A Study of Query Reformulation for Patent Prior Art Search

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 shorter queries representing (partial) titles, abstract, or claims to help inventors assess the patentability of their ideas prior to writing a full application. Hence, in this paper, we focus on both helping inventors to assess the patentability of their ideas and patents examiners to assess the patentability of a given patent application. Specifically, we propose a deep analysis of query reformulation methods that are targeted for patent prior art search. We also propose query reformulation methods that exploit the specific structure of patent documents as well as methods aimed to improve recall with a limited set of query expansion terms. 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 [12]: (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 shorter 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. However prior art search with partial applications is much different than queries with a full application – namely because the queries are much shorter and represent only parts of a patent application.

To assess the difficulty of querying with partial patent applications (such as the title, abstract, claims, or description sections), we refer to Figure 1. Here we show an analysis of the average Jaccard similarity (fraction of overlapping terms after removing patent-specific stopwords) between different queries (representing the title, abstract, or claims of a patent application) and the labeled relevant (all) and irrelevant documents (top 10 non-relevant documents ranked by BM25). We show results for the top 100 and bottom 100 queries of CLEP-IP 2010 evaluated according to 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. While these results suggest the description section may be the best part of a partial patent application to use for a query, there is still significant room for improving the overlap of query terms with the relevant documents, which suggests an investigation of *query reformulation* [3] methods.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

(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.

Query Reformulation is the process, which consists 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 [10]* reduces the query such that useless information is removed, while *query expansion [6]* enhance the query with additional terms likely to occur in relevant documents. These are the tasks we evaluate in this paper.

In summary, our contributions are the following:

1. A deep analysis of general methods of query expansion for patent 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 introduce our approaches for diverse patent query reformulation, and we discuss the experiments in Section 3. We present the related work in Section 4, and we conclude and provide some future directions in Section 5.

2. QUERY REFORMULATION FOR PATENTS

In this section, we first present the requirements of a query expansion method in Sections 2.1, then, we introduce a novel term selection method for query expansion in Section 2.2. Next, in Section 2.3 we introduce the motivations behind the benefit of query reduction, and a new approach of query reduction in Section 2.4.

2.1 General Framework for Query Expansion

Query expansion (QE) [6] is an approach that (automatically) adds terms to an initial query in order to improve retrieval performance. In exploring QE for patent search with partial patent applications, there are many configuration options and associated questions that we can consider:

Query type: We consider a partial patent application to consist of either the title, the abstract, the claims of the description section and allow one to query with each[1]. Critical questions are: (i) what part of a partial application an inventor should write to obtain the best search results? (ii) what part of a patent application a patent examiner should use to make a patent prior art search?

Query expansion source: We can consider the title, abstract, claims, and description section as different QE term sources and ask which section offers the best source of expansion terms? E.g., are the title words of particularly high value as expansion terms?

Relevance model: For initial retrieval of documents in the *pseudo-relevant* feedback set (PRF) — often used to generate the terms for QE — and subsequent re-retrieval with an expanded term set, there are various options for the relevance ranking model. In this work, we explore a probabilistic approach represented by the popular BM25 [18] algorithm as well as a vector space model approach as represented by TF-IDF [21]. A natural question is which relevance model works best for QE for patent prior art search?

Term selection method: Once we have identified a query expansion source, we may consider different methods of selecting terms for expansion. A standard method for term selection is based on the Rocchio [20] approach, but in the next subsection, we introduce an alternate term selection method intended to address the high-recall nature of patent prior art search. Then a natural question to ask is which term expansion method works best, and with which expansion source and retrieval model?

Before we proceed to evaluate the above questions, we first define a novel term selection method to address a potential deficiency of Rocchio as used in practice for high-recall search that we term MMRQE.

2.2 MMR Query Expansion (MMRQE)

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 we are asking for then is a method of "diverse" term selection — something like the *maximal marginal relevance* (MMR) [5]

[1]A query is always evaluated against the full content (title, abstract, claims, description) of granted patent applications since it is sensible to make use of all available content.

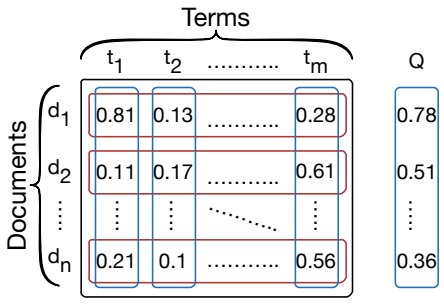


Figure 2: Notation used in MMR QE/QR.

algorithm for result set diversification, but rather than for diverse document selection as typically used, we intend to use it here for diverse term selection.

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 simply 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 General Framework for Query Reduction

A patent application is composed of the following sections: title, abstract, description, and claims, which are of progressively increasing length. While the title is usually between three and ten terms in length, the other section are longer, ranging from ten to thousand terms in length. Hence, query expansion may be not useful for these sections, and we propose to investigate query reduction as an alternative. Query reduction (QR) [10] attempts to reduce the query such that useless information is removed.

Table 2.3 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) [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 that if they are 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 2.3 shows the summary upper-bound performance for precision, recall, MAP, 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 [18] run using the Lucene search engine [16] and the original long query. "Oracle" refers to the situation when 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.

Similarly, in exploring QR for patent search with partial

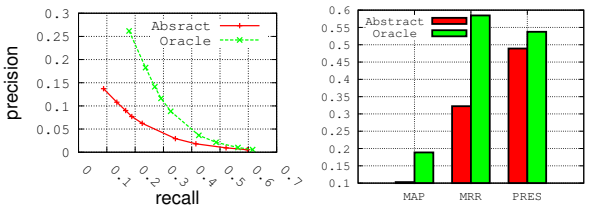


Figure 3: The utility of query reduction for 1304 abstract queries from the CLEF-IP 2010 data collection.

patent applications, there are many configuration options and associated questions that we can consider:

Query type: What part of a patent application is best suited to use with a QR method?

