Modeling Rich Interactions in Session Search – Georgetown University at TREC 2014 Session Track

Jiyun Luo, Xuchu Dong and Grace Hui Yang
Department of Computer Science
Georgetown University
Introduction

• Session search
  – Document retrieval for an entire search session.
• TREC Session Track provides log data which records
  – A sequence of query changes $q_1, q_2, \ldots, q_{n-1}, q_n$
  – The ranked list for each past query
  – Document clicked information and dwell time.
• TREC 2014 Session Track:
  – RL1 using the last query of a session
  – RL2 using any information in current session
  – RL3 using information from other sessions
• We use:
  – ClueWeb12 Category A as our corpus
Outline

• Introduction
• **Methods and Approaches**
  – Ad-hoc Retrieval Model (Ad-hoc)
  – Query Change Retrieval Model (QCM)
  – Weighted QCM
  – User-Click Model
  – Clustering
  – Session Performance Prediction and Replacement
• Submissions
• Evaluation Result
• Conclusion
Ad-hoc Retrieval Model (Ad-hoc)

• Multinomial Language Modeling + Dirichlet Smoothing.

• Term weight $P(t|d)$ as:

$$P(t|d) = \frac{TF(t,d) + \mu P(t|C)}{\text{length}(d) + \mu}$$

• $\mu$ is the Dirichlet smoothing parameter, and is set = 5000.
Query Change Retrieve Model (QCM)

• **Idea:** *Query Change is an important form of user feedback*
  

• Defining query change $\Delta q_i$ as the syntactic editing changes between two adjacent queries:
  
  \[ \Delta q_i = q_i - q_{i-1} \]

• **Added term** $+\Delta q_i$; **Removed term** $-\Delta q_i$; **Theme term** $q_{theme}$

**Table 1 A example of Query Change**

<table>
<thead>
<tr>
<th>Session</th>
<th>Queries</th>
<th>Query Change</th>
<th>$Q_{theme}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>$Q_1$ = hydropower efficiency</td>
<td>$+\Delta q_2$ = environment</td>
<td>hydropower</td>
</tr>
<tr>
<td></td>
<td>$Q_2$ = hydropower environment</td>
<td>$-\Delta q_2$ = efficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$Q_3$ = hydropower damage</td>
<td>$+\Delta q_3$ = damage</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-\Delta q_3$ = environment</td>
<td></td>
</tr>
</tbody>
</table>
The relevance score between one query \( q_i \) and a document \( d \) is calculated by:

\[
\text{Score}(q_i, d) = \log P(q_i \mid d) + \alpha W_{\text{Theme}} - \beta W_{\text{Add},\text{In}} + \varepsilon W_{\text{Add},\text{Out}} - \delta W_{\text{Remove}}
\]
Query Change Retrieve Model (QCM)

• The relevance score between one query $q_i$ and a document $d$ is calculated by:

$$Score(q_i,d) = \log P(q_i \mid d) + \alpha W_{Theme} - \beta W_{Add,In} + \epsilon W_{Add,Out} - \delta W_{Remove}$$

• The QCM model combines all queries in a session with a discount factor $\gamma$:

$$Score_{qcm}(q_1..n,d) = \sum_{i=1}^{n} \gamma^{n-i} \cdot Score(q_i,d)$$
Weighted QCM

- Weighted QCM combines queries based on query quality which is indicated by user click
  - **Strong SAT-Click** a clicked document with dwelled time $\geq 30$ seconds
  - **Weak SAT-Click** a clicked document with dwell time $\geq 10$ seconds and $< 30$ seconds
Weighted QCM

• Weighted QCM combines queries based on query quality which is indicated by user click
  – **Strong SAT-Click** a clicked document with dwelled time \( \geq 30 \) seconds
  – **Weak SAT-Click** a clicked document with dwell time \( \geq 10 \) seconds and \(< 30 \) seconds

\[
Score_{wqcm}(q_{1..n}, d) = \sum_{q_i \in Q_{good}} Score(q_i, d) + \omega \sum_{q_j \in Q_{bad}} Score(q_j, d)
\]

The good query set: Queries bringing at least one SAT-Click + the current query

The bad query set: Queries bringing no SAT-Click
User-Click Model

• We boost a document’s ranking score, if it is SAT-Clicked by users
  – Session Level User-Click Model for RL2

\[
Score_{session-click}(q_{1..n}, d) = Score_{qcm}(q_{1..n}, d) + Score_{session-boost}(q_{1..n}, d)
\]

score from QCM model

boost from Session level User-Click model

\[
Score_{session-boost}(q_{1..n}, d) = \frac{\psi|StrongSATClicks_d| + \theta|WeakSATClicks_d|}{\sum_{d_i \in session}(\psi|StrongSATClicks_{d_i}| + \theta|WeakSATClicks_{d_i}|)}
\]

Ψ points for a Strong SAT-Click, θ points for a Weak SAT-Click, sum up for the whole session

normalization to (0,1)
User-Click Model

– Topic Level User-Click Model for RL3

\[
Score_{\text{cluster-click}}(q_{1..n}, d) = Score_{qcm}(q_{1..n}, d) + Score_{\text{cluster-boost}}(q_{1..n}, d)
\]

\[
Score_{\text{cluster-boost}}(q_{1..n}, d) = \frac{\psi |\text{StrongSATClicks}_d| + \theta |\text{WeakSATClicks}_d|}{\sum_{d_i \in \text{Cluster}}(\psi |\text{StrongSATClicks}_{d_i}| + \theta |\text{WeakSATClicks}_{d_i}|)}
\]

boost from **Topic** level User-Click model

similar to session level User-Click model, however calculation is done for the **whole session cluster**

