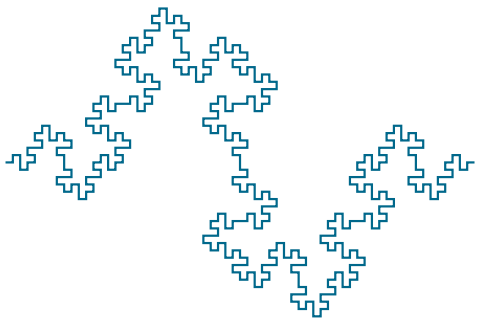
A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications

Mohamed Reda Bouadjenek
INRIA, France
Gabriela Ferraro and Scott Sanner
NICTA and Australian National University



Outline

- Patent prior art search
- Query reformulation for patents
- Diversification methods for query reformulation
- 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, Classification codes, etc.
- **Predefined document structure**: Title, Abstract, Description and 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 (lawyers, 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?

- **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)

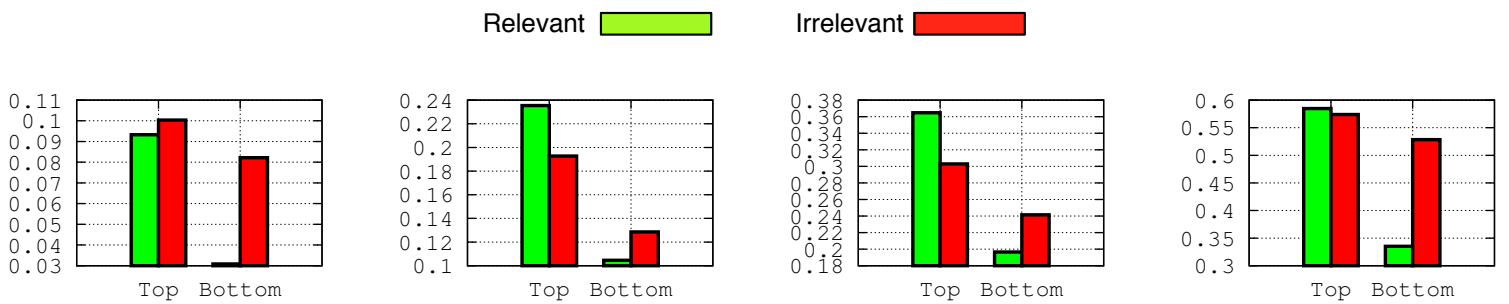
PATENT PRIOR ART SEARCH WITH PARTIAL APPLICATIONS

Writing a full patent application is time-consuming and costly

- We proposed to do patent prior art search with **partial (incomplete) patent applications**

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 (Robertson et al., 1993)
- 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) (Kumaran and Carvalho, 2009): reduces the query such that superfluous information is removed.
- **Query Expansion** (QE) (Efthimiadis, 1996): 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 and Vector Space Model, TF-IDF (Salton et al., 1975)
  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 (Salton, 1971), MRR QE/QR
  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 (Salton, 1971)

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

What is missed in Rocchio?

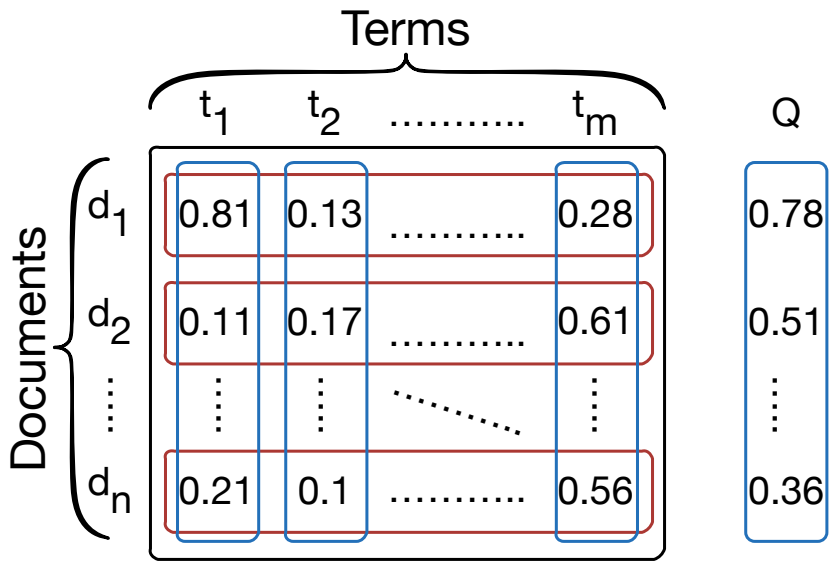
With a limited budget of k expansion terms, there is no guarantee that these terms cover all documents in the pseudo-relevant set.

