Overview of the NTCIR-14 OpenLiveQ-2 Task

Makoto P. Kato, Takehiro Yamamoto (Kyoto University), Sumio Fujita, Akiomi Nishida, Tomohiro Manabe (Yahoo Japan Corporation)
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 that can satisfy many REAL users in Yahoo! Chiebukuro (a CQA service)

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

Real users

Ranked lists of questions from participants’ systems are **INTERLEAVED**, presented to real users, and evaluated by their clicks.
Multileaving-based Evaluation

- Multileaved comparison methods are used in the online evaluation

Therefore, submitted runs are evaluated by REAL users’ clicks
<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>~1M</td>
<td>~1M</td>
</tr>
<tr>
<td><strong>Clickthrough data</strong> (with user demographics*)</td>
<td>Data collected for 3 months</td>
<td>Data collected for 3 months</td>
</tr>
</tbody>
</table>

The SECOND Japanese dataset for learning to rank (to the best of our knowledge) (basic features also available, i.e. language-independent)
In NTCIR-13 OpenLiveQ-1, only 10/83 runs were evaluated online. Afraid that each run cannot receive enough number of clicks.

OpenLiveQ-2 will use an interleaving method that selectively finds differences among runs. Guarantee significant differences between teams, while do not guarantee those within teams.

Most Runs Are Evaluated Online
Our Expectation: Further Improvement over Current State-of-the-arts

Yahoo

Best baseline

Credit

Baseline (as is)
SLOQ
OKSAT
KUIDL
TUA1
Baseline (#Ans)
cdlab
ORG
YJS
Erlc

Your runs
<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 2018</td>
<td>Data set release</td>
</tr>
<tr>
<td>May-Aug 2018</td>
<td>Offline evaluation phase</td>
</tr>
<tr>
<td>Aug 31, 2018</td>
<td>Run submission and Registration due</td>
</tr>
<tr>
<td>Sep-Nov 2018</td>
<td>Online evaluation phase</td>
</tr>
<tr>
<td>Jan 10, 2018</td>
<td>Evaluation result release</td>
</tr>
</tbody>
</table>

Looking forward to your participation. Thanks!