Università della Svizzera italiana

Faculty of Informatics

Query Driven Mining of Citation Networks

for Patent Citation Retrieval and Recommendation

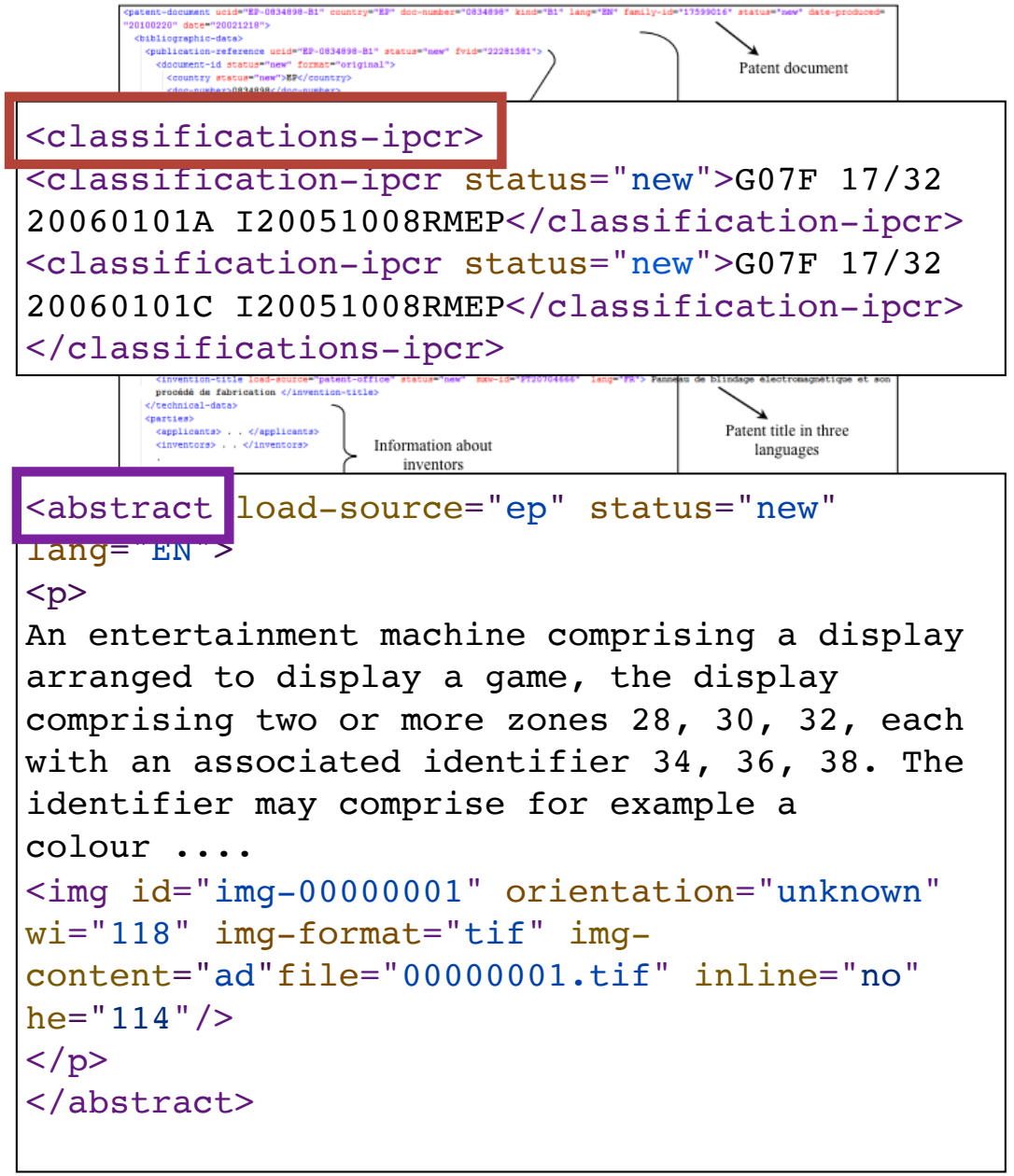
Parvaz Mahdabi †‡ **Fabio Crestani** †

University of Lugano, Switzerland †
Idiap Research Institute, Martigny, Switzerland ‡

CIKM 2014

Patent Document

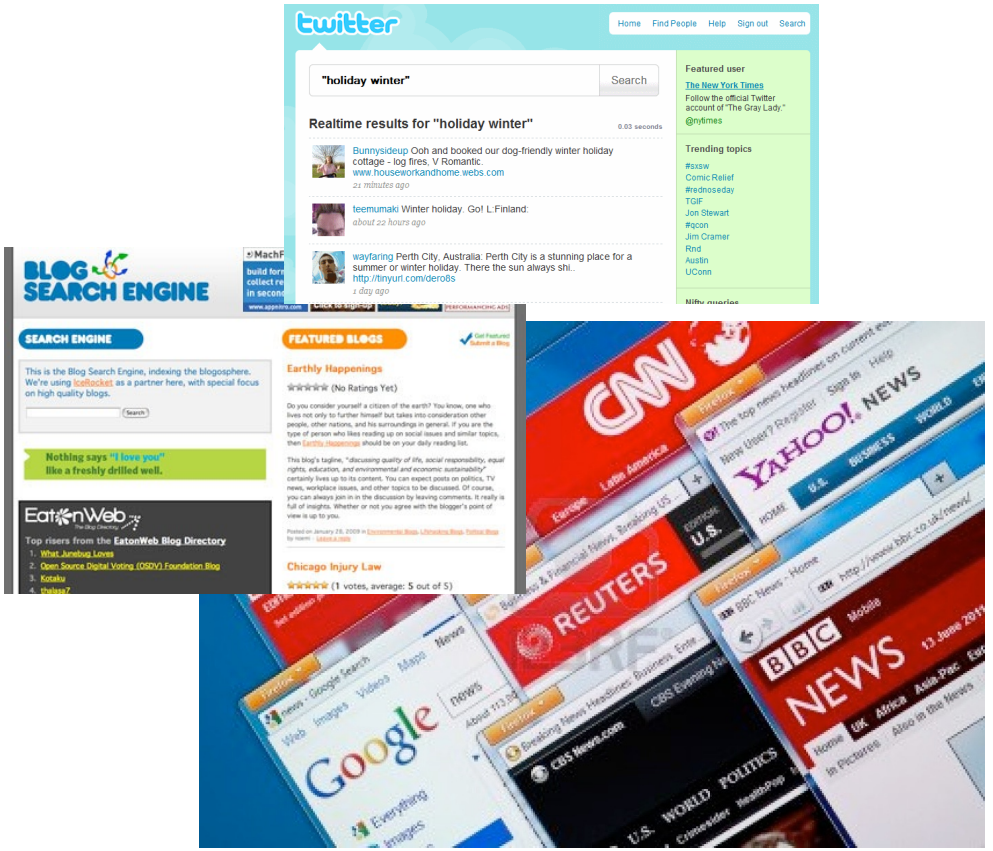
- Patent classifications
- Inventor information
- Title
- Abstract
- Description
- Claims