DIVERSE TERM SELECTION

We proposed a term selection method that takes into account diversity

Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) is a result set diversification algorithm, 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 EXPANSION BASELINES

- General QE method
  - **Rocchio** (Salton, 1971)
- Patent specific QE methods
  - **IPC** (Mahdabi et al., 2013) used the text definitions of the codes assigned to a patent application as a source for expansion.
  - **WSynSet** (Magdy and Jones, 2011) used the probability associated with the SynSet entries as a weight for each expanded term in the query.
  - **USynSet** (Magdy and Jones, 2011) used uniform weighting for all synonyms of a given term

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

Other QE methods

- Magdy et al. (Magdy and Jones, 2011) classic techniques of query expansion: WordNet
- Bashir et al. (Bashir and Rauber, 2010) with SRF set, used a machine learning approach by picking terms that may have a potential positive impact on the retrieval effectiveness.
- Verma and Varma (Verma and Varma, 2011): used IPC codes as queries, which are expanded using the citation network.

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*.

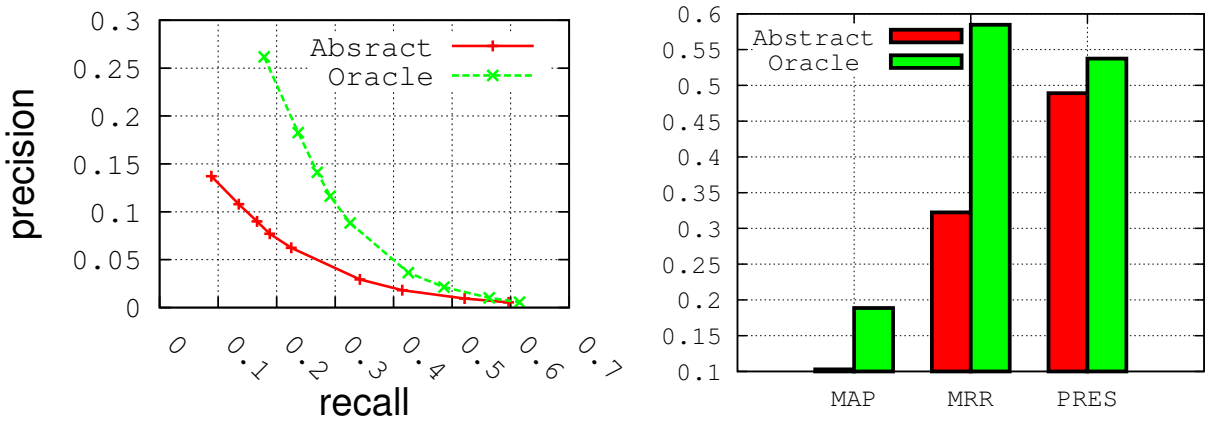


Figure: Sample of terms removed from the abstract section

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

QUERY REDUCTION BASELINES

- General QR method
  - **RocchioQR** use the lower Rocchio score for for query pruning
- Patent specific methods
  - **LMQR** (Ganguly et al., 2011): (i) computes Language Modeling similarities by calculating the probability of generating each segment from the top ranked documents; (ii) remove the least similar terms.
  - **IPC**: (i) rank the terms of the query according to both their frequency in the class code definition, and their frequency in the query. (ii) remove bottom terms of this ranking.

