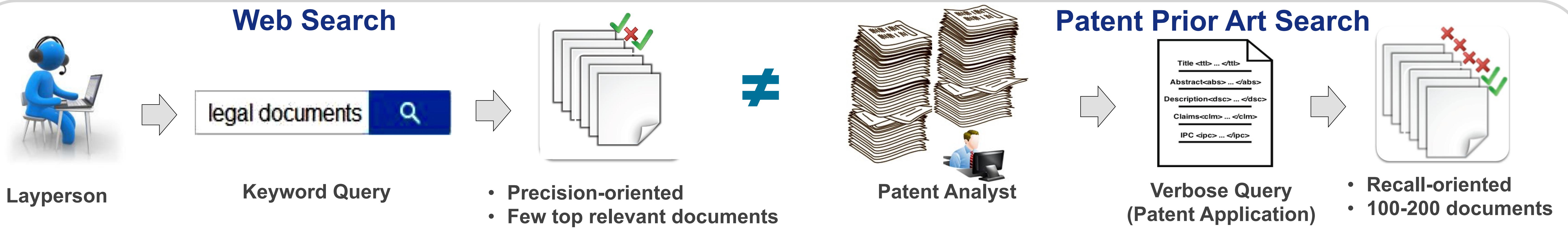
On Term Selection Techniques for Patent Prior Art Search

Mona Golestan Far[1,2], Scott Sanner[2,1], Mohamed Reda Bouadjenek[3], Gabriela Ferraro[2,1], David Hawking[1,4]

ANU[1], NICTA[2], INRIA & LIRMM[3], Microsoft (Bing)[4]

Introduction



Web Search

Layperson

Keyword Query

- Precision-oriented
- Few top relevant documents

≠

Patent Prior Art Search

Patent Analyst

Verbose Query
(Patent Application)

- Recall-oriented
- 100-200 documents

Patent Prior Art Search: Finding previously granted patents or any published work relevant to new patent application.

Problem Definition: **Generic IR techniques (e.g., different query reformulation) are ineffective for patent prior art search!**

Previous Work: Mainly focused on different query expansion techniques to cope with significant term mismatch [1].

Data Collection: English subset of CLEF-IP 2010 [2].

PATATRAS: Highly engineered system and top CLEF-IP 2010 competitor [3].

Oracular Term Selection

1. Relevance Feedback Score

Relevance feedback (RF) score for each term:

$$\mathbf{RF(t,Q) = Rel(t,Q) - Irr(t,Q)} \quad \mathbf{(1)}$$

$$\mathbf{t \in \{top - 100\ retrieved\ documents\}}$$

where

$$\mathbf{Rel(t) \rightarrow Avg.\ Term\ Frequency\ in\ Rel.\ Docs.}$$

$$\mathbf{Irr(t) \rightarrow Avg.\ Term\ Frequency\ in\ Irr.\ Docs.}$$

2. Oracular Query Formulation

Formulate two oracular queries:

① Oracular Query = $\{\mathbf{t} \in \mathbf{top-100} | \mathbf{RF(t,Q)} > \tau\}$

② Oracular Patent Query = $\{\mathbf{t} \in \mathbf{Q} | \mathbf{RF(t,Q)} > \tau\}$

