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

No Author Given

No Institute Given

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 of query reformulation for patent prior art search with partial patent applications, 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 (a) exploit patent structure and (b) leverage techniques for diverse term selection in query reformulation. 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.

Keywords: Query Reformulation, Patent Search.

1 Introduction

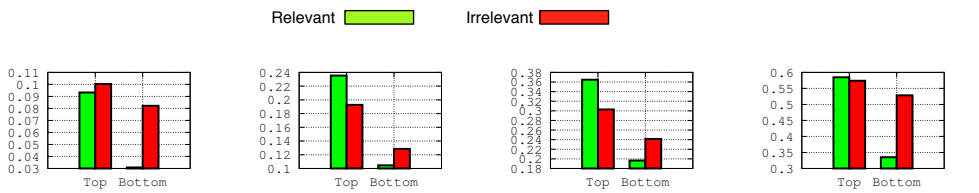
Patents are used by entities to legally protect their inventions and represent a multi-billion dollar industry of licensing and litigation. In 2012, 276,788 patent applications were approved in the US alone a number that has doubled in the past 15 years. Hence, helping both inventors and patent examiners 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 (partial) 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. Another important characteristic about patent patent prior art search is that, in contrast to scientific and technical writers, patent writers tend to generalize and maximize the scope of what is protected by a patent, and try to make sure that finding any relevant prior work by the patent examiner is a hard job.

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.

To assess the difficulty of querying with partial patent applications, we refer to Figure **??**. Here we show an analysis of the average Jaccard similarity[1] between different queries (representing the title, abstract, claims, or descriptions intended to represent a partial patent application) and the labeled relevant (all) and irrelevant documents (top 10 irrelevant documents ranked by BM25 [13]). We show results for the top 100 and bottom 100 queries (100 queries that perform the best, and 100 queries that perform the worst) of CLEP-IP 2010 evaluated according to Mean Average Precision (MAP). Note that, while the title section is usually composed by an average of six terms, the other sections are longer, ranging from ten to thousands of terms.



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

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

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

[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 [10].

set for any set of queries is less than one in four terms. Therefore, we suggest an investigation of *query reformulation* [1] methods as a means for improving the term overlap between queries that represent partial patent applications and relevant documents, with the objective of assessing not only the performance of standard query reformulation methods, but also the effectiveness of query reformulation methods that exploit patent-specific characteristics.

The rest of the paper is organized as follows: in Section we present the related work, in Section **??** we present query reformulation frameworks; in Section 5 we present our evaluation framework and results analysis; and in Section 6 we conclude with possible directions for future work.

2 Related work

Gaby said: this section should be shorter

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* (QR) [6] reduces the query such that superfluous information is removed, while *Query Expansion* (QE) [3] enhance the query with additional terms likely to occur in relevant documents.

Classical query expansion methods has been used for prior art search by Magdy et al. [9], 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 synonym sets for terms, and use them as a source of expansion terms, which showed significant improvement with respect to the baseline. Also, Bashir et al. [2] propose a query expansion with pseudo-relevance feedback. Query expansion terms are selected using a using a machine learning approach, by picking terms that may have a potential positive impact on the retrieval effectiveness. However, this approach can be computational expensive, since the presented features are complicated to compute, e.g. Pair-wise Terms Proximity features. Verma and Varma [17] propose a different approach, which instead of using the patent text to query, use its International Patent Classification (IPC) codes, which are expanded using the citation network. The formed query is used to perform an initial search. The results are then re-ranked using queries constructed from patent text. Throughout our experiments, we concluded that relying on non-patent terms to expand a query, leads to poor retrieval quality. Lastly, a more recent work by Mahdabi et al. [11] propose to combine query reduction and xpansion method for prior art search. For query reduction, they shorten the query by taking only the first claim since it contains the core of the invention. For the query expansion, they built a query-specific patent lexicon based on the definitions of the IPC. Then, the patent lexicon is used to select expansion terms that are focused on the reduced query. Lastly, a more recent work by Mahdabi et al. [11] 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.

