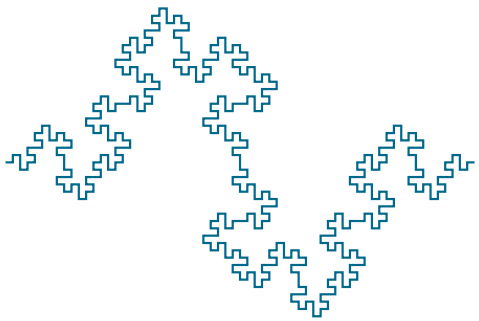
A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications

Mohamed Reda Bouadjenek
Gabriela Ferraro
Scott Sanner



Outline

- Prior art search
- Query reformulation for patents
- General Query reformulation methods
-
- Experiments
- Results and discussion
- Conclusion and future work

WHAT PATENTS ARE?

Patents are legal documents to protect an invention.

- **Rich meta**: Inventor, Author, Company, Country, Publication year, etc.
- **Predefined document structure**: Title, Abstract, Description, Claims.

Patent Applications vs. **Granted Patents**

What is patent prior art search?

Finding previously granted patents relevant for a patent application

- Patent examiners
- Patent authors/inventors

Challenges and data sets:

- NTCIR (since 2002)
- TREC-Chem (2007)
- CLEF-IP (2010/2011)*

PATENT PRIOR ART SEARCH

Why patent prior art search is different to standard Information Retrieval (IR)?

- **Queries** are full patent applications (hundreds of words organized into several sections)
- **Recall-oriented** task (retrieve all relevant documents at early ranks) while text and web search are **precision-oriented** (retrieve a subset of relevant documents)

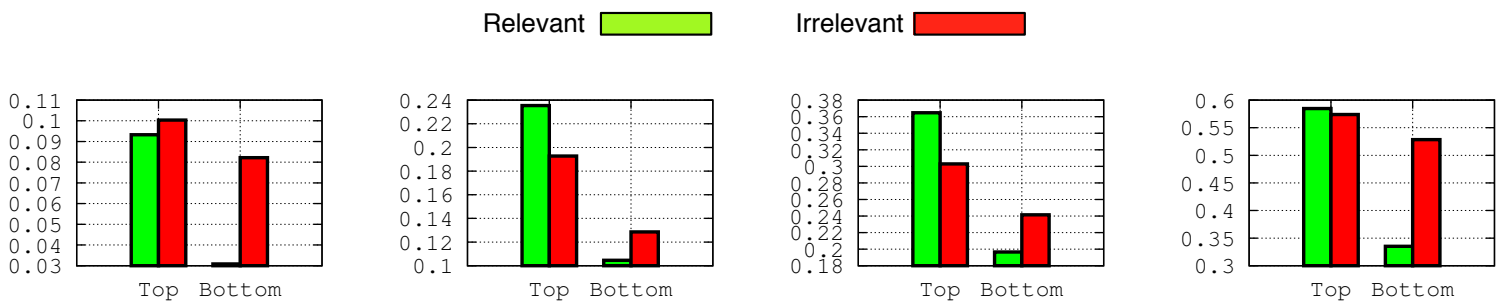
Patent Prior Art Search with partial applications

Writing a full patent application is time-consuming and costly

- We proposed to do patent prior art search with **partial (incomplete) patent applications**
- Intensive study about query reformulation for patent prior art search with partial patent applications

QUERYING WITH PARTIAL PATENT APPLICATIONS

- Term overlap (Jaccard Coefficient) of (ir)relevant documents with the result sets for different queries
- Top 100 and bottom 100 queries
- Top 10 irrelevant documents ranked by BM25 [?]
- CLEP-IP 2010



(a) Title query (b) Abs. query (c) Claims query (d) Desc. query

QUERYING WITH PARTIAL PATENT APPLICATIONS

There are 3 notable trends:

(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;

(iii) the best overlap for any relevant document set for any set of queries is less than one in four terms.

We investigate Query Reformulation methods

Query Reformulation

Query reformulation *is the process of transforming an initial query* Q *to another query* Q'*.*

- **Query Reduction** (QR) [**?**]: reduces the query such that superfluous information is removed.
- **Query Expansion** (QE) [**?**]: enhance the query with additional terms likely to occur in relevant documents.

