Improving Web Page Retrieval using Search Context from Clicked Domain Names

Rongmei Li
Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions
Motivation

In the setting of ad-hoc Web page retrieval

Common problems: \( \Rightarrow \) query ambiguity

- short keyword query for specific topic of interest
- different vocabularies for the same topic

Traditional Solution:

- explicit user feedback
  - understand search context
  - bridge semantic gap
- implicit user feedback
  - pseudo relevance feedback
Motivation

In the setting of ad-hoc Web page retrieval

Common problems: ⇒ query ambiguity

• short keyword query for specific topic of interest
• different vocabularies for the same topic

Traditional Solution:

• explicit user feedback
  - understand search context
  - bridge semantic gap

• implicit user feedback ⇒ query log
  - pseudo relevance feedback
Motivation

Query log contains users’ search history:

- query terms: Prostate cancer treatments
- retrieved documents:
  - http://appliedresearch.cancer.gov/accessibility/
- clicked documents:
  - http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- user identifier: 2178
Motivation

Query log contains users’ search history:

- specifying query terms: DEXA 09 Sommerhaus
- retrieved documents:
  - http://appliedresearch.cancer.gov/accessibility/
- clicked documents:
  - http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- an anonymous identifier: 2178
Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions
Related Work

Language modeling framework:
- combine ranking results of past queries
- build query model as the average of all past query models
- smooth the current query using past queries and clicked document summaries
- smooth the current query using past queries, retrieved documents, and clicked documents

Query expansion with terms from click-throughs:
- top-ranked documents of the initial retrieval result
- explicitly judged documents
Related Work

Re-ranking top ranked documents:

- re-score documents by statistical distribution of similar queries within a search session or all sessions
- adjust original rank with ranks generated by log queries that contain or are associated with the original query
- learn new rank from user interaction features
Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions
• **query terms**: Prostate cancer treatments

• retrieved documents:
  http://appliedresearch.cancer.gov/areas/monitoring.html
  http://appliedresearch.cancer.gov/accessibility/
  http://cancercontrol.cancer.gov/hcirb/ceccr/

• **clicked documents**:  
  http://appliedresearch.cancer.gov/accessibility/

• document ranks

• date and time of search action: 2006-03-19 16:33:54

• an anonymous identifier: 2178
Research Questions

- Can we extract the common topical context from such data for a query?
- Can we use this knowledge to improve retrieval performance effectively?
Standard Language Models

\[
P(t_1, \ldots, t_n | D) = \prod_{i=1}^{n} P(t_i | D), \quad P_{ml}(t_i | D) = \frac{tf(t_i, D)}{|D|}
\]

Jelinek-Mercer Smoothing:

\[
P(t | D) = \lambda P_{ml}(t_i | D) + (1 - \lambda) P_{ml}(t_i | C)
\]
Parsimonious Language Models

Use **EM estimator** to compute $P(t_i|D)$

**E-step:**
$$e_t = tf(t, D) \cdot \frac{\mu P(t|D)}{\mu P(t|D) + (1 - \mu) P_{ml}(t|C)}$$

**M-step:**
$$P(t|D) = \frac{e_t}{\sum_t e_t}, \text{ i.e. normalizing+prunning process}$$

Application: select topical terms from restored Web pages for query expansion
Document Ranking by Cross-entropy Score

\[ P_{ml}(t_i | Q) = \frac{tf(t_i, Q)}{|Q|} \]

\[ P'(t_i | D) = \lambda P_{ml}(t_i | D) + (1 - \lambda) P_{ml}(t_i | C) \]

\[ \text{Score}(D) = \sum_{i=1}^{l} [P_{ml}(t_i | Q) \cdot \log(P'(t_i | D))] \]
Query Log Modeling (1)

- query terms: Prostate cancer treatments
- retrieved documents:
  - http://appliedresearch.cancer.gov/accessibility/
- clicked documents:
  - http://appliedresearch.cancer.gov/accessibility/
- document ranks
- date and time of search action: 2006-03-19 16:33:54
- an anonymous identifier: 2178
Query Log Modeling (2)

Choose subsets of Web page collection from the Internet
- a large open Web directory
- a target collection whose Web pages will be ranked

Reconstruct URLs from domain names at 3 different levels
- **domain level:**
Query Log Modeling (2)

Choose subsets of Web page collection from the Internet

- a large open Web directory
- a target collection whose Web pages will be ranked

Reconstruct URLs from domain names at 3 different levels

- **domain level:**
  
  \[
  \text{http://www.cancer.gov} \Rightarrow \text{http://seer.cancer.gov/***}
  \]

- **server level:**
  
  \[
  \text{http://www.cancer.gov} \Rightarrow \text{http://www.cancer.gov/***}
  \]
Query Log Modeling (2)

Choose subsets of Web page collection from the Internet
- a large open Web directory
- a target collection whose Web pages will be ranked

Reconstruct URLs from domain names at 3 different levels
- **domain level:**
- **server level:**
- **URL level:**
Query Log Modeling (3)

Our strategies:

- **strategy 1**: promote the restored Web pages to the top of the ranking list.

  a b
  b d
  c a
  d c

- **strategy 2 & 3**: extract topical terms from restored Web pages or top ranked Web pages of strategy 2 to expand the original query.
Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions
Experiments - data

- Query log: 3 month search record containing 10 million queries and 20 million clicked domain names