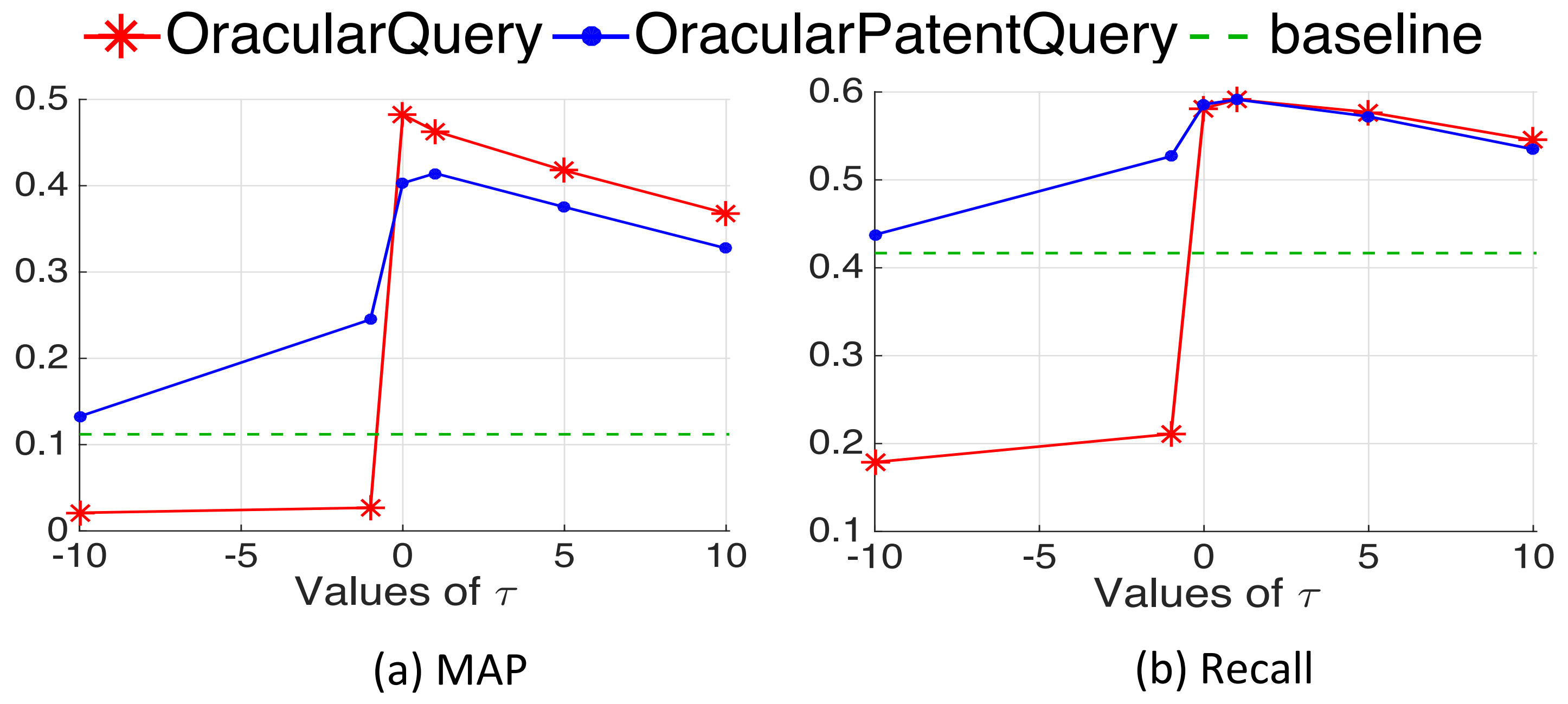
Take Home Message

- Sufficiency of terms in baseline query
- Over-sensitivity of IR models to inclusion of negative terms ($\tau < \mathbf{0}$)
- Need for precise methods to eliminate poor query terms (query reduction)

3. Baseline vs. Oracular Query

Table.1: Performance for the Baseline Query, two variants of the Oracular Query, and PATATRAS.

		Baseline	PATATRAS	Oracular Query	Oracular Patent Query
LM	MAP	0.112	0.226	**0.482**	**0.414**
	Recall	0.416	0.467	0.582	0.591
BM25	MAP	0.123	0.226	**0.492**	**0.424**
	Recall	0.431	0.467	0.584	0.598

