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

ABSTRACT

Patents are used by entities to legally protect their inventions and represent a multi-billion dollar industry of licensing and litigation. In 2012, 276,788 patent applications were approved in the US alone – a number that has doubled in the past 15 years. 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 to help inventors to assess the patentability of their ideas prior to writing a full application. In this paper, we carry out an intensive study about query reformulation for patent prior art search with partial patent application, 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. We also propose new query reformulation methods that exploit the specific structure of patents and carry out the trade-off between diversification and similarity to the query. We demonstrate that our methods improve both general (MAP) and patent-specific (PRES) evaluation metrics for prior art search performance on standardized datasets of CLEF-IP, with respect to both general and specific query reformulation methods.

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

General Terms: Algorithms, Experimentation.

Keywords: Query Reformulation, Patent Search.

1. INTRODUCTION

Patents are used by entities to legally protect their inventions and represent a multi-billion dollar industry of licensing and litigation. In 2012, 276,788 patent applications were approved in the US alone a number that has doubled in the past 15 years. Hence, helping both inventors and patents examiners to assess the patentability of a given patent application through a patent prior art search is a critical 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 [8]: (i) queries are full patent applications, which consist of documents with hundreds of words organized into several sections, while queries in text and web search constitute only a few words; (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 relevant documents.

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. Prior art search with queries that represent unfinished patent applications is certainly desirable, since writing a full application is time-consuming and costly, especially if lawyers are hired to assist.

A patent application is organized in, at least, four sections: title, abstract, claims and description. We assumed that a partial application consist in one of the mentioned sections. 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[1] between different queries (representing the title, abstract, or claims of a patent application) and the labeled relevant (all) and irrelevant documents (top 10 irrelevant documents ranked by BM25 [15]). 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). 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) the best overlap for any relevant document set for any set of queries is less than one in four terms.

[1]The Jaccard similarity is used to measure the term overlap between two sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. Before applying the Jaccard similarity, patent-specific stopwords were removed, as suggested by [11].

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

(c) Claims query. (d) Description query.

Figure 1: Average Jaccard similarity of (ir)relevant documents with the result sets for different queries.

While these results suggest the description section is the best part of a partial patent application to use as query, they also point out that the term overlap between the queries and the relevant documents can be very low. Therefore, we suggest an investigation of *query reformulation* [1] methods as a means for improving the term overlap between queries and relevant documents.

Query Reformulation is the process of transforming an initial query Q to another query Q'. This transformation may be either a reduction or an expansion of the query. *Query reduction* [6] reduces the query such that useless information is removed, while *query expansion* [4] enhance the query with additional terms likely to occur in relevant documents. In this paper, we carry out an intensive study about query reformulation for patent prior art search with partial patent application, 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, the contributions of this paper are the following:

1. A deep analysis of general methods of query expansion for patent prior-art search, as well as patent-specific query expansion methods;
2. A new method of query expansion based on diversification in terms selection;
3. A deep analysis of general methods of query reduction for patent search, as well as patent-specific query reduction methods;
4. A new method of query reduction based on diversification in terms selection;
5. A strong performance evaluation on standardized datasets of CLEF-IP using different configurations.

The rest of this paper is organized as follows: in Section 2 we present query reformulation frameworks. Section 3 is dedicated to present and discuss the experiments. Finally, conclusion and future directions are provided in Section 4.

2. QUERY REFORMULATION FOR PATENTS

During the exploration of query reformulation for patent search with partial patent applications, there are many configuration options and associated questions that we can consider:

Query type: We considered that a query of a partial patent application consist of either the title, the abstract, the claims or the description section. Critical questions is: what part of a partial application an inventor should write to obtain the best search results?

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 [15] algorithm, as well as a vector space model approach, TF-IDF [18]. A natural question is which relevance model works best for query reformulation for patent prior art search?

Query expansion source: We can consider the title, abstract, claims, and description sections as different term sources to determine which section offers the best source of expansion terms, e.g., are the title words of particularly high value as expansion terms? Note that this only applies to query expansion methods.

Term selection method: We considered different term selection methods for query reformulation. We evaluate the performance of term selection using Rocchio [17] and new term selection methods that we propose in the next sections. Then a natural question is, which term selection method works best, and with which configuration, i.e. query type, retrieval model, and term source for query expansion methods?

Before we proceed to evaluate the above questions, we first define in Section 2.1 a novel term selection method for QE hat we term MMRQE. This method is expected to address a potential deficiency of Rocchio as used in practice for high-recall search. Then, in Section 2.2 we present MMRQR, a novel QR method, which is expected to rebuild the query from scratch by selecting diversified terms.

2.1 Query Expansion Frameworks (MMRQE)

As already mentioned, Query Expansion (QE) [4] is an approach that (automatically) adds terms to an initial query in order to improve retrieval performance. The utility of QE for patent prior art search is motivated by the term overlap analysis depicted in Figure 1, which illustrates that there is a big term mismatched between queries an relevant documents. This term mismatch might be alleviated by QE methods.