Hierarchical Structure of IPC classes

H	ELECTRICITY
H03	BASIC ELECTRONIC CIRCUITRY
H03B	GENERATION OF OSCILLATIONS, DIRECTLY OR BY FREQUENCY-CHANGING, BY CIRCUITS EMPLOYING ACTIVE ELEMENTS WHICH OPERATE IN A NON-SWITCHING MANNER; GENERATION OF NOISE BY SUCH CIRCUITS ...
H03B 5/04	· · Modifications of generator to compensate for variations in physical values, e.g. power supply, load, temperature

Patent Retrieval Versus Standard Information Retrieval



Web Search

User

- normal user

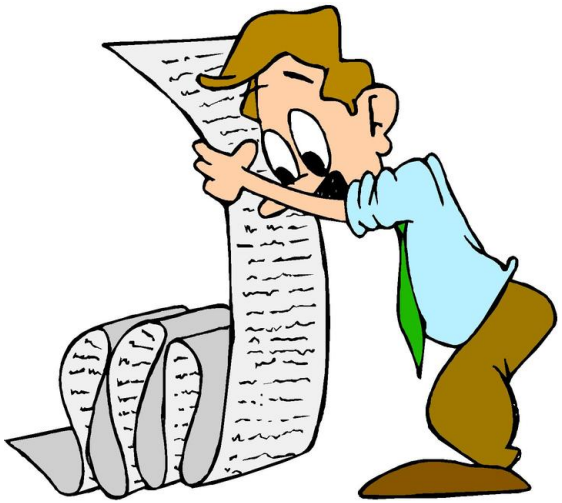
Query

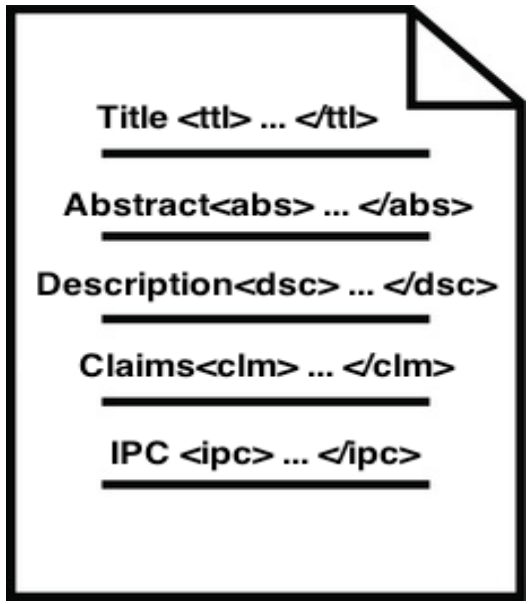
- Short (2, 3 words)

Goal of search

- Precision oriented
- The first page is examined

Prior-art Search





User

- Expert user (Patent Examiner)

Query

- Full application (hundreds of words)

Goal of search

- Recall oriented
- A long list is examined

Check Novelty

Challenges of Prior-art Search

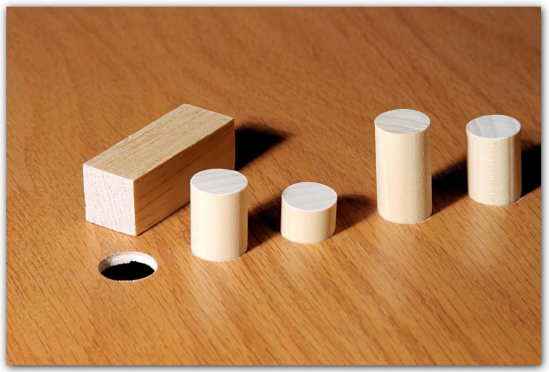
- A full patent application instead of a keyword query
- Term mismatch
  - Non-standardized acronyms: invented by authors
  - Homonyms: bus (1- motor vehicle, 2- within a computer system)
  - Synonyms: signal and wave
- Incorporating different relevance evidences that come together with a patent application

Research Questions

- How can we distinguish relevant docs from non-relevant docs
  - using the contextual similarity information diffused over the patent citation network
- Can this information be leveraged for formulating a better query?

