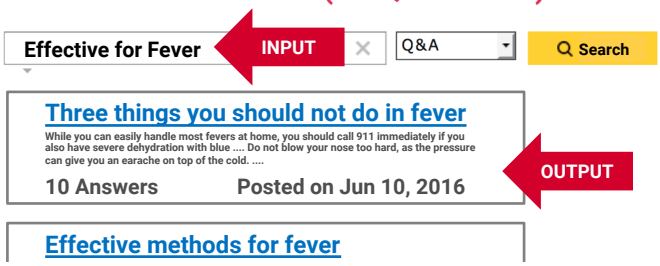
Overview of the NTCIR-14 OpenLiveQ-2 Task

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

Task

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



Data

	Training	Testing
Queries	1,000	1,000
Documents (or questions)	986,125	985,691
Clickthrough data (with user demographics)	Data collected for 3 months	Data collected for 3 months
Relevance judges	N/A	For 100 queries

The second Japanese dataset for learning to rank

The first one? It's the OpenLiveQ-1 dataset!

Evaluation Methodology

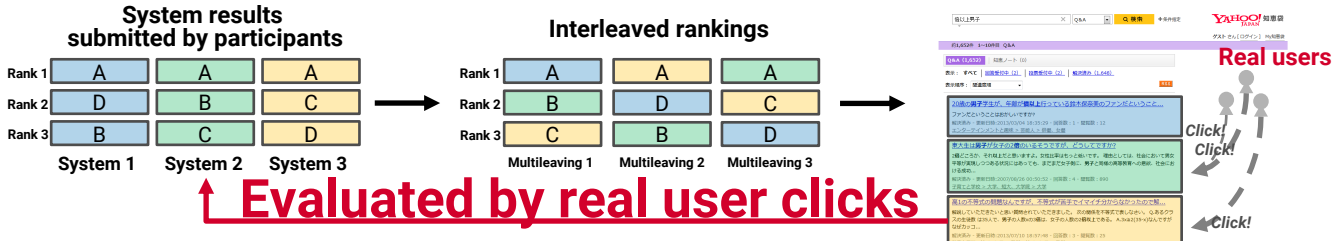
Offline Evaluation

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

Online Evaluation

Unlike OpenLiveQ-1, all the runs were evaluated online with the two-phase strategy (see below)

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

***Pairwise Preference Multileaving (PPM)* was used**

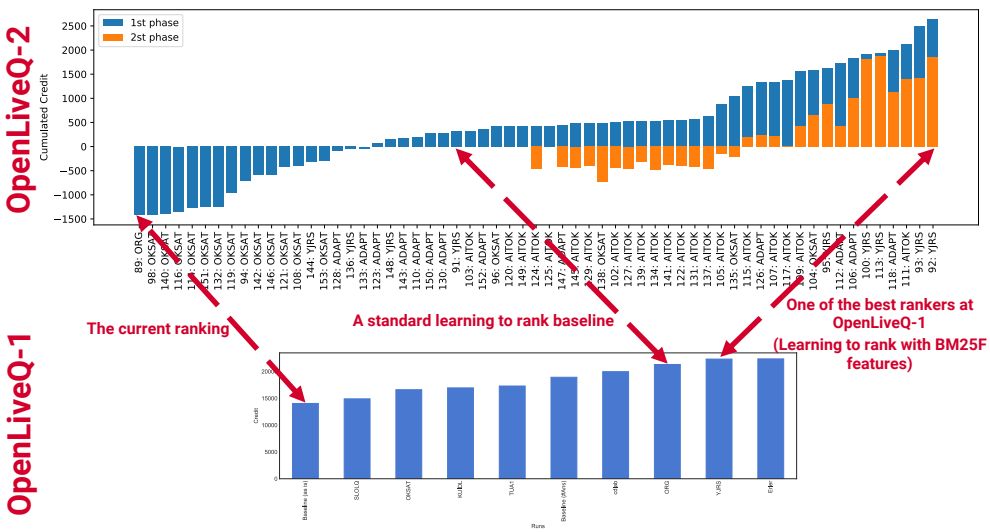
Oosterhuis,de Rijke :Sensitive and Scalable Online Evaluation with Theoretical Guarantees. In: CIKM. pp. 77–86 (2017)

To deal with a relatively large number of runs, we employed the two-phase strategy proposed in our recent work.

(Kato et al. Challenges of Multileaved Comparison in Practice: Lessons from NTCIR-13 OpenLiveQ Task, CIKM 2018)

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

Evaluation Results



Findings

- **The top performer in OpenLiveQ-1 also worked well in OpenLiveQ-2**
- **The differences of some ranker pairs were reproduced**
- **Quite different from the offline evaluation results (Confirmed the importance of evaluating all the runs online)**
- **Pairwise preferences at the 1st and 2nd phases are slightly different**