QUERY REFORMULATION FOR PATENTS

- **Query type**: title, abstract, claims, description.
  What part of a partial application an inventor should write to obtain the best search results?
- **Relevance model**: BM25, vector space model (VSM): TF-IDF [**?**]
  Which relevance model works best for query reformulation for patent prior art search?
- **Query expansion source**: title, abstract, claims, description
  Are the title words of particularly high value as expansion terms?
- **Term selection method**: Rocchio [**?**], MRRQR
  Which is the best selection method? and with which query type, retrieval model, and term source?

Query Expansion (QE) Frameworks

QE aims to alleviate the term mi mismatch between queries an relevant documents.

Rocchio

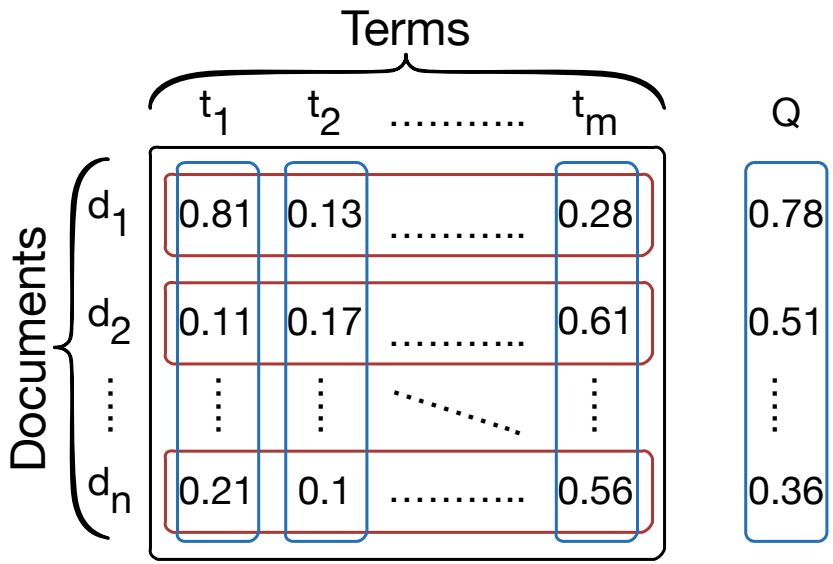
Derives a score for each potential query expansion term and in practice, the top-k scoring terms (often for $k \ll 200$) are used to expand the query and are weighted according to their Rocchio score during the second stage of retrieval.

What is missed in Rocchio?

DIVERSE TERM SELECTION

Maximal Marginal Relevance (MMR), a result set diversification algorithm (MMR) [**?**], usually used for **diverse document selection** (e.g., multi-document summarization)

NOTATION USED IN MMR QE/QR



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

QUERY REDUCTION FRAMEWORKS

QR aims to short long queries

We investigate the impact of QR methods when querying with long sections such as *abstract*, *claims* or *description*.

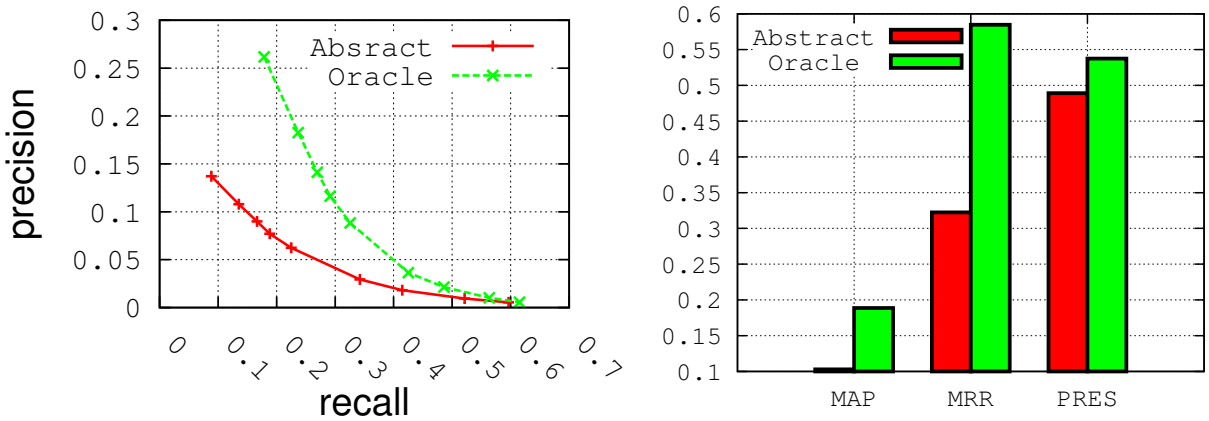


Figure: Sample of terms removed from the abstract section

MAP: Mean Average Precision; **MRR**: Mean Reciprocal Rank; **PRES**: Patent Retrieval Evaluation Score