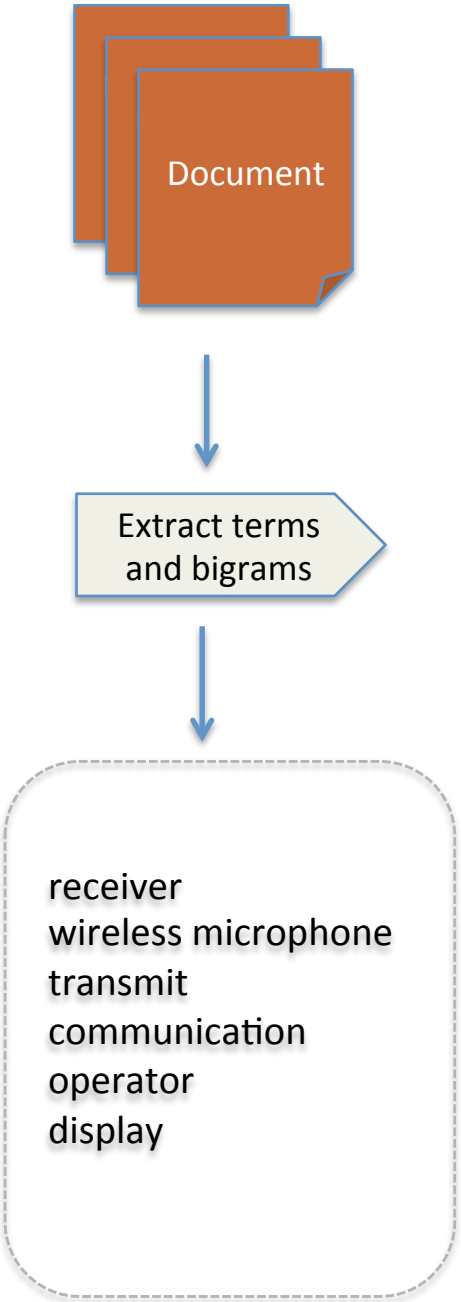
Hypothesis

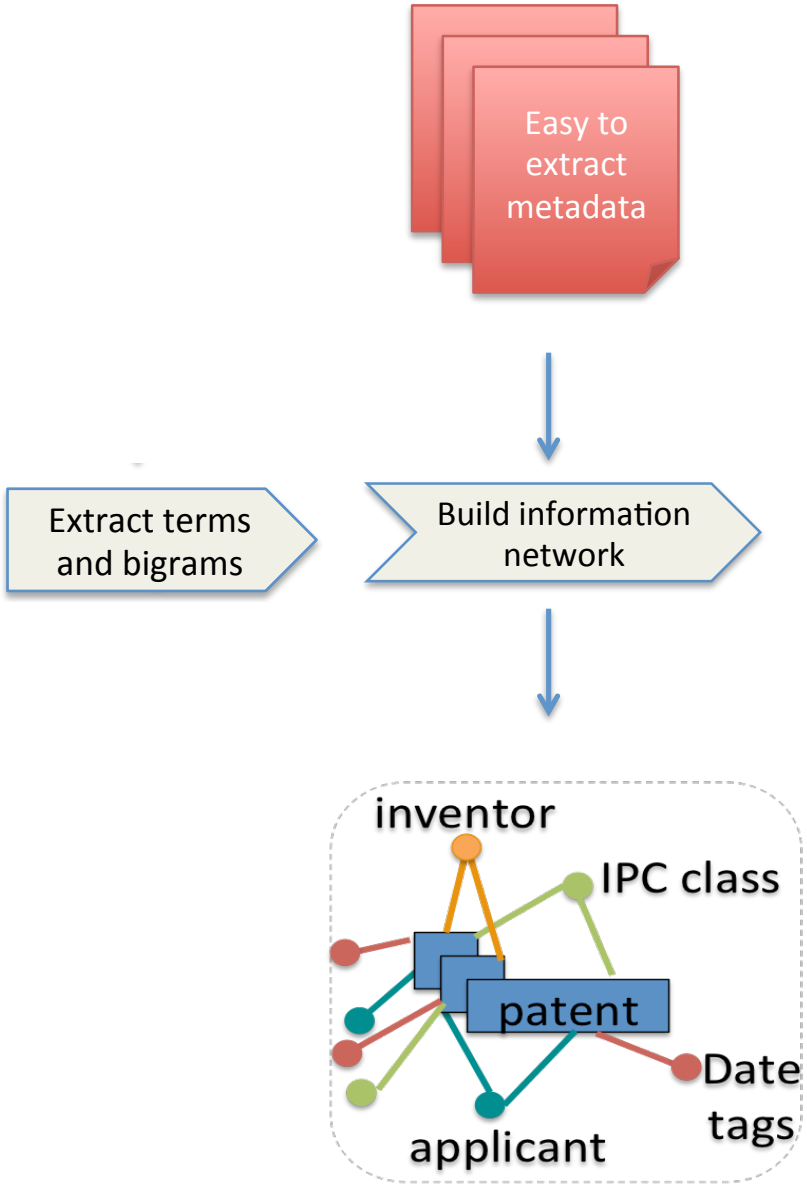
- Term distribution of influential documents in the citation graph might help to mitigate the term mismatch between the query and the relevant documents

