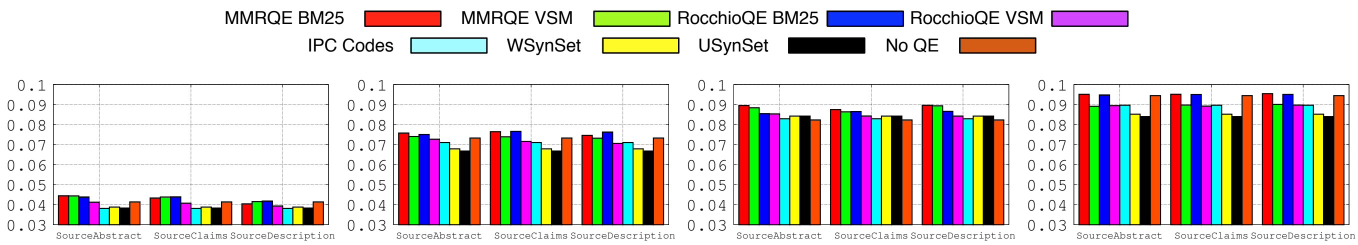
(a) Query Title. (b) Query Abstract. (c) Query Ext. Abstract. (d) Query Description.

Figure 4: MAP for QE methods on CLEF-IP 2010.



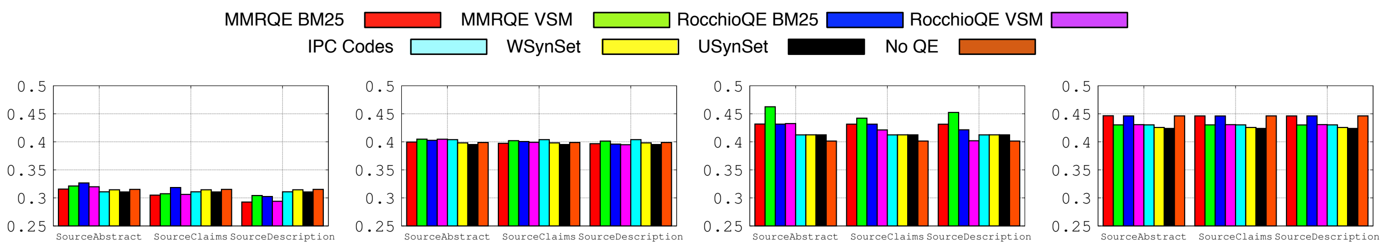
(a) Query Title. (b) Query Abstract. (c) Query Ext. Abstract. (d) Query Description.

Figure 5: PRES for QE methods on CLEF-IP 2010.



(a) Query Title. (b) Query Abstract. (c) Query Ext. Abstract. (d) Query Description.

Figure 6: MAP for QE methods on CLEF-IP 2011.



(a) Query Title. (b) Query Abstract. (c) Query Ext. Abstract. (d) Query Description.

Figure 7: PRES for QE methods on CLEF-IP 2011.

Table 2: Samples of queries extracted from CLEF-IP 2011, where QE improves the performance (P: Precision, R: Recall, RR: Reciprocal Rank, AP: Average Precision, PRES: Patent Retrieval Evaluation Score). MMRQE improves the two first examples, while Rocchio improves the third.