Relevance model: What is the relevance model that works best for QR for patent prior art search, i.e we investigate the probabilistic BM25 model [18], as well as the vector model [21]?

Term selection method: Many approaches of general and patent specific QR has been proposed. So, a natural question to ask is: which QR method works best, and with which retrieval model?

In the next section we propose a new method for query reduction based on MMR that we term MMRQR.

2.4 MMR Query Reduction (MMRQR)

Following the same motivations than those which led us to propose MMRQE, 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 throughout this section.

3. EXPERIMENTAL EVALUATION

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] [17, 19] with the default stemming and stop-word removal. We removed patent-specific stop-words as described in [12]. The

[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/`

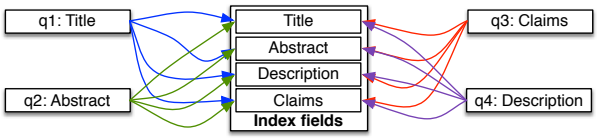


Figure 4: Model of the index fields matching with queries .

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 application (title, abstract, claims, and description) is indexed in a separate field. Hence, when a query is processed, all fields in the index are targeted as shown in Figure 4. 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 [19, 11]. Finally, we report Mean Average Precision (MAP), and Patent Retrieval Evaluation Score (PRES) [13], 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 [15], 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 [14], 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 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.

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 us-

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

ing few documents in the PRF set as it was also reported in [14] (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.

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*...}

TO DO

Figures 6 and 7 show the performance across different queries and sources of expansion respectively 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). From these results, we observe the best section to use for *both* querying and the source of query expansion terms is the claims section (see the bottom line of Figures 6 and 7). 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. Very similar overall results are obtained for CLEF-IP 2011 and for space reasons we cannot show them here.

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

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 (see the fourth column from Figures 6 and 7). Finally, we observed that using the IPC definitions as a source of expansion, as suggested by [15], gave poor performance (see Class Code curve along the Figures).

Regarding the best term selection method, we conclude that Rocchio is better in retrieving patents since it gives best performance for PRES in general, while MMRQE is better for ranking since it gives better performance for MAP.

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 the approach as **RocchioQR**.

Regarding patent specific QR methods, we implemented the approach proposed in [7]. 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 (unfrequent in the patent application). We denote this approach **IPC Codes**.

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

3.3.2 *Discussion*

In this section, we discuss the results of the evaluation performed on the QR methods described above. As recommended in [7] 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.

TO DO

Next, we carry out comprehensive experiments along the dimensions outlined in Section 2.3 with the following specific options:

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

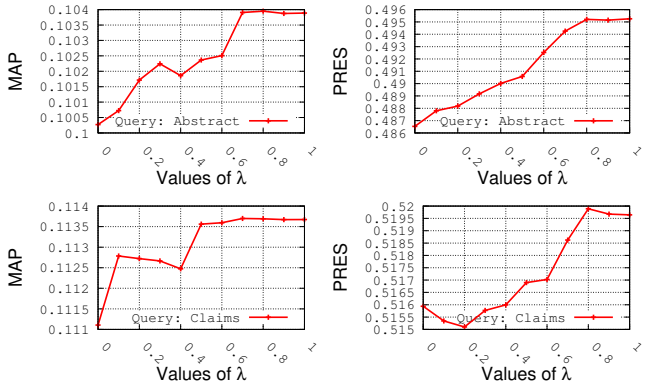


Figure 5: Impact of the diversity parameter λ on the performance of MMRQR.

4. RELATED WORK

The major contributions in patent search has focused on query formulation, where the objective is to find the best terms to represent a patent application as a query to achieve high retrieval effectiveness by retrieving all possible relevant documents at high ranks [12].

The scenario of patent prior art search consists of manually form queries by selecting high frequency terms from patent application. Hence, following this methodology, some algorithms of patent query reduction have been proposed to select only useful terms from patent application [7, 8]. We used the approach propose by Ganguly et al. [7] for QR as a baseline, and we showed that the performance of MMRQR outperform this approach in many cases.

Similarly, query expansion has been widely explored for patent search. Magdy et al. [14] 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. Therefore, they introduce a novel approach that automatically generates candidate synonyms sets for terms, and use them as a source of expansion terms, which showed significant improvement with respect to the baseline. We also used this approach as a baseline in our experiments, and we showed that MMRQE provides better performance in many cases. Also, Bashir et al. [4] 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 [22] 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 [22] doesn't guarantee to obtain good performance. Lastly, a more recent work by Mahdabi et al. [15] propose both a query reduction and expansion method. For the query reduction process, they just shorten the query document by taking the first claim since it contains the core of the invention. For the expansion process, they propose to 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.

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]. Reproducibility of search means that the retrieval system will give the same results for the same query each time. Hence, Kim et al. [9] propose an approach, which generate boolean queries by exploiting decision trees learned from pseudo-labeled documents and rank the suggested queries using query quality predictors.

5. 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.

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. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd 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)*, 31:121–187, 1996.

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

[8] 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.

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

[10] 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.

[11] 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.

[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. 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.

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

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

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

[17] 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.