<table>
<thead>
<tr>
<th>URL</th>
<th>Query</th>
<th>Date/Time</th>
<th>Page</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.kbb.com">http://www.kbb.com</a></td>
<td>kbb</td>
<td>2006-03-19 16:33:54</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><a href="http://www.remax.com">http://www.remax.com</a></td>
<td>remax</td>
<td>2006-03-21 10:32:18</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- Reconstruction source: DMOZ (Open Directory Project) or GOV2 collection

- Test data
  - terabyte collection of Web pages with .gov domain name (GOV2 collection)
  - 150 associated queries
Experiments - pre-processing results

Web page restoration at domain level

<table>
<thead>
<tr>
<th>Stripped URLs</th>
<th>URLs of GOV2 Web Pages</th>
<th>URLs of DMOZ Web Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>whitehouse.gov</td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
<td><a href="http://www.whitehouse.gov">www.whitehouse.gov</a></td>
</tr>
</tbody>
</table>

Motivation

31st August 2009

Rongmei Li
Experiments - topical terms

- original query terms: prostate cancer treatment
- expansion terms

<table>
<thead>
<tr>
<th>topical term</th>
<th>term probability (EM)</th>
<th>term probability (ML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer</td>
<td>0.0532</td>
<td>0.0248</td>
</tr>
<tr>
<td>patient</td>
<td>0.0150</td>
<td>0.0070</td>
</tr>
<tr>
<td>para</td>
<td>0.0125</td>
<td>0.0059</td>
</tr>
<tr>
<td>trial</td>
<td>0.0103</td>
<td>0.0048</td>
</tr>
<tr>
<td>cell</td>
<td>0.0100</td>
<td>0.0047</td>
</tr>
<tr>
<td>therapi</td>
<td>0.0097</td>
<td>0.0046</td>
</tr>
<tr>
<td>studi</td>
<td>0.0094</td>
<td>0.0044</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
## Performance of strategy 1

<table>
<thead>
<tr>
<th>models/levels</th>
<th>performance metrics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BPREF</td>
<td>improvement</td>
</tr>
<tr>
<td>baseline (JM)</td>
<td>0.4625</td>
<td>-</td>
</tr>
<tr>
<td>url</td>
<td>0.4625</td>
<td>-</td>
</tr>
<tr>
<td>server</td>
<td>0.4814</td>
<td>+3.93%</td>
</tr>
<tr>
<td>domain</td>
<td>0.4768</td>
<td>+3.00%</td>
</tr>
</tbody>
</table>

Motivation

Related Work

Our Work

Experiments

Conclusions
## Performance of strategy 2

<table>
<thead>
<tr>
<th>models/levels</th>
<th>performance metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
</tr>
<tr>
<td>baseline (JM)</td>
<td>0.3294</td>
</tr>
<tr>
<td>url.ML</td>
<td>0.3484</td>
</tr>
<tr>
<td>server.ML</td>
<td>0.3729</td>
</tr>
<tr>
<td>domain.ML</td>
<td>0.3714</td>
</tr>
<tr>
<td>url.EM</td>
<td>0.3509</td>
</tr>
<tr>
<td>server.EM</td>
<td><strong>0.3849</strong></td>
</tr>
<tr>
<td>domain.EM</td>
<td>0.3842</td>
</tr>
</tbody>
</table>
Performance of strategy 3

<table>
<thead>
<tr>
<th>models/levels</th>
<th>performance metrics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>improvement</td>
<td>P@10</td>
</tr>
<tr>
<td>baseline (JM)</td>
<td>0.3470</td>
<td>-</td>
<td>0.4909</td>
</tr>
<tr>
<td>url.ML</td>
<td>0.3821</td>
<td>+9.19%</td>
<td>0.5818</td>
</tr>
<tr>
<td>server.EM</td>
<td>0.4226</td>
<td>+17.89%</td>
<td><strong>0.6818</strong></td>
</tr>
<tr>
<td>domain.EM</td>
<td>0.4232</td>
<td>+18.01%</td>
<td>0.6455</td>
</tr>
<tr>
<td>url.EM</td>
<td>0.3959</td>
<td>+12.35%</td>
<td>0.5727</td>
</tr>
<tr>
<td>server.EM</td>
<td>0.4386</td>
<td>+20.88%</td>
<td>0.6727</td>
</tr>
<tr>
<td>domain.EM</td>
<td><strong>0.4391</strong></td>
<td>+20.97%</td>
<td>0.6455</td>
</tr>
</tbody>
</table>
Improvement on individual queries

Motivation

Related Work

Our Work

Experiments

Conclusions

31st August 2009

Rongmei Li
# Implicit vs Explicit (true) Feedback

<table>
<thead>
<tr>
<th>models/levels</th>
<th>performance metrics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>difference</td>
<td>P@10</td>
</tr>
<tr>
<td>baseline (JM11)</td>
<td>0.4734</td>
<td>-</td>
<td>0.8455</td>
</tr>
<tr>
<td>best run (11)</td>
<td>0.4391</td>
<td>7.25%</td>
<td>0.6455</td>
</tr>
<tr>
<td>baseline (JM29)</td>
<td>0.3366</td>
<td>-</td>
<td>0.7034</td>
</tr>
<tr>
<td>best run (29)</td>
<td>0.3020</td>
<td>10.28%</td>
<td>0.5172</td>
</tr>
</tbody>
</table>
Outline

- Motivation
- Related work
- Our work
- Experiments
- Conclusions
Conclusions

In this work, we:

- demonstrate how to restore stripped URLs to Web pages at three different levels from two collections
- present three strategies to integrate query and click-through information with the language modeling framework
- show that retrieval performance can be improved effectively
Thank you for your attention ... Questions and comments?