Query Expansion Methods for 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 shorter queries representing (partial) titles, abstract, or claims to help inventors assess the patentability of their ideas prior to writing a full application. In this paper, we focus on the latter task and specifically on query expansion methods that are targeted for patent prior art search. We propose query expansion 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.

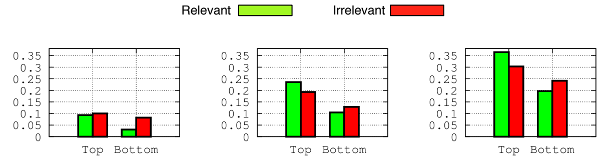
Categories and Subject Descriptors: H.3.3 [Information Systems]: Information Storage and Retrieval, Information Search and Retrieval

General Terms: Algorithms, Experimentation.

Keywords: Query Expansion, Patent Search.

1. INTRODUCTION

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 [4]: (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.



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

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

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, or claims 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 query to claims 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 claims 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 expansion* [3] methods. This is the task we evaluate in the rest of the paper.

2. QUERY EXPANSION FOR PATENTS

2.1 General Framework

Query expansion (QE) [3] 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, or the claims section[1] and allow one to query with each. A critical question is what part of a partial application an inventor should write to obtain the best search results?

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?

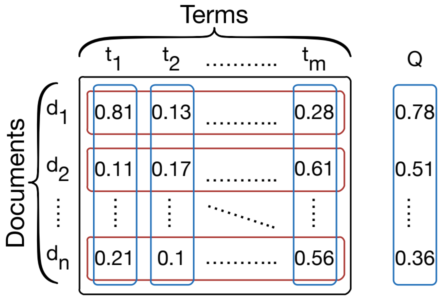
Relevance model: For initial retrieval of documents in the *pseudo-relevant* feedback set (PRF) — often used to generate the terms for QE — and subsequent re-retrieval with an expanded term set, there are various options for the relevance ranking model. In this work, we explore a probabilistic approach represented by the popular BM25 [8] algorithm as well as a vector space model approach as represented by TF-IDF [10]. 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 [9] 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) [2] 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.



	t_1	t_2	....	t_m	Q
d_1	0.81	0.13	....	0.28	0.78
d_2	0.11	0.17	....	0.61	0.51
⋮	⋮	⋮	⋱	⋮	⋮
d_n	0.21	0.1	....	0.56	0.36

Figure 2: Notation used in 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:

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

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.

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. To see if this is true (among other questions), we now proceed to empirical evaluation.

3. EXPERIMENTAL RESULTS

We used the Lucene IR System to index the English subset of CLEF-IP 2010 dataset[2] [7] with the default stemming, standard stop-word removal and patent-specific stop-

[1] We assume the description section has not been written since this is tantamount to writing a full patent application.

[2] http://www.ifs.tuwien.ac.at/ clef-ip/

words, as described in [4]. The English test set corresponds to 1303 topics. In our implementation, each section of a patent application (title, abstract, claims, and description) are indexed in separate fields. When a query is performed, all fields in the index are targeted. As recommended in [4] and confirmed in our own experimentation (not shown due to lack of space), best QE performance results are obtained when using few documents in the PRF set (in our case, the top five gave the best results).

We carry out comprehensive experiments along the four dimensions outlined in 2.1 with the following specific options:

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

e include one additional QE baseline motivated by [6], where the text definitions of the International Patent Classification (IPC) codes assigned to a patent application are used as a source for query expansion — this is denoted as **Class Code**.

The relevance model and term selection options give us four 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 both MMRQE and Rocchio, their parameters were fixed to their optimal values, which were estimated using the CLEF-IP training queries.

We report MAP and PRES on the top 1000 results. PRES [4] is a metric that combines recall with the quality of ranking and weights relevant documents lower in the ranking more highly than MAP.

Figures 3 and 4 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 3 and 4). 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 3 and 4). Finally, we observed that using the IPC definitions as a source of expansion, as suggested by [6], 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.