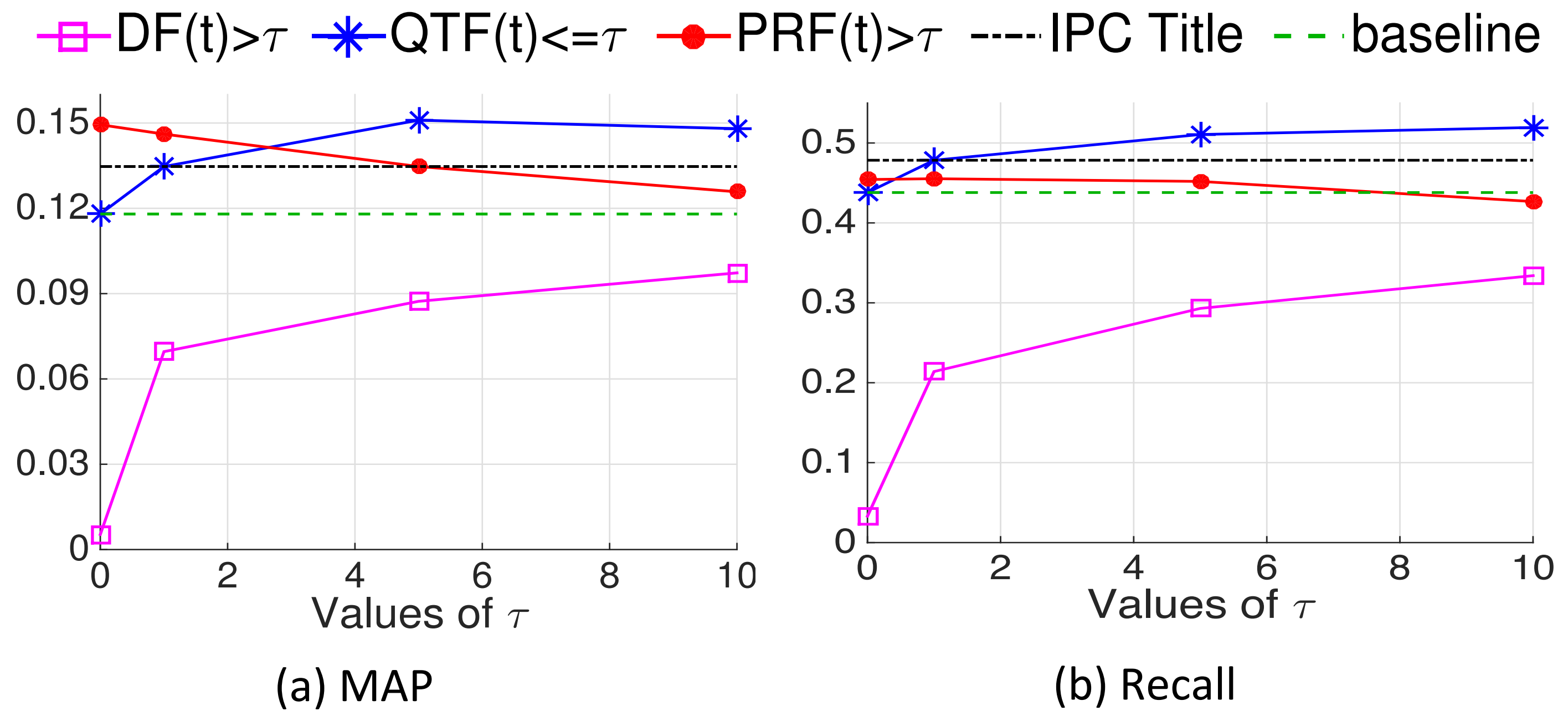


(a) MAP (b) Recall

Fig.1: Comparing the performance of two different oracular queries.

Query Reduction (QR): Approximating Oracular Query

1. Automated Reduction



(a) MAP (b) Recall

Fig.2: System performance vs. the threshold τ for four QR approaches.

QR Approaches:

1. Pruning document frequent (DF) terms ($DF(t) > \tau$).
2. Pruning query infrequent terms ($QTF(t) <= \tau$).
3. Pseudo relevance feedback term selection ($PRF(t) > \tau$).
4. Pruning IPC title general terms.

Take Home Message

- Automated QR methods **fail** to approximate oracular query.
- They cannot **discriminate** between positive and negative terms.

2. Semi-automated Interactive Reduction

Table.2: Performance of an Oracular Patent Query derived from only the top-k ranked relevant documents identified in the search results. We assume that the remaining documents in the top-100 are irrelevant.