1- Topic: EP-1921264-A2												
Abstract: An article of manufacture having a nominal profile substantially in accordance with Cartesian coordinate values of X, Y and Z set forth in a TABLE 1. Wherein X and Y are distances in inches which, when connected by smooth continuing arcs, define airfoil profile sections at each distance Z in inches. The profile sections at the Z distances being joined smoothly with one another to form a complete airfoil shape (22,23).												
Baseline performance:	**P@5:**	0.000	**P@10:**	0.000	**R@10:**	0.000	**RR:**	0.066	**AP:**	0.043	**PRES:**	0.777
MMRQE expanded terms: airfoil, rotor, blend, substanti, root, portion, includ, suction, form, tip												
MMRQE performance:	**P@5:**	0.000	**P@10:**	0.200	**R@10:**	0.666	**RR:**	0.142	**AP:**	0.124	**PRES:**	0.872
Rocchio expanded terms: airfoil, trail, edg, cool, **form**, blade, side, **portion**, **root**, lead												
Rocchio performance:	**P@5:**	0.000	**P@10:**	0.100	**R@10:**	0.333	**RR:**	0.142	**AP:**	0.100	**PRES:**	0.822
2- Topic: EP-1707587-A1												
Abstract: It is intended to provide a crosslinked polyrotaxane formed by crosslinking polyrotaxane moleculesvia chemical bonds which exhibits excellent optical properties in water or in an aqueous solution of sodium chloride; a compound having this crosslinked polyrotaxane; and a process for producing the same. The above object can be achieved by a crosslinked polyrotaxane having at least two polyrotaxane molecules, wherein linear molecules are included in a skewered-like state at the opening of cyclodextrin molecules and blocking groups are provided at both ends of the linear molecules, so as to prevent the cyclodextrin molecules from leaving, and cyclodextrin molecules in at least two polyrotaxane molecules being bonded to each other via chemical bond, characterized in that hydroxyl (-OH) groups in the cyclodextrin molecules are partly substituted with non-ionic groups.												
Baseline performance:	**P@5:**	0.400	**P@10:**	0.300	**R@10:**	0.600	**RR:**	1.000	**AP:**	0.477	**PRES:**	0.784
MMRQE expanded terms: bond, includ, thereof, convent, crosslink, plural, polyrotaxan, substanc, gelatin, fractur, realiz, uniform, chemic, physic, rotat, biodegrad, expans, resist, elast, entrop												
MMRQE performance:	**P@5:**	0.600	**P@10:**	0.300	**R@10:**	0.600	**RR:**	1.000	**AP:**	0.577	**PRES:**	0.797
Rocchio expanded terms: form, present, cyclodextrin, compris, molecul, polym, **includ**, **crosslink**, group, compound, relat, contact, water, monom, linear, composit, **thereof**, materi, **plural**, **bond**												
Rocchio performance:	**P@5:**	0.400	**P@10:**	0.200	**R@10:**	0.400	**RR:**	1.000	**AP:**	0.455	**PRES:**	0.770
3- Topic: EP-1754935-A1												
Abstract: The fire-rated recessed downlight includes a mantle. A radiating mouth (4) is defined in the mantle. A dilatable fireproof piece (5) is fixed in the radiating mouth (4). Radiating apertures (6 or 6') corresponding to the radiating mouth (4) is defined in the dilatable fireproof piece (5) or between edges of the dilatable fireproof piece (5) and edges of the radiating mouth (4). The radiating mouth (4) of the mantle and the dilatable fireproof piece (5) could help to radiate the heat in ordinary situation and the dilatable fireproof piece (5) will expand rapidly to close the radiating mouth (4) when on fire, therefore the fire inside the mantle will not spread to the outside.												
Baseline performance:	**P@5:**	0.200	**P@10:**	0.100	**R@10:**	0.111	**RR:**	0.250	**AP:**	0.086	**PRES:**	0.801
MMRQE expanded terms: mmateri, adapt, 2, hous, light, compris, result, form, support, includ, side, mount, 4, 3, 5, plural, fit, 1, extend, recess												
MMRQE performance:	**P@5:**	0.000	**P@10:**	0.100	**R@10:**	0.111	**RR:**	0.100	**AP:**	0.044	**PRES:**	0.767
Rocchio expanded terms:materi, 2, compris, light, adapt, support, form, 3, 1, surfac, **5**, **4**, **side**, **recess**, **hous**, fire, 10, **mount**, resist, wall												
Rocchio performance:	**P@5:**	0.400	**P@10:**	0.200	**R@10:**	0.222	**RR:**	0.333	**AP:**	0.146	**PRES:**	0.821

the abstract, the extended abastract or the description sections. A critical question is what part of a partial application is best suited for QR? Note that we consider that there is no interest in reducing a title query since it already contains very few terms.

- **Relevance model:** We explore a probabilistic approach represented by the popular BM25 [23] algorithm, as well as a vector space model (VSM) approach, TF-IDF [26]. A natural question is which relevance model works best for query reduction for patent prior art search?

