# NTCIR-7 Patent Mining Experiments at RALI

Guihong Cao, Jian-Yun Nie and Lixin Shi
Department of Computer Science and Operations Research
University of Montreal, Canada

# Patent Mining as Classification

- Task
  - Classify research abstracts into IPC
- Possible solution
  - Skewed distribution over classes
  - Non-parametric classification approach: kNN
- Investigated issues
  - Possible vocabulary between paper abstracts and patents (different fields)
    - Term distillation
    - Use a subset of fields
  - Pseudo-relevance feedback
  - Effect of k

# Basic Classification Approach

- Finding k closest documents using information retrieval
  - Language modeling approach for information retrieval
  - Measuring relevance by query likelihood

$$\left.\begin{aligned} P(w \mid D) &= \lambda \frac{tf(w, D)}{\mid D \mid} + (1 - \lambda) P(w \mid C) \\ P(q \mid D) &= \prod_{q_i} P(q_i \mid D) \end{aligned}\right\} \text{Select } k \text{ documents}$$

$$score(c, q) = \sum_{i=1}^{K} \delta\big(ipc(d_i) = c\big) P(q \mid d_i)$$

# Term Distillation

- Some common words in research paper are not common words in patent description (e.g. paper, study, propose)
- Filtering out the common words from paper abstracts

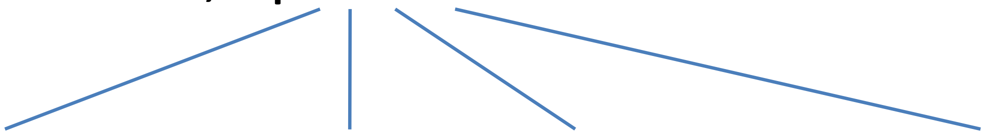
| | | | |
|---|---|---|---|
| e.g. | propose | prepare | shows |
| proposed | prepares | showing | |
| paper | based | preparing | |

# Query Expansion

- Pseudo-relevance feedback
  – More effective for short queries
  – Is it effective for the Patent Mining task (quite long query)?

# Using Patent Structures

- Do different fields have different impacts?
- Four main fields
  - Title, Abstract, Specification and Claim



Background, Description, Summary and Drawing

- Experiments:
  - Using some of the fields
  - Aggregating occurrence of query terms in different fields with linear interpolation (with equal weights)

# Term Distillation Results

| Model | P@30 | P@100 | MAP |
|---|---|---|---|
| Original | 0.0277 | 0.0047 | 0.1502 |
| Term Distillation | 0.0282 | 0.0046 | 0.1491 |

# Pseudo-Relevance Feedback (20 docs)

| #Exp. Terms | P@30 | P@100 | MAP |
|---|---|---|---|
| 0 | 0.0271 | 0.0047 | 0.1488 |
| 20 | 0.0274 | 0.0029 | 0.1470 |
| 40 | 0.0274 | 0.0030 | 0.1451 |
| 60 | 0.0277 | 0.0029 | 0.1447 |
| 80 | 0.0277 | 0.0030 | 0.1439 |
| 100 | 0.0276 | 0.0030 | 0.1456 |

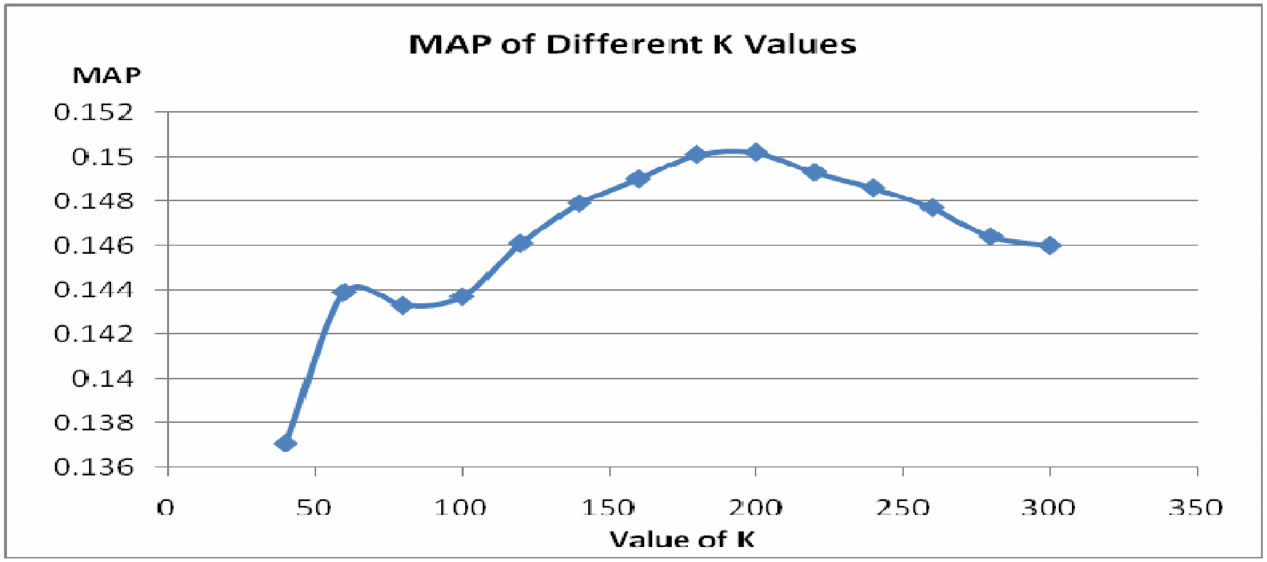
# The Impact of Different Fields

T: title A: abstract S: specification C: claim

B: background D: description M: summary R: drawing

| Fields | P@30 | P@100 | MAP |
|---|---|---|---|
| T+A+S+C | 0.0277 | 0.0047 | 0.1502 |
| T+A+B | 0.0270 | 0.0041 | 0.1470 |
| T+A+B+D | 0.0281 | 0.0049 | 0.1489 |
| T+A+B+D+M | 0.0276 | 0.0047 | 0.1495 |

# The Impact of K



# Observations

- Only the value of K has some impact on classification effectiveness
- The other factors do not seem to affect the classification accuracy:
  - Different fields
  - pseudo-relevance feedback
  - Term distillation
- Questions:
  - Exploiting more characteristics of patents?
  - Term relationships?