THE UTILITY OF QUERY REDUCTION FOR 1304 ABSTRACT QUERIES OF THE CLEF-IP 2010 DATASET

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

Experiments Setup

- CLEF-IP 2010:
  - 2.6 million patent documents
  - 1303 English topics (queries)
- CLEF-IP 2011: 3 million patent documents
  - 2.6 million patent documents
  - 1351 English topics (queries)
- Lucene IR System
- LucQE: Rocchio method for Lucene
- Standard stop-words removal
- Patent-specific stop-words removal [**?**]
- Each patent section is indexed in a separate field
- Queries target all the fields in the index
- Filtering using the International patent Classification (IPC) of the queries [**?**, **?**]
- Evaluation on the top 1000 results

Query Expansion Baselines

Other QE methods

Magdy et al. [**?**] classic techniques of query expansion: WordNet Bashir et al. [**?**] with pseudo-relevance feedback.
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, which features??
Verma and Varma [**?**]: used International Patent Classification (IPC) codes as queries, which are expanded using the citation network.

PSEUDO RELEVANT FEEDBACK (PRF) SIZE

Effect of PRF set with various numbers of feedback documents on the CLEF-IP 2010 dataset.

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

* 20 terms are used for query expansion

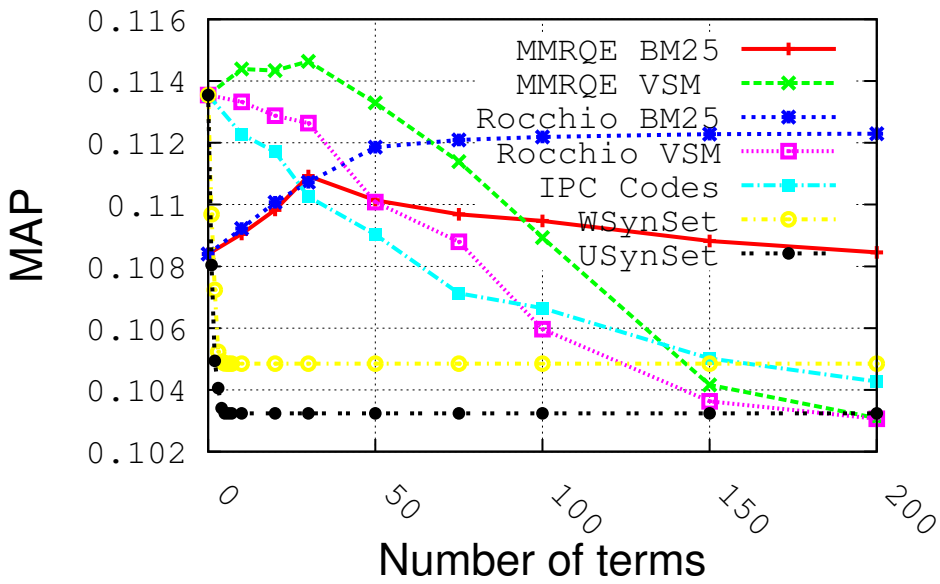
EXPERIMENTS FOR QE

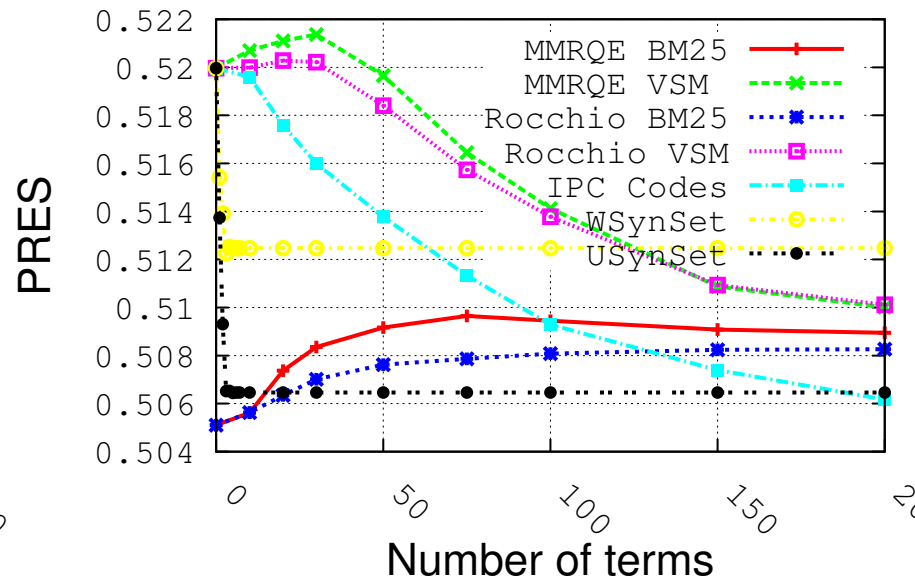
Experiments options:

