Overview of the NTCIR-14 OpenLiveQ-2 Task

Makoto P. Kato (University of Tsukuba), Takehiro Yamamoto (University of Hyogo), Sumio Fujita, Akiomi Nishida, Tomohiro Manabe (Yahoo Japan Corporation)
• Task Design (4 slides)
• Data (5 slides)
• Evaluation Methodology (11 slides)
• Evaluation Results (4 slides)
Goal

Improve the **REAL performance** of question retrieval systems in a production environment

Performance evaluated by **REAL users**

**Yahoo! Chiebukuro** (a CQA service of Yahoo! Japan)
Task

- Given a query, return a ranked list of questions
  - Must satisfy many REAL users in Yahoo! Chiebukuro (a CQA service)

**Effective for Fever**

**Three things you should not do in fever**
While you can easily handle most fevers at home, you should call 911 immediately if you also have severe dehydration with blue .... Do not blow your nose too hard, as the pressure can give you an earache on top of the cold. ....

10 Answers  Posted on Jun 10, 2016

**Effective methods for fever**
Apply the mixture under the sole of each foot, wrap each foot with plastic, and keep on for the night. Olive oil and garlic are both wonderful home remedies for fever. 10) For a high fever, soak 25 raisins in half a cup of water.

2 Answers  Posted on Jan 3, 2010
OpenLiveQ provides an OPEN LIVE TEST ENVIRONMENT

Ranked lists of questions from participants’ systems are **INTERLEAVED**, presented to real users, and evaluated by their clicks
• **Differences**
  – A new document (question) collection
  – New clickthrough data
  – New online evaluation techniques

• **While we kept**
  – The task design
  – The topic set
  – The relevance judgments
  – The offline evaluation methodology
Data at OpenLiveQ-2

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Queries*</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Documents (or questions)</td>
<td>986,125</td>
<td>985,691</td>
</tr>
<tr>
<td>Clickthrough data</td>
<td>Data collected for 3 months</td>
<td>Data collected for 3 months</td>
</tr>
<tr>
<td>Relevance judges*</td>
<td>N/A</td>
<td>For 100 queries</td>
</tr>
</tbody>
</table>

The **second** Japanese dataset for learning to rank
(to the best of our knowledge)
(* indicates “the same as that in OpenLiveQ-1”)

Do you know the first one?
## Data at OpenLiveQ-1

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Queries</strong></td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td><strong>Documents</strong> (or questions)</td>
<td>984,576</td>
<td>982,698</td>
</tr>
<tr>
<td><strong>Clickthrough data</strong></td>
<td>Data collected for 3 months</td>
<td>Data collected for 3 months</td>
</tr>
<tr>
<td><strong>Relevance judges</strong></td>
<td>N/A</td>
<td>For 100 queries</td>
</tr>
</tbody>
</table>

The first Japanese dataset for learning to rank (to the best of our knowledge)
2,000 queries sampled from a query log

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Query</th>
<th>English Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLQ-0001</td>
<td>バイオハザード</td>
<td>Bio Hazard</td>
</tr>
<tr>
<td>OLQ-0002</td>
<td>チベット</td>
<td>Tibet</td>
</tr>
<tr>
<td>OLQ-0003</td>
<td>ぶどう</td>
<td>Grape</td>
</tr>
<tr>
<td>OLQ-0004</td>
<td>プリウス</td>
<td>Prius</td>
</tr>
<tr>
<td>OLQ-0005</td>
<td>twice</td>
<td>twice</td>
</tr>
<tr>
<td>OLQ-0006</td>
<td>割り勘</td>
<td>separate checks</td>
</tr>
<tr>
<td>OLQ-0007</td>
<td>gta5</td>
<td>gta5</td>
</tr>
</tbody>
</table>

