On Term Selection Techniques for Patent Prior Art Search

Mona Golestan Far

Accepted paper at **SIGIR 2015**, August 09-13

Mona Golestan Far (ANU & NICTA)
Scott Sanner (NICTA & ANU)
Reda Bouadjenek (INRIA & LIRMM)
Gabriela Ferraro (NICTA & ANU)
David Hawking (Microsoft (Bing) & ANU)

Outlines

NICTA

- Introduction
- Challenges
- Previous Work
- Baseline IR Framework
- Formulating Oracular Queries Based on Relevance Feedback
- Approximate Oracular Queries by Query Reduction Techniques:
  - Automated
  - Semi-automated (Interactive)

Introduction

NICTA

Patents

Legal Documents to Protect an Invention.

Patent Prior Art Search

Finding all (Patent) Documents, which

- May Invalidate the Novelty of a Patent Application, or
- Have Common Parts with Patent Application and Should Be Cited.

Users:

Patent Analysts

Web Search

User

- Layperson

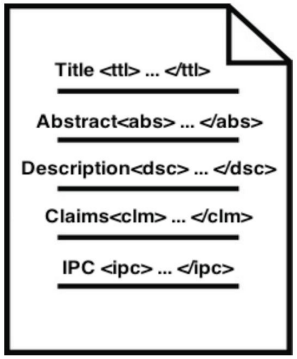
Query

- Keywords
- 2-3 words
- Short

Goal of search

- Precision-oriented
- Few top relevant documents that satisfy query intent

Prior Art Search



User

- Patent analyst

Query

- Patent document
- 1000 of words
- Long

Goal of search

- Recall-oriented
- Top 100-200 documents are examined.

Research Question

NICTA

Why Do Standard IR Techniques fail for Patent Prior Art Search?

It Is **Too Difficult** to Get Improved Over the **Baseline**!

Previous Work

NICTA

- Mentioned **Term Mismatch** as the main cause of Low Effectiveness

  [Roda et al., 2010][Lupu et al., 2011][Magdy 2012][Mahdabi, 2013]

- Query Reformulation Techniques
  - Query Expansion
  - Query Reduction
- Reported Little Improvement