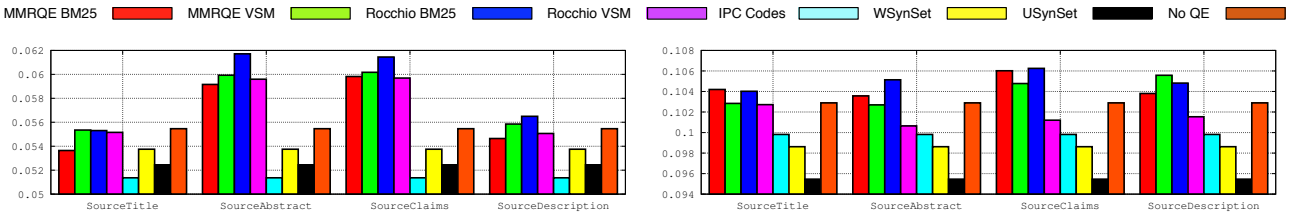
[18] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-2. In *TREC*, pages 21–34, 1993.

[19] 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.

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

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

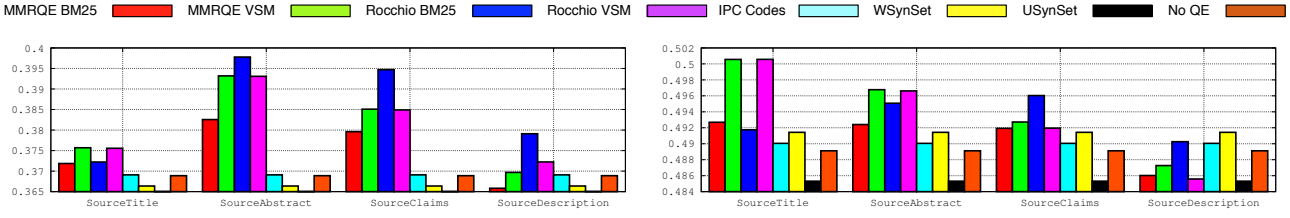
[22] M. Verma and V. Varma. Patent search using ipc classification vectors. In *PaIR*, 2011.



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

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

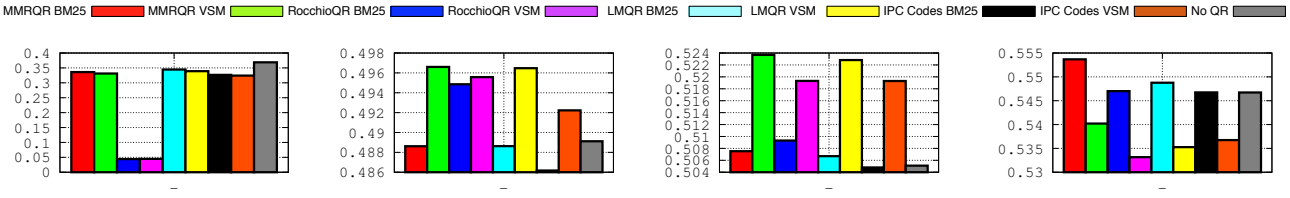
Figure 6: Mean Average Precision (MAP) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



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

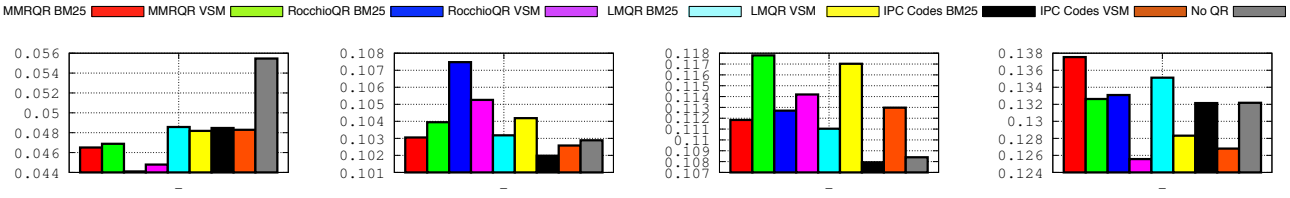
(c) Query Claims. (d) Query Description.

Figure 7: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



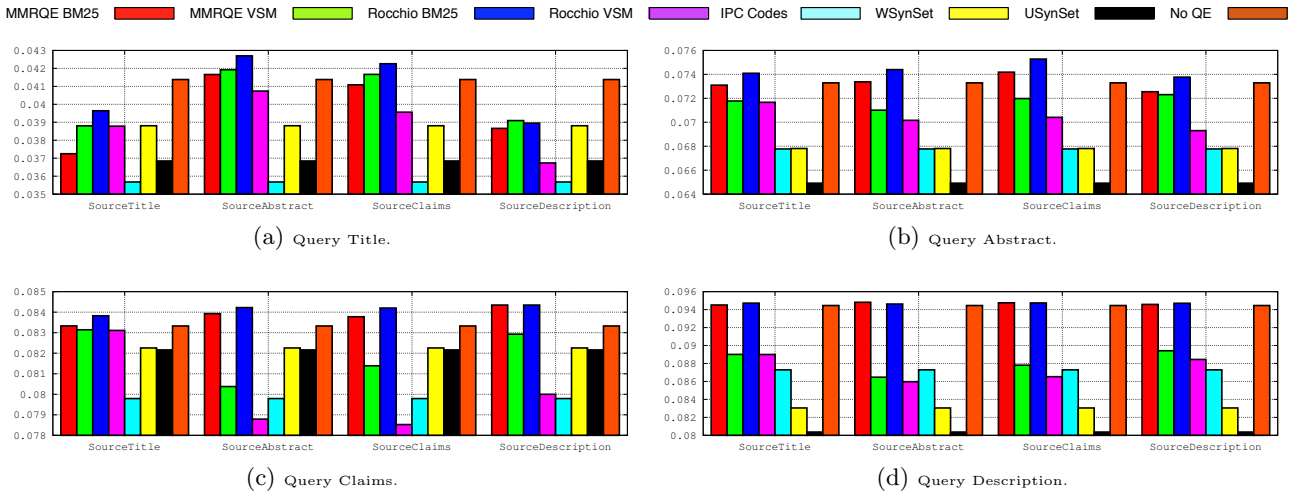
(a) Query Title. (b) Query Abstract. (c) Query Claims. (d) Query Description.