Filtered out
- Time-sensitive queries
- X-rated queries
- Related to any of the ethic, discrimination, or privacy issues
<table>
<thead>
<tr>
<th>Query ID</th>
<th>Rank</th>
<th>Question ID</th>
<th>Title</th>
<th>Snippet</th>
<th>Status</th>
<th>Timestamp</th>
<th># answers</th>
<th># views</th>
<th>Category</th>
<th>Body</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLQ-0001</td>
<td>1</td>
<td>q13166161098</td>
<td>バイオハザードって設定に無い…</td>
<td>にゃいけどピアノ弾いたことない人は楽…</td>
<td>Solved</td>
<td>2016/11/13 3:35</td>
<td>1</td>
<td>42</td>
<td>エンターテイメントと趣味＞ゲーム</td>
<td>バイオハザードって…レベッカもジルも弾けなかった場合</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>2</td>
<td>q14166076254</td>
<td>バイオハザードって設定にむ…</td>
<td>タックルしたり足で蹴りまくればこわせる…</td>
<td>Solved</td>
<td>2016/11/10 3:47</td>
<td>1</td>
<td>18</td>
<td>エンターテイメントと趣味＞ゲーム</td>
<td>バイオハザードって…なので、バイオハザードアウトプレイシリーズ…</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>3</td>
<td>q11166238681</td>
<td>バイオハザードのゲームの…</td>
<td>バイオハザード4が好きで30週くらいして…</td>
<td>Solved</td>
<td>2016/11/21 3:29</td>
<td>3</td>
<td>19</td>
<td>エンターテイメントと趣味＞ゲーム</td>
<td>バイオハザードのゲーム…</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLQ-2000</td>
<td>998</td>
<td>q11137434581</td>
<td>夫婦生活で一番疲れる…</td>
<td>が自分に隠していることがあった…</td>
<td>Solved</td>
<td>2014/10/28 15:14</td>
<td>6</td>
<td>0</td>
<td>生き方と恋愛、人間関係の悩み＞…</td>
<td>夫婦生活で一番疲れる事は何ですか…自分が相手に…</td>
</tr>
<tr>
<td>OLQ-2000</td>
<td>999</td>
<td>q1292632642</td>
<td>夫婦生活について教えて下さい…</td>
<td>主人とセックスをしておらず二年半に…</td>
<td>Solved</td>
<td>2012/9/3 9:51</td>
<td>5</td>
<td>701</td>
<td>生き方と恋愛、人間関係の悩み＞…</td>
<td>夫婦生活について教えて下さい。…</td>
</tr>
<tr>
<td>OLQ-2000</td>
<td>1000</td>
<td>q1097950260</td>
<td>旦那との今後の夫婦生活、旦那…</td>
<td>結婚して来年の1月で2年になります…</td>
<td>Solved</td>
<td>2012/12/5 10:01</td>
<td>4</td>
<td>640</td>
<td>生き方と恋愛、人間関係の悩み＞…</td>
<td>旦那との今後の夫婦生活、…</td>
</tr>
</tbody>
</table>
## Clickthrough Data

<table>
<thead>
<tr>
<th>Query ID</th>
<th>Question ID</th>
<th>Rank</th>
<th>CTR</th>
<th>Male</th>
<th>Female</th>
<th>0s</th>
<th>10s</th>
<th>20s</th>
<th>30s</th>
<th>40s</th>
<th>50s</th>
<th>60s</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLQ-0001</td>
<td>q10165187300</td>
<td>1</td>
<td>0.059</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q11164148731</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q11166231691</td>
<td>1</td>
<td>0.023</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q11166372256</td>
<td>1</td>
<td>0.036</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q13161212253</td>
<td>1</td>
<td>0.051</td>
<td>0.909</td>
<td>0.091</td>
<td>0</td>
<td>0.091</td>
<td>0.364</td>
<td>0.182</td>
<td>0.182</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q13166161098</td>
<td>1</td>
<td>0.021</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q14164350104</td>
<td>1</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q14164384744</td>
<td>1</td>
<td>0.188</td>
<td>0.909</td>
<td>0.091</td>
<td>0</td>
<td>0.091</td>
<td>0.364</td>
<td>0.182</td>
<td>0.182</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q14165651359</td>
<td>1</td>
<td>0.048</td>
<td>0.909</td>
<td>0.091</td>
<td>0</td>
<td>0.091</td>
<td>0.364</td>
<td>0.182</td>
<td>0.182</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q11166278091</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q11166476886</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q12164569302</td>
<td>2</td>
<td>0.037</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q12165573687</td>
<td>2</td>
<td>0.083</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q10162841855</td>
<td>3</td>
<td>0.036</td>
<td>0.909</td>
<td>0.091</td>
<td>0</td>
<td>0.091</td>
<td>0.364</td>
<td>0.182</td>
<td>0.182</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q12164050757</td>
<td>3</td>
<td>0.06</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q12164687517</td>
<td>3</td>
<td>0.049</td>
<td>0.909</td>
<td>0.091</td>
<td>0</td>
<td>0.091</td>
<td>0.364</td>
<td>0.182</td>
<td>0.182</td>
<td>0.091</td>
<td>0.091</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q12165837862</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q14158395769</td>
<td>3</td>
<td>0.027</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Evaluation Methodology