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

EXPERIEMENTS RESULTS FOR QE

Query: Claims section Expansion source: Abstract section Date set: CLEF-IP 2010 dataset





Discussion about QR

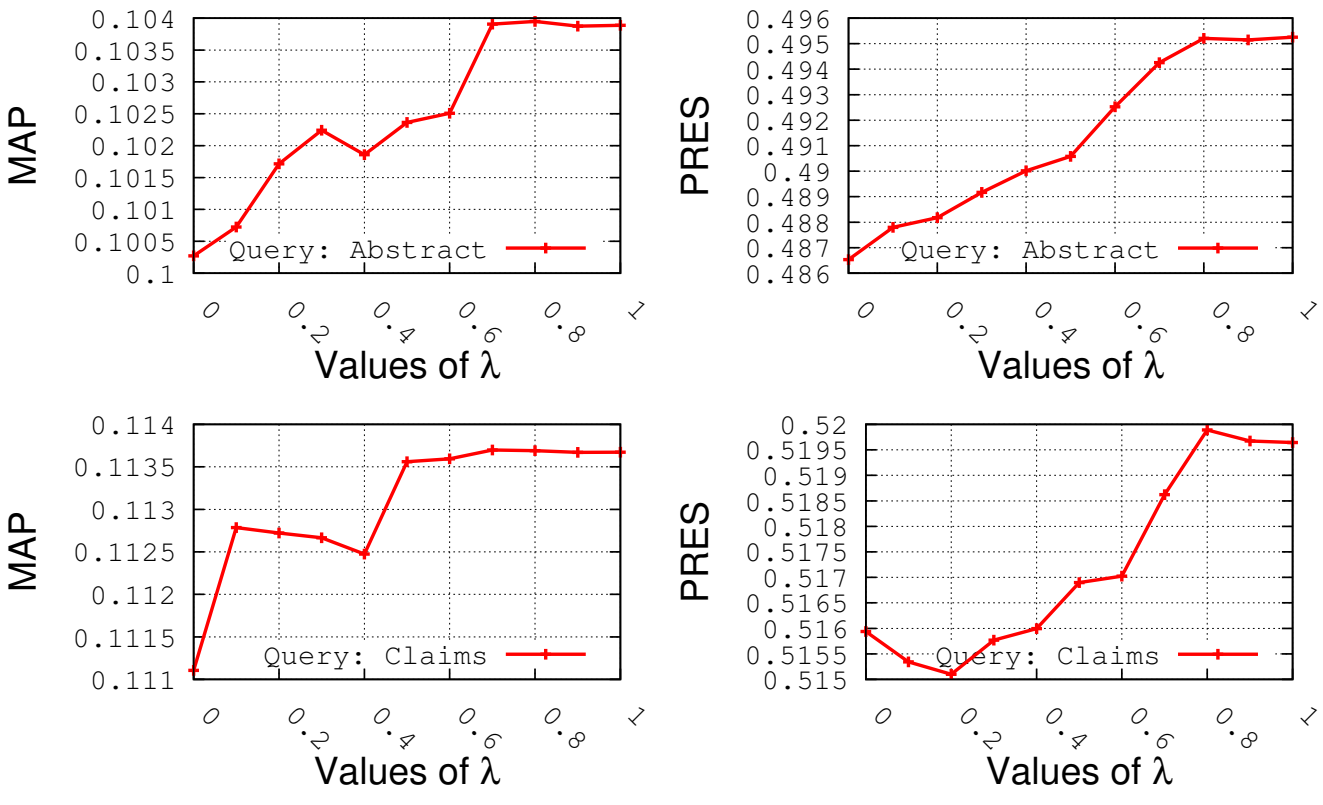
Samples of queries (CLEF-IP 2011) where QE improves the performance

