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)





Agenda

- **Task Design** (4 slides)
- **Data** (5 slides)
- **Evaluation Methodology** (11 slides)
- **Evaluation Results** (4 slides)

Goal

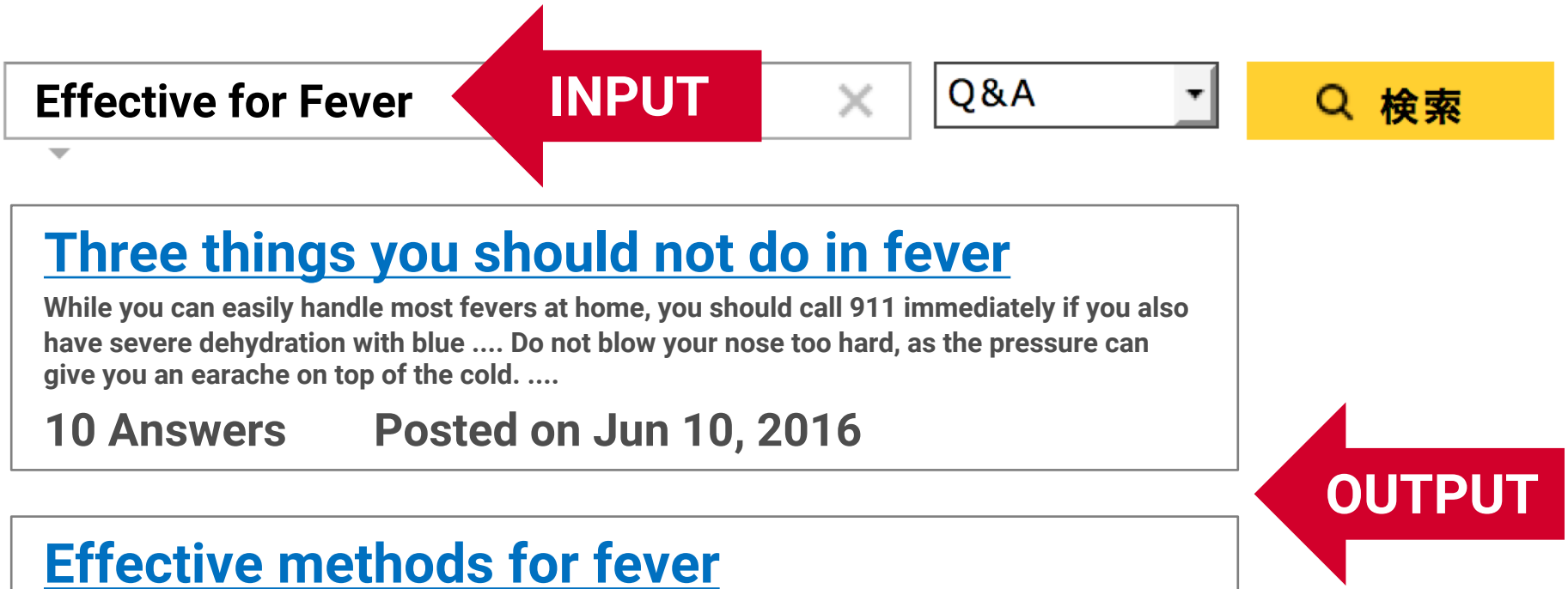
Performance evaluated by REAL users