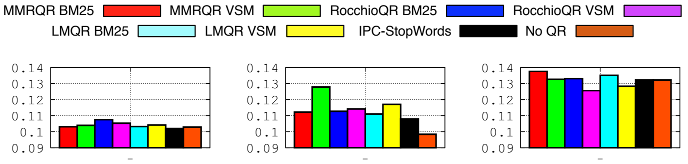
- **Term selection method:** We consider the different query reduction methods described in Section 2, i.e. RocchioQR, MMRQR, LMQR, IPC-StopWords and ask what is the best QR method for patent search? Further, how do these results compare to QE for the same queries?

To summarize all the results obtained over all the above configurations, Figures 8, 9, 10 and 11 show the respective MAP and PRES performance obtained for all QR methods (on CLEF-IP 2010 and CLEF-IP 2011), when selecting the optimal number of terms removed from the original queries. From these results, we make the following observations:

1. The best performing QR methods show benefits vs. No QR for all queries (i.e., abstract, extended abstract and description).

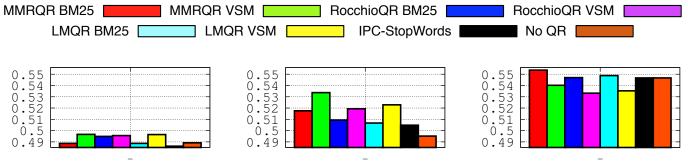
2. The term selection methods that provide the best performance are, in general, MMRQR followed by RocchioQR.

3. When dealing with medium-length queries (i.e., abstract and extended abstract), VSM performs better



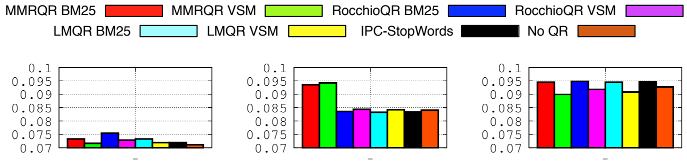
(a) Q. Abstract. (b) Q. Ext. Abstract. (c) Q. Description.

Figure 8: MAP for QR methods on CLEF-IP 2010.



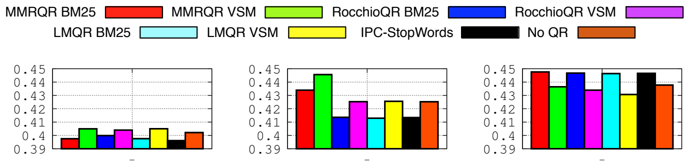
(a) Q. Abstract. (b) Q. Ext. Abstract. (c) Q. Description.

Figure 9: PRES for QR methods on CLEF 2010.



(a) Q. Abstract. (b) Q. Ext. Abstract. (c) Q. Description.

Figure 10: MAP for QR methods on CLEF-IP 2011.



(a) Q. Abstract. (b) Q. Ext. Abstract. (c) Q. Description.

Figure 11: PRES for QR methods on CLEF 2011.

than BM25, while for very long queries (i.e., description), BM25-based QR methods perform better than VSM-based QR methods.

4. In comparison to the MAP and PRES results for QE from Figure 4 and Figure 5, the best QE and QR methods perform comparably for abstract queries, whereas for extended abstract and description queries, the best QR method slightly outperforms the best QE method and No QR. Hence, the best overall retrieval result in this work in terms of both MAP and PRES comes from a description query with a generic (non-patent specific) QR method.

Finally, to give an insight of the effect of MMRQR and LMQR over the performance, Table 3 shows some queries where QR methods are helpful. Terms in bold are terms removed by MMRQR, whereas terms underlined are terms removed by Rocchio. Terms removed by the two methods are both in bold and underlined. First, we notice that even there are common terms removed from the original queries by both MMRQR and LMQR, the terms removed by MMRQR tend to be similar between them (e.g., *laser, light, interferometer*, in 1-Topic), which favor retaining diverse relevant terms in the query. However, for the third topic, MMRQR removed the main terms from the query (*motor*, and thermal *load*), which probably decreases the quality of the query.

4. RELATED WORK

We believe the outlined patent-specific query reformulation methods described in Section 2 circumscribe a range of patent-specific approaches spanning synonym lexicons, specially derived language models, and IPC code resources; hence our evaluation supported the objective of identifying general query reformulation methods from the novel perspective of partial patent application prior art search that may be deserving of further investigation in future work.