Query Reduction baselines

General QR method: Rocchio method for query pruning [**?**]: This 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. This approach is denoted **LMQR**.
IPC codes for query reduction: (i) For each patent application, we take the definitions of the IPC codes which are associated to it. (ii) rank the terms of the query according to both their frequency in the class code definition, and their frequency in the query. (iii) remove bottom terms of this ranking from the query (i.e. good terms are terms that occur a lot in the query, and few in the class code definition, whereas bad terms are those that occur few in the query, and a lot in the class code definition). The intuition is that, terms in the IPC code definition may represent "stopwords", especially if they are rare

QR DISCUSSION

Best QR performance results are also obtained when using few documents in the PRF set (top 5).
Impact of the diversity parameter λ on the performance of MMRQR on the CLEF-IP 2010 dataset



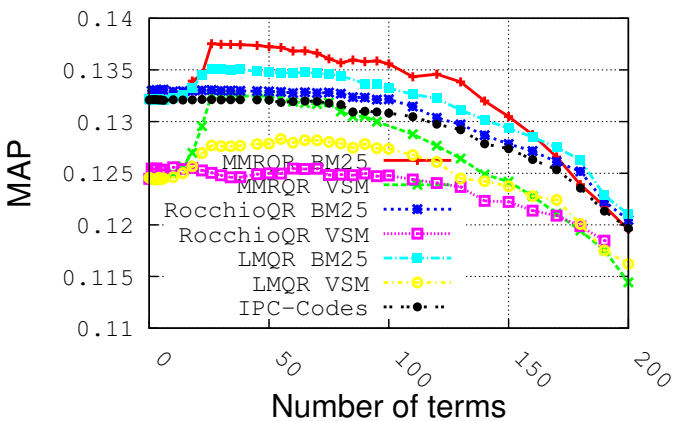
QR EXPERIMENTS

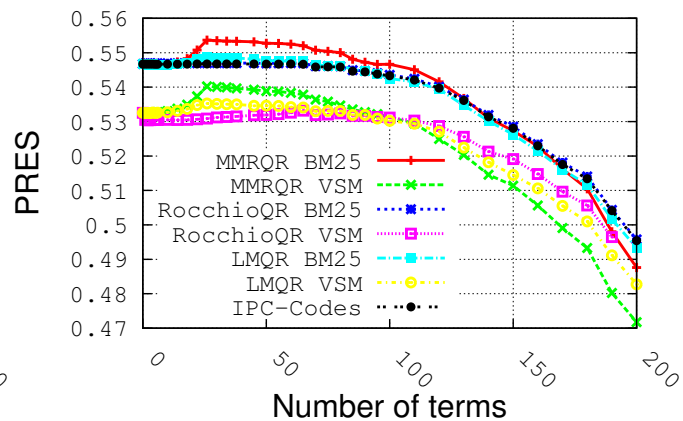
Experiment options:

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

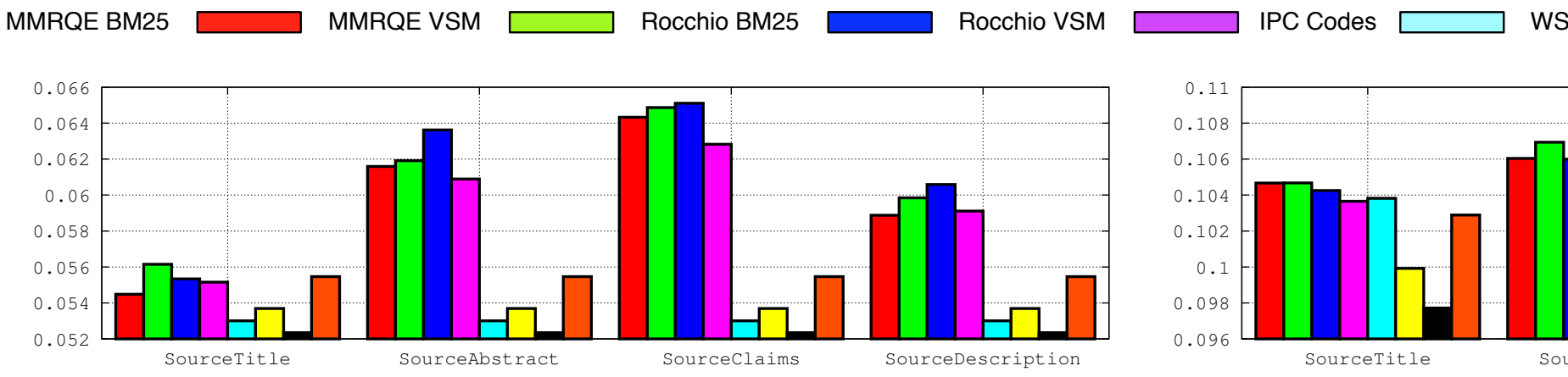
Sample table!!!