- **Offline evaluation** (July 25, 2018 – Sep 15, 2018)
  - Evaluation with relevance judgment data
    - Similar to that for a traditional ad-hoc retrieval tasks

- **Online evaluation** (Sep 28, 2018 - Jan 6, 2019)
  - Evaluation with real users
    - All the systems were evaluated online
    - Background
      Only the best run from each team in the offline evaluation was invited to the online evaluation at OpenLiveQ-1.
      This wasn’t so good. They do not always agree!
Offline Evaluation

• **Relevance judgments**
  – Crowd-sourcing workers report all the questions on which they want to click

• **Evaluation Metrics**
  – **Q-measure (primary measure)**
    • A kind of MAP for graded relevance
  – **nDCG** (normalized discounted cumulative gain)
    • Ordinary metrics for Web search
  – **ERR** (expected reciprocal rank)
    • Users stop the traverse when satisfied

• **Accept submission once per day via CUI**
5 assessors were assigned for each

- Relevance ≡ # assessors who want to click
Submission

- Submission by CUI
  
curl http://www.openliveq.net/runs -X POST
  > -H "Authorization:KUIDL:ZUEE92xxLAKL1WX2Lxqy"
  > -F run_file=@data/your_run.tsv

- Leader Board (anyone can see the performance of participants)

  - 65 submissions from 5 teams
Participants

- AITOK  Tokushima University
- YJRS  Yahoo Japan Corporation
- OKSAT  Osaka Kyoiku University
- DCU-ADAPT  Dublin City University
- ORG  Organizers
• AITOK achieved the best performances among five teams
• A concern about overfitting on test queries
Online Evaluation

- Multileaved comparison methods are used in the online evaluation
  - Schuth, Sietsma, Whiteson, Lefortier, de Rijke: Multileaved comparisons for fast online evaluation, CIKM2014.

- **Pairwise Preference Multileaving (PPM) was used**
  - SOTA in interleaved comparison

- Sep 28, 2018 - Jan 6, 2019 (~ 3 months)
  - # impressions: 313,454
    - **NOTE:** we did not use all the impressions at Yahoo Chiebukuro for this evaluation
Interleaving: an alternative to A/B testing

- Evaluation based on user feedback on the ranking generated by interleaving multiple rankings

- 10-100 times as efficient as A/B testing

- Multileaving = Interleaving for 3+ rankings
Given multiple rankings $\mathcal{R}$, PPM generates \textbf{interleaved rankings} such that
- A document at $k$-th rank is selected from documents at $1, \ldots, k$-th rank in $\mathcal{R}$
- A document can be selected only once

\textbf{Example of Ranking $\alpha$}
- Rank 1: 1 $\sim \{1, 4\}$, Rank 2: 4 $\sim \{2, 4, 5\}$, Rank 3: 3 $\sim \{2, 3, 5, 6\}$
• Given a query from a user, an interleaved ranking is selected randomly and presented to the user
• Observe his/her clicks on the interleaved ranking
A ranking receives a positive score if it agrees with pairwise prefs. indicated by the clicks.

A negative score is given as the ranking disagrees with the prefs.

- A positive score is given as the ranking agrees with the prefs.
Hard to find statistically significant differences with 65 rankings (or 2,080 pairs)

**Two-phase Strategy**

1. **Identifying top-k rankings with a half of impressions**
   - 164,478 impressions were allocated to find top-30 rankings
2. **Comparing only the top-k rankings with the rest of impressions**
   - 148,976 impressions were allocated to find differences among the top-30 rankings

Online Evaluation Result

- **Blue bar**: the cumulated score at the 1st phase
- **Red bar**: the cumulated score at the 2nd phase
- Runs are sorted by that at the 1st phase

Baseline
(the current ranking)
• **Quite different from the offline evaluation results**
  – Confirmed the importance of evaluating all the runs online
• **YJRS** achieved the best performance, while no sig. diff. from the top eight runs
Progress from OpenLiveQ-1

- The differences were reproduced
  - Should have submitted a paper to CENTRE?
- The top performer in OpenLiveQ-1 also did a good job in OpenLiveQ-2
Conclusions

• **OpenLiveQ brought online evaluation into NTCIR**
  – Real needs, real users, and real clicks

• **The 1\textsuperscript{st} and 2\textsuperscript{nd} Japanese datasets for learning to rank**
  – With demographics of searchers

• **Evaluation results showed**
  – A large difference between offline and online evaluation
  – The performance of the two-phase strategy for interleaving
  – Some results in OpenLiveQ-1 were reproduced in OpenLiveQ-2