4. RELATED WORK

There are a variety of existing query expansion methods that use synonyms (both from WordNet and automatically generated) [5], by supervised learning [1], by IPC codes (a variant on our baseline approach that additionally uses the citation network of patents) [11], and a query-specific patent lexicon based on the definitions of the IPC [6]. While our intention in this paper was to comprehensively evaluate very general methods for QE using *partial patent applications*; it would be interesting future work to comprehensively evaluate all of these patent-specific QE methods with our generic methods for partial patent application queries.

5. CONCLUSION

In this paper we analyzed general QE methods for patent prior art search with incomplete patent applications. We demonstrated that QE methods are critical for short queries, that claims are the ideal section content 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.

6. REFERENCES

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

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

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

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

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

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

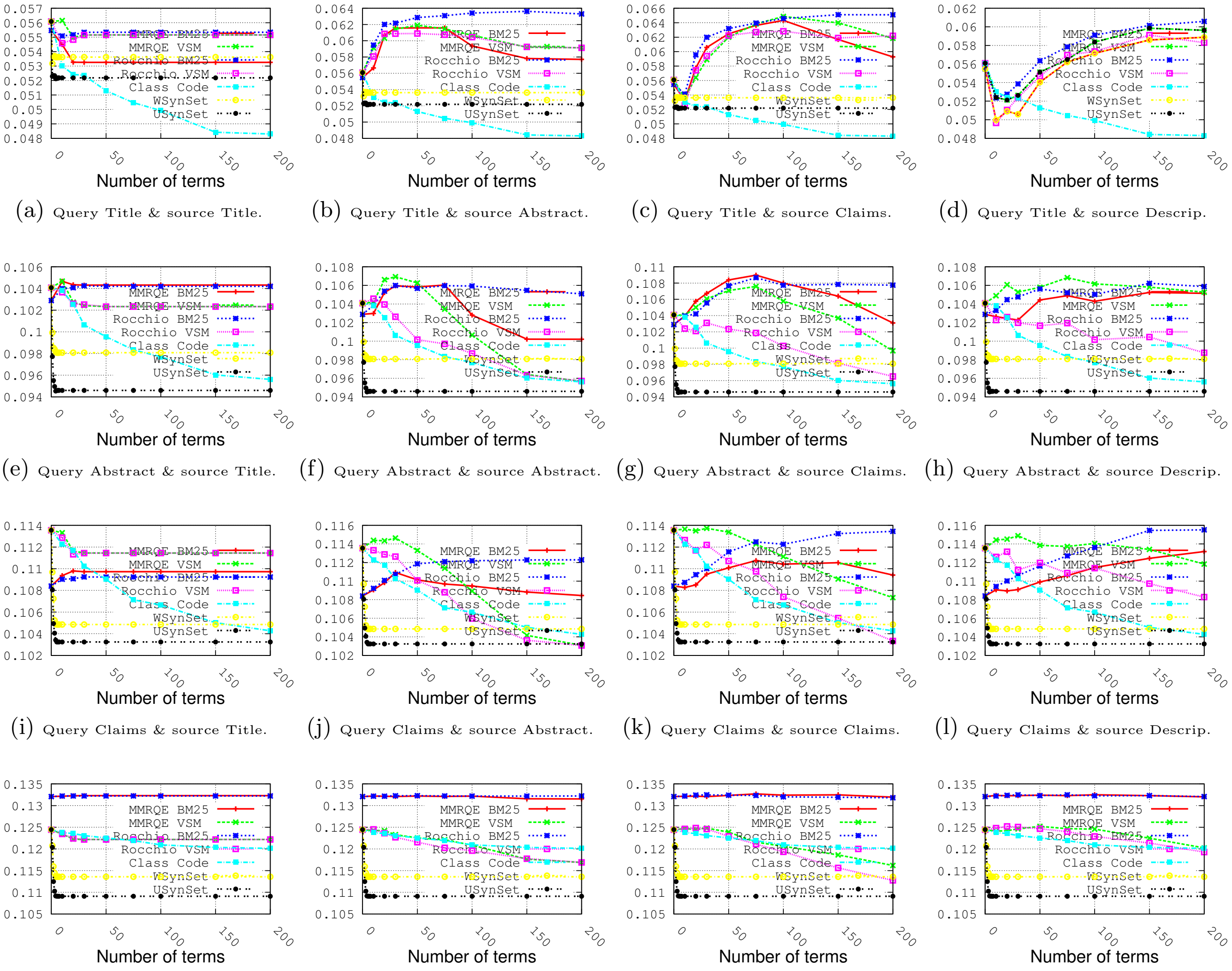
[7] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