QR WHILE USING THE DESCRIPTION SECTION FOR QUERYING (CLEF-IP 2010)

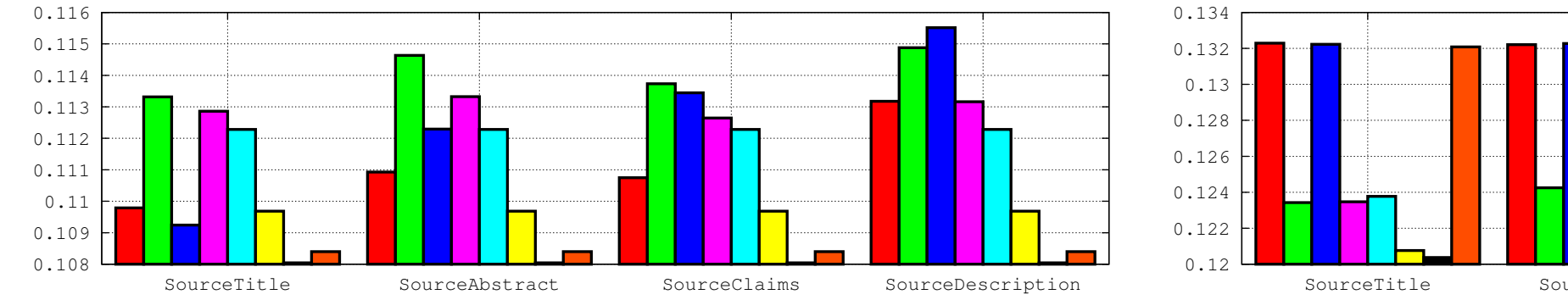




MAP AND PRES FOR QE METHODS ON CLEF-IP 2010



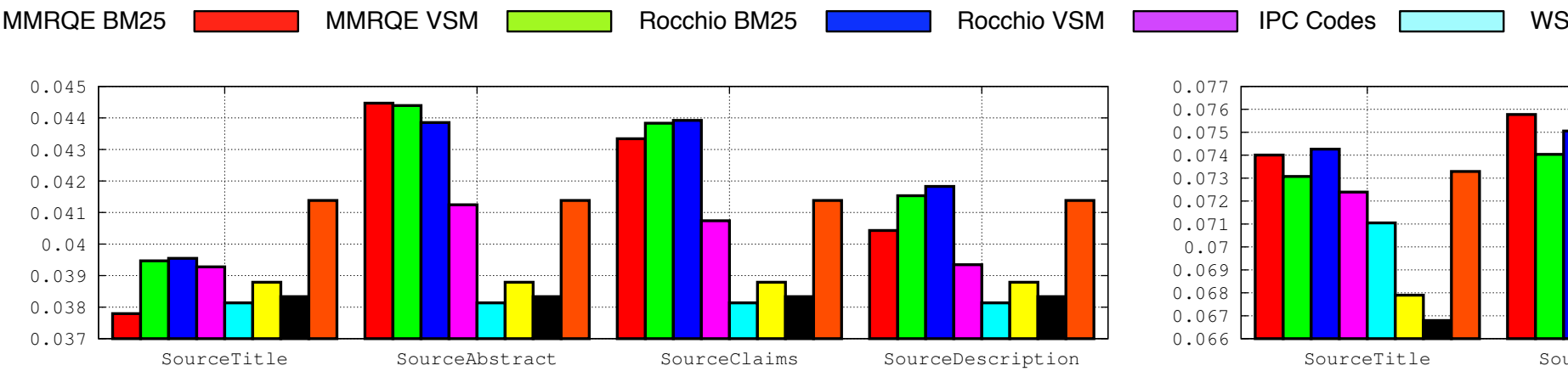
(a) Query Title.