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

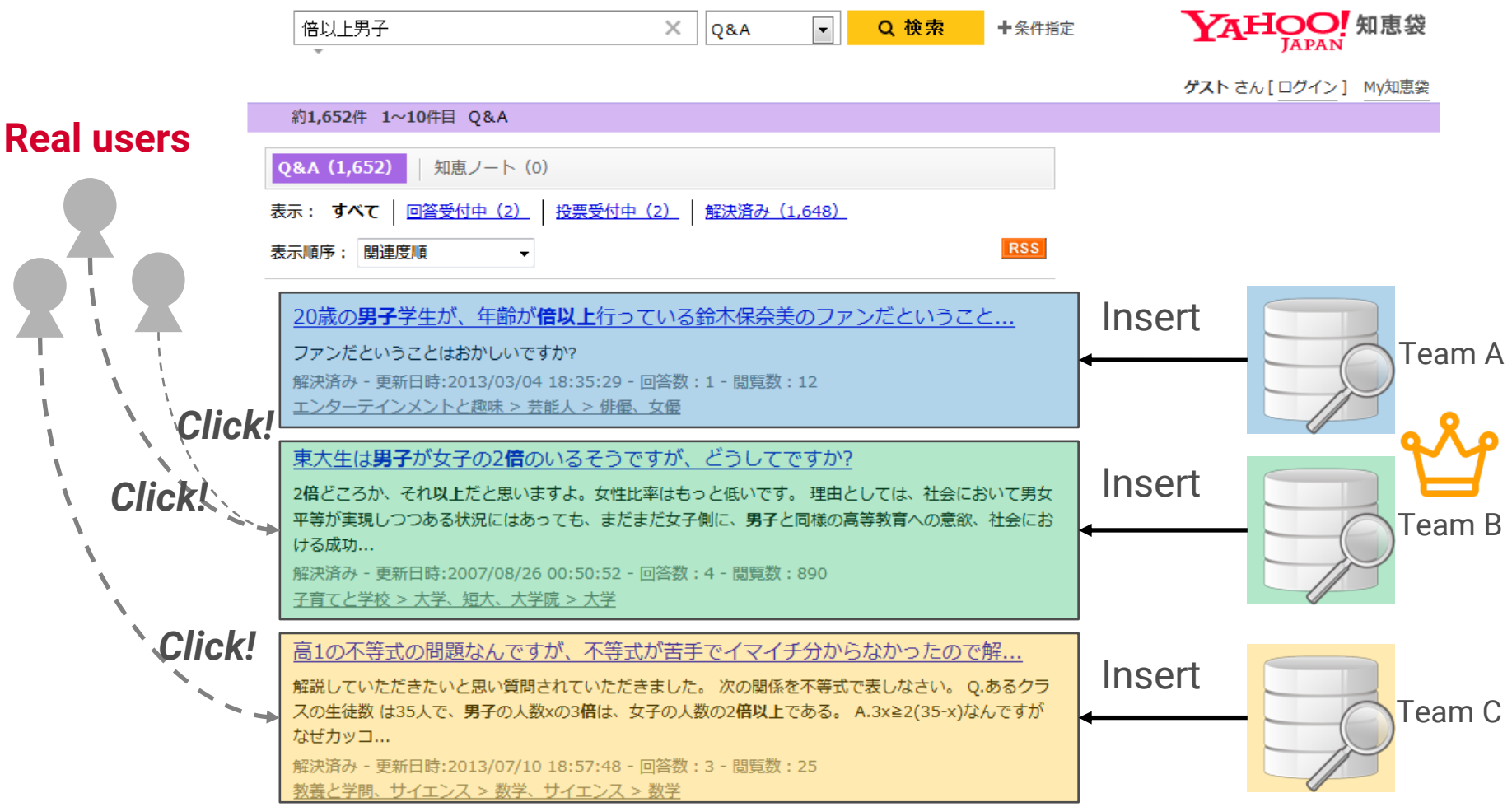
Yahoo! Chiebukuro
(a CQA service of Yahoo! Japan)

Task

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



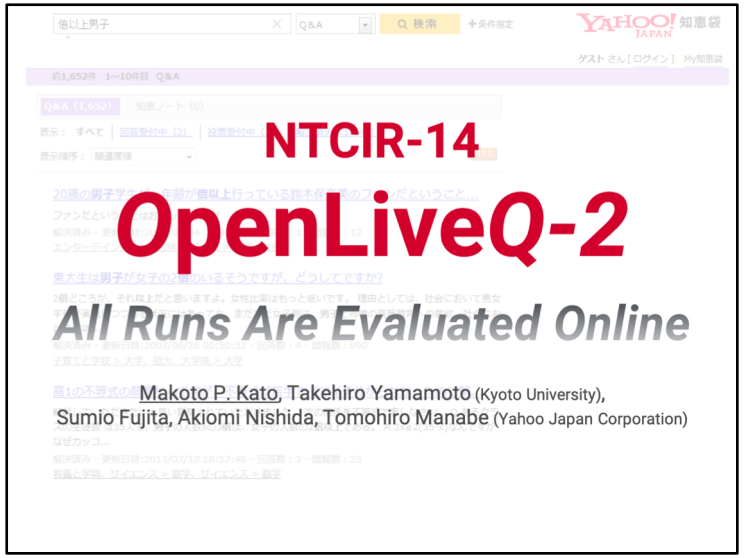
OpenLiveQ provides an OPEN LIVE TEST ENVIRONMENT