While space precludes a full discussion, we remark that as a term selection method in QE, Rocchio derives a score for each potential query expansion term and in practice, the top-k scoring terms (often for $k \ll 200$) are used to expand the query and are weighted according to their Rocchio score during the second stage of retrieval. The caveat of this approach is that given a limited budget of k expansion terms, there is no inherent guarantee that these terms "cover" all documents in the pseudo-relevant set. It seems that what

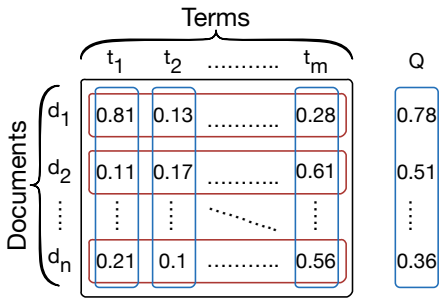


Figure 2: Notation used in MMR QE/QR.

we are asking for then is a method of "diverse" term selection — such as the *Maximal Marginal Relevance* (MMR) [3] algorithm for result set diversification. But rather than use MMR for diverse document selection (as typically used), we intend to use it here for diverse term selection. In what follows, we present a novel term selection method inspired by MMR, to address the deficiency of Rocchio, that we call MMR Query Expansion (MMRQE).

We begin our formal description of MMRQE by first defining some necessary notation. MMRQE takes as input a pseudo-relevant feedback set of n documents (PRF), which is obtained after a retrieval for the initial query. From the PRF set, we build a document-term matrix of n documents and m terms as shown in Figure 2, which uses a TF-IDF weighting for each document vector (row d_i for $1 \leq i \leq n$). However, as we will see shortly, the view that will be important for us in this work is instead the term vector (column t_j for $1 \leq j \leq m$). To represent the query Q column vector in Figure 2 having a numerical entry for every document d_i, we found that computing the BM25 or TF-IDF score between each document d_i and the query provided the best performance (in our experiments, the score used is given by the indicated relevance model).

Given a query representation Q, we aim to select an optimal subset of k terms $T_k^* \subset D$ (where $|T_k^*| = k$ and $k \ll |m|$) 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 the MMR diverse selection criterion:

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

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 it's cosine similarity with any currently selected term in T_{k-1}^*. The parameter $\lambda \in [0, 1]$ trades off relevance and diversity and we found $\lambda = 0.5$ to generally provide the best results in our experiments on the CLEF-IP training dataset collection.

The key insight we want to conclude this section with 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.

2.2 Query Reduction Frameworks (MMRQR)

As mentioned earlier, Query Reduction (QR) [6] attempts to reduce the query such that useless information is removed. The patent sections: title, abstract, claims and description are of progressively increasing length. 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 when querying with long sections such as abstract, claims and description.

Table 2.2 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 score (PRES) [9] 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, we increase PRES of the original long query.

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 2.2 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 [15] run using the Lucene search engine [13] 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 target in this second part of our work for all query sections, i.e. title, abstract, claims, and description.

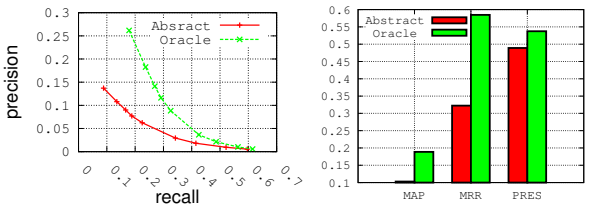


Figure 3: The utility of query reduction for 1304 abstract queries from the CLEF-IP 2010 data collection.

Following the same motivations than those which led us to propose MMRQR, we propose to greedily rebuild the query from its terms, while choosing diversified terms. Formally, given a query representation Q, we aim 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 as given in Equation 1. Note that the we used all the sections of the patent documents of the PRF set to built the document-term matrix of n documents and m terms shown in Figure 2. Here, we found that $\lambda = 0.8$ provides generally the best results in our experiments on the CLEF-IP training dataset collection.

In the next section, we propose a deep evaluation of QE and QR methods for patent search, and we attempt to answer the questions we asked in the beginning of this section.

3. EXPERIMENTAL EVALUATION

In this section we first explain our experimental setup for evaluating the effectiveness of the different methods. Then, we discuss the results of the QE and QR methods.

3.1 Experimental Setup

We used the Lucene IR System[2] to index the English subset of CLEF-IP 2010 and CLEF-IP 2010 datasets[3] [14, 16] with the default stemming and stop-word removal. We removed patent-specific stop-words as described in [8]. CLEF-IP 2010 contains 2.6 million patent documents and CLEF-IP 2011 consists of 3 million patent documents. The English test sets of CLEF-IP 2010 and CLEF-IP 2011 correspond to 1303 and 1351 topics respectively. 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. But, when a query is processed, all fields in the index are targeted, since it is sensible to use all available content.