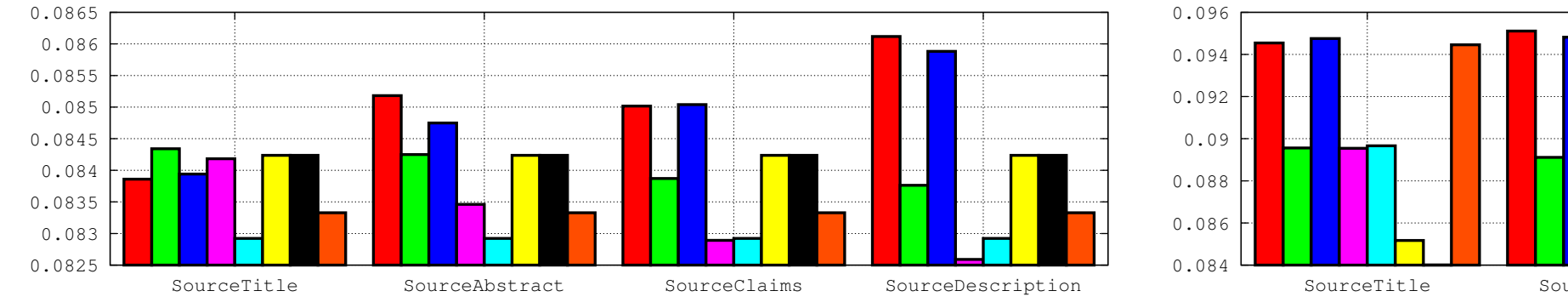
(c) Query Claims.

MAP AND PRES FOR QE METHODS ON CLEF 2011

for MMRQE $\lambda = 0.5$



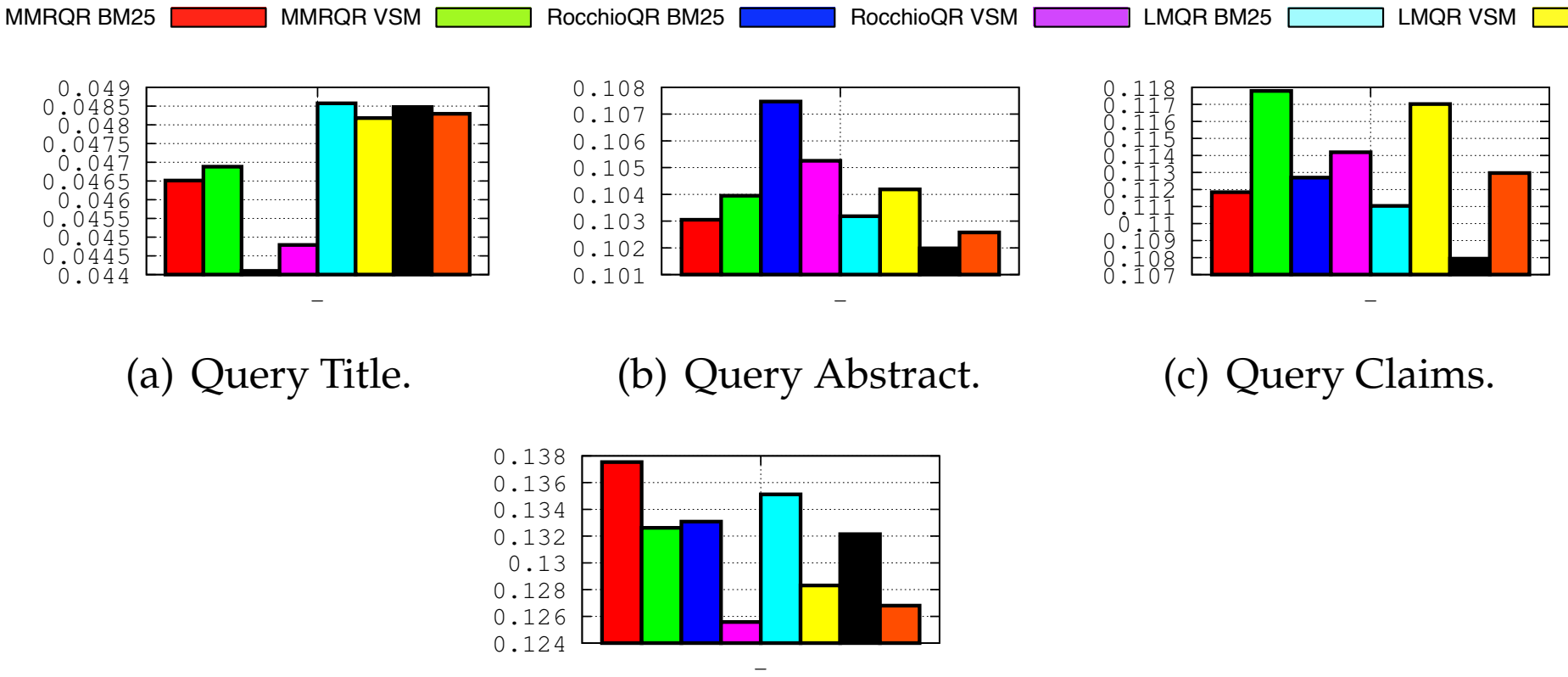
(a) Query Title.