Ranked lists of questions from participants' systems are INTERLEAVED, presented to real users, and evaluated by their clicks

Differences from NTCIR-13 OpenLiveQ-1

- **Differences**
  - A new document (question) collection
  - New clickthrough data
  - New online evaluation techniques



A slide used at the NTCIR-13 conf.

- **While we kept**
  - The task design
  - The topic set
  - The relevance judgments
  - The offline evaluation methodology

Data at OpenLiveQ-2

	Training	Testing
Queries*	1,000	1,000
Documents (or questions)	986,125	985,691
Clickthrough data	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

(to the best of our knowledge)

(* indicates "the same as that in OpenLiveQ-1")

Do you know the first one?

Data at OpenLiveQ-1

	Training	Testing
Queries	1,000	1,000
Documents (or questions)	984,576	982,698
Clickthrough data	Data collected for 3 months	Data collected for 3 months
Relevance judges	N/A	For 100 queries

The first Japanese dataset for learning to rank
(to the best of our knowledge)

Queries

- **2,000 queries sampled from a query log**

OLQ-0001	バイオハザード	Bio Hazard
OLQ-0002	チベット	Tibet
OLQ-0003	ぶどう	Grape
OLQ-0004	プリウス	Prius
OLQ-0005	twice	twice
OLQ-0006	割り勘	separate checks
OLQ-0007	gta5	gta5

- **Filtered out**
  - Time-sensitive queries
  - X-rated queries
  - Related to any of the ethic, discrimination, or privacy issues

Questions

answers & # views

Query ID	Rank	Question ID	Title	Snippet	Status	Timestamp	# answers	# views	Category	Body	Best answer
OLQ-0001	1	q13166161098	バイオハザードって設定に無…	弾いたけどピアノ弾いたことない人は楽譜…	Solved	2016/11/13 3:35	1	42	エンターテインメントと趣味 > ゲーム	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
OLQ-0001	2	q14166076254	バイオハザードって設定にむ…	タックルしたり足で蹴りまくればこわせる…	Solved	2016/11/10 3:47	1	18	エンターテインメントと趣味 > ゲーム	バイオハザードって…	なので、バイオハザードアウトブレイクシリーズ…
OLQ-0001	3	q11166238681	バイオハザードのゲームの…	バイオハザード4が好きで30週くらいして…	Solved	2016/11/21 3:29	3	19	エンターテインメントと趣味 > ゲーム	バイオハザードのゲーム…	個人的には… BIOHAZARD REVELATIONS UNVEILED EDITION …
…	…	…	…	…	…	…	…	…	…	…	…
OLQ-2000	998	q11137434581	夫婦生活で一番疲れる…	自分が相手に隠していることがあった…	Solved	2014/10/28 15:14	6	0	生き方と恋愛、人間関係の悩み >…	夫婦生活で 一番疲れる事は何ですか…	自分が相手に隠していることがあったら、どんなこと…
OLQ-2000	999	q1292632642	夫婦生活について教えて下さい…	主人とセックスをしておらず二年半に…	Solved	2012/9/3 9:51	5	701	生き方と恋愛、人間関係の悩み > …	夫婦生活について教えて下さい。 …	
OLQ-2000	1000	q1097950260	旦那との今後の夫婦生活、旦那…	結婚して来年の1月で2年になります。…	Solved	2012/12/5 10:01	4	640	生き方と恋愛、人間関係の悩み > …	旦那との今後の夫婦生活、…	