Also query reduction methods have been applied in order to deal with long queries, which are composed by full patent applications [4,5]. Even these methods are interesting, they are not suited to deal with short queries (partial patent applications). [4] technique decomposes a query (a patent section) into constituent text segments and computes 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. Recently, [11] 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 has not been implemented as its authors already point out its poor performance.

3 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 considered:

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

Relevance model: We explore a probabilistic approach represented by the popular BM25 [13] algorithm, as well as a vector space model (VSM) approach, TF-IDF [16]. A natural question is which relevance model works best for query reformulation for patent prior art search?

Query expansion source: We 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? In addition we included two other patent specific QE methods as baselines. Motivated by [11], 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 two variants of the QE approach proposed in [9], which automatically generates candidate synonyms sets (SynSet) for terms, and use it as a source of expansion terms, which are denoted as **WSynSet**) and **USynSet**). Note that this only applies to query expansion methods.

Term selection method: We consider different term selection methods for query reformulation. We evaluate the performance of term selection using Rocchio [15] 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?

Query reduction method: 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**. 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 **StopIPC Codes**.

In summary, ...

- **Query type:** {Title, Abstract, Claims, Description}
- **Relevance model:** {BM25, Vector-Space Model}
- **Query expansion source:** {Title, Abs., Claims, Description, IPC, WSynSet, USynSet}
- **Query reformulation method:** {Rocchio, RocchioQR, LMQR, StopIPC}

For all methods, their parameters were fixed to their optimal values, which were estimated using the CLEF-IP training queries.

In the next section, we propose a deep evaluation of query reformulation methods...

4 Experimental Setup

We used the Lucene IR System[2] to index the English subset of CLEF-IP 2010 and CLEF-IP 2010 datasets[3] [12,14] 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 [7,14]. Finally, we report MAP, and PRES, 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.

Gabi said: a review from ckim commented that the ipc filter could be too restricted taking into account that patent applications usually dont contained revised ipc codes. But, if I remembered correctly, when you turn off the filter the results were no significant different and you decided to keep the filter on, since the systems works faster. Is that true? . If yes, we should commmented in the paper.

5 Experimental Evaluation

In this section, we discuss the results of the evaluation performed on the query reformulation methods described above.

Figure 2 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 (VSM and BM25), 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).

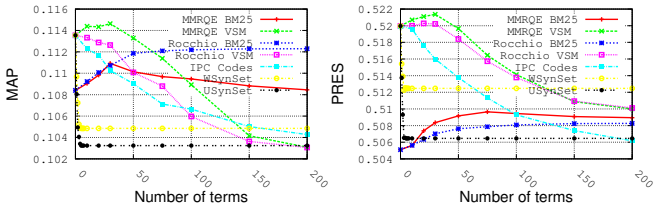


Fig. 2: Results obtained while using the claims for querying and the abstract as source of query expansion on the CLEF-IP 2010 dataset.

To summarize all the results obtained over all the above configurations, Figures 4, 5, 6, and 7 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 for each method). From these results, we first observe that the best section to use for querying is the description section (see Figures ??, ??, ??, and ??). We attribute this to the fact that the description section has more content along with relevant terms that define the invention since a detailed summary of the invention is described therein.

According to our experiments, the **best source for query expansion is the claims section**. We attribute this to the fact that, the claims contained not only relevant, but also, specific terminology, since the scope of the invention is described therein. However, when querying using the claims, other sources of query expansion provided better performance. This may be because claims contained a lot of repetition and consequently, lack of the diversity necessary to capture similar inventions that are described using synonyms or more specific or abstract terms. It is interesting to notice that the description is not either a good source for expansion, since its content is too broad, therefore, it contains many irrelevant terms that hurt the performance.

As expected, we observed that query expansion is not useful for very long queries (i.e. description), indicating that in advanced stages of the patent application process, QE is not relevant. We also notice that using external sources for expansion such as the IPC code definitions, synsets from Wordnet gave poor performance.

