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

- 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

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

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

It is **Too Difficult** to Get Improved Over the Baseline!
Previous Work

- 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

• 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

Data Collection
- CLEF-IP 2010
- English Subset

IR Model
- TF-IDF
- BM25
- LM

IPC filter

Patent Query
- IPC code
- Title
- Abs.
- Description
- Claims

Relevance Feedback
Pseudo Relevance Feedback

Top-100
- Doc1  rel
- Doc2  irr
- Doc3  irr
- Doc99 rel
- Doc100 irr
Oracular Relevance Feedback System

• 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
We define Relevance Feedback (RF) score for each term as follows:

\[
RF(t, Q) = \text{Rel}(t, Q) - \text{Irr}(t, Q)
\]  

where

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

\[
\text{Irr}(t) \rightarrow \text{Avg. Term Frequency in Irr. Docs.}
\]
Oracular Query Formulation

• We Formulate Two Oracular Queries:

1. Oracular Query

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

2. Oracular Patent Query

\[ \{ t \in Q | RF(t, Q) > \tau \} \]
Baseline vs. Oracular Query

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>PATATRAS</th>
<th>Oracular Query</th>
<th>Oracular Patent Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM</td>
<td>MAP</td>
<td>0.112</td>
<td><strong>0.482</strong></td>
<td><strong>0.414</strong></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.416</td>
<td>0.582</td>
<td>0.591</td>
</tr>
<tr>
<td>BM25</td>
<td>MAP</td>
<td>0.123</td>
<td><strong>0.492</strong></td>
<td><strong>0.424</strong></td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>0.431</td>
<td>0.584</td>
<td>0.598</td>
</tr>
</tbody>
</table>

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

- OracularQuery
- OracularPatentQuery
- baseline

Values of $\tau$:

-10 -5 0 5 10

MAP:
0 0.1 0.2 0.3 0.4 0.5

(1) System is Oversensitive to Noisy Terms.

(2) High MAP by Selecting Useful Terms in Patent Query.
Query Reduction (QR)

• We Need to Reduce Query to Get Improved.
Approximating Oracular Query

- 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

1. Pruning Document Frequent (DF) Terms
   - Remove Terms with High Avg. Term Frequency in Top 100 ($DF(t) > \tau$)

2. Pruning Query Infrequent Terms
   - ($QTF(t) \leq \tau$)

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.
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)

- $DF(t) > \tau$
- $QTF(t) \leq \tau$
- $PRF(t) > \tau$

--- IPC Title

--- baseline

<table>
<thead>
<tr>
<th>Values of $\tau$</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>5</td>
<td>0.06</td>
</tr>
<tr>
<td>10</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Anecdotal Example

(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.


QTF Terms: starch: -3.5, enzym: 29.5, enz: -5.5, hydrid: -5.5, reaction: -2, reagent: 1.2, ether: 2, enzym: -5.5


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

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

<table>
<thead>
<tr>
<th></th>
<th>1st Rel. Patent (k=1)</th>
<th>1st Three Rel. Patents (k=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.289</td>
<td>0.369</td>
</tr>
<tr>
<td>Avg. Recall</td>
<td>0.484</td>
<td>0.547</td>
</tr>
</tbody>
</table>

1. MAP Doubles Over the Baseline (0.112 $\rightarrow$ 0.289)
2. Outperforms PATATRAS (0.226 $\rightarrow$ 0.289)
Minimum Effort

- Baseline Returns Top Rel. Patent
  - 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

Thank you

mona.golestnfar@anu.edu.au
References and Further Reading


References