Other work: (Mahdabi et al., 2013) short queries by taking only the first claim of a patent application.

EXPERIMENTS SETUP

- CLEF-IP 2010:
  - 2.6 million European patent documents
  - 1303 English topics (queries)
- CLEF-IP 2011: 3 million patent documents
  - 2.6 million European patent documents
  - 1351 English topics (queries)
- Lucene IR System
- LucQE: Rocchio method for Lucene
- Standard stop-words removal
- Patent-specific stop-words removal (Magdy, 2012)
- 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 (Lopez and Romary, 2009; Roda et al., 2009)
- Evaluation on the top 1000 results

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*...}

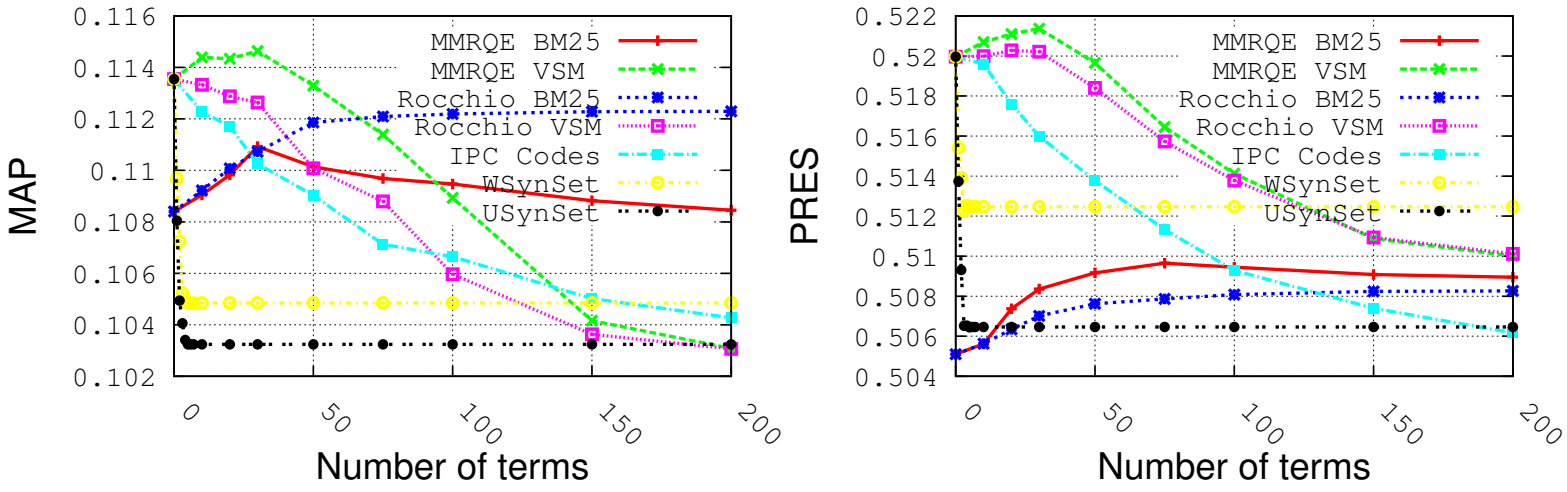
PSEUDO RELEVANT FEEDBACK (PRF) SIZE