Figure 8: Mean Average Precision (MAP) on CLEF-2010 (for MMRQR $\lambda = 0.8$).



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

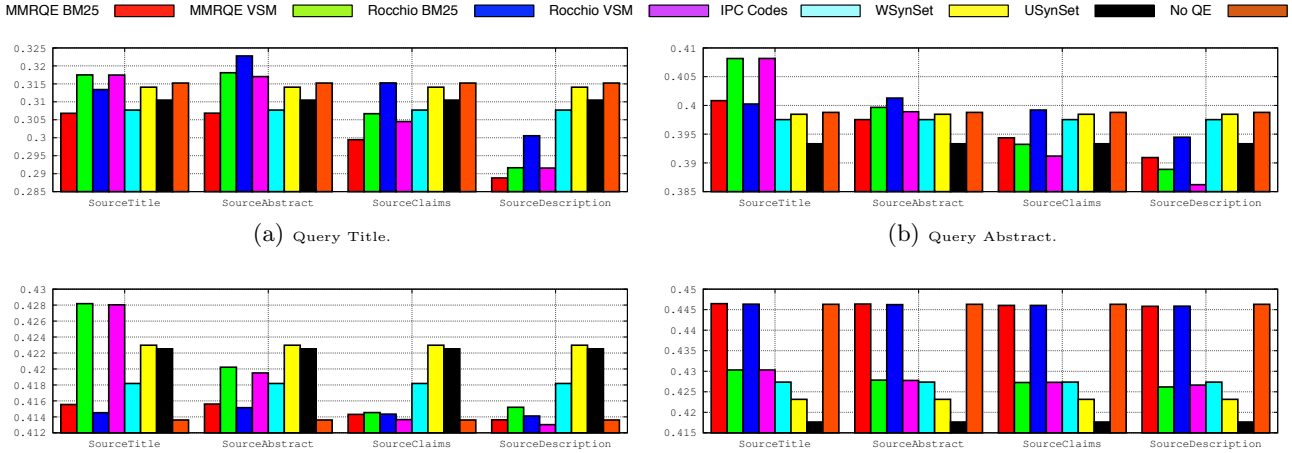
Figure 9: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQR $\lambda = 0.8$).



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

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

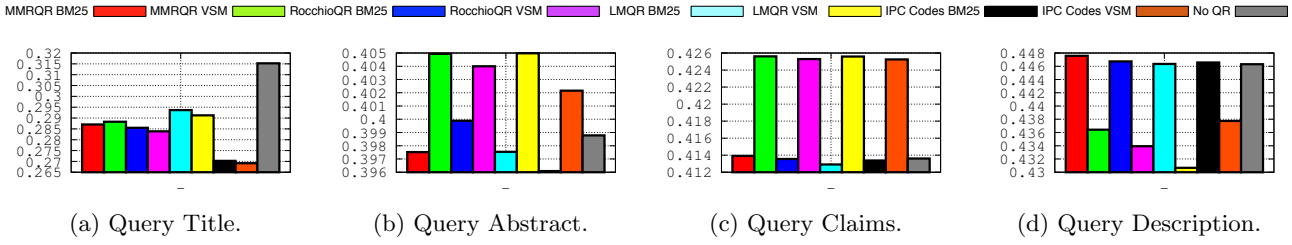
Figure 10: Mean Average Precision (MAP) on CLEF-2011 (for MMRQE $\lambda = 0.5$).



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

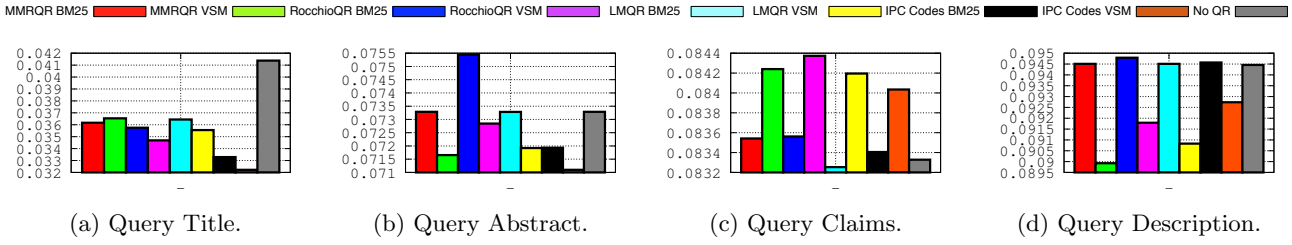
(c) Query Claims. (d) Query Description.

Figure 11: Patent Retrieval Evaluation Score (PRES) on CLEF-2011 (for MMRQE $\lambda = 0.5$).