[2]We used the LucQE module, which provides an implementation of the Rocchio QE method for Lucene.
`http://lucene-qe.sourceforge.net/`
[3]`http://www.ifs.tuwien.ac.at/~clef-ip/`

We also used the patent classification (IPC) for filtering the results by constraining them to have common classifications with the patent topic as suggested in previous works [7, 16]. Finally, we report Mean Average Precision (MAP), and Patent Retrieval Evaluation Score (PRES) [9], which combines Recall with the quality of ranking and weights relevant documents lower in the ranking more highly than MAP. We report the evaluation metrics on the top 1000 results.

3.2 Experimental Results for QE

3.2.1 *Query Expansion Baselines*

In addition to the general Rocchio approach for QE, we included two other patent specific QE methods as baselines. Motivated by [12], we used the text definitions of the International Patent Classification (IPC) codes assigned to a patent application as a source for query expansion — this is denoted as **IPC Codes**. We also implemented the QE approach proposed in [10], which automatically generates candidate synonyms sets (SynSet) for terms, and use it as a source of expansion terms. This approach has two variants: (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**). The combination of QE methods, relevance model and term selection options give us seven QE algorithms to evaluate. When MMRQE is used in combination with the VSM, the additional terms use the weights provided by the Rocchio method, whereas when using MMRQE and Rocchio with BM25, there is no need to weight the terms. For all methods, their parameters were fixed to their optimal values, which were estimated using the CLEF-IP training queries.

Other QE methods have been also explored for patent prior art search. For example, Magdy et al. [10] experiment a set of classic techniques of query expansion, which rely on pseudo-relevance feedback and WordNet as source of expansion terms. However, none of these approaches were able to achieve a significant improvement over the baseline/queries without expansion. Also, Bashir et al. [2] propose a query expansion with pseudo-relevance feedback. They proposed to select relevant terms for the expansion process 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. Verma and Varma [19] propose a different approach, which instead of using the patent text to query, they use its International Patent Classification (IPC) codes as query 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. Therefore, the approach proposed in [19] doesn't guarantee to obtain good performance. Lastly, a more recent work by Mahdabi et al. [12] propose a query expansion method that build a query-specific patent lexicon based on the definitions of the IPC. Then, this patent

lexicon is used to select expansion terms that are focused on the query topic.

3.2.2 Discussion

In this section, we discuss the results of the evaluation performed on the QE methods described above. But before, we first discuss the effect of the size of the PRF set on the performance. Table 2 shows the impact of the PRF size on the performance for the two QE algorithms Rocchio and MMRQE. These results are shown on the CLEF-IP 2010 training queries, which consists of 196 topics. We observe that the best QE performance results are obtained when using few documents in the PRF set as it was also reported in [10] (in our case, the top five gave the best results). This is certainly due to the fact that a large PRF set set will include too much irrelevant documents, whose the terms may negatively affect the quality of the expanded query.

Table 2: Effect of PRF with varying numbers of feedback documents on prior-art patent search. 20 terms are used for query expansion.

Query/Source	Metric	Method	5	10	20
Query: Abstract	MAP	Rocchio	0.074	0.072	0.070
	BL=0.073	MMRQE	0.074	0.071	0.071
Source: Claims	PRES	Rocchio	0.409	0.409	0.409
	BL=0.403	MMRQE	0.411	0.411	0.410
Query: Claims	MAP	Rocchio	0.083	0.080	0.079
	BL=0.081	MMRQE	0.082	0.080	0.080
Source: Claims	PRES	Rocchio	0.443	0.445	0.446
	BL=0.433	MMRQE	0.445	0.444	0.442

Next, we carry out comprehensive experiments along the dimensions outlined in Section 2.1 with the following specific options:

- **Query type:** {Title, Abstract, Claims, Description}
- **Query expansion source:** {Title, Abs., Claims, Descrip.}
- **Relevance model:** {BM25, Vector-space Model (VSM)}
- **Term selection method:** {Rocchio, MMRQE, *etc*...}

Figure 4 shows the results obtained in terms of MAP and PRES for CLEF-IP 2010 for different numbers of expanded terms k on the x-axis (with $k = 0$ using no QE, just the baseline retrieval model). For lack of space we show only the results of queries extracted from the claims and the abstract used as source of query expansion. From these results, we make the following observations: (i) for the two retrieval models, MMRQE provides the best performance for both MAP and PRES (except for MAP, where Rocchio BM25 provides better performance than MMRQE BM25), (ii) for both MMRQE and Rocchio, the best performance is obtained while adding no more than 50 terms to the original queries (adding more terms may have no effect, or decrease the performance), and (iii) exploiting external sources for query expansion provides poor performance (IPC code definition and SynSets).

