Overview of the NTCIR-13 OpenLiveQ Task

Makoto P. Kato, Takehiro Yamamoto (Kyoto University), Sumio Fujita, Akiomi Nishida, Tomohiro Manabe (Yahoo Japan Corporation)
Agenda

• Task Design (3 slides)
• Data (5 slides)
• Evaluation Methodology (12 slides)
• Evaluation Results (6 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)
• 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 EVIRONMENT

Ranked lists of questions from participants’ systems are INTERLEAVED, presented to real users, and evaluated by their clicks
The first Japanese dataset for learning to rank
(to the best of our knowledge)
(basic features also available, i.e. language-independent)

<table>
<thead>
<tr>
<th>Data</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>984,576</td>
<td>982,698</td>
</tr>
<tr>
<td>Clickthrough data (with user demographics*)</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>
• **2,000 queries sampled from a query log**

<table>
<thead>
<tr>
<th>OLQ-0001</th>
<th>OLQ-0002</th>
<th>OLQ-0003</th>
<th>OLQ-0004</th>
<th>OLQ-0005</th>
<th>OLQ-0006</th>
<th>OLQ-0007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bio Hazard</td>
<td>Tibet</td>
<td>Grape</td>
<td>Prius</td>
<td>twice</td>
<td>separate checks</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>
<th>Best answer</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…</td>
<td>弾いたけどピアノ弾いたこ…</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…</td>
<td>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>タックルしたたり足で蹴りま…</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…</td>
<td>なので、バイオハザードアウトプレイヤーのシリーズ…</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>
<td>個人的には… BIOHAZARD REVELATION S UNVEILED EDITION …</td>
</tr>
<tr>
<td>OLQ-2000</td>
<td>998</td>
<td>q11137434581</td>
<td>夫婦生活で一番疲れると…</td>
<td>自分が相手に隠していること…</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…</td>
<td>自分が相手に隠していることがあって…</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…</td>
<td>Solved</td>
<td>2012/9/3 9:51</td>
<td>5701</td>
<td>0</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…</td>
<td>どんなこと…</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…</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>4640</td>
<td>0</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>
<td></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>0</td>
<td>1</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>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q14165651359</td>
<td>1</td>
<td>0.048</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>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>1</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</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>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>OLQ-0001</td>
<td>q12164687517</td>
<td>3</td>
<td>0.049</td>
<td>0.833</td>
<td>0.167</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.167</td>
<td>0.667</td>
<td>0.167</td>
<td>0</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>
• **The current ranking of Yahoo CQA**
  – Outperforming this baseline may indicate room for providing better services for users

• **Several learning to rank (L2R) baselines**
  – Features
    • Features listed in Tao Qin, Tie-Yan Liu, Jun Xu, Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval, Information Retrieval, Volume 13, Issue 4, pp. 346-374, 2010. + # answers + # views
  – Algorithm: a linear feature-based model
Evaluation Methodology

• **Offline evaluation** (Feb 2017 – Apr 2017)
  – Evaluation with relevance judgment data
    • Similar to that for a traditional ad-hoc retrieval tasks

• **Online evaluation** (May 2017 – Aug 2017)
  – Evaluation with real users
    • 10 systems were selected by the results of the offline test
• **Relevance judgments**
  – Crowd-sourcing workers report all the questions on which they want to click

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

• **Accept submission once per day via CUI**
• **5** assessors were assigned for each
  – **Relevance** ≡ # assessors who want to click

---

**Search query: grape**

あなたは今Yahoo!知恵袋を訪れており、上記のキーワードで検索を行ったとします。
以下の質問の全てに目を通して、「あなたがクリックしたいと思う質問を全て選んでください。」

タイトルをクリックしても色のついた四角い部分をクリックしても選択できます。

----

ぶどうの甘みは房子か先の方どっちが甘いのかね？

ぶどうはイチゴと逆で上の方が甘いです。
解決済み - 更新日時: 2015/10/21 21:26:34 - 回答数：1 - 閲覧数：3
地域、旅行、お出かけ > 国内 > 季節のおでかけ

プドウ糖についてお尋ねします。調剤薬局で貰ったぶどう糖を親しい人から貰...