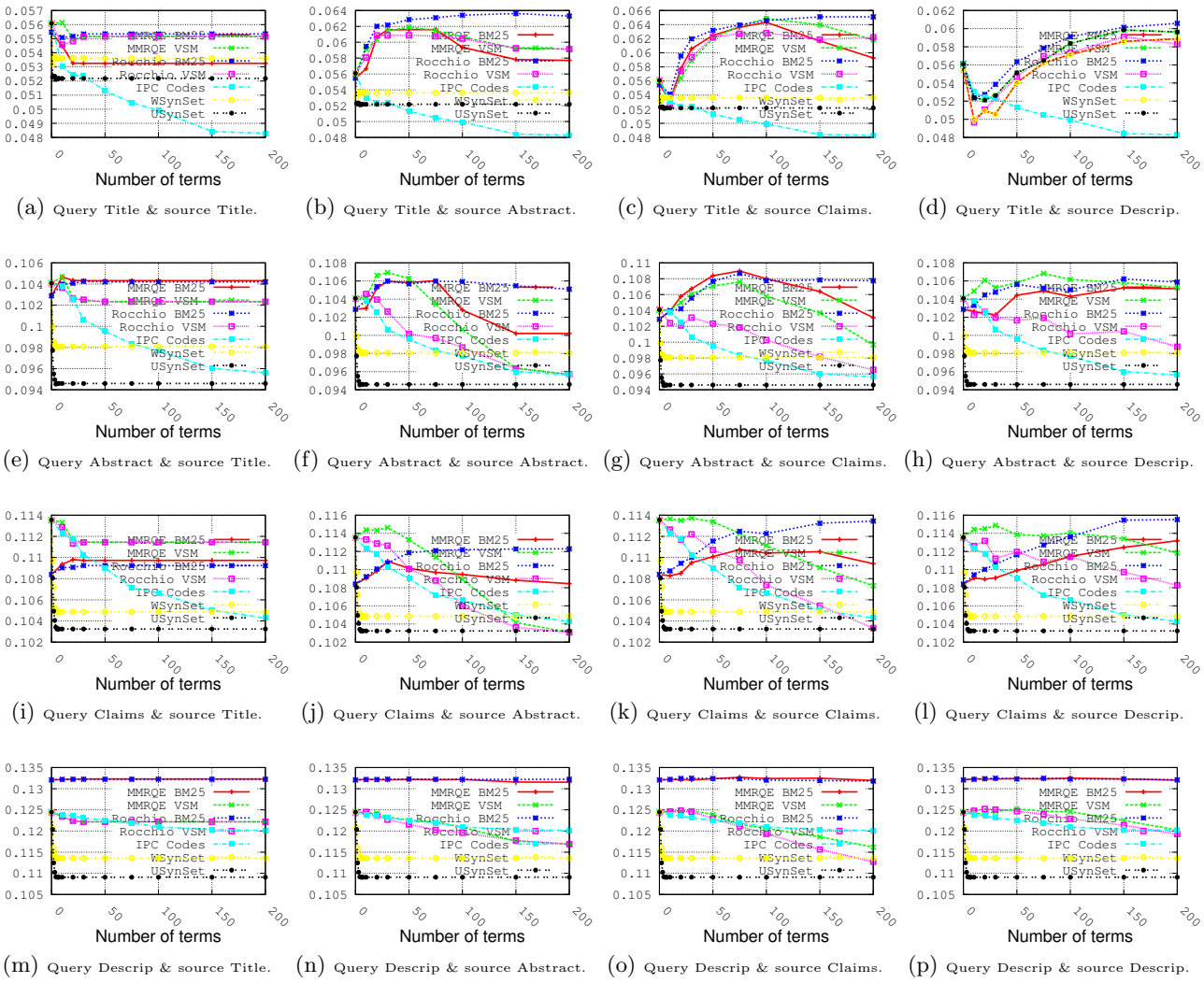
(a) Query Title. (b) Query Abstract. (c) Query Claims. (d) Query Description.

Figure 12: 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 13: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQR $\lambda = 0.8$).



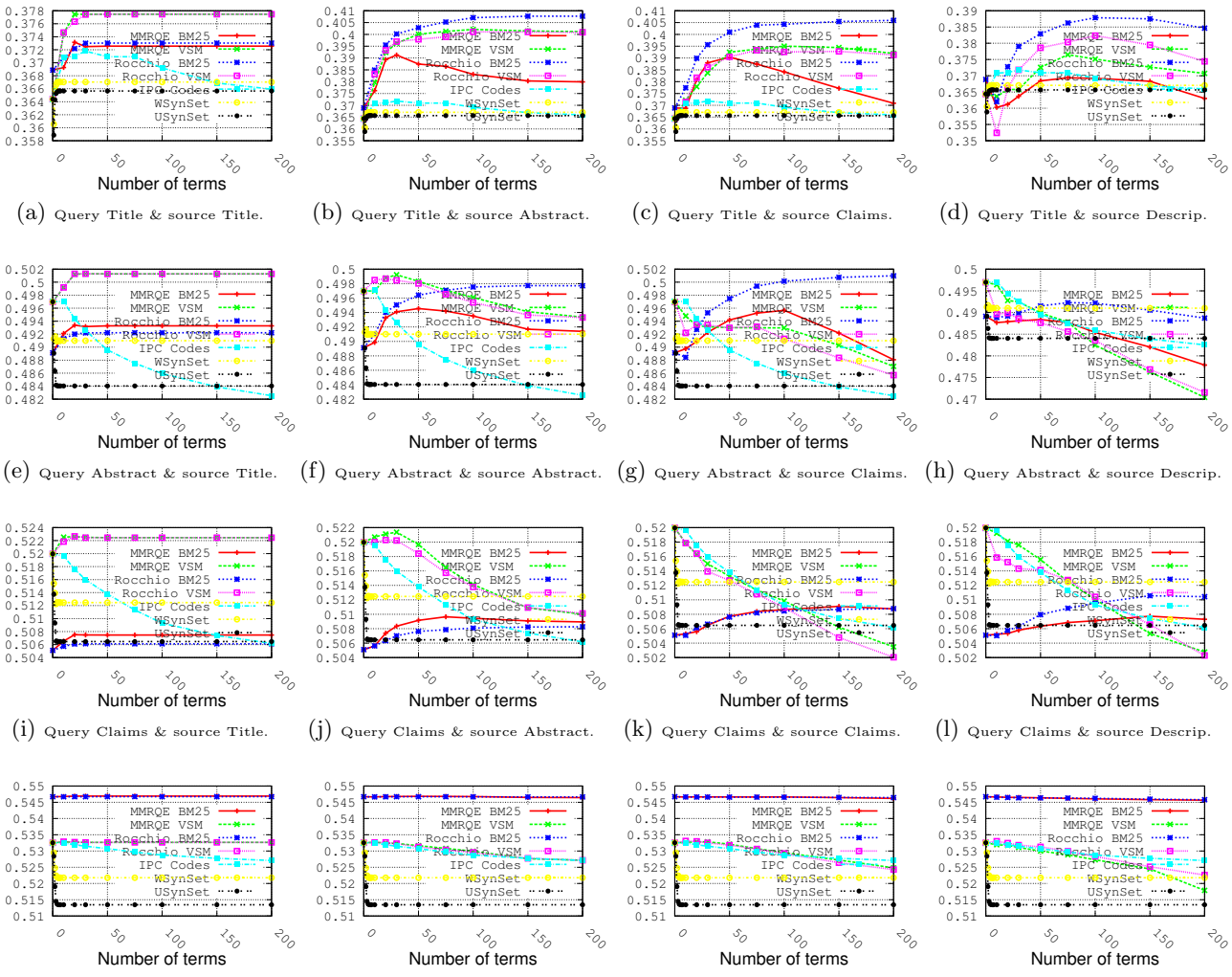
(a) Query Title & source Title. (b) Query Title & source Abstract. (c) Query Title & source Claims. (d) Query Title & source Descrip.