Discussion As recommended in [4] 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 3 shows the results obtained in terms of MAP and PRES for CLEF-IP 2010 for different numbers of removed terms k on the x-axis (with $k = 0$ using no QR, just the baseline retrieval model). For lack of space we show only the results of queries extracted from the description. These results tell us mainly two things: (i) for the two retrieval models, MMRQR provides the best performance for both MAP and PRES, and (ii) for almost all methods, the best performance is obtained when removing about 30 terms from the original queries (in the case where the description is used for querying). Removing more terms will decrease significantly the performance.

To summarize all the results obtained over all the above configurations, Figures 8, 9, 10, and 11 show the performance obtained for all the QR methods, when selecting the optimal number of terms removed from the original queries (number of terms removed that maximizes the performance for each method).

From these results, we make the following observations: (i) query reduction is very often not useful for short queries (i.e. title), since no QR method outperforms significantly the baseline (i.e. No QR), (ii) when dealing with very long

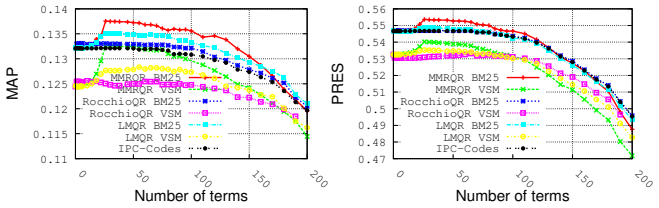


Fig. 3: Results obtained for QR while using the description for querying on the CLEF-IP 2010 dataset.

query (i.e. description), BM25 based QR methods perform better than VSM based QR methods

6 Conclusion

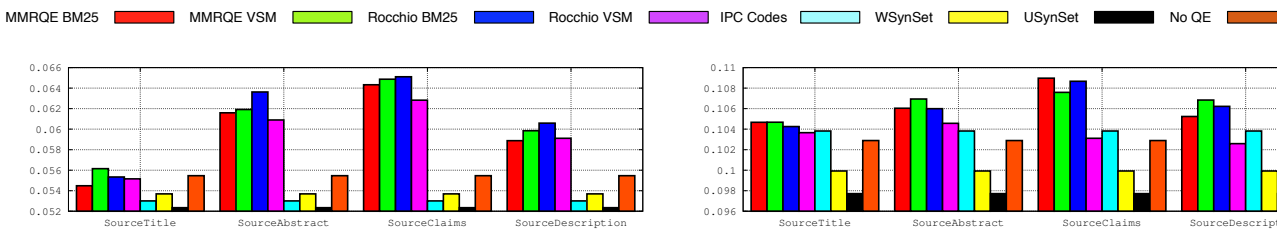
In this paper we analyzed general and specific QE and QR methods for patent prior art search for partial (incomplete) patent applications on two patent retrieval corpora, namely CLEF-IP 2010 and CLEF-IP 2011. We demonstrated that QE methods are critical for short queries, i.e. title, abstract, and claims, but useless for very long queries, i.e. the description section. We also showed that claims is the best section that works with QE both to query with and to use as a source of query expansion terms, suggesting that claims should be written at early stages of the patent application drafting so that they can be use to performed patent prior art search. We also demonstrate that the novel MMRQE method 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 extent for claims and description sections, which are considered the longest sections in a patent application. We also demonstrated that our new QR method MMRQR improve both recall and precision in many cases. Future work may consist of exploiting query quality predictors to identify useless terms in a query using machine learning methods.

References

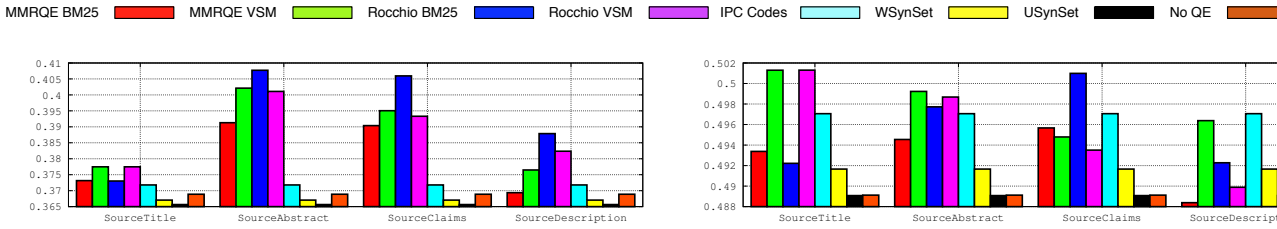
1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2010.
2. S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, 2010.
3. E. N. Efthimiadis. Query expansion. *Annual Review of Inf. Systems and Technology (ARIST)*, 31:121–187, 1996.