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

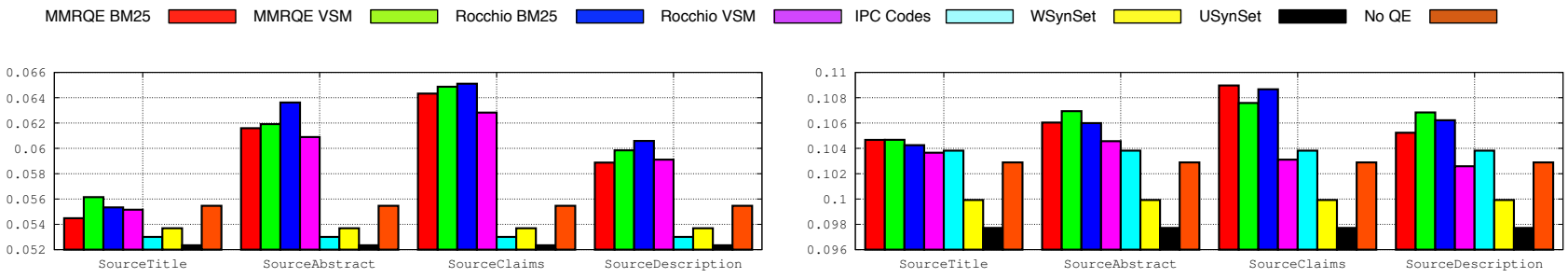
EXPERIEMENTS RESULTS FOR QE



Query: Claims, **Date set**: CLEF-IP 2010 , **Expansion source**: Abstra

- (i) for VSM and BM25, MMRQE provides the best performance for MAP and PRES (except for MAP, where Rocchio BM25 provides better performance than MMRQE BM25);
- (ii) adding more than 50 terms hurts the performance of MMRQE and Rocchio;
- (iii) exploiting external sources provides poor performance (IPC code definition and SynSets).

MAP FOR QE METHODS ON CLEF-IP 2010



(a) Query Title.

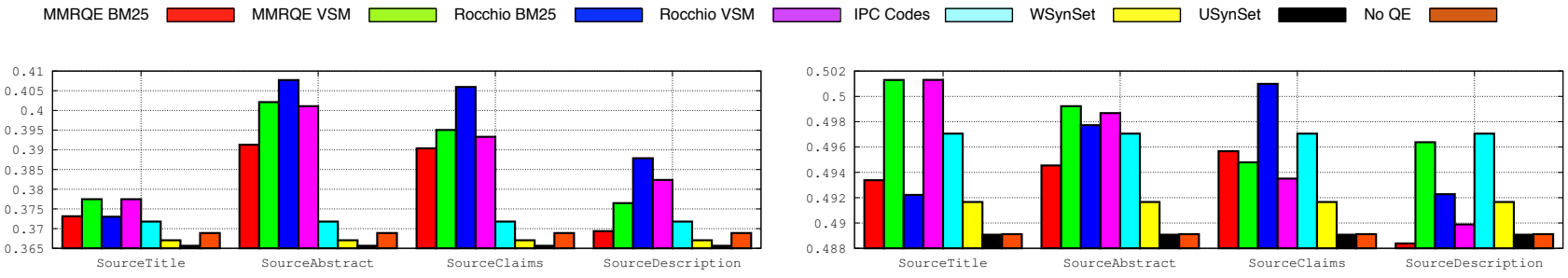
(b) Query Abstract.

(c) Query Claims.

(d) Query Description.

*for MMRQE $\lambda = 0.5$

PRES FOR QE METHODS ON CLEF-IP 2010



(e) Query Title.