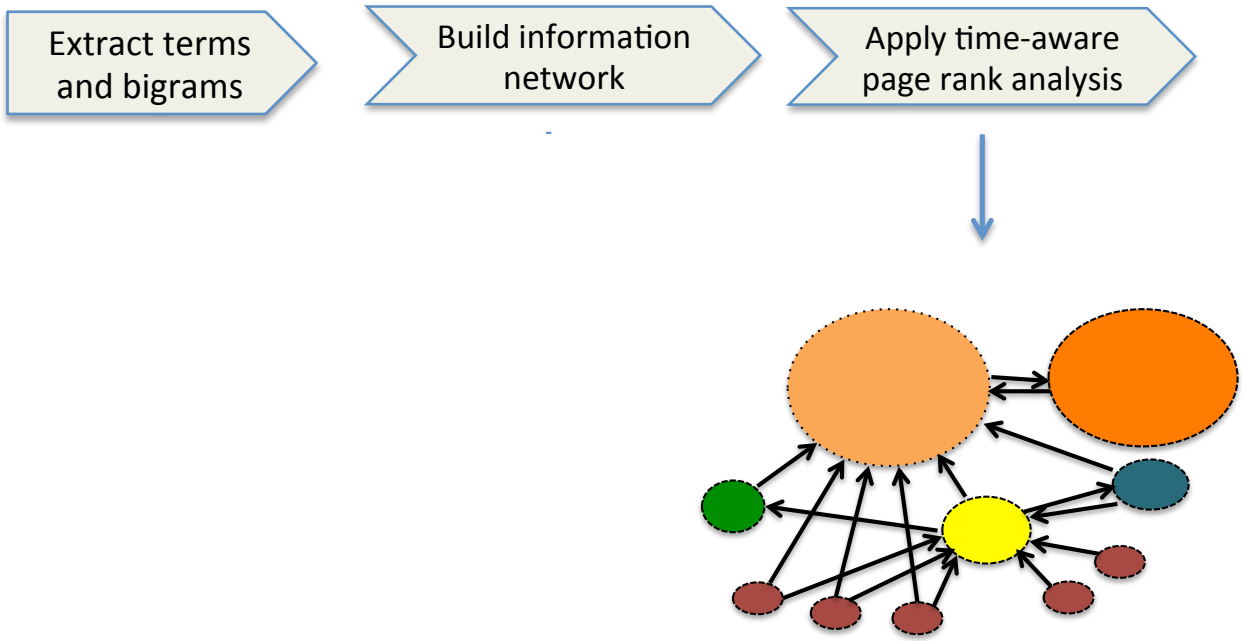


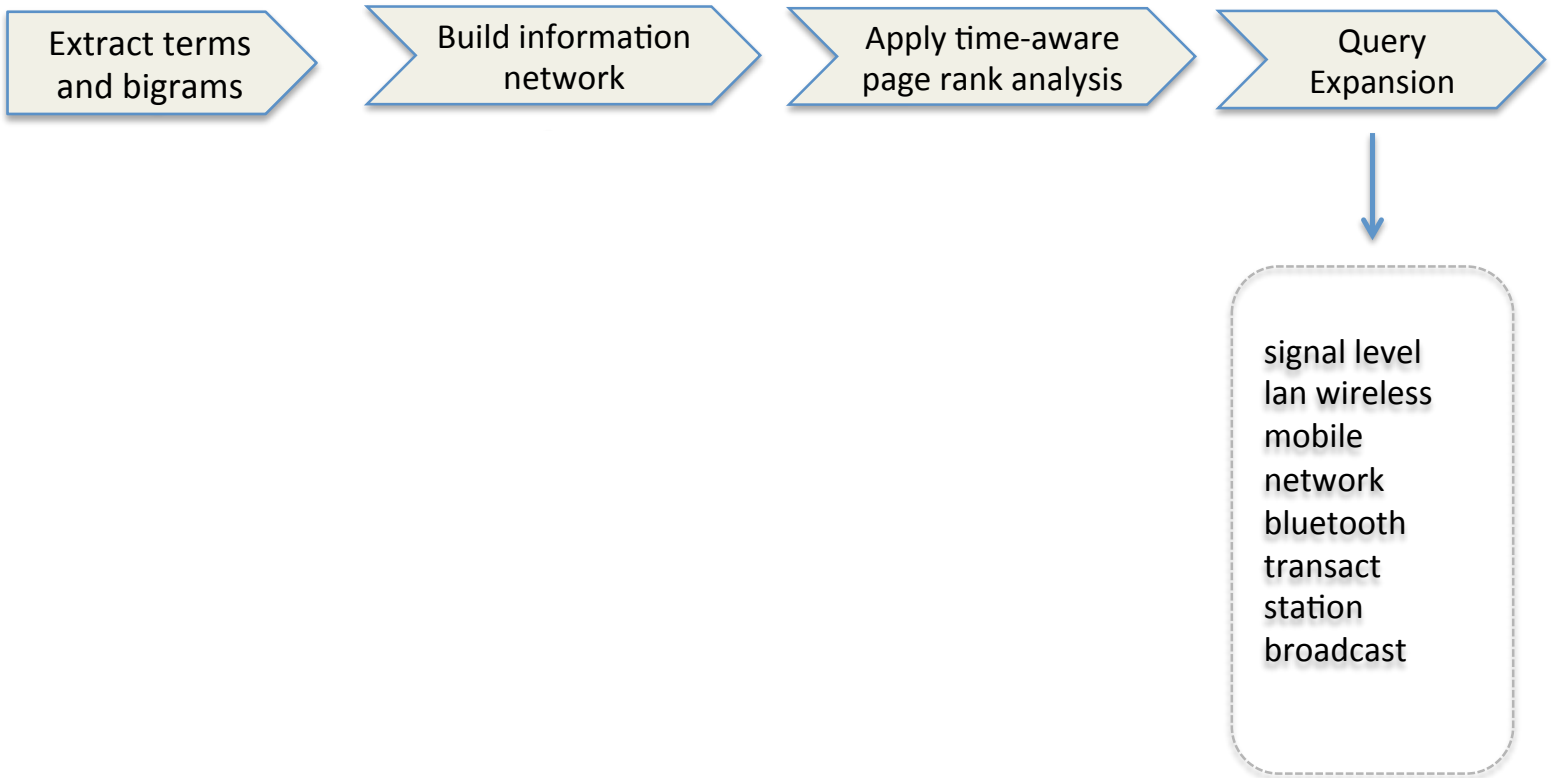
Related Work

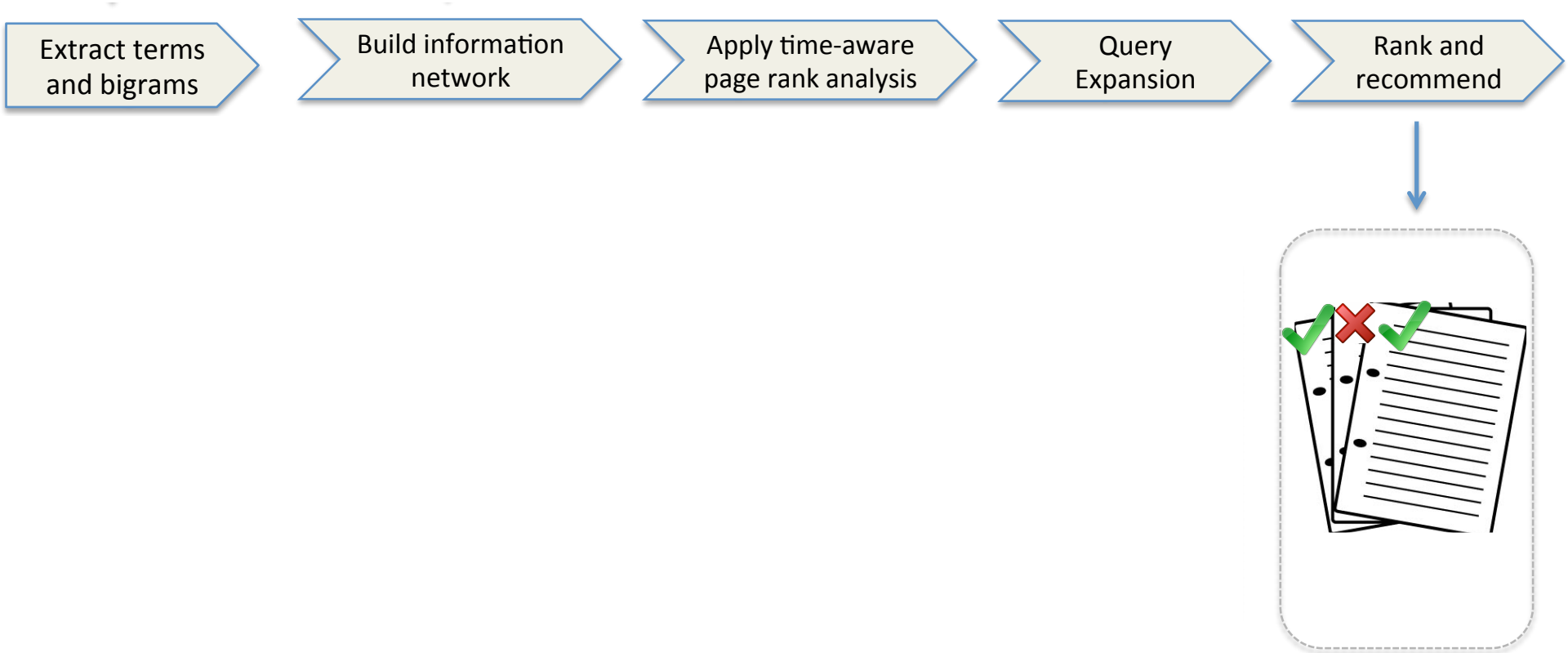
- Textual and categorical similarity
  - Lopez and Romary (CLEF 2010)
  - Magdy and Jones (CLEF 2010)
  - D'hondt et. al (CLEF 2011)
  - Bashir and Rauber (ECIR 2010)
  - Mahdabi et. al (SIGIR 2012, SIGIR 2013)
- Citation Network Analysis
  - Fujii (SIGIR 2007)
- Network analysis derived from interacting companies and inventors
  - Yang et. al (CIKM 2011)
  - Tang et. al (SIGKDD 2012)
  - Wu et. al (WSDM 2013)
- Temporal Citation Analysis
  - Wang et. al (CIKM 2014)
  - Zhang et. al (CIKM 2014)


