Combining Document Representations for Known Item Retrieval

Multiple sources of evidence are becoming more common
- Structured documents
- Linked documents
- Form document representations from these different sources
- Flat text of the document
- Text from documents that reference the document
- Representations using structural information about the document

Goal: combine the document representations in a way that will improve results

Old Idea
- Bayesian Inference Networks can accommodate multiple document representations (Inquery)
- Most often done by using different query representations using techniques similar to meta-search methods

Meta-Search Hypotheses [Croft 2000]
Adapted to Combining Representations

1. Scores/ranks across representations must be compatible
   - Same range – it makes sense to combine them
2. Representations must be high quality
3. Scores/ranks across representations should agree
   - Lower variance for correct documents than incorrect documents

Existing Meta-Search Approaches

- Ranks
  - Few assumptions about the representations
  - Ranks are “on the same scale”
  - Borda (sum of n - rank)
  - Condorcet
  - Reciprocal Rank (sum of 1/rank)

- Scores
  - More information in scores
  - May need normalization to make the scores compatible
  - CombSUM (sum of score)
  - CombMNZ (number of scores != 0 * sum of score)

Combining Representations is Different from Meta-Search

We can:
- choose the ranking algorithms used on the document representations
- create score normalization functions tailored to the ranking algorithms
- create models that combine information on the term level, rather than post-retrieval

Another Approach to Combining Reps – A Mixture-Based Language Model

- A straightforward extension of traditional language models in IR
- Combines information on the term level
- Estimate a new language by combining the language models estimated from each representation
  \[
  p(w|D) = \sum_{D_i}^{1} \lambda_i p(w|D_i)
  \]
  where \(D\) is a document, \(D_i\) is the document’s \(i\)th representation
- Different representations can receive different weights (\(\lambda\)), based on our belief of the quality of the representation
- Document is ranked by the generative probability of the new language model
  \[
  p(q|D) = \prod_{i=1}^{n} p(q_i|D_i)
  \]
Known Item Finding

- User has a specific document in mind
- The user can provide a good, terse description of the document
- Search engine's goal is to return the document as high in the ranking as possible

Evaluation Testbeds

- TREC 10 Homepage Finding
  - 80 Training topics (used to empirically set $\lambda$)
  - 145 Test Topics
  - WT10G Corpus - 1.7 million HTML documents
- TREC 11 Named-Page Finding
  - 150 Test Topics
  - .GOV Corpus
  - 1 million HTML documents
  - ¼ million other documents

Experimental Setup

<table>
<thead>
<tr>
<th>Base Representations</th>
<th>Ranking Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full document</td>
<td>Okapi</td>
</tr>
<tr>
<td>In-link</td>
<td>Traditional Generative Language Models</td>
</tr>
<tr>
<td>Title</td>
<td>Mixture-based Generative Language Model</td>
</tr>
<tr>
<td>META tags</td>
<td></td>
</tr>
<tr>
<td>Modified fonts</td>
<td></td>
</tr>
<tr>
<td>Image ALT tags</td>
<td></td>
</tr>
</tbody>
</table>

Performance of Individual Document Representations

<table>
<thead>
<tr>
<th>Language Models</th>
<th>OKAPI</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Homepage</td>
<td>Named-Page</td>
<td>Homepage</td>
</tr>
<tr>
<td>FULL</td>
<td>0.239</td>
<td>0.578</td>
<td>0.300</td>
</tr>
<tr>
<td>LINK</td>
<td>0.548</td>
<td>0.438</td>
<td>0.515</td>
</tr>
<tr>
<td>TITLE</td>
<td>0.345</td>
<td>0.371</td>
<td>0.332</td>
</tr>
<tr>
<td>ALT</td>
<td>0.141</td>
<td>0.158</td>
<td>0.186</td>
</tr>
<tr>
<td>FONT</td>
<td>0.164</td>
<td>0.146</td>
<td>0.155</td>
</tr>
<tr>
<td>META</td>
<td>0.067</td>
<td>0.107</td>
<td>0.115</td>
</tr>
</tbody>
</table>

Experimental Results: Hypothesis 1 - Score Compatibility

- X axis cumulative (+full is 3 representations: link, title, and full)
- Appropriate score normalization is important
- A MSE measure can give a prediction on the ordering of score normalization methods

Experimental Results: Hypothesis 2 - Representation Quality

- Graph suggests that only high quality representations help
- However: combining the three worst representations yields a MRR of 0.371! (Best of the three is 0.194)
- Best algorithms are robust to the inclusion of bad representations
- Preconditions for successful combination are not clear
### Experimental Results: Hypothesis 3 - Variance

- The variance of the correct document is usually **higher** than those of incorrect documents.
- This is different from meta-search.
- Not surprising given the nature of the document representations:
  - Correct documents: we expect that a query may be highly ranked for a couple of the structurally formed representations, but not all.
  - Incorrect documents: the query does not match any of the representations well, so the scores and ranks are closer to each other across the representations.

### Conclusions on Combining Document Representations for Known-Item Finding

- Score normalization important
  - Can be tuned to the ranking algorithm
- Not clear on how important the quality of representations is
  - Best algorithms are robust
- The score/rank variance of correct documents across representations is **higher** than for incorrect documents
- Can effectively combine representations at the term level
- Language models an effective tool for combining document representations
- Combining document representations is a distinct problem from meta-search
- Structural information is very common in documents (HTML, XML, ...), so combining representations is an important problem
- We should work toward developing techniques that leverage the unique characteristics of combining representations