To summarize all the results obtained over all the above configurations, Figures 7, 8, 11, and 12 show the performance obtained for all the QE methods, while selecting the

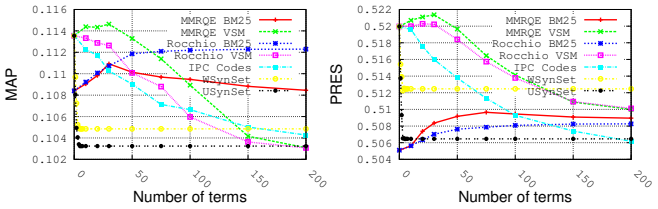


Figure 4: Results obtained while using the claims for querying and the abstract as source of query expansion on the CLEF-IP 2010.

optimal number of terms used for the expansion (number of terms that maximizes the performance). From these results, we first observe that the best section to use for querying is the description section (see Figures 7(d), 8(d), 7(d), and 8(d)). We attribute this to the fact that the description section has more content along with more terms relevant to specific details that describe the patent, since the core of the invention is described therein.

Secondly, regarding the source of query expansion, we observe that the best source is the claims. We attribute this to the fact that the claims section has more content along with more terms relevant to specific details of the patent, since the core of the invention is described therein. However, when querying using the claims, other sources of query expansion may provide better performance. This may be because the query needs more general terms than the technical terms already in the claims section. It is interesting to notice that the description is not either a good source for expansion, since it may contain more general terms that may hurt the performance.

Thirdly, we observed that query expansion is not useful for very long queries (i.e. description), indicating that in an advanced stages of the patent application process, QE is not relevant. We also notice that when dealing with more complex queries such as abstract or claims, MMRQE is more effective than Rocchio, which suggest that diverse term selection is not crucial for short queries.

Finally, we observed that using the IPC code definitions (as suggested by [12]) and SynSet (method of [10]) as a source of expansion , gave poor performance (see IPC Codes and SynSet bars along the Figures).

Regarding the best term selection method, we conclude that in general MMRQE provide better performance than Rocchio. To to give an insight of the effect of MMRQE and Rocchio over the performance, Table 3 shows some queries where QE methods improved the performance. First of all, it is interesting to notice that even if there is common terms selected to expand the queries by both MMRQE and Rocchio, the lists of MMRQE contain more diversified terms in (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, 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

Table 3: 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

irrelevant patents.

3.3 Experimental Results for QR

3.3.1 Query Reduction Baselines

As a general QR method, we proposed to adapt the Rocchio method for query pruning. Basically, the idea is once we have computed the Rocchio modified query vector, we take only terms of the initial query that appear in this vector and rank them using the Rocchio score. Then, we remove n terms with the lower score. We refer to this approach as **RocchioQR**.

Regarding patent specific QR methods, we implemented the approach proposed in [5]. This technique decomposes a query (a patent section) into constituent text segments and compute the Language Modeling (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. This approach is denoted **LMQR**. Finally, we also proposed a baseline method that use IPC codes for query reduction as follows: (i) For each patent application, we take the definitions of the IPC codes which are associated to it. Then, (ii) we rank the terms of the query according to both their frequency in the class code definition, and their frequency in the query. Finally, (iii) we 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). The intuition is that, terms in the IPC code definition may represent "stopwords", especially if they are rare (infrequent in the patent application). We denote this approach **IPC Codes**. The combination of QR method with the relevance model and

term selection options give us eight QR algorithms to evaluate. Again, the parameters of all methods were fixed to their optimal values, which were estimated using the CLEF-IP training queries.

Recently, [12] proposed as reduction process to short the query by taking only the first claim of a patent application since it contains the core of the invention. This approach hasn't been implemented because of its poor performance, which has been already pointed out by its authors.

3.3.2 *Discussion*

In this section, we discuss the results of the evaluation performed on the QR methods described above. As recommended in [5] and confirmed in our own experimentation (not shown due to lack of space), best QR performance results are also obtained when using few documents in the PRF set (in our case, the top five gave the best results).

Figure 5 shows the impact of the diversity parameter λ on the performance of MMRQR. The results are shown using BM25 retrieval model, and using abstract and claims for querying. Throughout our experiments, we concluded that the best value of λ is 0.8, which indicates that few diversification in term selection can provide some improvement. It is clear that if we consider only diversification to select terms ($\lambda = 0$), the overall performance are significantly degraded. This is certainly due to the fact that if we consider only diversified terms in the query, there is a loss in the meaning of the query, and thus we increase the probability of retrieving irrelevant patents.

