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

Mohamed Reda Bouadjenek, **Gabriela Ferraro**, Scott Sanner

June 2013

Outline

- Motivation
- Query reformulation for patents
- General Query reformulation methods
-
- Experiments
- Results and discussion
- Conclusion and future work

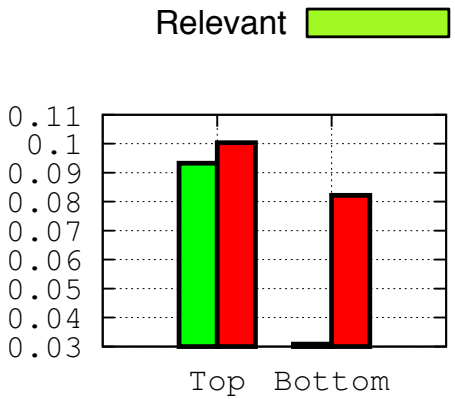
Patent Prior Art Search

- Patents are legal documents that ....
  - Patent sections
  - Title, Abstract, Description and Claims
- Finding previously granted patents relevant for a patent application
- Work has been devoted to perform patent prior art search with complete patent applications, CLEF-IP 2010 and 2011
- Writing a full patent application is time-consuming and costly.
- We proposed to do patent prior art search with partial (incomplete) patent applications

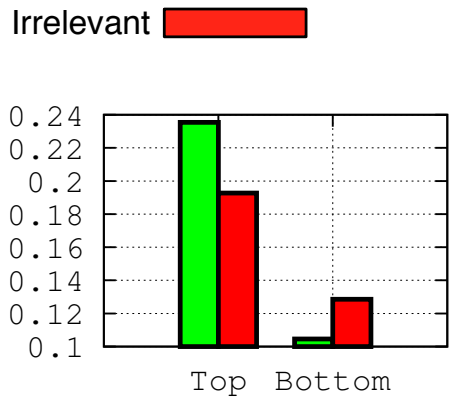
Why patent prior art search is different to standard Information Retrieval (IR)

- Queries are full patent applications (hundreds of words organized into several sections)
- patent prior art search is a **recall-oriented** task 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.

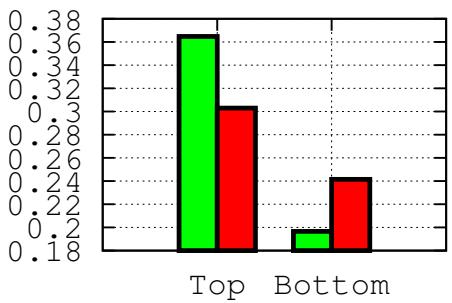
Average Jaccard similarity of (ir)relevant documents with the result sets for different queries



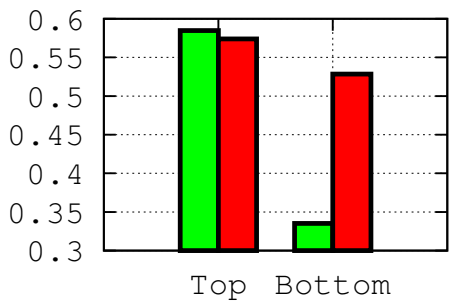
(a) Title query.



(b) Abstract query.



(c) Claims query.



(d) Description query.

Term overlap between (ir)relevant documents with the results sets for different queries

top 10 irrelevant documents ranked by BM25 [**?**]
Top 100 and bottom 100 queries (100 queries that perform the best, and 100 queries that perform the worst) CLEP-IP 2010
There are 3 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;

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

We investigate query formulation methods

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
Intensive study about query reformulation for patent prior art search with partial patent applications
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.

Query Reformulation for patents

uery type: title, abstract, claims or the description section What part of a partial application an inventor should write to obtain the best search results?

ace model: For initial retrieval of documents in the *pseudo-relevant* feedback set (PRF) and subsequent re-retrieval, a probabilistic approach represented by the popular BM25 [**?**] and vector space model (VSM) approach, TF-IDF [**?**]. Which relevance model works best for query reformulation for patent prior art search?

on source: title, abstract, claims, and description sections as different term sources Are the title words of particularly high value as expansion terms? * Note that this only applies to query expansion methods.

n method: Rocchio [**?**] and new term selection methods that we propose in the next sections Which term selection method works best, and with which configuration, i.e. query type, retrieval model, and term source for query expansion methods?

Query Expansion Frameworks

large term mismatch between queries an relevant documents. This term mismatch may be alleviated by QE methods. 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.

Maximal Marginal Relevance Query Expansion

Result set diversification algorithm *Maximal Marginal Relevance* (MMR) [**?**], usually used for **diverse document selection** (multi-document summarization)
diverse term selection to address the deficiency of Rocchio

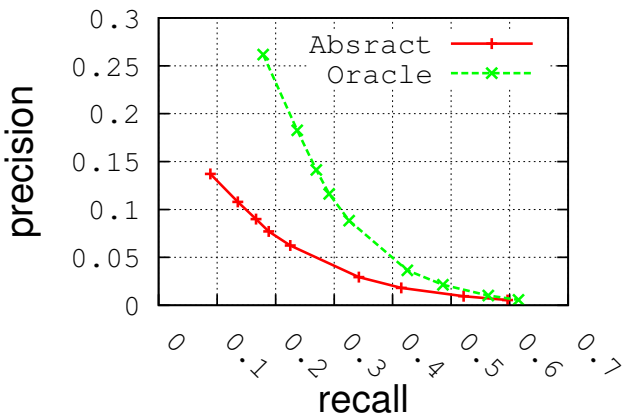
Notation used in MMRQE

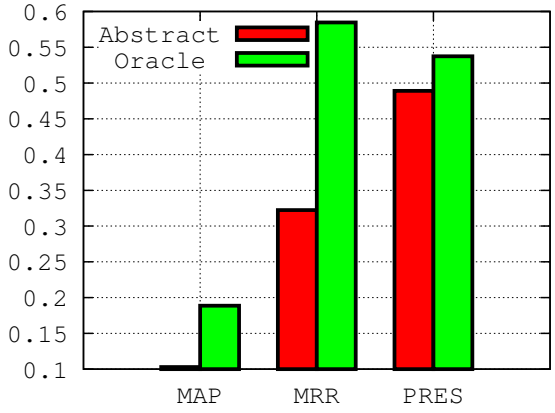
(1)

Query Reduction Frameworks

While the title is usually composed by an average of six terms, the other sections are longer, ranging from ten to thousands of terms. Therefore, we investigate the impact of query reduction methods when querying with long sections such as abstract, claims or description.

Sample of terms removed from the abstract section





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