(e) Query Abstract & source Title. (f) Query Abstract & source Abstract. (g) Query Abstract & source Claims. (h) Query Abstract & source Descrip.

(i) Query Claims & source Title. (j) Query Claims & source Abstract. (k) Query Claims & source Claims. (l) Query Claims & source Descrip.

(m) Query Descrip & source Title. (n) Query Descrip & source Abstract. (o) Query Descrip & source Claims. (p) Query Descrip & source Descrip.

Figure 14: Mean Average Precision (MAP) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



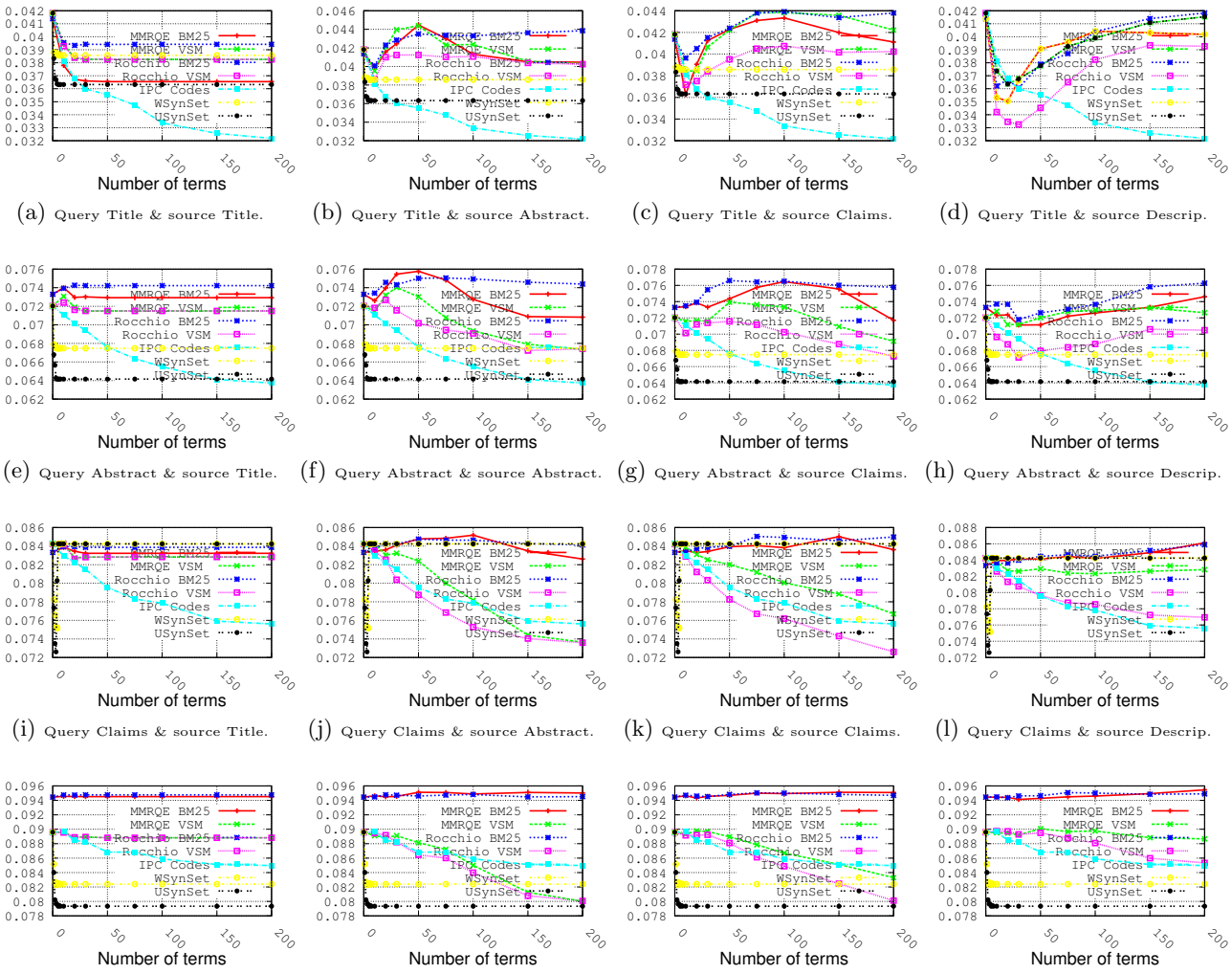
(a) Query Title & source Title. (b) Query Title & source Abstract. (c) Query Title & source Claims. (d) Query Title & source Descrip.