4. D. Ganguly, J. Leveling, W. Magdy, and G. J. Jones. Patent query reduction using pseudo relevance feedback. In *CIKM*, 2011.
5. 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.
6. G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR*, 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. 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. A study on query expansion methods for patent retrieval. In *PaIR*, 2011.
10. P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *SIGIR*, pages 505–514, New York, NY, USA, 2012. ACM.
11. P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *SIGIR*, 2013.
12. 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.
13. S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-2. In *TREC*, pages 21–34, 1993.
14. G. Roda, J. Tait, F. Piroi, and V. Zenz. Clef-ip 2009: Retrieval experiments in the intellectual property domain. In C. Peters, G. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Penas, and G. Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, volume 6241 of *Lecture Notes in Computer Science*, pages 385–409. Springer, 2009.
15. G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.
16. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Comm. ACM*, 18(11):613–620, nov. 1975.
17. M. Verma and V. Varma. Patent search using ipc classification vectors. In *PaIR*, 2011.



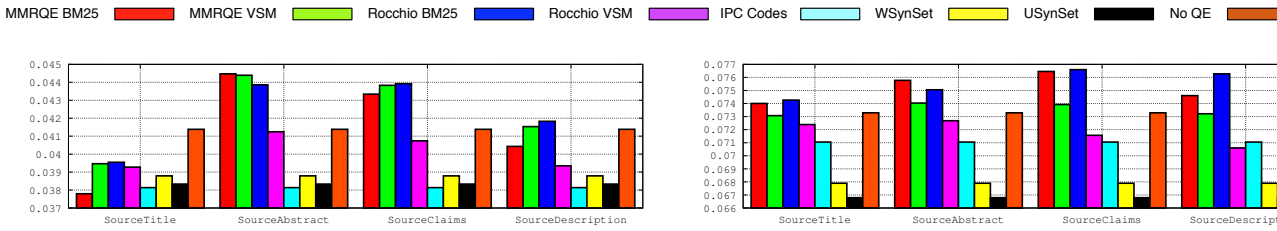
(a) Query Title. (b) Query Abstract.

Fig. 4: Mean Average Precision (MAP) for QE methods on CLEF-IP 2010 (for MMRQE $\lambda = 0.5$).



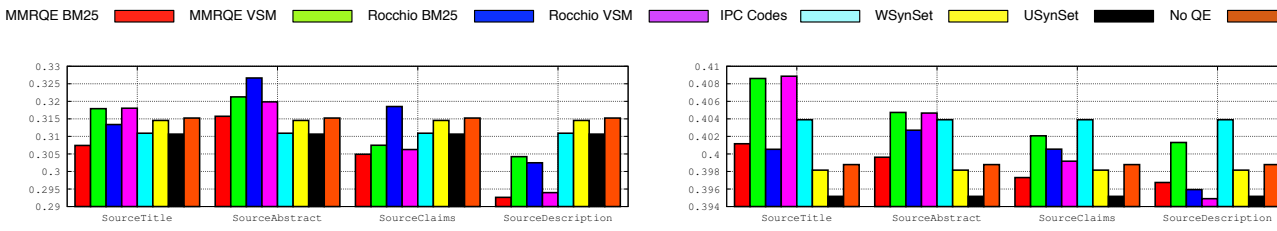
(a) Query Title. (b) Query Abstract.

Fig. 5: Patent Retrieval Evaluation Score (PRES) for QE methods on CLEF-IP 2010 (for MMRQE $\lambda = 0.5$).



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

Fig. 6: Mean Average Precision (MAP) for QE methods on CLEF-IP 2011 (for MMRQE $\lambda = 0.5$).