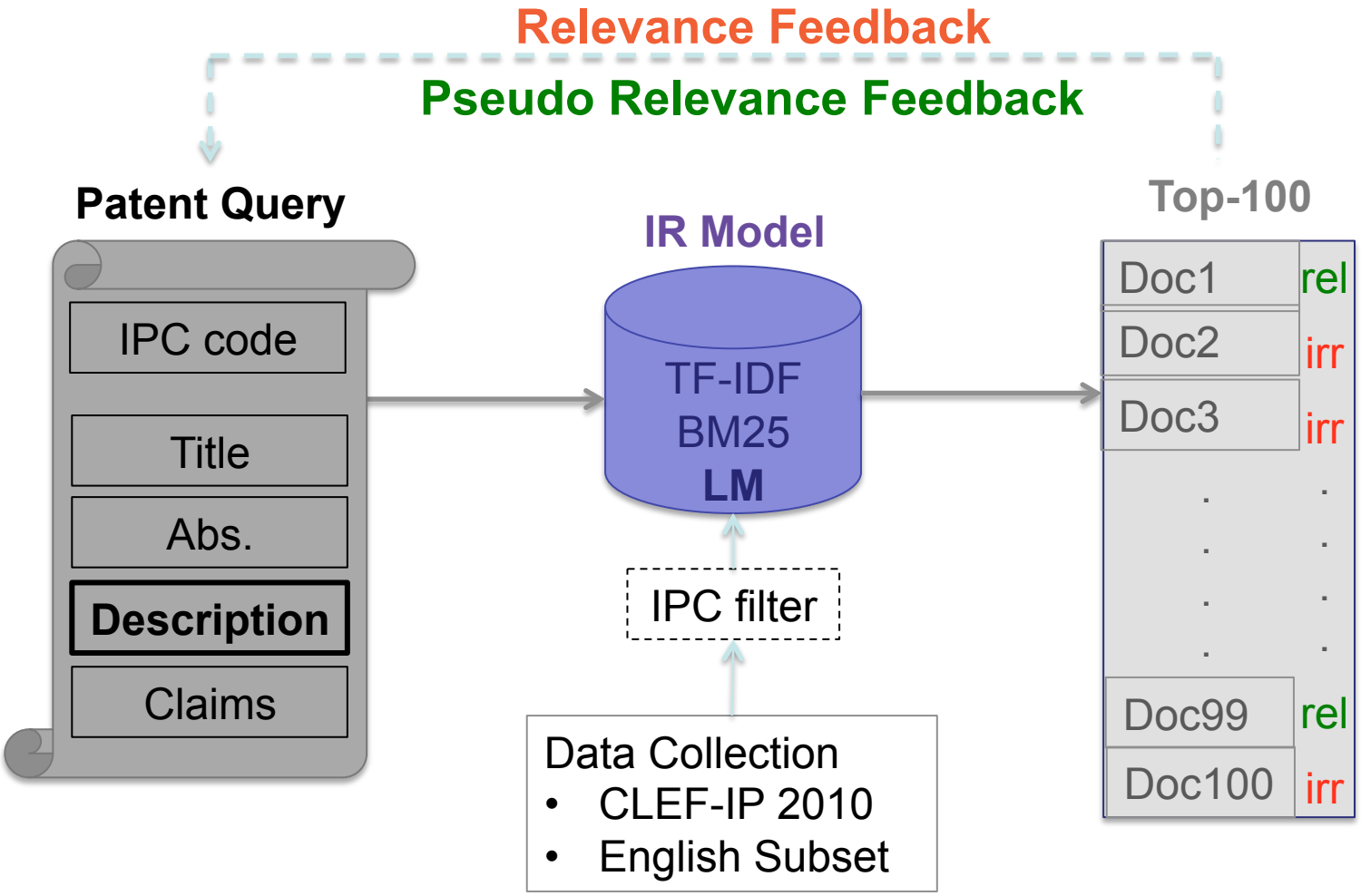
Previous Work

NICTA

- PATATRAS [Lopez et al., 2010]
  - Top CLEF-IP 2010 Competetor
  - Highly Engineered
  - Used Multiple Retrieval Models
  - Used Patent Metadata
  - Used Citation Structure

MAP = 0.226 Recall=0.467

Baseline IR Framework



Oracular Relevance Feedback System

NICTA

- Extract Terms from Judged Relevant Documents to Understand:
  1. The Adequacy of the Baseline Patent Query
  2. An Upper-bound on Performance
  3. The Sufficiency of Terms in the Original Patent Query

Relevance Feedback Score

NICTA

- We Define Relevance Feedback (RF) Score for Each Term as Follows:

$$t \in \{\text{top-100}\}$$

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

where

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

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

Oracular Query Formulation

NICTA

- We Formulate Two Oracular Queries:

1. Oracular Query

$$\{t \in top - 100 | RF(t, Q) > \tau\}$$

2. Oracular Patent Query

$$\{t \in Q | RF(t, Q) > \tau\}$$

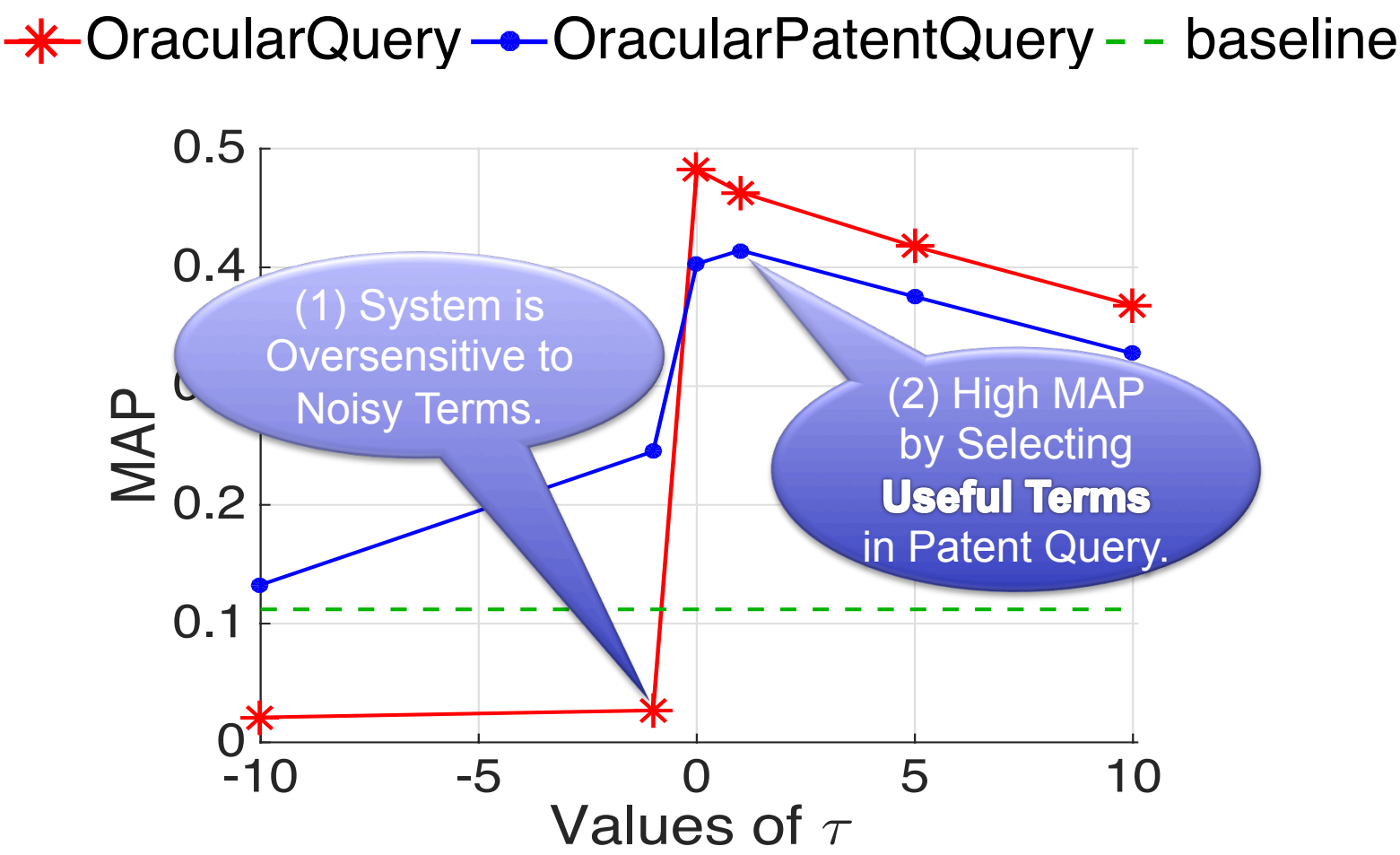
Baseline vs. Oracular Query

NICTA

		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

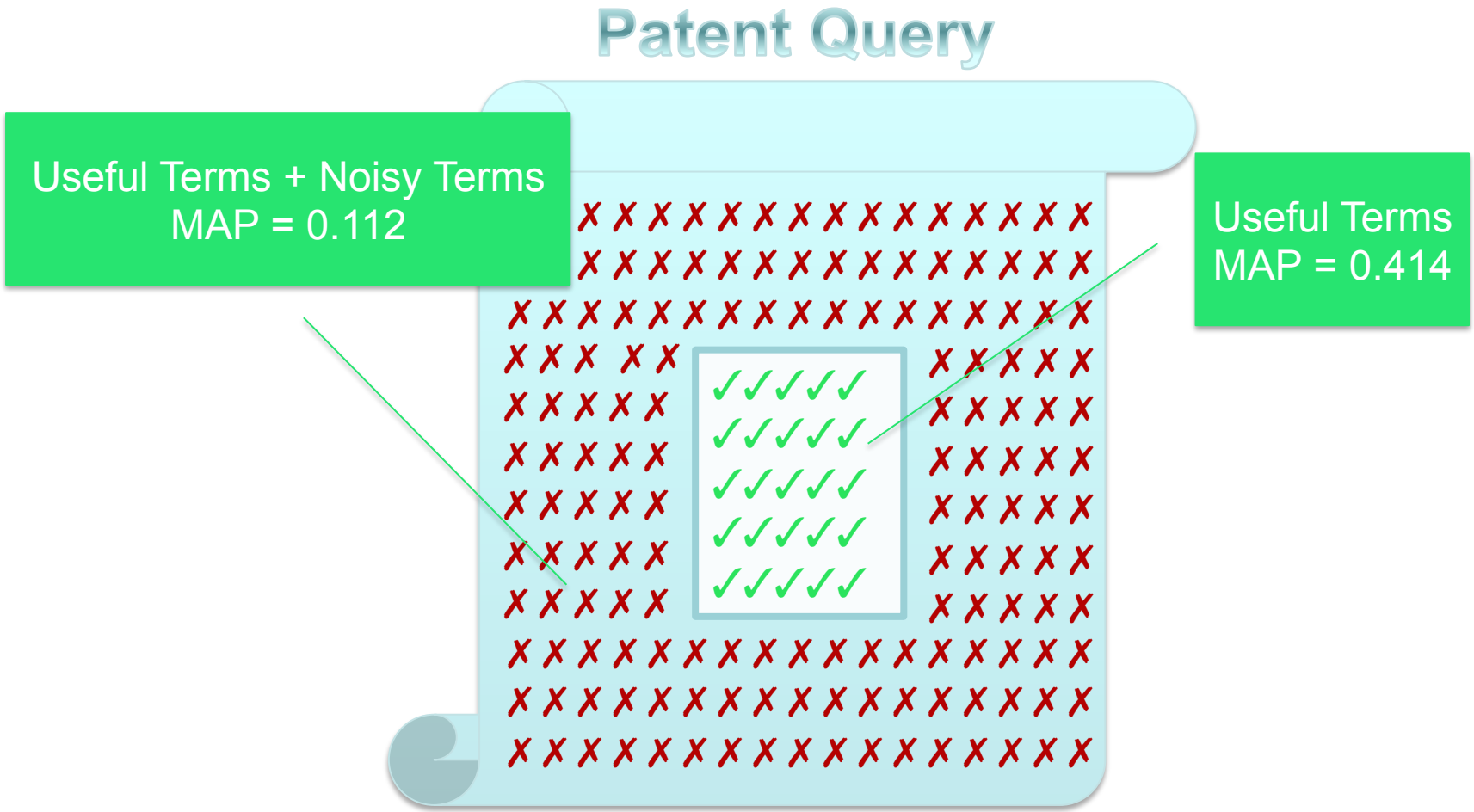
- Oracular Queries
  - **Outperform** the **Baseline**
  - Perform **Twice** as Well on MAP as **PATATRAS**
  - An Upper-bound Performance

Compare Oracular Queries (MAP)



Query Reduction (QR)

- We Need to Reduce Query to Get Improved.



Approximating Oracular Query

NICTA

- Gain Achieved for **Oracular Patent Query** Motivates Us to Approximate It Using:

1. Fully Automated Reduction Techniques

2. Semi-automated Interactive Reduction Techniques

Automated Reduction

NICTA

1. Pruning Document Frequent (DF) Terms
   - Remove Terms with High Avg. Term Frequency in Top 100 ($DF(t) > \tau$)
2. Pruning Query Infrequent Terms
   - ($QTF(t) <= \tau$)
3. Pruning General Terms in IPC Code Title.
   - **Titles of IPC Codes** Indicate the Intended Content of Patents Classified Under That Code.
   - We Assume General Terms in IPC Code Title as **Stop-words**.