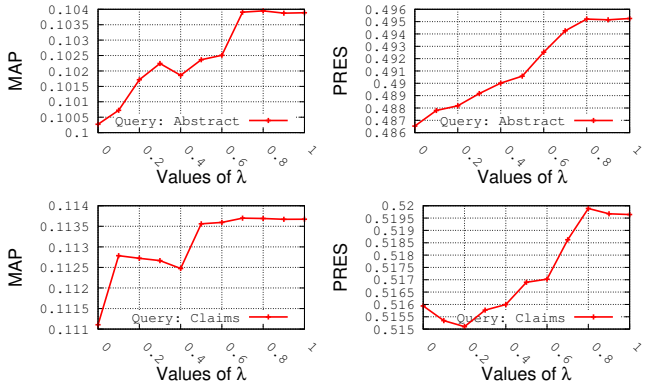


Figure 5: Impact of the diversity parameter λ on the performance of MMRQR.

Next, we carry out experiments along the dimensions outlined in Section 2.2 with the following specific options:

- **Query type:** {Title, Abstract, Claims, Description}
- **Relevance model:** {BM25, Vector-space Model (VSM)}
- **Term selection method:** {RocchioQR, MMRQR, *etc...*}

Figure 6 shows the results obtained in terms of MAP and PRES for CLEF-IP 2010 for different numbers of expanded terms k on the x-axis (with $k = 0$ using no QE, just the baseline retrieval model). For lack of space we show only the results of queries extracted from the claims and the abstract used as source of query expansion. From these results, we make the following observations: (i) for the two retrieval models, MMRQE provides the best performance for both MAP and PRES (except for MAP, where Rocchio BM25 provides better performance than MMRQE BM25), (ii) for both MMRQE and Rocchio, the best performance is obtained while adding no more than 50 terms to the original queries (adding more terms may have no effect, or decrease the performance), and (iii) exploiting external sources for query expansion provides poor performance (IPC code definition and SynSets).

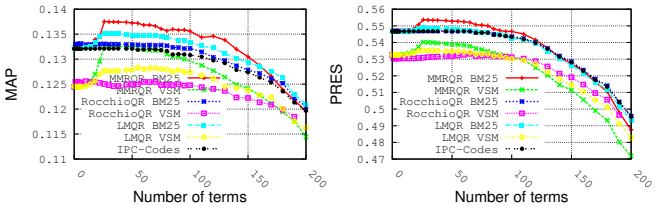


Figure 6: Results obtained for QR while using the description for querying on the CLEF-IP 2010 dataset.

To summarize all the results obtained over all the above configurations, Figures 7, 8, 11, and 12 show the performance obtained for all the QE methods, while selecting the optimal number of terms used for the expansion (number of terms that maximizes the performance). From these results, we first observe that the best section to use for querying is the description section (see Figures 7(d), 8(d), 7(d), and 8(d)). We attribute this to the fact that the description section has more content along with more terms relevant to specific details that describe the patent, since the core of the invention is described therein.

Secondly, regarding the source of query expansion, we observe that the best source is the claims. We attribute this to the fact that the claims section has more content along with more terms relevant to specific details of the patent, since the core of the invention is described therein. However, when querying using the claims, other sources of query expansion may provide better performance. This may be because the query needs more general terms than the technical terms already in the claims section. It is interesting to notice that the description is not either a good source for expansion, since it may contain more general terms that may hurt the performance.

Thirdly, we observed that query expansion is not useful for very long queries (i.e. description), indicating that in an advanced stages of the patent application process, QE is not relevant. We also notice that when dealing with more complex queries such as abstract or claims, MMRQE is more effective than Rocchio, which suggest that diverse term selection is not crucial for short queries.

Finally, we observed that using the IPC code definitions (as suggested by [12]) and SynSet (method of [10]) as a source of expansion , gave poor performance (see IPC Codes and SynSet bars along the Figures).

Regarding the best term selection method, we conclude that in general MMRQE provide better performance than Rocchio. To to give an insight of the effect of MMRQE and Rocchio over the performance, Table 3 shows some queries

Table 4: Samples of queries extracted from CLEF-IP 2011, where MMRQR improves the performance.

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

where QE methods improved the performance. First of all, it is interesting to notice that even if there is common terms selected to expand the queries by both MMRQE and Rocchio, the lists of MMRQE contain more diversified terms in (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, 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.

4. CONCLUSION

In this paper we analyzed general QE and QR methods for patent prior art search with incomplete patent applications on two patent retrieval corpora, namely CLEF-IP 2010 and CLEF-IP 2011. We demonstrated that QE methods are critical for short queries, i.e. title, abstract, and claims, but useless for very long queries, i.e. the description section. We also showed that claims are the best section that works with QE both to query with and to use as a source of query expansion terms, and that a new method MMRQE improves QE results in many cases. Future work can look at more patent-specific methods of QE for prior art search with partial patent applications and how they can be integrated with methods like MMRQE. Regarding QR methods, we showed that these techniques are effective to some extend for claims and description sections, which are considered as the longest sections in a patent application. We also demonstrated that our new QR methods MMRQR may improve both recall and precision in many cases. For this second part, future work may consist in exploiting query quality predictors to identify useless terms in a query using machine learning methods.

5. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2010.

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

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998.