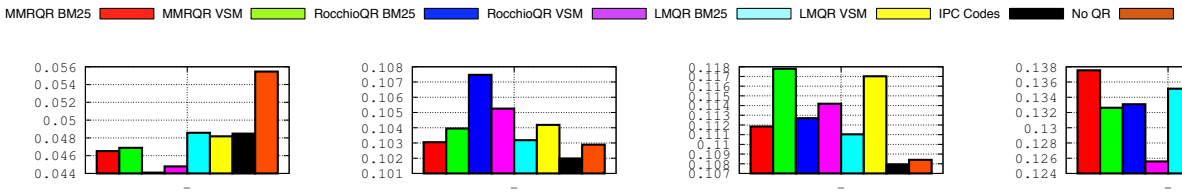
(a) Query Title.

(b) Query Abstract.

(c) Query Claims.

(d) Query Description.

Fig. 7: Patent Retrieval Evaluation Score (PRES) for QE methods on CLEF 2011 (for MMRQE $\lambda = 0.5$).



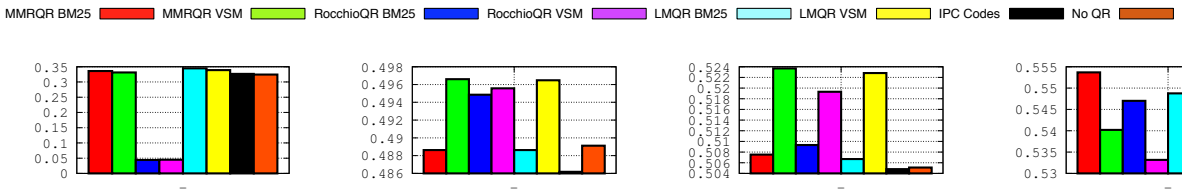
(a) Query Title.

(b) Query Abstract.

(c) Query Claims.

(d) Query Description.

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



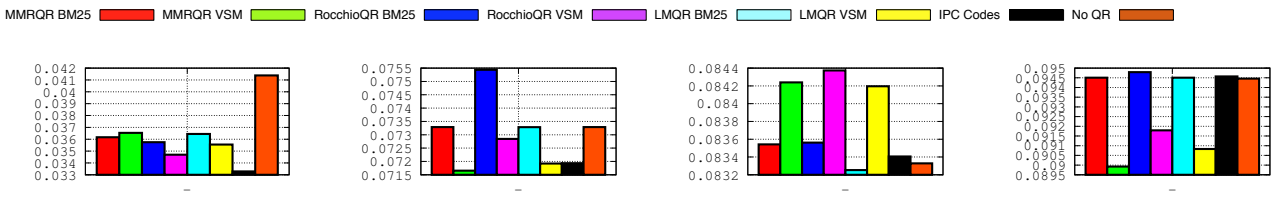
(a) Query Title.

(b) Query Abstract.

(c) Query Claims.

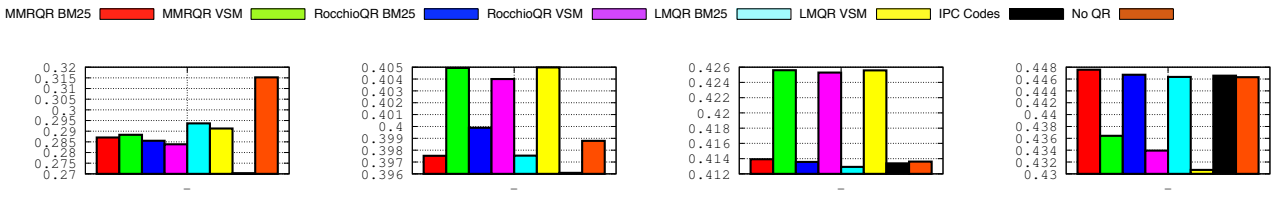
(d) Query Description.

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



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

Fig. 10: Mean Average Precision (MAP) for QR methods on CLEF-IP 2011 (for MMRQR $\lambda = 0.8$).



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

Fig. 11: Patent Retrieval Evaluation Score (PRES) for QR methods on CLEF 2011 (for MMRQR $\lambda = 0.8$).