However, some more complex patent-specific methods have also been explored for general patent prior art search. The scenario of patent prior art search consists of manually form queries by selecting high frequency terms from patent application. Hence, in [9], authors proposed a new term selection method using different term frequencies depending on the genre in the NTCIR-3 Patent Retrieval Task.

Also, Xue and Croft [28] advocates the use of the full patent application as the query to reduce the burden on patent examiners. They conducted a series of experiments in order to examine the effect of different patent fields, and concludes with the observation that the best Mean Average Precision (MAP) is achieved using the text from the description section of the query patent with raw term frequencies. Also, Fuji [7] showed that retrieval effectiveness can be improved by combining IR methods with the result of citation extraction.

Bashir et al. [4] propose a query expansion with pseudo-relevance feedback. Query expansion terms are selected using a using a machine learning approach, by picking terms that may have a potential positive impact on the retrieval effectiveness. However, this approach can be computational expensive, since the presented features are complicated to compute, e.g. Pair-wise Terms Proximity features. Verma and Varma [27] propose a different approach, which instead of using the patent text to query, use its International Patent Classification (IPC) codes, which are expanded using the citation network. The formed query is used to perform an initial search. The results are then re-ranked using queries constructed from patent text. Throughout our experiments, we concluded that relying on other terms to form a query rather than those in the patent application, leads to poor retrieval quality. Lastly, a more recent work by Mahdabi et al. [16] propose a unified framework for query expansion which incorporates bibliographic information, IPC classifications, and temporal features to improve the initial query built from the query patent. They used the link-based structure of the

Table 3: Samples of queries extracted from CLEF-IP 2011, where MMRQR improves the performance. (P: Precision, R: Recall, RR: Reciprocal Rank, AP: Average Precision, PRES: Patent Retrieval Evaluation Score). MMRQR improves the two first examples, while LMQR improves the third.

1- Topic: EP-1424597-A2												
Abstract: Measurements of an interferometric measurement system are corrected for variations of atmospheric conditions such as pressure, temperature and turbulence using measurements from a second harmonic interferometer (10). A ramp, representing the dependence of the SHI data on path length, is removed before use of the SHI data. The SHI may use a passive Q-switched laser (11) as a light source and Brewster prisms (142,144) in the receiver module. Optical fibers may be used to conduct light to the detectors (145-147). A mirror reflecting the measurement beams has a coating of a thickness selected to minimize the sensitivity of the SHI data to changes in coating thickness.												
Baseline performance:	**P@5:**	0.000	**P@10:**	0.000	**R@10:**	0.000	**RR:**	0.037	**AP:**	0.022	**PRES:**	0.648
MMRQR removed terms: temperatur, detector, path, laser, light, interferometr, brewster, sensit, repres, sourc												
MMRQR performance:	**P@5:**	0.000	**P@10:**	0.100	**R@10:**	0.166	**RR:**	0.111	**AP:**	0.053	**PRES:**	0.761
LMQR removed terms: minim, conduct, variat, shi, turbul, condit, pressur, remov, ramp, thick												
LMQR performance:	**P@5:**	0.000	**P@10:**	0.000	**R@10:**	0.000	**RR:**	0.076	**AP:**	0.036	**PRES:**	0.724
2- Topic: EP-1498393-A1												
Abstract: In methods for recovering and recycling helium and unreacted chlorine from a process for manufacturing optical fiber an exhaust gas is recovered typically from a consolidation furnace and is separated into helium-rich and chlorine-rich gas streams. The helium-rich stream is typically dried and blended with make-up helium and the chlorine-rich stream is typically purified and blended with make-up chlorine so that both may be reused in the optical fiber production process.												
Baseline performance:	**P@5:**	0.200	**P@10:**	0.100	**R@10:**	0.125	**RR:**	0.200	**AP:**	0.060	**PRES:**	0.481
MMRQR removed terms: stream, rich, fiber, reus, product, dri, separ, exhaust, method, make												
MMRQR performance:	**P@5:**	0.200	**P@10:**	0.200	**R@10:**	0.250	**RR:**	0.250	**AP:**	0.106	**PRES:**	0.604
LMQR removed terms: dri, rich, process, **product, make, reus,** unreact, typic, blend, **method,**												
LMQR performance:	**P@5:**	0.200	**P@10:**	0.200	**R@10:**	0.250	**RR:**	0.200	**AP:**	0.097	**PRES:**	0.552
3- Topic: EP-1314594-A1												
Abstract: An air conditioner for air conditioning the interior of a compartment includes a compressor (C) and an electric motor (84). The compressor (C) compresses refrigerant gas and changes the displacement. The electric motor (84) drives the compressor (C). A motor controller (72) rotates the motor (84) at a constant reference speed. A detection device (92) detects information related to the thermal load on the air conditioner. A current sensor (97) detects the value of current supplied to the electric motor. A controller (72) controls the compressor based on the detected thermal load information and the detected current value. The controller (72) computes a target torque of the compressor based on the thermal load information. In accordance with the computed target torque, the controller (72) computes a target current value to be supplied to the electric motor. The controller (72) further controls the displacement of the compressor such that the detected current value matches the target current value.												
Baseline performance:	**P@5:**	0.600	**P@10:**	0.400	**R@10:**	0.307	**RR:**	1.000	**AP:**	0.301	**PRES:**	0.777
MMRQR removed terms: refer, motor, current, relat, condit, constant, suppli, compress, load, match												
MMRQR performance:	**P@5:**	0.400	**P@10:**	0.500	**R@10:**	0.384	**RR:**	0.500	**AP:**	0.221	**PRES:**	0.774
LMQR removed terms: compart, **suppli, current,** ga, **refer, compress, relat,** interior, thermal, **match,**												
LMQR performance:	**P@5:**	0.400	**P@10:**	0.400	**R@10:**	0.307	**RR:**	1.000	**AP:**	0.266	**PRES:**	0.802