Document
Extract terms and bigrams
receiver
wireless microphone
transmit
communication
operator
display


Easy to extract metadata
Extract terms and bigrams
Build information network
inventor
IPC class
patent
Date
tags
applicant

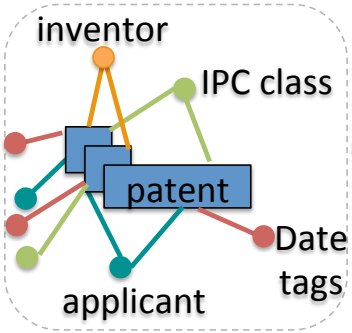

Extract terms and bigrams
Build information network
Apply time-aware page rank analysis

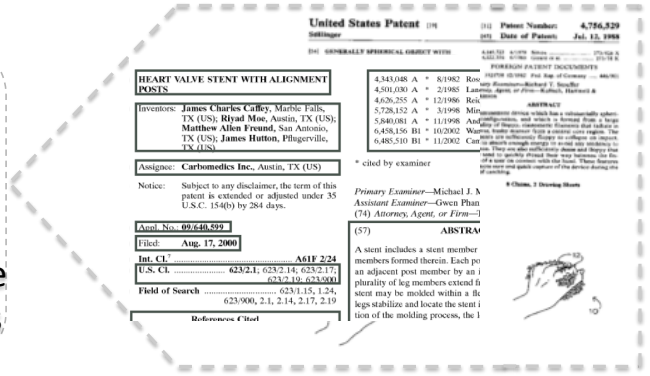

Extract terms and bigrams
Build information network
Apply time-aware page rank analysis
Query Expansion
signal level
lan wireless
mobile
network
bluetooth
transact
station
broadcast


Extract terms and bigrams
Build information network
Apply time-aware page rank analysis
Query Expansion
Rank and recommend

Building the Network

- Given a query patent, we retrieve an initial rank list using lexical similarity (*root set*)
- All documents that cite or are cited by a document in the root set are collected (base set)
- Used EPO web service to extract citation information





Problem Definition

- Directed weighted graph $G = (V, E)$
- V is a set of $|V| = N$ patent documents
- A set of citation relationships between patent documents $E \in V \times V$
- A set of attributes associated to a patent (applicant, inventor, conceptual tags)

$$X = \{x_1, x_2, \ldots, x_N\}$$

- Goal: suggest a ranked list of citations for a patent based on its content and attributes

Initial Query Model

- Estimate a query by quantifying the difference between the language model of the query document and the language model of the collection (the cross entropy)

Outline

- Introduction and challenges
- Related work and definitions
- Building the network
- Time aware network analysis
- Reranking and query expansion
- Experimental results
- Conclusions

Initial Probability

- The initial probability is exponentially discounted according to the age of the node

$$\rho_i = e^{\frac{-\text{age}}{\tau_d}}$$

- Compensating for the bias of the page rank algorithm against recent documents

Weights for Citation Relationships

- Number of common classification codes
- Number of common inventors
- Number of common applicants
- Lexical similarity (cosine similarity)
- Difference of the publication date
- Combination of different weights

Weighted PageRank

$$PR{-}W(u) = \lambda . \frac{O_u}{\sum_{p \in L_v} O_p} + (1-\lambda) . \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

Re-ranking and Query Expansion

- Identify and weigh terms in influential documents belonging to the citation graph
- Use weighted page rank values as document prior to guide term selection