[4] E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, 31:121–187, 1996.

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

[6] 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.

[7] P. Lopez and L. Romary. Patatras: retrieval model combination and regression models for prior art search. In *Proceedings of the 10th cross-language evaluation forum conference on Multilingual information access evaluation: text retrieval experiments*, CLEF'09, pages 430–437, Berlin, Heidelberg, 2009. Springer-Verlag.

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

[9] W. Magdy and G. J. Jones. PRES: A Score Metric for Evaluating Recall-oriented Information Retrieval Applications. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 611–618, New York, NY, USA, 2010. ACM.

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

[11] P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 505–514, New York, NY, USA, 2012. ACM.

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

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

[14] 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.

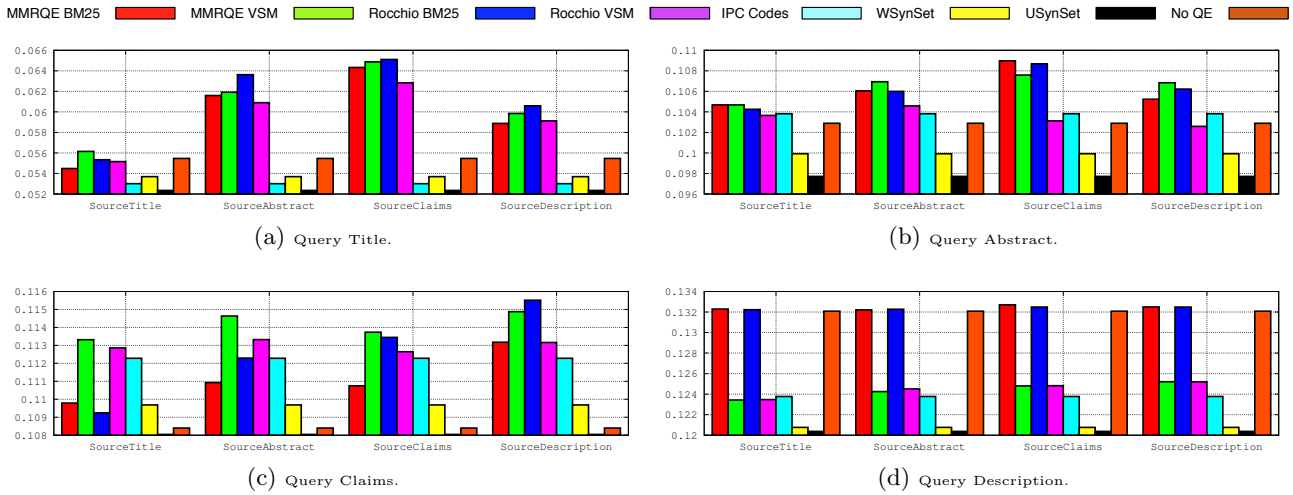
[15] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-2. In *TREC*, pages 21–34, 1993.

[16] 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 Berlin Heidelberg, 2009.

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

[18] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.

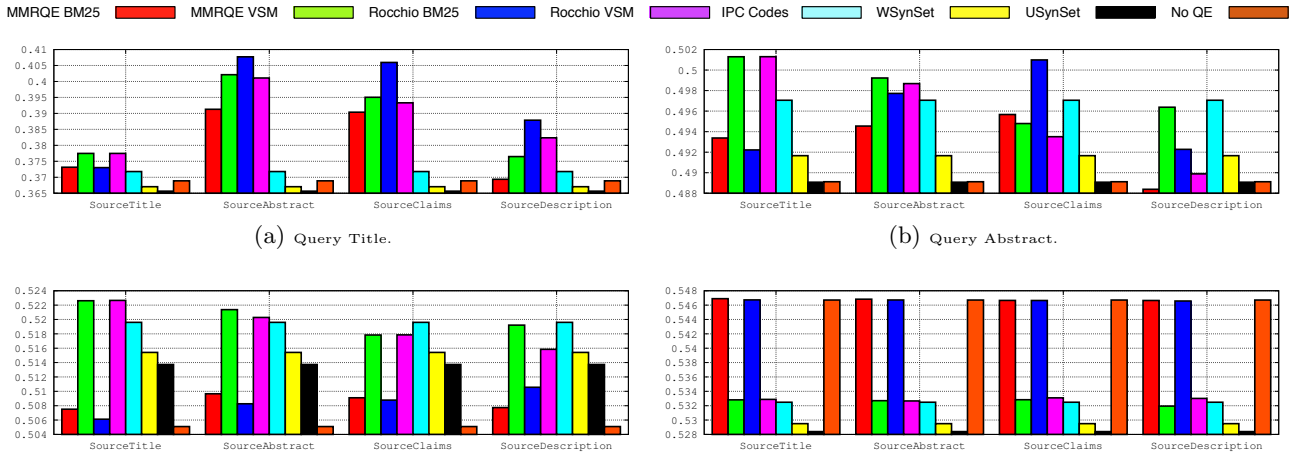
[19] M. Verma and V. Varma. Patent search using ipc classification vectors. In *PaIR*, 2011.