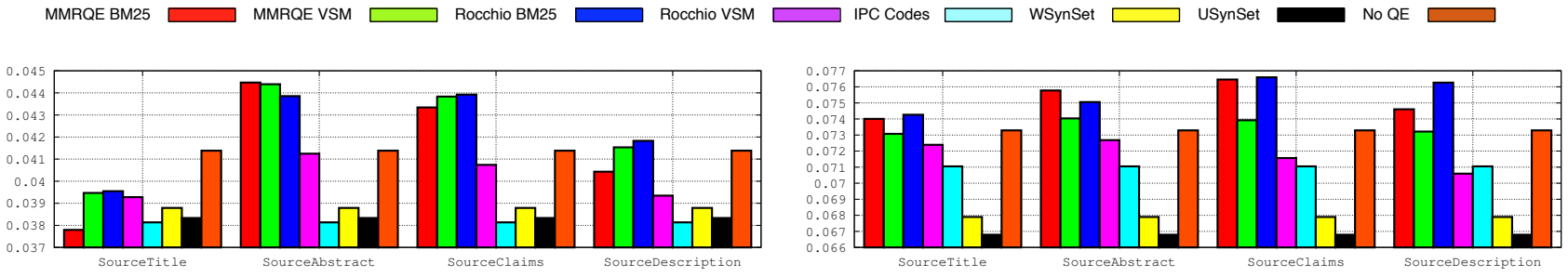
(f) Query Abstract.

(g) Query Claims.

(h) Query Description.

*for MMRQE $\lambda = 0.5$

MAP FOR QE METHODS ON CLEF 2011



(i) Query Title.

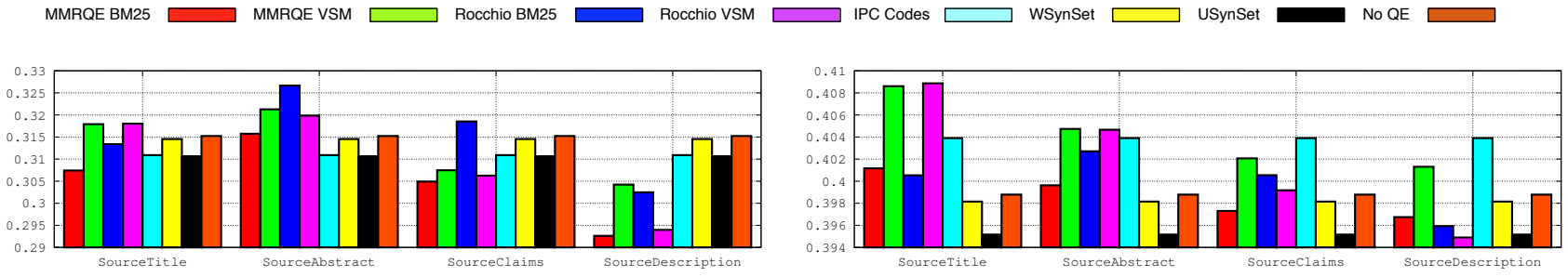
(j) Query Abstract.

(k) Query Claims.

(l) Query Description.

*for MMRQE $\lambda = 0.5$

PRES FOR QE METHODS ON CLEF 2011



(m) Query Title.

(n) Query Abstract.

(o) Query Claims.

(p) Query Description.