$$P(t|Q_{cit}) = Z_t \sum_{D \in G_{cit}} P(t|D)P(D)$$

- Interpolate citation query model with initial query model

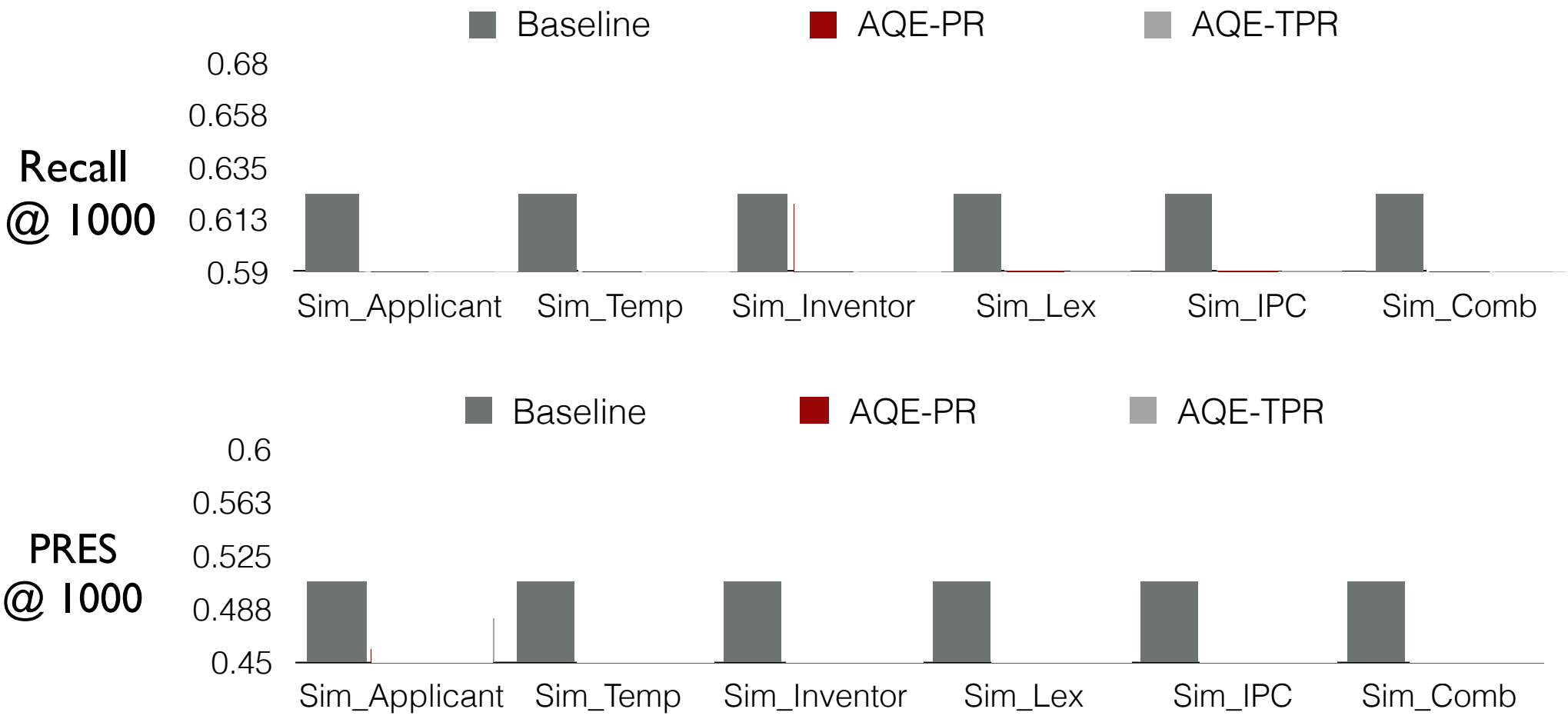
Experimental Settings

- Dataset
  - two patent collections: CLEF-IP 2010 and CLEF-IP 2011 (relevance judgement: citations)
  - size of each collection: 1.3 million patents
- Evaluation metric
  - PRES*: designed for recall-oriented application
- Baseline
  - second best participant of the CLEF-IP 2010 (no bias favoring applicant's initial citation list**)