Automated Reduction (cont.)

NICTA

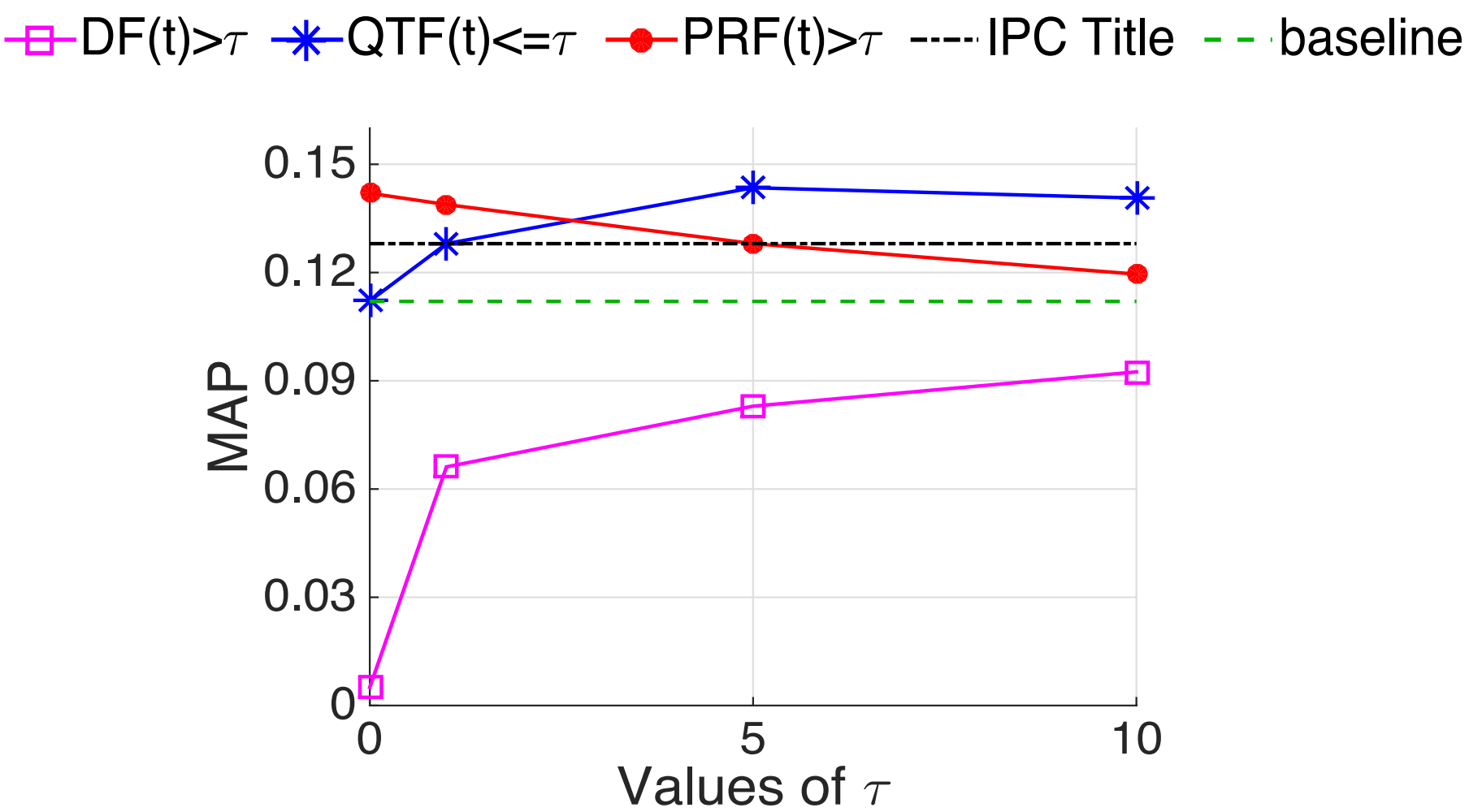
4. Pseudo Relevance Feedback (PRF) Term Selection

   – Calculate PRF Score the Same as RF Score.