*for MMRQE $\lambda = 0.5$

DISCUSSION ABOUT QE

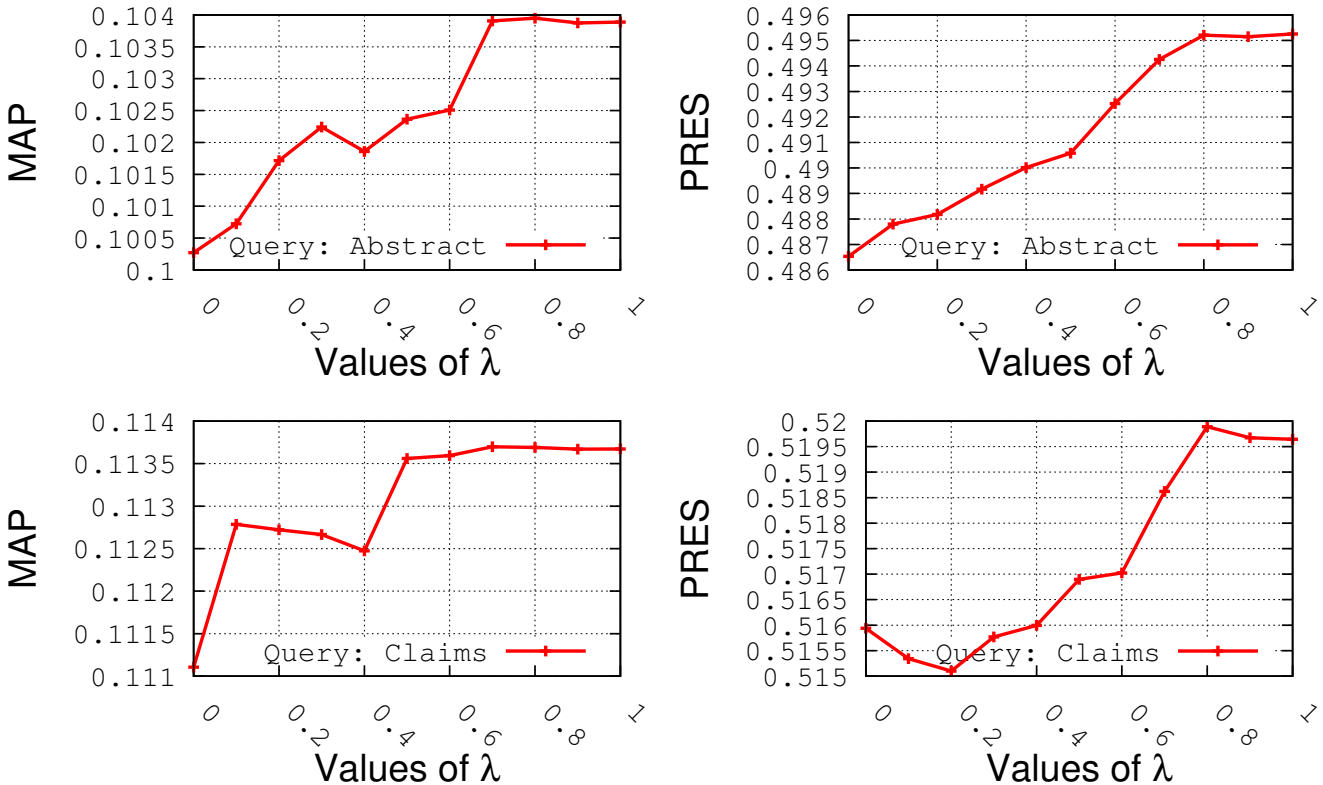
- **Querying** with the Description section gives the best performance;
- **Best source of expansion** are the Claims
- When we query with the Claims other source of expansion are better;
- Using IPC and synomym Synsets gives poor performance;
- MMRQE outperforms Rocchio in most of the cases.

QUERY REDUCTION EXPERIMENTS

Experiment options:

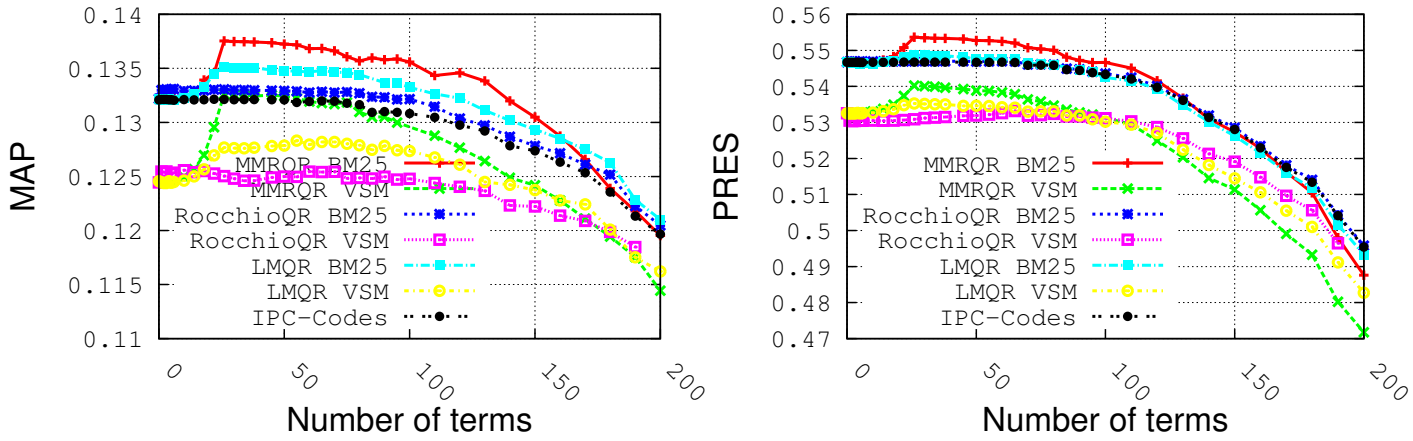
- **Query type:** {Title, Abstract, Claims, Description}
- **Relevance model:** {BM25, Vector-space Model (VSM)}
- **Term selection method:** {RocchioQR, MMRQR, *etc*...}

