Query and Document Models for Enterprise Search

Krisztian Balog, Katja Hofmann, Wouter Weerkamp, Maarten de Rijke

ISLA, University of Amsterdam
http://ilps.science.uva.nl

Outline

- Document search
- (what didn’t work) Document models
- (what did work) Query models
- Expert search
  - Try what worked for document search: query models
  - Wrap up, further directions

Task and Aims

- New task and collection
  - Task: Create an overview page for a given topic
  - Looking for documents that discuss the topic in detail
  - Topic definition includes example documents
- Aims
  - How does a web document search system perform on this task?
  - How to make use of example documents?

Approach

1. Rank documents according to query likelihood
   \[ p(d|q) \propto p(d)p(q|d) \]
2. Assume term independence and uniform document priors
   \[ p(d|q) \propto \prod_{t \in q} p(t|\theta_d)^{n(t,q)} \]
3. Perform computation in the log domain
4. Represent the model as a probability distribution over terms instead of a set of keywords
   \[ \log p(d|q) \propto \sum_{t \in q} p(t|q)p(t|\theta_d) \]

Document model

\[ p(t|\theta_d) = \sum_{i=1}^{n} \lambda_ip(t|d_i) + \lambda_c p(t) \]

- Mixture of components
  - Body, title, metadata, headings, anchor text
- How to set component weights?
  - Equal weights to all components
  - Train on TREC WEB Track 2004 test set
### Results

<table>
<thead>
<tr>
<th>Query model</th>
<th>TRECWEB04 weights</th>
<th>Document text only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>MAP 0.3163 P@5 0.4381 P@10 0.5119</td>
<td>MAP 0.3430 P@5 0.4786</td>
</tr>
<tr>
<td>Blind feedback</td>
<td>MAP 0.3549 P@5 0.4881 P@10 0.5119</td>
<td>MAP 0.4270 P@5 0.6214</td>
</tr>
<tr>
<td>Example documents</td>
<td>MAP 0.3613 P@5 0.4952 P@10 0.4381</td>
<td>MAP 0.3430 P@5 0.4786</td>
</tr>
<tr>
<td>Blind + Example</td>
<td>MAP 0.3923 P@5 0.5548 P@10 0.4381</td>
<td>MAP 0.4219 P@5 0.6095</td>
</tr>
<tr>
<td>Term based</td>
<td>MAP 0.3430 P@5 0.4786 P@10 0.4381</td>
<td>MAP 0.3430 P@5 0.4786</td>
</tr>
</tbody>
</table>

* Document text = Body + Title + Metadata

### Query model

- Baseline: \( p(t|q) = \frac{n(t, q)}{|q|} \)
- Approximate relevance-based query models (Lavrenko and Croft, 2001)
  - Blind feedback documents
  - Example documents
  - Combination
    - Document-based (example documents are added to the list of feedback documents)
    - Term-based combination of query models

### Topic-level comparison

- Baseline
- Blind feedback
- Example documents

### Topic-level comparison

#### #39: cane toads

<table>
<thead>
<tr>
<th>Topic</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.6764</td>
</tr>
<tr>
<td>Blind feedback</td>
<td>0.7223</td>
</tr>
<tr>
<td>Example documents</td>
<td>0.4239</td>
</tr>
</tbody>
</table>

#### #12: cancer risk

<table>
<thead>
<tr>
<th>Topic</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.6764</td>
</tr>
<tr>
<td>Blind feedback</td>
<td>0.7223</td>
</tr>
<tr>
<td>Example documents</td>
<td>0.4239</td>
</tr>
</tbody>
</table>

### Topic-level comparison

#### #39: cane toads

- Baseline
- Blind feedback
- Example documents

#### #12: cancer risk

- Baseline
- Blind feedback
- Example documents
Conclusions

- Document model
- Web settings do not carry over
- Using document text as the only component is "good enough"
- Query model
- Relevance-based query models improve
- Blind Feedback +5% MAP
- Example documents +24% MAP

Aims

- Finding relevant documents is vital to expert search
- This year the same set of topics is used for document and expert search
- Try to improve expert search by applying relevance based query models

Results

<table>
<thead>
<tr>
<th>Query model</th>
<th>Document search</th>
<th>Expert search</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>P910</td>
<td>MAP</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.3430</td>
<td>0.4786</td>
</tr>
<tr>
<td>Relevance based query models</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blind feedback</td>
<td>0.3613</td>
<td>0.5119</td>
</tr>
<tr>
<td>Example documents</td>
<td>0.4270</td>
<td>0.6214</td>
</tr>
<tr>
<td>Blind + Example</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document based</td>
<td>0.3787</td>
<td>0.5429</td>
</tr>
<tr>
<td>Term based</td>
<td>0.4219</td>
<td>0.6095</td>
</tr>
</tbody>
</table>

Approach

Document-based model

\[ p(q|d) \]

\[ q \rightarrow d \rightarrow c_{a1} \rightarrow p(c_{a1}|q) \]

\[ c_{a1} \rightarrow p(c_{a2}|q) \]

Example documents

Blind feedback

Dr. Michael Fenech: keeping our genes healthy

Expert search

Finding relevant documents is vital to expert search

This year the same set of topics is used for document and expert search

Try to improve expert search by applying relevance based query models
Conclusions

- Query model based on blind feedback hurts performance
- Topic drift?
- Query model constructed from example documents results in +26% MAP

Wrap-up

- Document models
- Mixture models don’t pay off
- Query models
  - Document search
  - Blind feedback +
  - Example documents ++
  - Combination (worse than example only)
  - Expert search
  - Blind feedback –
  - Example documents ++

Future directions

- Example documents
  - Richer representation of the user’s information need
  - Make further use of them
    - E.g., people associated with example documents for expert finding

Questions?

Krisztian Balog
kbalog@science.uva.nl
http://www.science.uva.nl/~kbalog
How does performance depend on the number of example documents?

![Graph showing performance vs number of example documents]

Document search - Mixture weights

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Equal w.</th>
<th>WEB04</th>
<th>CSIRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>0.3430</td>
<td>0.3229</td>
<td>0.3163</td>
<td>0.3529</td>
</tr>
<tr>
<td>P@5</td>
<td>0.4952</td>
<td>0.4952</td>
<td>0.4905</td>
<td>0.5286</td>
</tr>
<tr>
<td>P@10</td>
<td>0.4786</td>
<td>0.4524</td>
<td>0.4381</td>
<td>0.5024</td>
</tr>
<tr>
<td>Body</td>
<td>0.18</td>
<td>0.10</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>0.18</td>
<td>0.30</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Metadata</td>
<td>0.18</td>
<td>0.05</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Headings</td>
<td>0.18</td>
<td>0.10</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Anchors</td>
<td>0.18</td>
<td>0.40</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Doc.text (B+T+M)</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collection</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>0.10</td>
</tr>
</tbody>
</table>