   – Assume Top 5 Patents are Relevant and Remaining Patents are Irrelevant.

   – Formulate a Query by Selecting Terms Based on Their PRF Score ($PRF(t) > \tau$).

Compare QR Methods (MAP)



Anecdotal Example

NICTA

(PAC-1293) - Abstract: The invention relates to an emulsifier, a method for preparing said emulsifier, and to its use in various applications, primarily food and cosmetic applications.The invention also relates to the use of said emulsifier for the creation of an elastic, gelled foam. An emulsifier according to the invention is based on a starch which is enzymatically converted, using a specific type of enzyme, and modified in a specific esterification reaction.

DF Terms:star…20.1,oil:8.6,
disper:-8.7…roduct:-5.5,
slurri:-11,v…-2,food:-12,
agent:5,debr…par:-0.8,
hour:-5

QTF Terms:st…:-3.5,
enzym:29.5,e…rid:-5.5,
reaction:-2,…yl:0.06,
reagent:1.2,…0.3,wt:-4.6,
ether:2,enzym…5.5

PRF Terms:starch:14.6,encapsul:…,chees:-4,oil:8.6,
hydrophob:5.4,agent:5,casein:-2.2,degrad:17,deriv:12,
tablet:5.3,debranch:-11,imit:-1,viscos:7.8,oxid:6,
activ:6,osa:9.3,funnel:2.7,amylas:26,amylopectin:-7,
maiz:20.6

IPC Title Terms:cosmet:3.8,toilet:0.2,prepar:-0.8,
case:0.5,accessori:-0.01,store:-0.4,handl:0.07,
pasti:-0.2,amylos:-20,fibrou:-0.01,pulp:-1.3,
constitut:-0.06,paper:1.3,impregn:-0.1,emulsifi:6.7,
wet:-0.3,dispers:-9,saccharid:-12,produc:-0.6,agent:5

Proposed QR Methods Cannot **Discriminate** between **Useful** and **Noisy** Terms.