IMPACT OF THE DIVERSITY PARAMETER λ ON THE PERFORMANCE OF MMRQR (CLEF-IP 2010 DATASET)



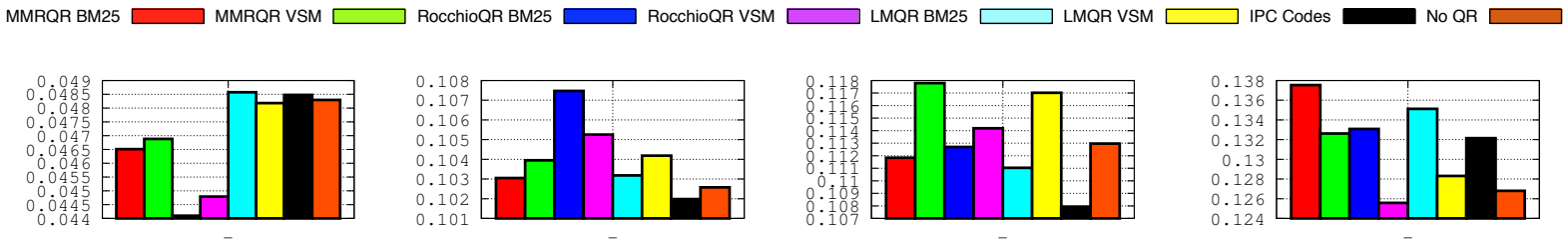
Best QR performance results are also obtained when using few documents in the PRF set (top 5).

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

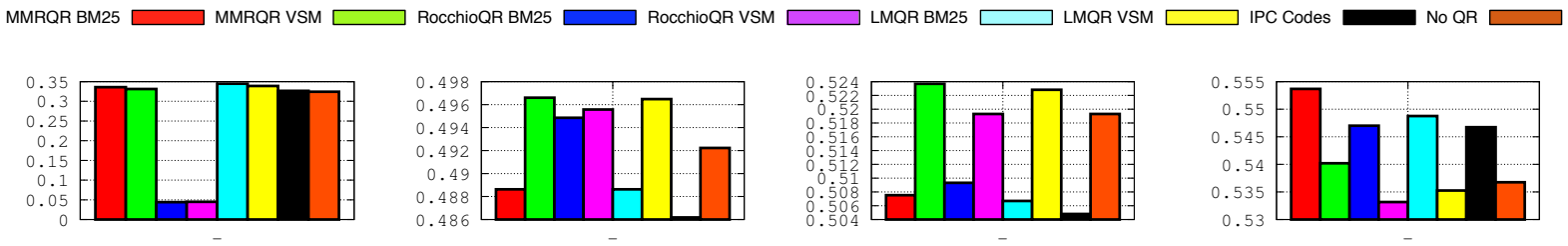


- (i) MMRQR provides the best performance for both MAP and PRES;
- (ii) for almost all methods, the best performance is obtained when removing about 30 terms from the original queries.