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

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

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

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



(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.

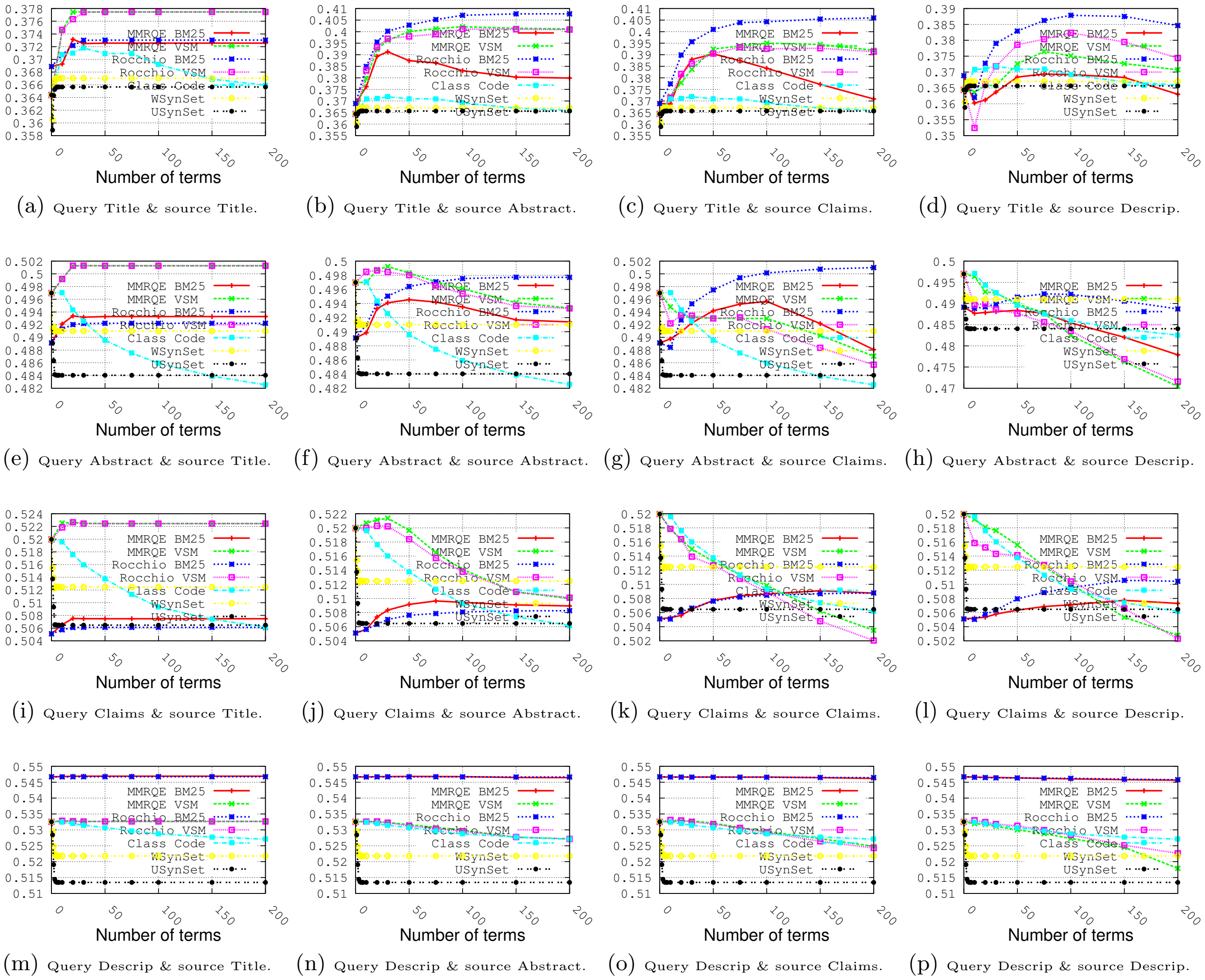
(m) Query Descrip & source Title.