• A session cluster is a set of sessions that sharing similar search topics
Clustering

• Topic ID is not obtainable in real search practice.
  – cluster sessions by comparing queries’ similarity

  ➢ Convert all queries in one session to a term vector
  ➢ Assign idf value as weight to each dimension
  ➢ Cluster sessions based on the Euclidean distance of these vectors

• We use K-means clustering algorithm and set K = 60
Session Performance Prediction and Replacement

- For sessions that share similar search topics
  - predict their performance
  - replace bad sessions’ results with good sessions’

- Predict session performance
  - Extract several features (n) from the sessions
  - Rank sessions by formula:

\[
score_e(s) = \sum_{i=1..n} \frac{1}{\text{# of sessions satisfying } F_i = \text{TRUE}} \ast I(F_i)
\]
Session Performance Prediction and Replacement

• Features Table

Table 2 Features Extracted for each Session

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>Search intent is comparison</td>
</tr>
<tr>
<td>$F_2$</td>
<td>No user-click in session $s$</td>
</tr>
<tr>
<td>$F_3$</td>
<td>$t_{dwell} \leq 5$ s. $t_{dwell}$ is the sum of dwell time in a session.</td>
</tr>
<tr>
<td>$F_4$</td>
<td># of unique terms in session $s \geq 20$.</td>
</tr>
<tr>
<td>$F_5$</td>
<td>$t_{dwell} &lt; \frac{t^{(3)}_{dwell_per_click}}{2}$</td>
</tr>
<tr>
<td>$F_6$</td>
<td>Session s does not contain the most frequent search term in $T(s)$.</td>
</tr>
<tr>
<td>$F_7$</td>
<td># of unique terms in session $s \leq 6$</td>
</tr>
<tr>
<td>$F_8$</td>
<td># of SAT clicks in session $s &lt; \frac{\sum_{s' \in T(s)} # of SAT clicks in session s'}{</td>
</tr>
</tbody>
</table>

* $T(s)$ means a session cluster including session $s$
Outline

• Introduction

• Methods and Approaches
  – Ad-hoc Retrieval Model (Ad-hoc)
  – Query Change Retrieval Model (QCM)
  – Weighted QCM
  – User-Click Model
  – Clustering
    – Session Performance Prediction and Replacement

• Submissions

• Evaluation Result

• Conclusion
## Our Submissions

<table>
<thead>
<tr>
<th>RL1</th>
<th>Ad-hoc Retrieval Model</th>
</tr>
</thead>
</table>
| RL2  | • Weighted QCM ($\omega=0.65$)  
      • Session Level User-Click Model |
| RL3  | • Weighted QCM ($\omega=0.65$)  
      • Topic Level User-Click Model |
|      | • Weighted QCM ($\omega=0.8$)  
      • Topic Level User-Click Model  
      • Topic Level User-Click Model  
      • Weighted QCM ($\omega=0.8$)  
      • Topic Level User-Click Model using topic ids  
      • Session Performance Prediction and Replacement |

- RL3 in RUN1 and RUN2 using session clusters based on query similarity
- RL3 in RUN3 using session cluster based on topic id
  - Why? similar queries leads to similar retrieval list in our system. Not useful when apply session replacement strategy
## Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>GUS14RUN1</th>
<th>GUS14RUN2</th>
<th>GUS14RUN3</th>
<th>Max</th>
<th>Med</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>nDCG@10</td>
<td>P@10</td>
<td>nDCG@10</td>
<td>P@10</td>
<td>nDCG@10</td>
</tr>
<tr>
<td>RL1</td>
<td>0.2053</td>
<td>0.378</td>
<td>0.2053</td>
<td>0.378</td>
<td>0.2053</td>
</tr>
<tr>
<td>RL2</td>
<td>0.2458</td>
<td>0.426</td>
<td>0.2482</td>
<td>0.427</td>
<td>0.2482</td>
</tr>
<tr>
<td>RL3</td>
<td>0.2443</td>
<td>0.423</td>
<td>0.2458</td>
<td>0.424</td>
<td>0.2580</td>
</tr>
</tbody>
</table>

- **2\textsuperscript{nd}** rank in task RL1, **1\textsuperscript{st}** rank in task RL2 and RL3
  - Adjusting term weight based on query change is effective
  - Combining queries in a session is useful for Session Track
  - User-Click is effective to predicate relevance

- A small performance drop from RL2 to RL3 in RUN1 and RUN2
  - cluster sessions based on query similarity may work, however need more work to refine it

- A small increase from RL2 to RL3 in RUN3
  - For sessions sharing same search topics, replacing poor sessions’ results using good sessions’ is practical.
Conclusion

• Achieve 20.9% increase from RL1 to RL2 by utilizing
  – query change feedback
  – user click feedback

• Achieve 4% increase from RL2 to RL3 by
  – Topic level User-Click Model
  – Session performance prediction and replacement
Thanks!

Jiyun Luo, Xuchu Dong and Grace Hui Yang
Department of Computer Science
Georgetown University