* Magdy and Jones: SIGIR 2010

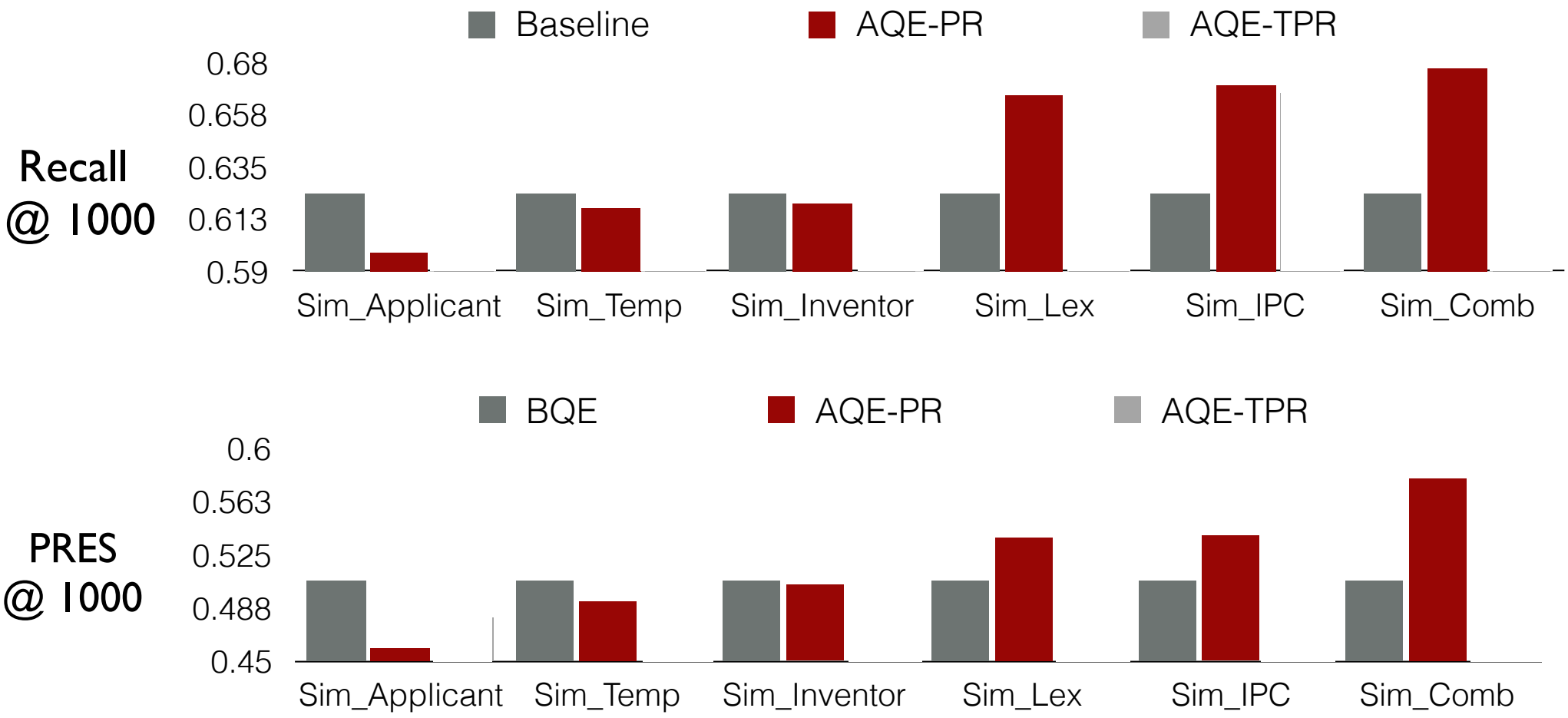
** Magdy, Lopez and Jones: ECIR 2010

Results on CLEF-IP 2010



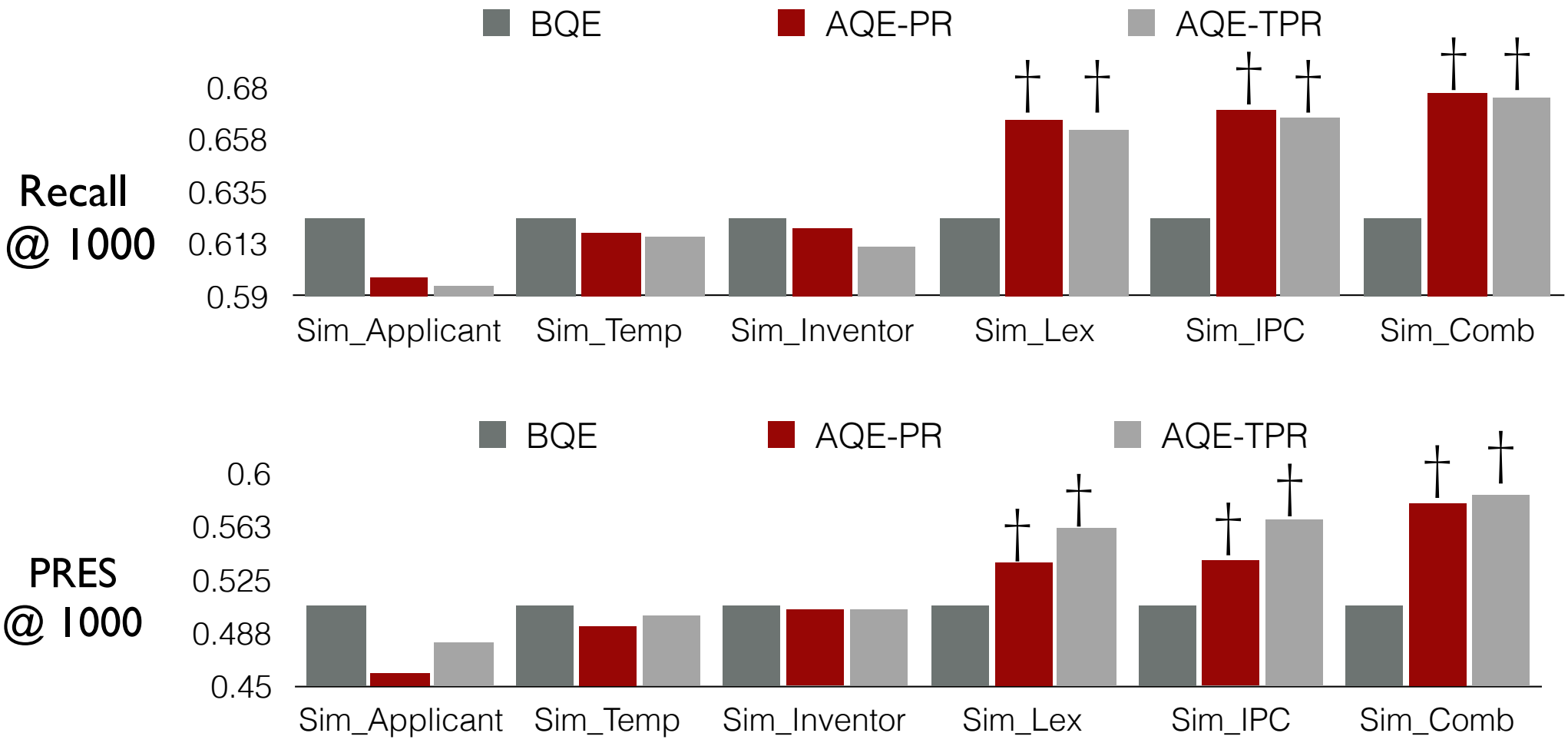
Baseline: before QE
AQE-PR: after QE - weighted PR
AQE-TPR: after QE - weighted temporal PR

Results on CLEF-IP 2010



Baseline: before QE
AQE-PR: after QE - weighted PR
AQE-TPR: after QE - weighted temporal PR

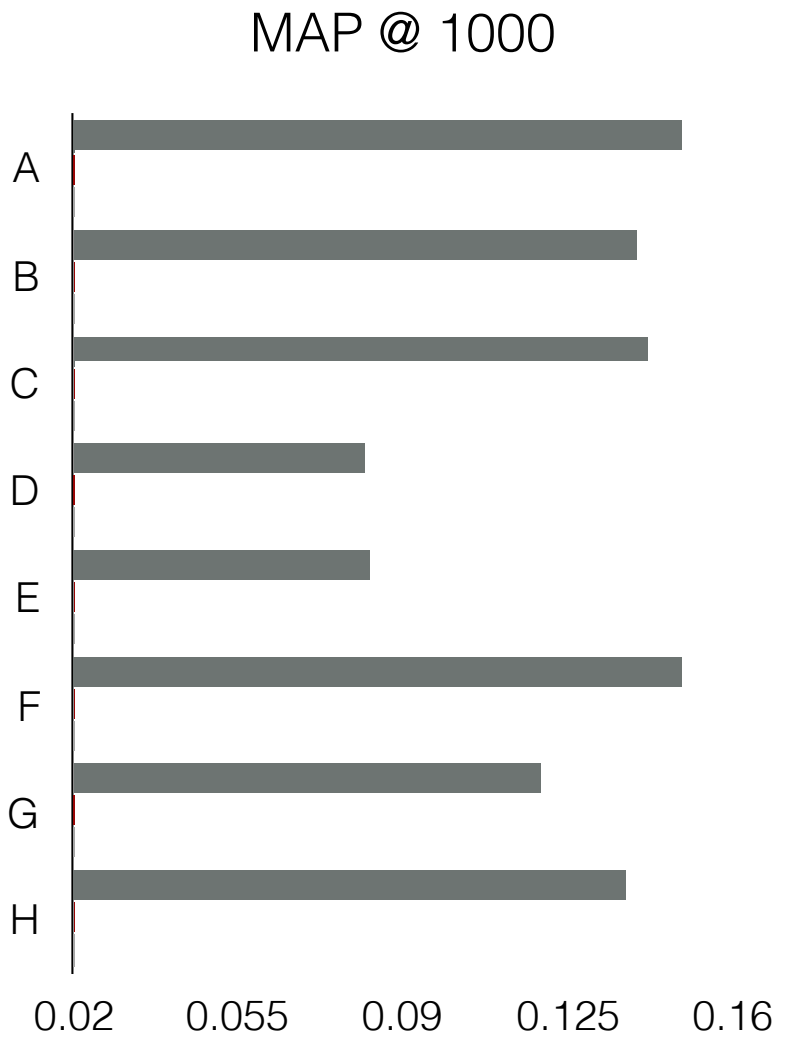
Results on CLEF-IP 2010



Baseline: before QE
AQE-PR: after QE - weighted PR
AQE-TPR: after QE - weighted temporal PR