(a) Query Title. (b) Query Abstract.

(c) Query Claims. (d) Query Description.

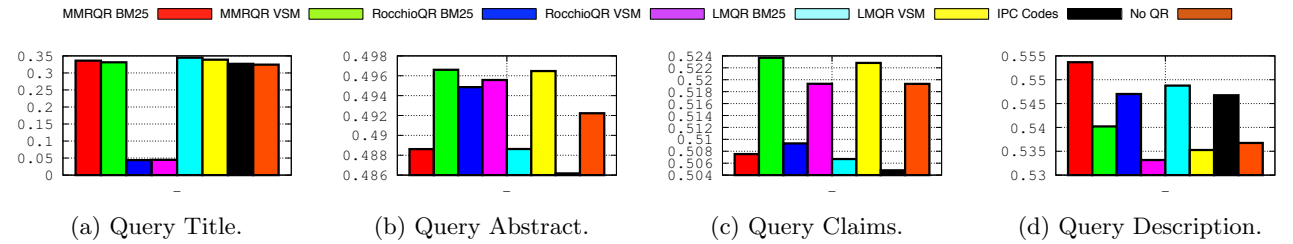
Figure 7: Mean Average Precision (MAP) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



(a) Query Title. (b) Query Abstract.

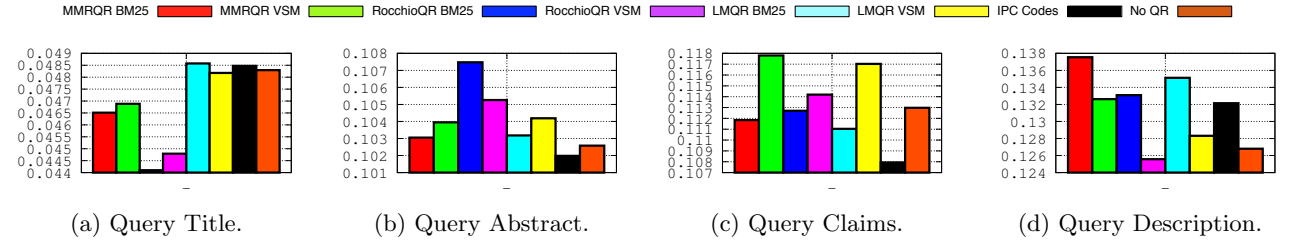
(c) Query Claims. (d) Query Description.

Figure 8: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



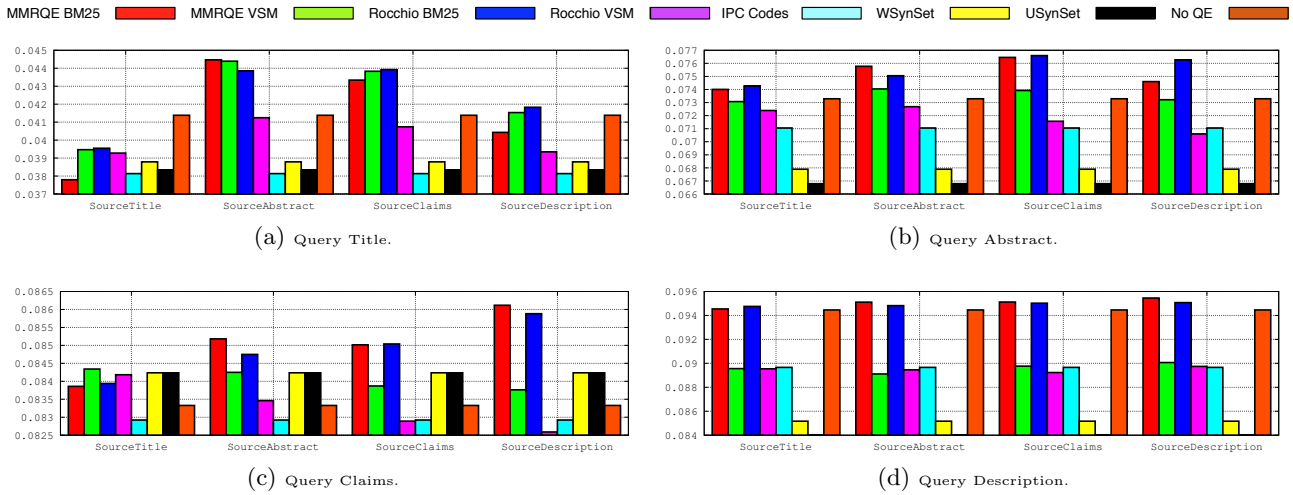
(a) Query Title. (b) Query Abstract. (c) Query Claims. (d) Query Description.

Figure 9: Mean Average Precision (MAP) on CLEF-2010 (for MMRQR $\lambda = 0.8$).



