Motivation:
- Queries are usually ambiguous and/or underspecified.
- Different users often have different intents for the same query.
To learn user’s search intent, subtopic mining plays an role in information retrieval problem.

Introduction

Assumption:
- Subtopics are the specification or reformations of the original query.
- Some subtopics are more likely than others.

Mining Subtopics:
- Index subtopics from logs, using Lucene.
  - Given a topic, search subtopics in across logs.
- Estimating the co-occurrence frequency of subtopics.
- Filtering subtopics using some rules
  - Removing duplicates that have similar sense.

Main Processing

Subtopic Selection:
- Given a topic, select subtopics using rules
  - The length of the subtopic, its Edit-Distance to the topic and some other features

Ranking:
- Estimate the rank of the subtopics
  - Choose the subtopics with high frequency.
  - If there is a tie, choose the subtopics with nearest Edit-Distance to the topic
  - And further, if there is also a tie, choose the subtopic with lexicographically smaller one.

Evaluation

Primary Evaluation Metric:
\[ D# - nDCG@k = y I - rec@k + (1 - y) D - nDCG@k \]

Results:

<table>
<thead>
<tr>
<th>Runs</th>
<th>I-Rec@10</th>
<th>D-nDCG@10</th>
<th>D#-nDCG@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEM12-S-E-1A</td>
<td>0.3780</td>
<td>0.4233</td>
<td>0.4007</td>
</tr>
<tr>
<td>SEM12-S-E-2A</td>
<td>0.3777</td>
<td>0.4250</td>
<td>0.4014</td>
</tr>
<tr>
<td>SEM12-S-E-3A</td>
<td>0.3403</td>
<td>0.3573</td>
<td>0.3488</td>
</tr>
<tr>
<td>SEM12-S-E-4A</td>
<td>0.3727</td>
<td>0.3471</td>
<td>0.3599</td>
</tr>
<tr>
<td>SEM12-S-E-5A</td>
<td>0.3659</td>
<td>0.3445</td>
<td>0.3552</td>
</tr>
</tbody>
</table>

Conclusion

- We demonstrated that co-occurrence and Edit-Distance features achieve better result for few topics.
- Query logs are utilized only, moreover, other resources i.e. Wikipedia or Search engine hits might have useful features.
- Our system has lack of benefits from subtopic clustering that we leave as future work.

Discussion

Result: Needs Improvement
- utilizing Wikipedia for disambiguating some subtopics,
- anchor text for aggregating more subtopics
- adopting semantic similarity measures
- clustering subtopics to filter duplicating intents or extract more useful intents