(e) Query Abstract & source Title. (f) Query Abstract & source Abstract. (g) Query Abstract & source Claims. (h) Query Abstract & source Descrip.

(i) Query Claims & source Title. (j) Query Claims & source Abstract. (k) Query Claims & source Claims. (l) Query Claims & source Descrip.

(m) Query Descrip & source Title. (n) Query Descrip & source Abstract. (o) Query Descrip & source Claims. (p) Query Descrip & source Descrip.

Figure 15: Patent Retrieval Evaluation Score (PRES) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



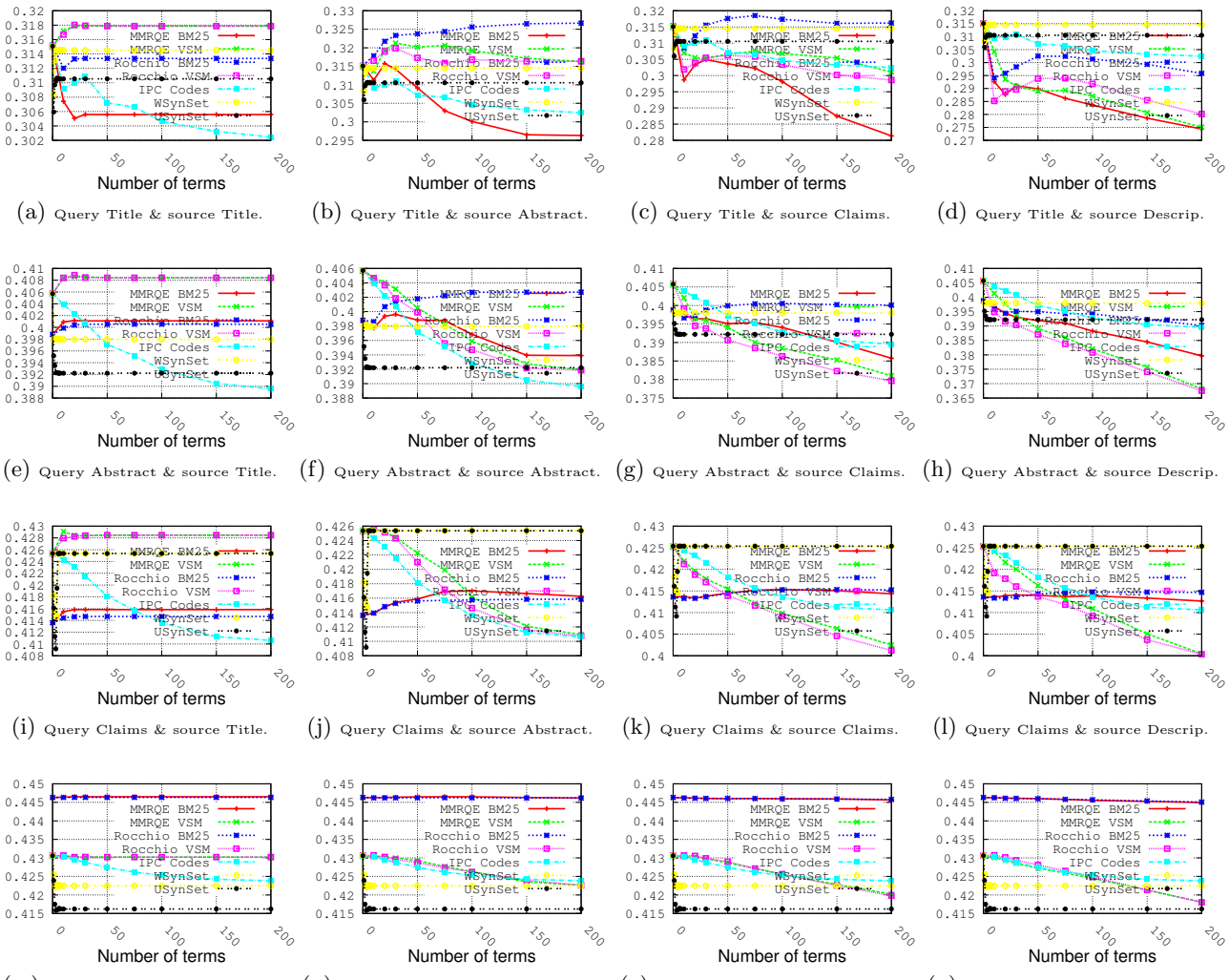
(a) Query Title & source Title. (b) Query Title & source Abstract. (c) Query Title & source Claims. (d) Query Title & source Descrip.

(e) Query Abstract & source Title. (f) Query Abstract & source Abstract. (g) Query Abstract & source Claims. (h) Query Abstract & source Descrip.

(i) Query Claims & source Title. (j) Query Claims & source Abstract. (k) Query Claims & source Claims. (l) Query Claims & source Descrip.