citation graph together with the term distribution of cited documents and built a query model from the citation graph. They also used the publication dates associated with the patents to adapt the query model to the change of vocabulary over time. The results showed the advantage of using the term distribution of the cited documents together with the publication dates. In [18] authors propose to build a topic dependent citation graph, starting from the initially retrieved set of feedback documents and utilizing citation links of feedback documents to expand the set. They identify the important documents in the topic dependent citation graph using a citation analysis measure. Then, they use the term distribution of the documents in the citation graph to estimate a query model by identifying the distinguishing terms. Then, they use these terms to expand the original query. Finally, in [17] authors propose a method based on a random walk in a network of patent citations, to find influential documents in the citation network of a query patent, which can serve as candidates for drawing query terms and bigrams for query refinement.

Finally, other works investigated query suggestion for patent prior art search, which reflect real-life scenario of examiners, who form reproducible boolean queries [1, 2, 10].

5. CONCLUSIONS

In this paper, we analyzed various query strategies for patent prior art search with partial (incomplete) applications along with generic and patent-specific query reformulation (expansion and reduction) methods. Hence, in this scenario of partial patent application, we considered only sections, which are more likely to be written by the inventor (i.e., the title, the abstract, the description section, and an extended abstract). We performed a comprehensive comparative evaluation of these methods on the CLEF-IP patent corpus for prior art search.

We showed that the description is the best partial application section to query with, followed by the extended abstract, the abstract, and lastly the title section. However, the largest boost in performance (about 165% for MAP) comes when switching from a title query to an abstract query or extended abstract; smaller relative boosts are given by querying instead with the full description (about 10% to 30% for MAP). This is a critical insight since it is substantially easier for the patent inventor to draft an abstract or an extended abstract rather than a full patent description and in doing so, still manage to retrieve the majority of prior art that would have been retrieved with the full description.

We observed that query expansion (QE) methods are useful for short to medium length queries (i.e., title, abstract, and extended abstract), but useless for very long queries (i.e., the description section). We also showed that the description section does not provide the best source of expansion terms for QE, rather the claims or the abstract tend to offer better candidate terms for QE. In the same vein, we also found traditional IR methods like Rocchio or variations to work just as well for QE (and generally better) in comparison with patent-specific methods that used specialized expansion sources such as synonym lexicons or IPC code definitions (at least for the methods that we evaluated). For QE, future work should investigate how can we exploit patent-specific meta-data such as inventor and citation networks to better exploit specialized domains of discourse relevant to patent subfields.