貰いました。「疲れたときに良い」とのことでしたのでウォーキングの時や疲れた時などに二杯ほど飲みていました。
昨日、調剤薬局に行ったのでプドウ糖のことを聞くと「糖尿の人には良いですが・・・」と言われ...
解決済み - 更新日時: 2015/04/03 18:53:24 - 回答数：1 - 閲覧数：20
健康、美容とファッション > 健康、病気、病院 > 病気、症状

ぶどうとお酒と法律についてぶどう酒を作るのは法律で禁止というのを見たの...
Submission

• Submission by CUI

```bash
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)

- 85 submissions from 7 teams
Participants

- **YJRS**: additional features and weight optimization
- **Erler**: Topic inference based Translation Language Model
- **SLOLQ**: A neural network based document model + similarity and diversity-based rankings
- **TUA1**: Random Forests
- **OKSAT**: integration of careful designed features
Offline Evaluation Results

nDCG@10

Best baseline

Yahoo

ERR@10

Best baseline

Q

Best baseline
Similar results. The top performers are OKSAT, cdlab, and YJRS
Different results. The top performers are YJRS and Erler

Turned out to be more consistent with the online evaluation
• Multileaved comparison methods are used in the online evaluation

• Optimized multileaving (OM) was used
  – OM is one of the interleaving methods for evaluating multiple rankings
  – Found the best in our experiments:
    Manabe et al. A Comparative Live Evaluation of Multileaving Methods on a Commercial cQA Search, SIGIR 2017

• May 2017 - August 2017 (~ 3 months)
  – # impressions: 410,812
A Comparative Live Evaluation of Multileave Methods on a Commercial cQA Search
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

- Multileaveling = Interleaving for $3 \geq$ rankings
Intuitive Explanation of Optimized Multileaving (OM)

- Interleaved rankings are shown to users with prob. $P(\alpha)$, $P(\beta)$, and $P(\gamma)$, respectively.
- Give a credit $1/k$ to each ranking if its document at rank $k$ is clicked.
- Evaluate rankers by the cumulated credits.
Bias in Interleaving

Interleaved rankings

- If $P(\alpha) = P(\beta) = P(\gamma) = 1/3$, it is likely that Ranking A > Ranking B
  - As top-ranked docs are more likely to be clicked
- OM optimizes the presentation probability to minimize this bias
  - More precisely, OM minimizes the difference of expected cumulated credits of rankers for rank-biased random clicks
Forcing Comparison of Rankings

Interleaved rankings

<table>
<thead>
<tr>
<th>Rank 1</th>
<th>Rank 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ranking α</strong></td>
<td><strong>Ranking β</strong></td>
</tr>
<tr>
<td>ID: 1</td>
<td>ID: 4</td>
</tr>
<tr>
<td>ID: 4</td>
<td>ID: 1</td>
</tr>
<tr>
<td><strong>Ranking γ</strong></td>
<td><strong>Ranking δ</strong></td>
</tr>
<tr>
<td>ID: 1</td>
<td>ID: 4</td>
</tr>
<tr>
<td>ID: 2</td>
<td>ID: 5</td>
</tr>
</tbody>
</table>

- (i) \( P(\alpha) = P(\beta) = 1/2, P(\gamma) = P(\delta) = 0 \), or
- (ii) \( P(\alpha) = P(\beta) = 0, P(\gamma) = P(\delta) = 1/2 \), can result in zero bias
  - But (ii) never force the user to compare documents from different rankings
  → Less chances to know the difference

- OM optimizes the presentation probability to maximize the chance of comparison
(Modified) Optimized Multileaving [Manabe et al., SIGIR2017]

The chance of comparison
= expected variance of cumulated credits for each ranker

\[
\min_{p_k} \left\{ \alpha \sum_{r=1}^{\lambda_r} + \sum_{k=1}^{\sigma_k^2} \right\}
\]

\[
\forall r, \forall j, j' - \lambda_r \leq \mathbb{E}[g(I_j, r)] - \mathbb{E}[g(I_{j'}, r)] \leq \lambda_r
\]

Bias = the difference of expected cumulated credits of rankers

• Slightly different from Schuth et al.’s
  – Theirs sometimes fails due to “no solution”
Erler and YJRS outperformed the best baseline (no sig. dif.)
How many days were necessary to find sig. dif. for X% of run pairs (with Bonferroni correction)

- 10 days: sig. dif. found for 82.2%
- 20 days: sig. dif. found for 91.1%
- 64> days: sig. dif. found for 93.3%
1. Some differences from the offline evaluation
   - Offline: OKSAT > cdlab ≒ YJRS > Erler
   - Online: Erler ≒ YJRS > cdlab > OKSAT

2. YJRS and Erler outperformed the best baseline in the online evaluation
   - Still room for improvement?

3. All the runs outperformed the current ranking
   - The current state-of-the-arts can improve the quality (or CTR) of the existing service
Conclusions

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

• **The first Japanese dataset for learning to rank**
  – With demographics of searchers

• **Demonstrated the capability of interleaving methods**

• **Discussions**
  – Which should we rely on, offline or online? (Especially when they are different)
  – Lack of reproducibility
NTCIR-14

OpenLiveQ-2

All Runs Are Evaluated Online

Makoto P. Kato, Takehiro Yamamoto (Kyoto University), Sumio Fujita, Akiomi Nishida, Tomohiro Manabe (Yahoo Japan Corporation)