Semi-automated Interactive Reduction

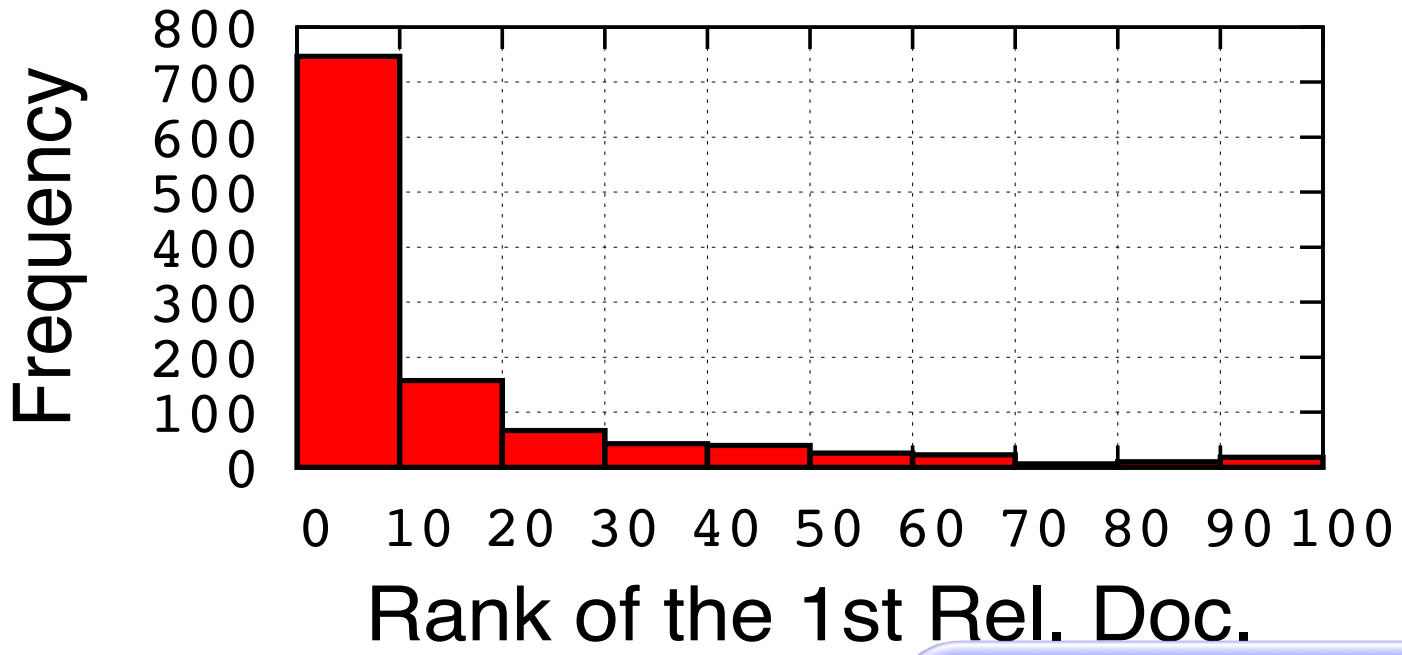
NICTA

- Identify Top *k* Rel. Patents in Initial Result Set.
- Calculate RF Score by Identified Rel. Patents.
- Select Query Terms Based on Their RF scores.

	1st Rel. Patent (k=1)	1st Three Rel. Patents (k=3)
MAP	**0.289**	**0.369**
Avg. Recall	0.484	0.547

1. MAP **Doubles** Over the **Baseline** (0.112 → 0.289)
2. **Outperforms PATATRAS** (0.226 → 0.289)

Minimum Effort



- Baseline Returns Top Rel. Paten
  - 80% of Time in Top 10 results, and
  - 90% of Time in Top 20.

Interactive Methods Offer a Promising Avenue for Simple but Effective Term Selection in Prior Art Search.

Questions

mona.golestnfar@anu.edu.au

References and Further Reading

[1] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In ECIR, 2010.

[2] M. R. Bouadjenek, S. Sanner, and G. Ferraro. A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications. In ICAIL, 2015.

[3] D. Ganguly, J. Leveling, W. Magdy, and G. J. Jones. Patent query reduction using pseudo relevance feedback. In CIKM, 2011.

[4] H. Itoh, H. Mano, and Y. Ogawa. Term distillation in patent retrieval. In ACL workshop on Patent corpus processing, 2003.

[5] H. Joho, L. A. Azzopardi, and W. Vanderbauwhede. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In IIiX, 2010.

[6] Y. Kim and W. B. Croft. Diversifying query suggestions based on query documents. In SIGIR, 2014.

[7] P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. In CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation, 2010.

[8] W. Magdy. Toward higher eectiveness for recall-oriented information retrieval: A patent retrieval case study. PhD thesis, Dublin City University, 2012.

Refrences

[9] 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.

[10] W. Magdy, P. Lopez, and G. J. Jones. Simple vs. sophisticated approaches for patent prior-art search. In Advances in Information Retrieval, pages 725-728. Springer, 2011.

[11] P. Mahdabi and F. Crestani. Patent query formulation by synthesizing multiple sources of relevance evidence. ACM Transactions on Information Systems, 2014.

[12] P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: query disambiguation using proximity information for patent retrieval. In SIGIR, 2013.

[13] K. T. Maxwell and W. B. Croft. Compact query term selection using topically related text. In SIGIR, 2013.

[14] F. Piroi. Clef-ip 2010: Prior art candidates search evaluation summary. Technical report, IRF TR, Vienna, 2010.

[15] M. Verma and V. Varma. Patent search using IPC classication vectors. In PaIR, 2011.

[16] X. Xue and W. B. Croft. Transforming patents into prior-art queries. In SIGIR, 2009.