(n) Query Descrip & source Abstract.

(o) Query Descrip & source Claims.

(p) Query Descrip & source Descrip.

Figure 3: Mean Average Precision (MAP) on CLEF-2010 (for MMRQE $\lambda = 0.5$).



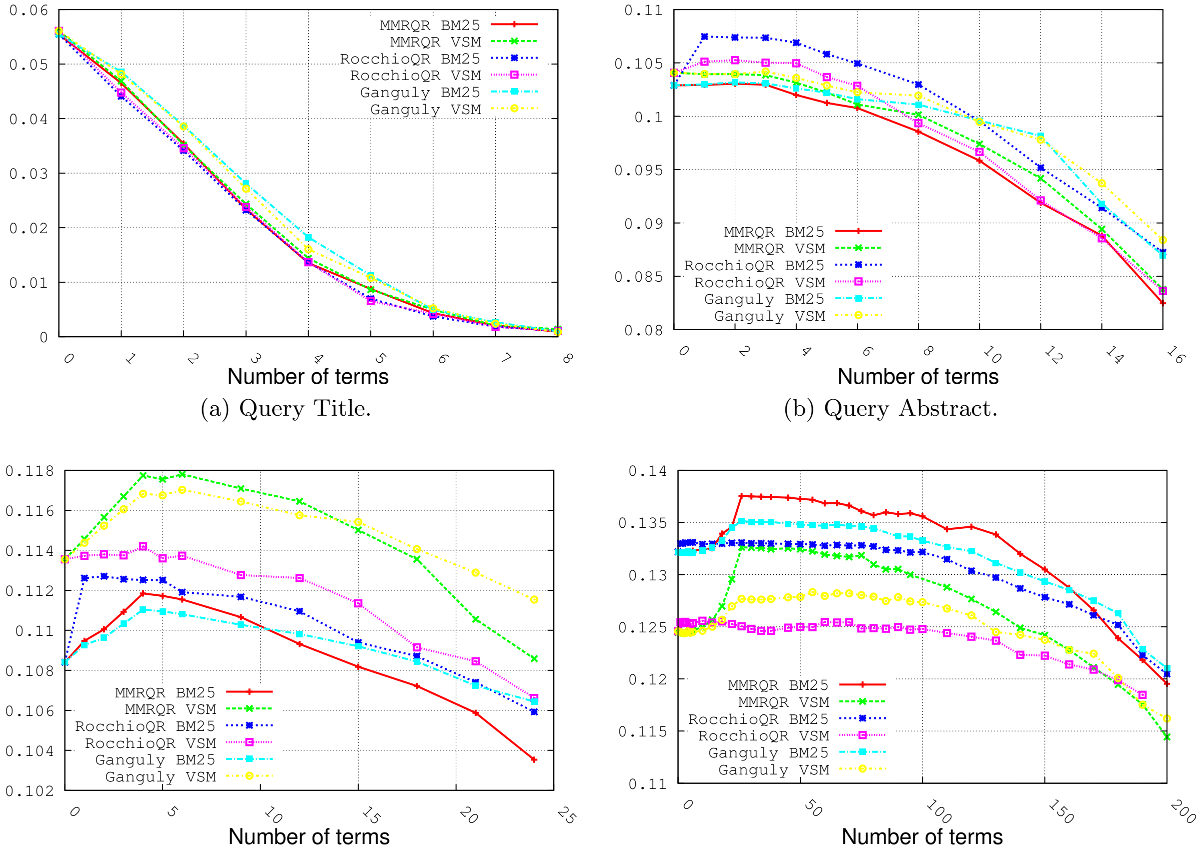
(a) Query Title & source Title. (b) Query Title & source Abstract. (c) Query Title & source Claims. (d) Query Title & source Descrip.

