Improving the Estimation of Relevance Models Using Large External Corpora

Don Metzler

Joint work with Fernando Diaz
Motivation

- Performance of pseudo-relevance feedback (PRF) depends critically on ability to find relevant material in collection.
- What if there is little or no relevant material in collection?
- Idea: expand against target collection plus one or more external collections.
Past Work

- **TREC 6**
  - “increasing the size of the database increases the likelihood of finding good expansion concepts” (Allan et. al., 1997)
  - “it is quite clear that ‘blind’ query modification is beneficial provided that a large enough database is available” (Walker et. al., 1997)

- **Local context analysis (LCA)**
  - LCA on larger, external collection led to 12% increase in 11-point average precision (Xu and Croft, 2000)

- **TREC Robust Track**
  - Web expansion
Our Goals

- Formalize multi-collection PRF
- Perform more detailed experiments
  - Does external expansion help for ad hoc retrieval on a web collection?
  - Is web expansion always the best or can we get by with expanding using something smaller?
  - How does external expansion compare to true feedback?
- Develop a better understanding of PRF
  - When and why does external expansion work?
Model

\[ Q = \lambda_1 \cdot C_1 + \lambda_2 \cdot C_2 + \ldots + \lambda_N \cdot C_N \]
Model

- Compute relevance model over each collection using top $M$ documents
- Mix models together
- Sample top $K$ terms from combined model and form expanded Indri query
  - $\#weight( \ P(\text{w}_1|R)\ w_1 \ ... \ P(\text{w}_K|R)\ w_K )$
- Mix original query with expanded query terms by forming Indri query
  - $\#weight( \ \lambda \ Q_{\text{ORIG}} \ (1-\lambda) \ Q_{\text{EXPANDED}} )$
Experimental Setup

- Experiments done using Indri
- Query-time stopping
- Krovetz stemming
- Evaluation uses 10-fold cross validation

<table>
<thead>
<tr>
<th>Collection</th>
<th>Documents</th>
<th>Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>trec12</td>
<td>469,949</td>
<td>150</td>
</tr>
<tr>
<td>robust</td>
<td>472,525</td>
<td>250</td>
</tr>
<tr>
<td>wt10g</td>
<td>1,692,096</td>
<td>100</td>
</tr>
<tr>
<td>bignews</td>
<td>6,422,629</td>
<td>-</td>
</tr>
<tr>
<td>gov2</td>
<td>25,205,179</td>
<td>-</td>
</tr>
<tr>
<td>web</td>
<td>19,200,000,000</td>
<td>-</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>BIGNEWS</th>
<th></th>
<th>GOV2</th>
<th></th>
<th>WEB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>QL</td>
<td>RM3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trec12</td>
<td>0.2502</td>
<td>0.3201</td>
<td>0.3204</td>
<td><strong>0.3319</strong></td>
<td>0.2709</td>
<td>0.3215</td>
</tr>
<tr>
<td>robust</td>
<td>0.2649</td>
<td>0.3214</td>
<td><strong>0.3501</strong></td>
<td><strong>0.353</strong></td>
<td>0.2748</td>
<td>0.3207</td>
</tr>
<tr>
<td>wt10g</td>
<td>0.1982</td>
<td>0.203</td>
<td><strong>0.2256</strong></td>
<td><strong>0.2331</strong></td>
<td>0.1999</td>
<td>0.1958</td>
</tr>
</tbody>
</table>

- **QL**: query likelihood
- **RM3**: original query + target RM
- **External**: original query + external RM
- **Mixture**: original query + mixture of target and external RM

- **Bold**: statistically significantly better than RM3
- **Italics**: statistically significantly worse than RM3
Pseudo vs. True Feedback

- Pseudo-relevance feedback using external collections yields strong effectiveness
- How does it compare to true feedback?

Experimental Setup
- Simulate feedback on top $K$ documents
- Construct ‘true’ relevance model from documents judged relevant
- Formulate query as mixture of original query and top terms from this model
Feedback Results

- trec12
  - both types of expansion equivalent to feedback on 5 documents

- robust
  - target expansion equivalent to feedback on 2 documents
  - external expansion equivalent to feedback on 5 documents

Legend

- dotted line = PRF using target collection
- dashed line = PRF using external collection
Collection Size Effects

- What is the relationship between external collection size and effectiveness?
- How large must the external collection be before we see diminishing returns?

Experimental Setup
- Randomly subsample documents from BIGNEWS
- Use sampled collection for external expansion
- Plot MAP vs. external collection size
Size Experiments
Size Experiments
Size Experiments

Graph showing the relationship between % of SIGNEWS and MFP.
Concept Density

- Can we predict the best collection to use for expansion?
- Collection size not only important factor
  - Expanding by GOV2 was not helpful
- Topic coverage important
  - Estimated RM likely to be poor for topics with poor coverage
  - *Concept density* measures how densely represented query concepts are in a collection
Concept Density

- Basic steps:
  - Extract concepts from query
  - Compute concept density for each collection
  - (optional) Construct RM from collection with highest concept density

- Extracting concepts
  - Use ‘dependence model’ concepts

- Computing density
  - Proportion of top $K$ documents that contain the concept
## Concept Density Example

**Query:** olympics training swimming

<table>
<thead>
<tr>
<th>Concept</th>
<th>robust density</th>
<th>bignews density</th>
</tr>
</thead>
<tbody>
<tr>
<td>olympics</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>training</td>
<td>0.88</td>
<td>1</td>
</tr>
<tr>
<td>swimming</td>
<td>0.98</td>
<td>1</td>
</tr>
<tr>
<td>#1( training swimming )</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>#1( olympics training )</td>
<td>0.1</td>
<td>0.24</td>
</tr>
<tr>
<td>#1( olympics training swimming )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#uw8( training swimming )</td>
<td>0.36</td>
<td>0.94</td>
</tr>
<tr>
<td>#uw8( olympics swimming )</td>
<td>0.62</td>
<td>0.98</td>
</tr>
<tr>
<td>#uw8( olympics training )</td>
<td>0.22</td>
<td>0.88</td>
</tr>
<tr>
<td>#uw12( olympics training swimming )</td>
<td>0.04</td>
<td>0.74</td>
</tr>
</tbody>
</table>

**overall density**: 0.43 0.66
Concept Density Results

change in density vs change in AvgP for trec12
Concept Density Results

robust04

change in AvgP

change in density
Concept Density Results

wt10g

change in density vs change in AvgP
Conclusions

**Contributions**
- Formal method for performing PRF using one or more external collections
- Comparison of PRF using large news collection and the web as external resources
- Used concept density to developing a better understanding of why PRF works/fails

**Future work**
- Use Wikipedia as external resource
- Further exploration of concept density