	Baseline	PATATRAS	Oracular Patent Query (k=1)	Oracular Patent Query (k=3)
MAP	0.112	0.226	**0.289**	**0.369**
Avg. Recall	0.416	0.467	0.484	0.547

- MAP **doubles** over the baseline (0.112 → 0.289)
- **Outperforms** PATATRAS (0.226 → 0.289)

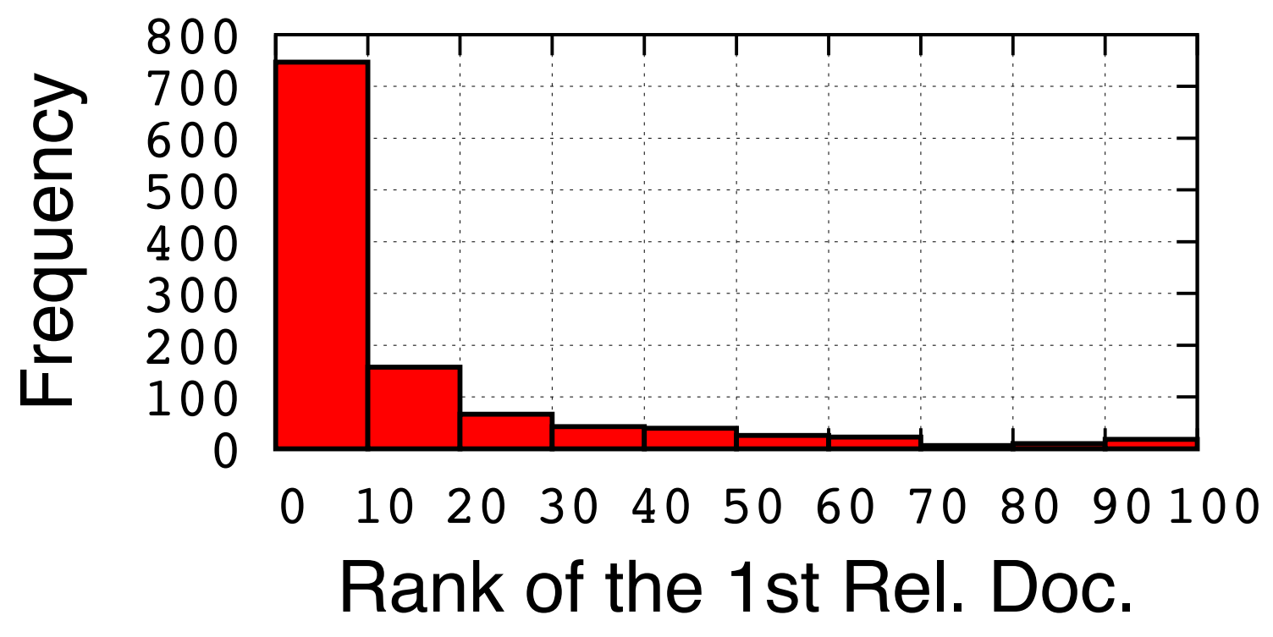


Fig.3: The distribution of the first relevant document rank over test queries.

- Baseline returns first rel. patent
  - ✓ 80% of time in top 10 results,
  - ✓ 90% of time in top 20.
- Minimal user effort

Take Home Message

- **Interactive methods** offer a promising avenue for simple but effective term selection in prior art search.

Acknowledgement

I would like to thank SIGIR for offering me a travel grant to attend the 2015 conference in Chile.

References

[1] W. Magdy and G. J. Jones. A study on query expansion methods for patent retrieval. In Proceedings of the 4th Workshop on Patent Information Retrieval, 2011.

[2] F. Piroi. CLEF-IP 2010: Prior art candidates search evaluation summary. Technical report, IRF TR, Vienna, 2010.

[3] P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. In CLEF 2010 LABs and Workshops, Notebook Papers, 2010.

Australian National University

NICTA

Inria INVENTEURS DU MONDE NUMÉRIQUE

Microsoft

You can find the paper, poster and the supplementary material at the authors' websites. Scanning the QR code on the right leads to the author's website.

mona.golestanfar@anu.edu.au, ssanner@gmail.com, reda.bouadjenek@inria.fr, gabriela.ferraro@nicta.com.au, david.hawking@acm.org