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



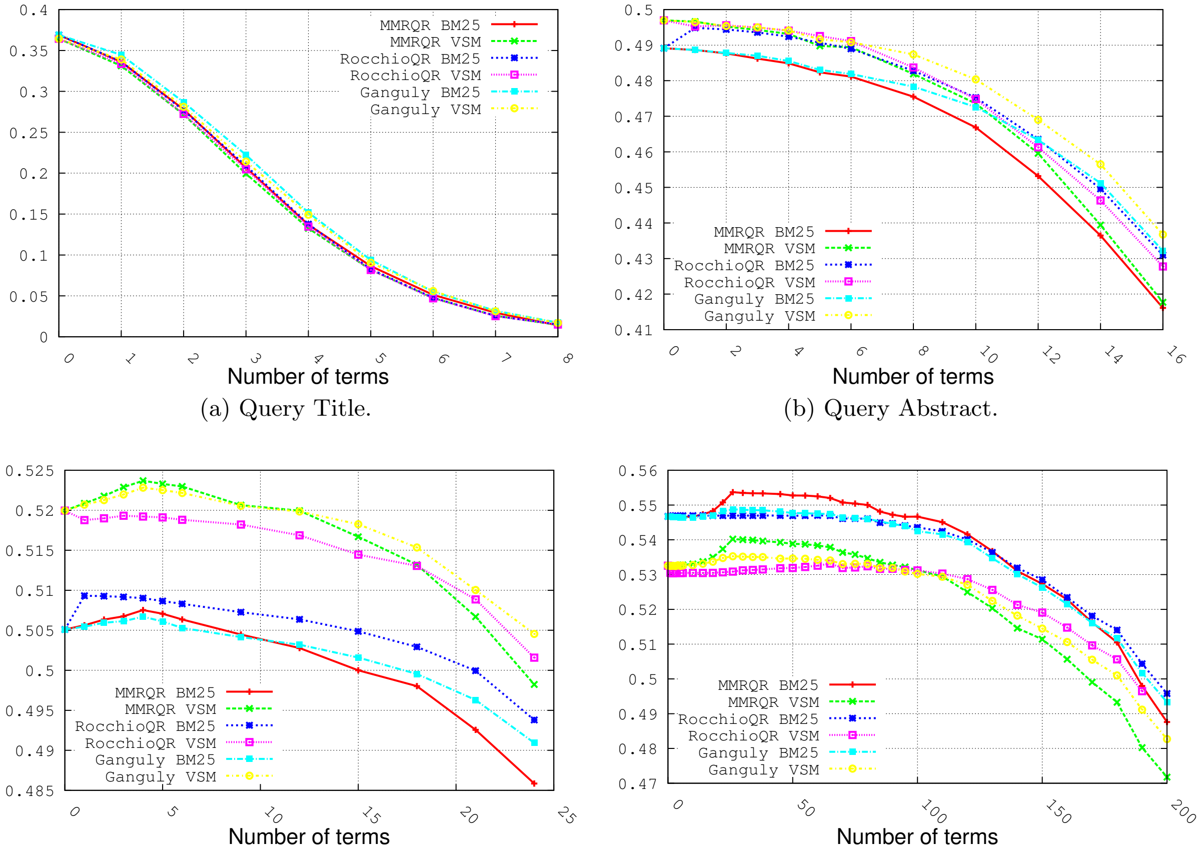
(a) Query Title.

(b) Query Abstract.

(c) Query Claims.

(d) Query Description.

Figure 5: 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 6: Patent Retrieval Evaluation Score (PRES) for MMRQR on CLEF-2010 (for MMRQE $\lambda = 0.8$).