(m) Query Descrip & source Title. (n) Query Descrip & source Abstract. (o) Query Descrip & source Claims. (p) Query Descrip & source Descrip.

Figure 16: Mean Average Precision (MAP) on CLEF-2011 (for MMRQE $\lambda = 0.5$).



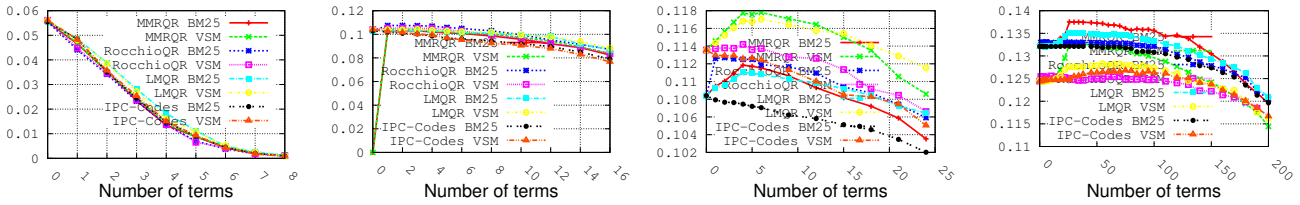
(a) Query Title & source Title. (b) Query Title & source Abstract. (c) Query Title & source Claims. (d) Query Title & source Descrip.

(e) Query Abstract & source Title. (f) Query Abstract & source Abstract. (g) Query Abstract & source Claims. (h) Query Abstract & source Descrip.

(i) Query Claims & source Title. (j) Query Claims & source Abstract. (k) Query Claims & source Claims. (l) Query Claims & source Descrip.

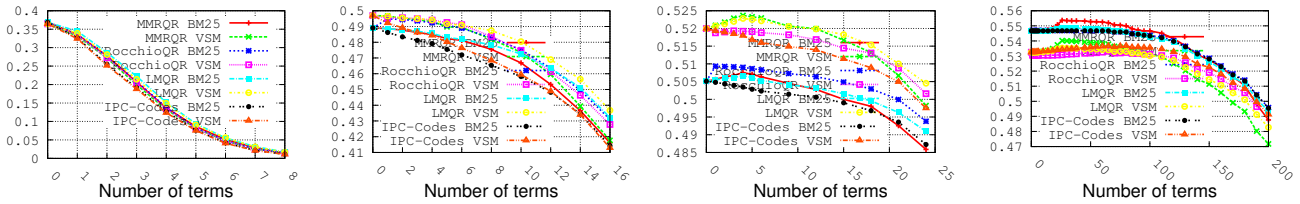
(m) Query Descrip & source Title. (n) Query Descrip & source Abstract. (o) Query Descrip & source Claims. (p) Query Descrip & source Descrip.

Figure 17: Patent Retrieval Evaluation Score (PRES) on CLEF-2011 (for MMRQE $\lambda = 0.5$).



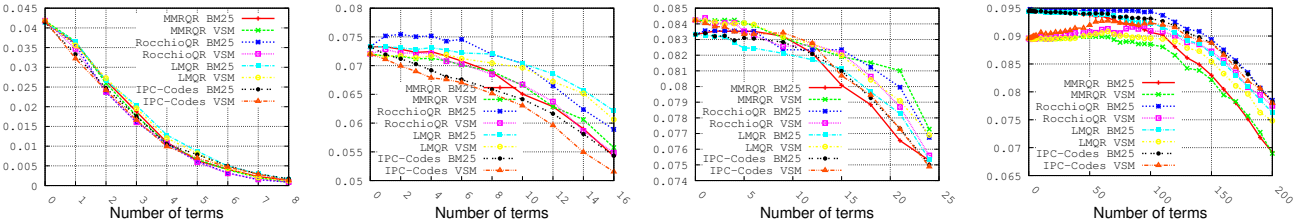
(a) Query Title. (b) Query Abstract. (c) Query Claims. (d) Query Description.

Figure 18: Mean Average Precision (MAP) for MMRQR on CLEF-2010 (for MMRQE $\lambda = 0.8$).



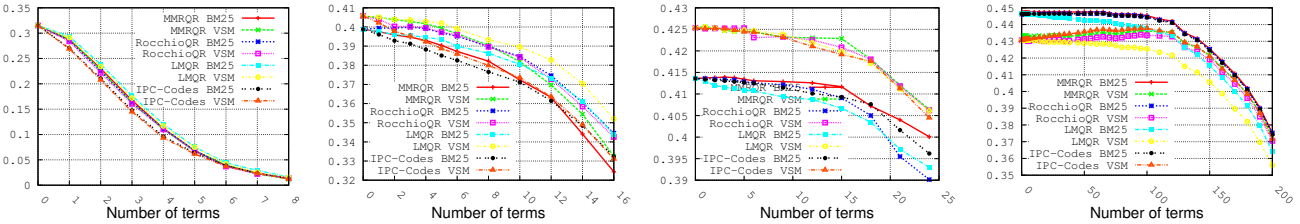
(a) Query Title. (b) Query Abstract. (c) Query Claims. (d) Query Description.

Figure 19: Patent Retrieval Evaluation Score (PRES) for MMRQR on CLEF-2010 (for MMRQE $\lambda = 0.8$).



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

Figure 20: Mean Average Precision (MAP) for MMRQR on CLEF-2011 (for MMRQE $\lambda = 0.8$).



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

Figure 21: Patent Retrieval Evaluation Score (PRES) for MMRQR on CLEF-2011 (for MMRQE $\lambda = 0.8$).