Effective Structured Query Formulation for Session Search

Dongyi Guan

Department of Computer Science
Georgetown University
dg372@georgetown.edu

Grace Hui Yang

nDCG@10 for TREC 2011 runs

nazli@cs.georgetown.edu

Problem
- A search session contains
  - Interactions with Previous Queries (Snippets, Search Results, Clicked Information)
  - Current Query
- Given a session, the tasks are to retrieve relevant documents to the current query.
- Four subtasks
  - RL1: only use the current query \( q_s \)
  - RL2: uses the previous queries \( q_1, q_2, ..., q_n \) and the current query \( q_s \)
  - RL3: provides top retrieved documents for previous queries
  - RL4: and information about which top results are clicked by users

Structure Query Formulation
- Identify Nugget: substrings in a query, similar to phrase
  serving spinal cord paralysis \( \iff \) serving #1(spinal cord) paralysis
- Add positional information in snippets
  ...preservation uk spinal cord injury care in egypt cky cord
  1 2 3 4 5 6 7 8 9 10
  noats guitar viewsnowid power cord malfunction spinal cord
  11 12 13 14 15 16 17 18
  stimulator...injury spinal cord dell...extension cord
  19 20 21 22 23 24 25
  nylon...cord paralysis vocal vegas...
  26 27 28 29 30

- Strict Method
  \[
  \frac{\text{count}(w_i, \text{Snippet})}{\min(\text{count}(w_i, \text{Snippet}), \text{count}(w_i, \text{Snippet}))} \geq \theta \implies w_i \text{ is nugget}
  \]
- Relaxed Method
  \[
  \phi = \begin{cases} 
  \#1(w_i) & |x_i(w_i) - \bar{x}(w_i)| \leq 5 \\
  \#2(w_i) & 5 < |x_i(w_i) - \bar{x}(w_i)| \leq 10 \\
  \#3(w_i) & |x_i(w_i) - \bar{x}(w_i)| > 10 
  \end{cases}
  \]
  \[
  \bar{x}(w_i) = \frac{1}{k} \sum_{j=1}^{k} \frac{x_{j}(w_i)}{\text{count}(w_i, S_j)}
  \]

Re-ranking by Dwell Time
Dwell time: the elapsed time that user stays in the page
\[
\Delta t = t_{end} - t_{start}
\]
Clicked documents: \( \{c_1, c_2, ..., c_k\} \)
Associated dwell time: \( \{\Delta t_1, \Delta t_2, ..., \Delta t_k\} \)
Re-ranking the returned documents \( \{d_j\} \) by the score:
\[
s(d_j) = \sum_{i=1}^{k} \text{Sim}(d_j, c_i) \cdot \Delta t_i
\]

Conclusions
- Terms from previous queries can boost the search performance significantly
- Removing duplicated queries improves the search performance
- Formulating structured queries by grouping terms into nuggets is very effective for TREC2011 data.

Query Expansion
- with previous queries
  \[
  \sum_{\text{weight}(\lambda_i \# \text{combine}(\text{nugget}_{i1}, \text{nugget}_{i2}, ..., \text{nugget}_{in}), w_1, w_2, ..., w_n)}
  \]
  \[
  \lambda_k = \begin{cases} 
  \lambda_p & k = 1, 2, ..., n - 1 \\
  1 - \lambda_p & k = n 
  \end{cases}
  \lambda_p = 0.4
  \]
- with anchor texts
  - Generate anchor log by harvestlinks in Lemur toolkit
  - Extract the top 5 frequent anchor text in the previous results
  - Weights are proportional to the frequency of anchor text
  - Append to the structured query expanded with previous queries
  \[
  \beta_{w_1} \# \text{combine}(e_1) \beta_{w_2} \# \text{combine}(e_2) \cdots \beta_{w_n} \# \text{combine}(e_n)
  \]
  frequency anchor text
e.g. serving spinal cord consequences \( \iff \) #weight(1.0 #1(spinal cord) 0.6 consequences 0.4 paralysis 1.0 serving 0.38 #combine(type of paralysis) 0.0048 #combine(quadriplegia paraplegia) 0.0048 paraplegia 0.0048 #combine(spinal cord injury) 0.0024
  #combine(quadriplegia tetrarplegia ))

Duplicated Queries
- Duplicates between a previous query and the current query, use the current query only
- Duplicates among previous queries, remove the duplicated queries

Results

<table>
<thead>
<tr>
<th>Queries</th>
<th>Without removing duplicate</th>
<th>Removing duplicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>shoulder joint pain</td>
<td>Rweight(1.4 joint 0.4 rns 1.4 pain 1.4 shoulder 0.036 #1(shoulder pain) )</td>
<td>0.55</td>
</tr>
<tr>
<td>shoulder joint pain nh</td>
<td>#1(frozen shoulder) 0.01 #1(shoulder pain causes) 0.006</td>
<td>0.64 (16.36%)</td>
</tr>
<tr>
<td>shoulder joint pain</td>
<td>#1(bursitis) 0.006 #1(painful shoulder conditions)</td>
<td></td>
</tr>
</tbody>
</table>

nDCG@10 for TREC 2012 runs

<table>
<thead>
<tr>
<th>Method</th>
<th>nDCG@10</th>
<th>%Chg</th>
<th>nDCG@10</th>
<th>%Chg</th>
<th>nDCG@10</th>
<th>%Chg</th>
<th>nDCG@10</th>
<th>%Chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL1</td>
<td>0.47</td>
<td>38.24%</td>
<td>0.47</td>
<td>38.24%</td>
<td>0.43</td>
<td>33.33%</td>
<td>0.43</td>
<td>33.33%</td>
</tr>
<tr>
<td>RL2</td>
<td>0.48</td>
<td>41.18%</td>
<td>0.45</td>
<td>32.35%</td>
<td>RL4</td>
<td>0.45</td>
<td>32.35%</td>
<td>0.45</td>
</tr>
<tr>
<td>RL3</td>
<td>0.44</td>
<td>29.41%</td>
<td>N/A</td>
<td>0.45</td>
<td>RL4</td>
<td>0.45</td>
<td>32.35%</td>
<td>0.45</td>
</tr>
<tr>
<td>RL4</td>
<td>0.46</td>
<td>35.30%</td>
<td>0.44</td>
<td>29.41%</td>
<td>0.44</td>
<td>29.41%</td>
<td>0.44</td>
<td>29.41%</td>
</tr>
</tbody>
</table>

Method
- original query: remote
- re-ranking by dwell time (RL4 only): 0.44
- remove duplicated query (RL4 only): 0.44
- nDCG@10 for TREC 2012 runs
- Best run: 0.46
- Median run: 0.44

Removal of duplicate query
- nDCG@10 for TREC 2012 runs
- Best run: 0.46
- Median run: 0.44