Overview of the NTCIR-13 OpenLiveQ Task

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

Task

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

Effective for Fever — **INPUT** — ✕ | Q&A | 🔍 Search

OUTPUT

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

Data

	Training	Testing
Queries	1,000	1,000
Documents (or questions)	÷ 1,000,000	÷ 1,000,000
Clickthrough data (with user demographics*)	Data collected for 3 months	Data collected for 3 months
Relevance judges	N/A	For 100 queries

* User demographics: Gender and age of search users

The first Japanese dataset for learning to rank

Evaluation Methodology

Offline Evaluation
DCG, ERR, and Q-measure were used with questions judged by crowd-sourcing workers

Online Evaluation
The best run from each team in the offline evaluation was invited to the online evaluation

Multileaving was used in the online evaluation: ranked lists of questions from participants' systems are **merged**, presented to real users, and evaluated by their clicks
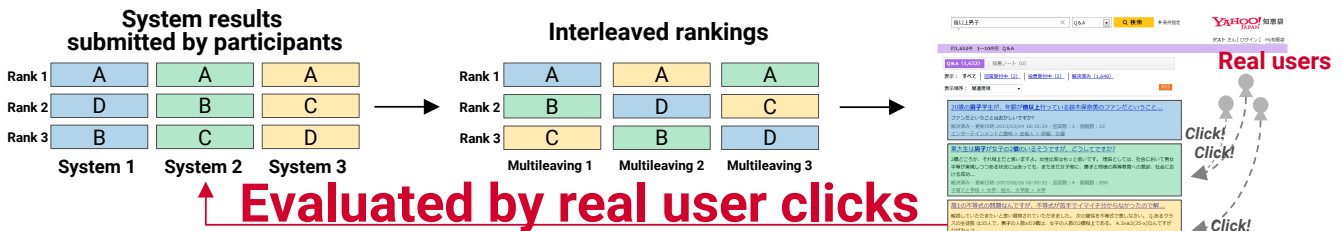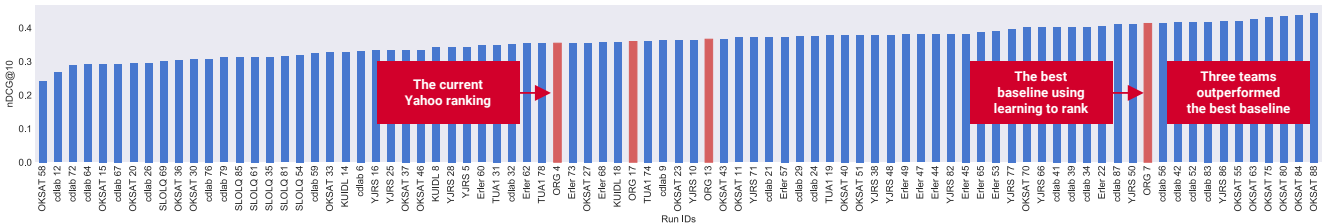

Evaluated by real user clicks

Optimized Multileaving **was employed based on our SIGIR paper:**
T. Manabe et al. (2017) A Comparative Live Evaluation of Multileaving Methods on a Commercial cQA Search, SIGIR 2017.

Evaluation Results

Offline Evaluation (nDCG@10)

The current Yahoo ranking

The best baseline using learning to rank

Three teams outperformed the best baseline

Online Evaluation

The online evaluation at Yahoo! Chiebukuro was conducted between May 9, 2017 and August 8, 2017 **(approx. 100 days)**. The total number of impressions used for the online evaluation is **410,812**.

1. Some differences from the offline evaluation

2. YJRS and Erler outperformed the best baseline (no sig. dif.)

3. All the runs outperformed the current ranking

10 days: sig. dif. found for 82.2%

20 days: sig. dif. found for 91.1%

64> days: sig. dif. found for 93.3%

How many days were necessary to find sig. dif. for X% of run pairs

Interested? Join us at NTCIR-14! Follow @NTCIROpenLiveQ and visit http://www.openliveq.net/