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

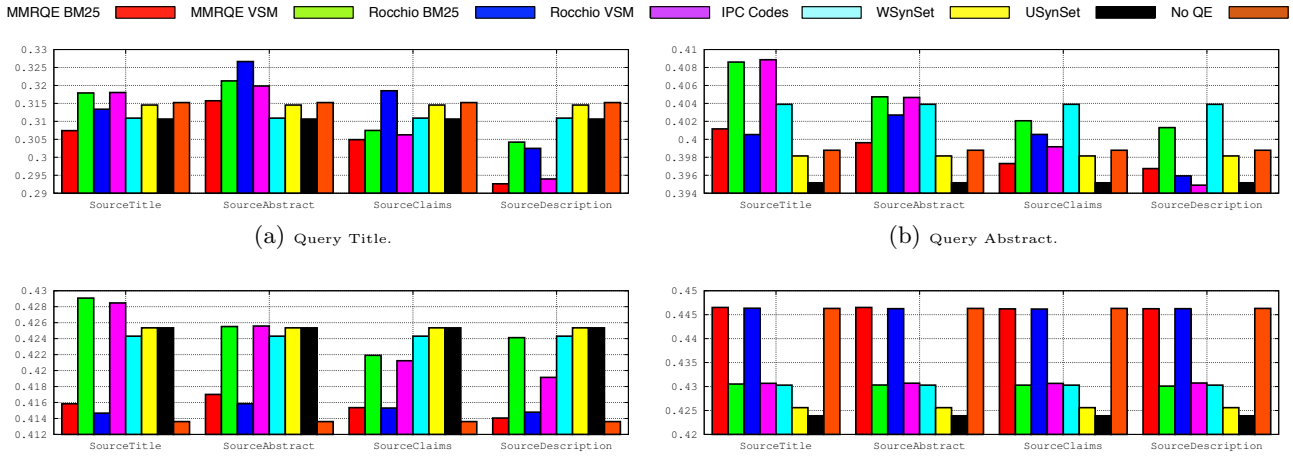
Figure 10: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQR $\lambda = 0.8$).



(a) Query Title. (b) Query Abstract.

(c) Query Claims. (d) Query Description.

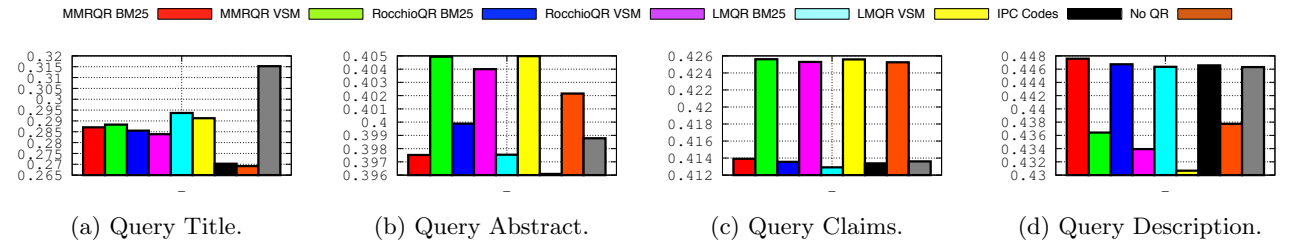
Figure 11: Mean Average Precision (MAP) on CLEF-2011 (for MMRQE $\lambda = 0.5$).



(a) Query Title. (b) Query Abstract.

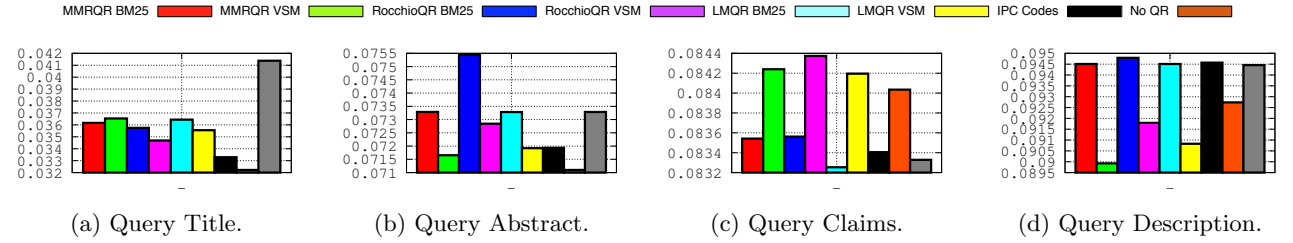
(c) Query Claims. (d) Query Description.

Figure 12: Patent Retrieval Evaluation Score (PRES) on CLEF-2011 (for MMRQE $\lambda = 0.5$).



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

Figure 13: Mean Average Precision (MAP) on CLEF-2010 (for MMRQR $\lambda = 0.8$).



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

Figure 14: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQR $\lambda = 0.8$).