(c) Query Claims.

MAP AND PRES FOR QR METHODS ON CLEF 2010

for MMRQR $\lambda = 0.8$



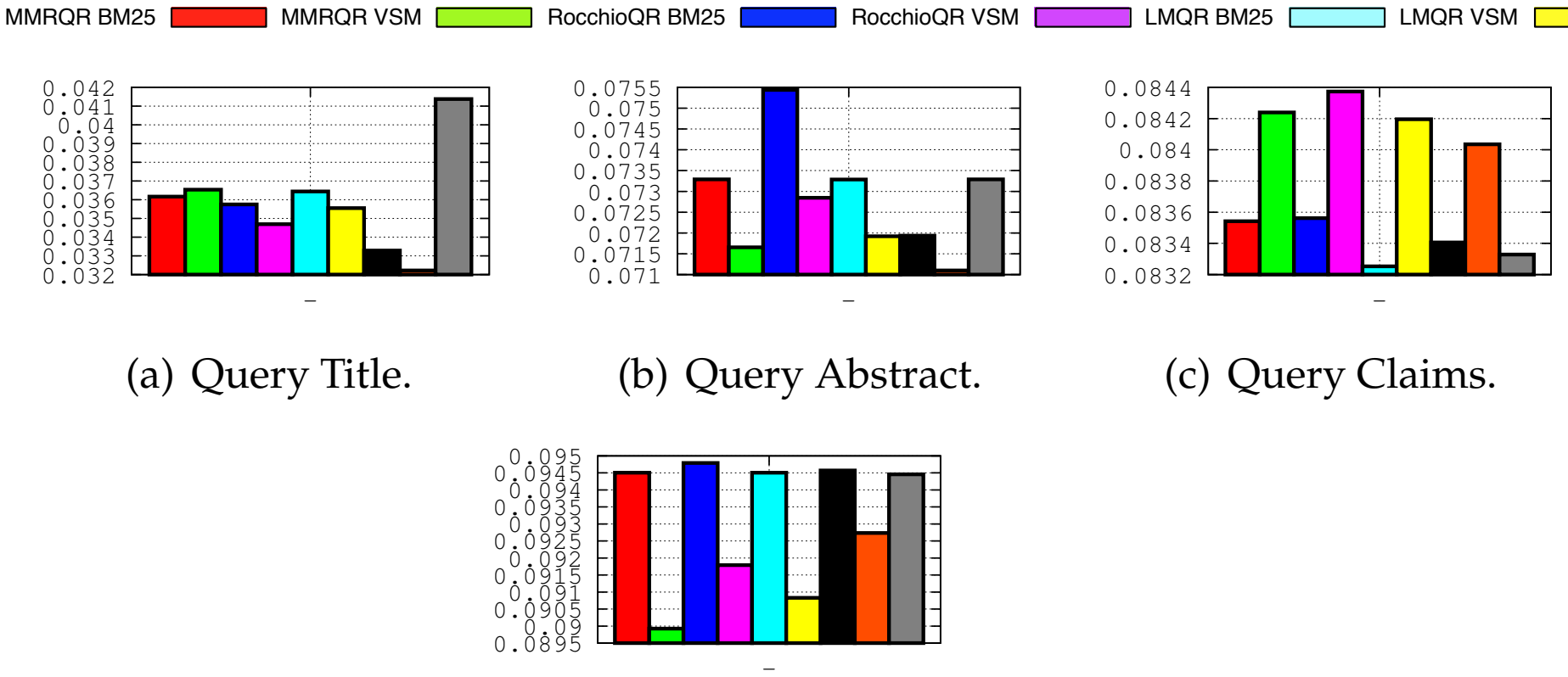
(a) Query Title. (b) Query Abstract. (c) Query Claims.

(d) Query Description.

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

MAP AND PRES FOR QR METHODS ON CLEF 2011

for MMRQR $\lambda = 0.8$



(a) Query Title.

(b) Query Abstract.

(c) Query Claims.

(d) Query Description.

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

Contributions are the following:

1. Novel contributions for query expansion and reduction that leverage (a) patent structure and (b) term diversification techniques.
2. A thorough comparative analysis of existing and novel methods for query expansion and reduction in patent prior-art search on standardized datasets of CLEF-IP.