MAP AND PRES FOR QR METHODS ON CLEF 2010



(q) Query Title. (r) Query Abs. (s) Query Claims. (t) Query Descr.



(u) Query Title. (v) Query Abs. (w) Query Clai. (x) Query Descr.

*for MMRQR $\lambda = 0.8$

DISCUSSION ABOUT QR

- When dealing with very long query (i.e. description), BM25 based QR methods perform better than VSM based QR methods;
- In general, MMRQR provides better performance than the other methods.
- In MMRQR, $\lambda = 0.8$ indicates that few diversification in term selection can provide some improvement.

Contributions are the following:

1. Novel contributions for query expansion and reduction that leverage:
   - **patent structure**;
   - **a term diversification technique**.
2. A thorough comparative analysis of existing and novel methods for **query expansion and reduction** in **patent prior-art search with partial applications** on standardized datasets of CLEF-IP.

CONCLUSIONS

- We analyzed general and specific **QE** and **QR** methods for patent prior art search for partial (incomplete) patent applications (CLEF-IP 2010, 2011);
- The **claims should be written at early stages of the patent application drafting** (the best section that works with QE/QR (to query with and to use as a source of query expansion/reduction terms)
- The novel **MMR QE/QR** methods improves results in many cases.

Future work

Look at more patent-specific methods of and how they can be integrated with methods like MMRQE.

Thank you!

Shariq Bashir and Andreas Rauber. Improving retrievability of patents in prior-art search. In *ECIR*, 2010. ISBN 3-642-12274-4, 978-3-642-12274-3. doi: 10.1007/978-3-642-12275-0_40. URL http://dx.doi.org/10.1007/978-3-642-12275-0_40.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, 1998. ISBN 1-58113-015-5. doi: 10.1145/290941.291025. URL http://doi.acm.org/10.1145/290941.291025.

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

Debasis Ganguly, Johannes Leveling, Walid Magdy, and Gareth J.F. Jones. Patent query reduction using pseudo relevance feedback. In *CIKM*, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063863. URL http://doi.acm.org/10.1145/2063576.2063863.

Giridhar Kumaran and Vitor R. Carvalho. Reducing long

queries using query quality predictors. In *SIGIR*, SIGIR '09, pages 564–571, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572038. URL http://doi.acm.org/10.1145/1571941.1572038.

Patrice Lopez and Laurent 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. ISBN 3-642-15753-X, 978-3-642-15753-0. URL http://dl.acm.org/citation.cfm?id=1887364.1887426.

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

Walid Magdy and Gareth J.F. Jones. A study on query expansion methods for patent retrieval. In *PaIR*, 2011. ISBN

978-1-4503-0955-4. doi: 10.1145/2064975.2064982. URL http://doi.acm.org/10.1145/2064975.2064982.

Parvaz Mahdabi, Shima Gerani, Jimmy Xiangji Huang, and Fabio Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In *SIGIR*, 2013. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484056. URL http://doi.acm.org/10.1145/2484028.2484056.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-2. In *TREC*, pages 21–34, 1993.

Giovanna Roda, John Tait, Florina Piroi, and Veronika Zenz. Clef-ip 2009: Retrieval experiments in the intellectual property domain. In Carol Peters, GiorgioMaria Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Penas, and Giovanna 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. ISBN 978-3-642-15753-0. doi: 10.1007/978-3-642-15754-7_47. URL http://dx.doi.org/10.1007/978-3-642-15754-7_47.

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

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL http://doi.acm.org/10.1145/361219.361220.

Manisha Verma and Vasudeva Varma. Patent search using ipc classification vectors. In *PaIR*, 2011. ISBN 978-1-4503-0955-4. doi: 10.1145/2064975.2064980. URL http://doi.acm.org/10.1145/2064975.2064980.