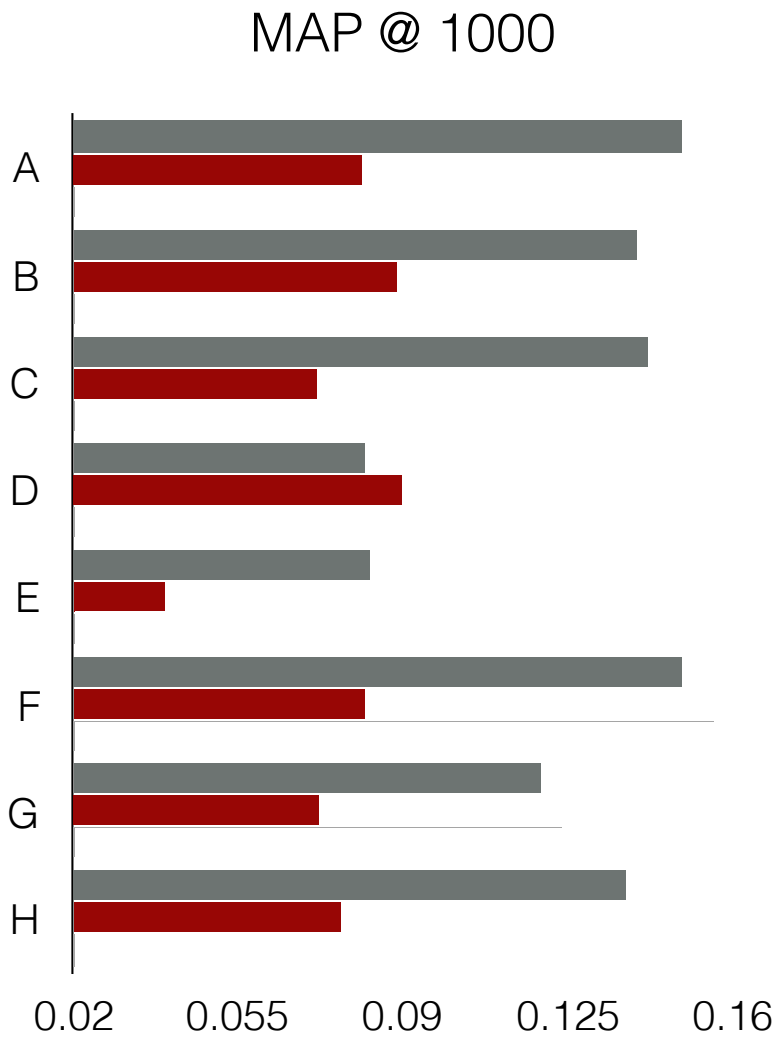
†: $p < 0.01$ over the baseline

Results on Different Technological fields



A: Human necessities
B: Performing operations and transporting
C: Chemistry and Metallurgy
D: Textiles and Papers
E: Fixed Constructions
F: Mechanical Engineering
G: Physics
H: Electricity

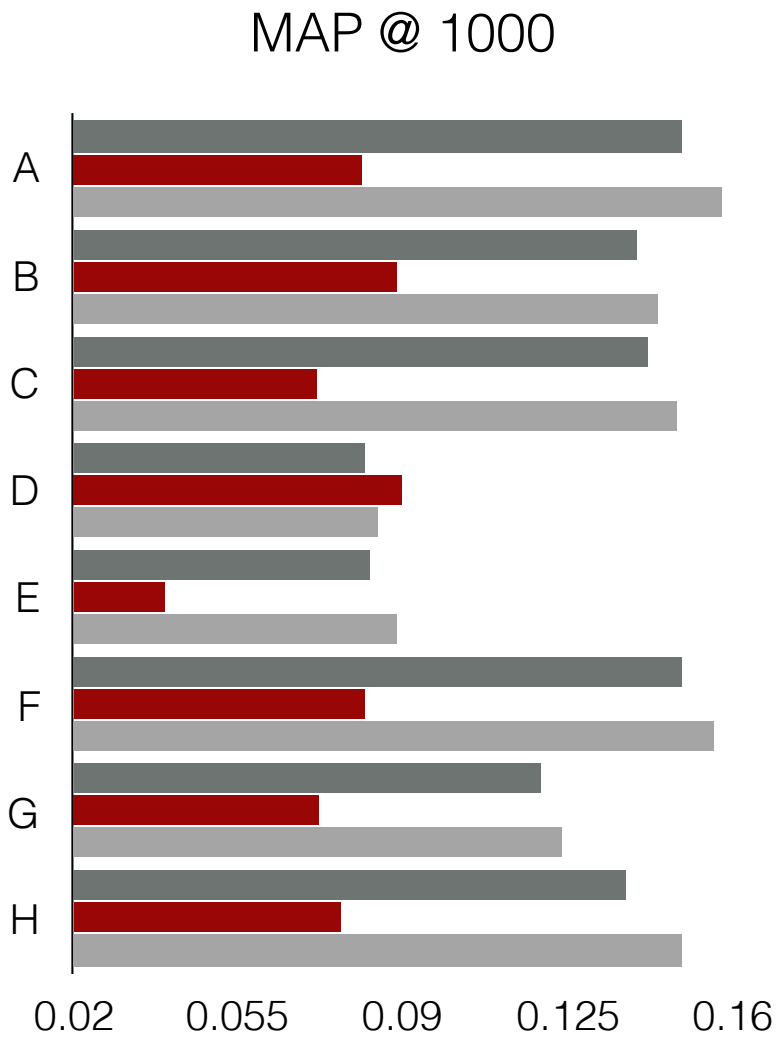
Results on Different Technological fields



A: Human necessities
B: Performing operations and transporting
C: Chemistry and Metallurgy
D: Textiles and Papers
E: Fixed Constructions
F: Mechanical Engineering
G: Physics
H: Electricity

Baseline
AQE-PR
AQE-TPR

Results on Different Technological fields



A: Human necessities
B: Performing operations and transporting
C: Chemistry and Metallurgy
D: Textiles and Papers
E: Fixed Constructions
F: Mechanical Engineering
G: Physics
H: Electricity

Conclusions

- Built a directed weighted graph of citations, encoding contextual similarities as edge weights: common IPC classes, applicants and inventors and temporal distance
- Found influential documents using time aware weighted page rank on the citation network
- Candidate documents are used to draw terms for query expansion
- Lexical and categorical similarities work best for recall, temporal information improves the MAP

Università
della
Svizzera
italiana

**Faculty
of Informatics**

Query Driven Mining of Citation Networks

for Patent Citation Retrieval and Recommendation

Parvaz Mahdabi †‡ **Fabio Crestani** †

University of Lugano, Switzerland †
Idiap Research Institute, Martigny, Switzerland ‡

CIKM 2014