Regarding query reduction (QR) methods, we showed these techniques are generally most effective compared to QE for the extended abstract and the description sections (the two longest sections used as a partial application query). Albeit by a slim margin over No QR, the overall best retrieval performance results in this work are achieved with generic (non-patent specific) QR methods for description queries. Future work may consist of exploiting query quality predictors to identify useless terms in a query using machine learning methods.

In conclusion, we return to our initial objective to aid the patent inventor in identifying an effective pre-application prior art search strategy. Our evaluation reveals the critical insight that while querying with a full description, perhaps combined with generic query reduction methods, yields strong overall retrieval performance. Nonetheless, we also find that querying with an abstract or an extended abstract and using generic query reformulation methods represents the best trade-off in terms of writing effort vs. retrieval efficacy (i.e., querying with the description sections only lead to marginal improvements).

Finally, we believe that future work should investigate whether QE methods for abstract of extended abstract queries can rival the best methods for description queries — if such a result were possible, it would significantly reduce the effort required on behalf of the patent inventor to identify potentially invalidating prior art for a new patent idea.

6. REFERENCES

[1] S. Adams. A practitioner's view on PaIR. In *Proceedings of the 4th workshop on PaIR*, 2011.

[2] L. Azzopardi, W. Vanderbauwhede, and H. Joho. Search system requirements of patent analysts. In *SIGIR*, 2010.

[3] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 2 edition, 2010.

[4] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, 2010.

[5] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *SIGIR*, 1998.

[6] E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, 1996.

[7] A. Fujii. Enhancing Patent Retrieval by Citation Analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 793–794, New York, NY, USA, 2007. ACM.

[8] D. Ganguly, J. Leveling, W. Magdy, and G. J. F. Jones. Patent query reduction using pseudo relevance feedback. In *CIKM*, 2011.

[9] 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*, 2003.

[10] Y. Kim, J. Seo, and W. B. Croft. Automatic boolean query suggestion for professional search. In *SIGIR*, 2011.

[11] 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*, SIGIR '09, pages 564–571, New York, NY, USA, 2009. ACM.

[12] W. Magdy. *Toward Higher Effectiveness for Recall-Oriented Information Retrieval: A Patent Retrieval Case Study*. PhD thesis, Dublin City University School of Computing, 2012.

[13] W. Magdy and G. J. F. Jones. PRES: A Score Metric for Evaluating Recall-oriented Information Retrieval Applications. In *SIGIR*, pages 611–618, New York, NY, USA, 2010. ACM.

[14] W. Magdy and G. J. F. Jones. A study on query expansion methods for patent retrieval. In *PaIR*, 2011.

[15] P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *SIGIR*, pages 505–514, New York, NY, USA, 2012. ACM.

[16] P. Mahdabi and F. Crestani. Patent Query Formulation by Synthesizing Multiple Sources of Relevance Evidence. *ACM Trans. Inf. Syst.*, 32(4):16:1—-16:30, 2014.

[17] P. Mahdabi and F. Crestani. Query-Driven Mining of Citation Networks for Patent Citation Retrieval and Recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1659–1668, New York, NY, USA, 2014. ACM.

[18] P. Mahdabi and F. Crestani. The Effect of Citation Analysis on Query Expansion for Patent Retrieval. *Inf. Retr.*, 17(5-6):412–429, 2014.

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

[20] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[21] M. McCandless, E. Hatcher, and O. Gospodnetic. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA, 2010.

[22] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz. CLEF-IP 2011: Retrieval in the Intellectual Property Domain. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[23] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-2. In *TREC*, pages 21–34, 1993.

[24] G. Roda, J. Tait, F. Piroi, and V. Zenz. CLEF-IP 2009: Retrieval Experiments in the Intellectual Property Domain. In C. Peters, G. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Penas, and G. Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, volume 6241 of *Lecture Notes in Computer Science*, pages 385–409. Springer, 2009.

[25] G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[26] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, 1975.

[27] M. Verma and V. Varma. Patent search using IPC classification vectors. In *PaIR*, 2011.

[28] 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*, SIGIR '09, pages 808–809, New York, NY, USA, 2009. ACM.