THUIR at NTCIR-10
INTENT-2 Task

IR Group of Tsinghua University

Yufei Xue, Fei Chen, Aymeric Damien, Cheng Luo, Shuai Huo, Min Zhang, Yiqun Liu, Shaoping Ma
Overview

• THUIR@INTENT2: three subtasks
  – English Subtopic Mining
  – Chinese Subtopic Mining
  – Document Ranking
English Subtopic Mining

- External resources v.s. Top search results

Original Query

- External resources
  - Linear Combination
- Top search results
  - Clustering & Keyword Extraction

Post process and combination
English Subtopic Mining

- External Resource Based Subtopic Mining
  - Subtopic candidate generation
    - Query Completion (Google, Bing, Yahoo)
    - Query Suggestion (Google, Bing, Yahoo)
    - Google Insights / Google Keywords Generator
    - Wikipedia (Disambiguation items)
  - Post process: Remove candidates without any query keywords
  - Linear combination
    - Google Insights: 0.15; Google Keywords Generator: 0.75; Query Suggestion / Completion: 0.05
English Subtopic Mining

- Top Results Based Subtopic Mining
  - Result document description
    - Search result snippets
    - Important fields of result documents ("h1", anchor, ...)
    - BM25 scores are calculated for each word
  - Result clustering
    - PAM (Partitioning Around Medoids) algorithm
    - Without assigning the number of clusters
  - Keyword extraction for each cluster
    - Select the most frequent word and extend it to an n-gram.
  - Rank keywords by their clusters.
English Subtopic Mining

- Combination of subtopics
  - Linear combination
  - Duplication removing with WordNet
  - Normalization and re-ranking.
Chinese Subtopic Mining

- **Candidate subtopic generation**
  - Query suggestions collected from Google, Sogou, Baidu and Bing
  - Disambiguation items
  - collected from Hudong.com and Wikipedia
  - Keywords extracted from LDA topics generated on clicked snippets

- **Candidate ranking**
  - Credibility of external resources (e.g. Wikipedia = 2, Google = 2, Hudong = 1, …)
  - Number of common words
  - Length of the subtopic
  - Number of words in clicked snippets
Chinese Subtopic Mining

- Clicked snippets & user intent
  - User clicks a result => user is interested in the snippets of the results
  - Click-through information: SogouQ
- LDA on clicked snippets
  - 10 implicit topics for each query.
Chinese Subtopic Mining

- Result comparisons
  - Snippet click-through information helps improve candidate ranking
  - Candidates generated by LDA on snippets are not so effective

<table>
<thead>
<tr>
<th></th>
<th>I-rec@10</th>
<th>D-nDCG@10</th>
<th>D#-nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Query suggestion</td>
<td>0.3792</td>
<td>0.4739</td>
</tr>
<tr>
<td>2</td>
<td>1 + Snippet</td>
<td>0.3786</td>
<td><strong>0.5028</strong></td>
</tr>
<tr>
<td>3</td>
<td>2+LDA</td>
<td><strong>0.3839</strong></td>
<td>0.4843</td>
</tr>
</tbody>
</table>
Document Ranking

- **Selective Diversification**
  - Informational query:
    - IA-Select according to the D#-nDCG value of a document

- **Query Type Identification**
  - nCS(q):
    - While performing a navigational type search request, user tend to click a small number of URLs
    - Session of q that involves less than n clicks) / (session of q)
  - nRS(q):
    - While performing a navigational type search request, user tend to click only the first few URLs.
    - (Session of q that involves clicks only on top n results) / (Session of q)
  - CD(q):
    - (Click on the most popular result of q) / (Click on all results of q)

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigational</td>
<td>NAV</td>
<td>NAV</td>
</tr>
<tr>
<td>Informational &amp; Transactional</td>
<td>91.07%</td>
<td>85.62%</td>
</tr>
<tr>
<td>Precision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>90.71%</td>
<td>86.18%</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.91</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Document Ranking

- Diversify Results Based on Novelty

Set $S = d_0$

$|S| < k \quad \Rightarrow \quad d_i = \text{argmax } f(d_i, S)$

Add $d_0$ to $S$

$|S| < k \quad \Rightarrow \quad f(d_i, S) = \sum_{d_j \in S} (N - r_j) + \lambda \times \cos(w_i, w_j)$

End
## Document Ranking

### Experimental Results

<table>
<thead>
<tr>
<th></th>
<th>I-rec@10</th>
<th>D-nDCG@10</th>
<th>D#-nDCG@10</th>
<th>DIN-nDCG@10</th>
<th>P+Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>0.7247</td>
<td><strong>0.4207</strong></td>
<td>0.5727</td>
<td>0.2858</td>
<td>0.2653</td>
</tr>
<tr>
<td><strong>Selectively</strong></td>
<td>0.6731</td>
<td>0.3587</td>
<td>0.5159</td>
<td>0.2611</td>
<td>0.2203</td>
</tr>
<tr>
<td><strong>Novelty based</strong></td>
<td><strong>0.7258</strong></td>
<td>0.4201</td>
<td><strong>0.5729</strong></td>
<td><strong>0.2865</strong></td>
<td><strong>0.2663</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thank you