Clickthrough Data

CTR

Gender

Age

Query ID	Question ID	Rank	CTR	Male	Female	0s	10s	20s	30s	40s	50s	60s
OLQ-0001	q10165187300	1	0.059	1	0	0	0	0	0	0	1	0
OLQ-0001	q11164148731	1	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q11166231691	1	0.023	1	0	0	0	0	0	1	0	0
OLQ-0001	q11166372256	1	0.036	1	0	0	0	1	0	0	0	0
OLQ-0001	q13161212253	1	0.051	0.909	0.091	0	0.091	0.364	0.182	0.182	0.091	0.091
OLQ-0001	q13166161098	1	0.021	0	0	0	0	0	0	0	0	0
OLQ-0001	q14164350104	1	0.1	0	0	0	0	0	0	0	0	0
OLQ-0001	q14164384744	1	0.188	1	0	0	0	0.5	0	0	0.5	0
OLQ-0001	q14165651359	1	0.048	1	0	0	0	0	0	1	0	0
OLQ-0001	q11166278091	2	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q11166476886	2	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q12164569302	2	0.037	0	0	0	0	0	0	0	0	0
OLQ-0001	q12165573687	2	0.083	0	0	0	0	0	0	0	0	0
OLQ-0001	q10162841855	3	0.036	1	0	0	0.5	0.5	0	0	0	0
OLQ-0001	q12164050757	3	0.06	1	0	0	0	0	0	1	0	0
OLQ-0001	q12164687517	3	0.049	0.833	0.167	0	0	0	0.167	0.667	0.167	0
OLQ-0001	q12165837862	3	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q14158395769	3	0.027	0	0	0	0	0	0	0	0	0

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**

Relevance Judgments

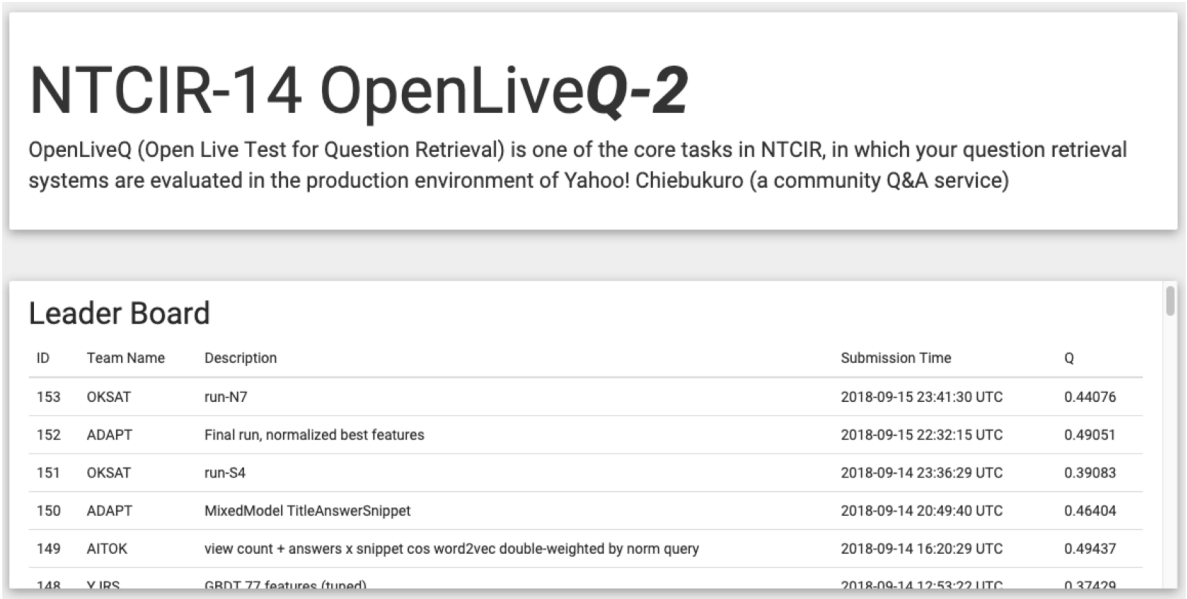
- **5** assessors were assigned for each
  - **Relevance** ≡ # assessors who want to click

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



NTCIR-14 OpenLive*Q-2*

OpenLiveQ (Open Live Test for Question Retrieval) is one of the core tasks in NTCIR, in which your question retrieval systems are evaluated in the production environment of Yahoo! Chiebukuro (a community Q&A service)

Leader Board

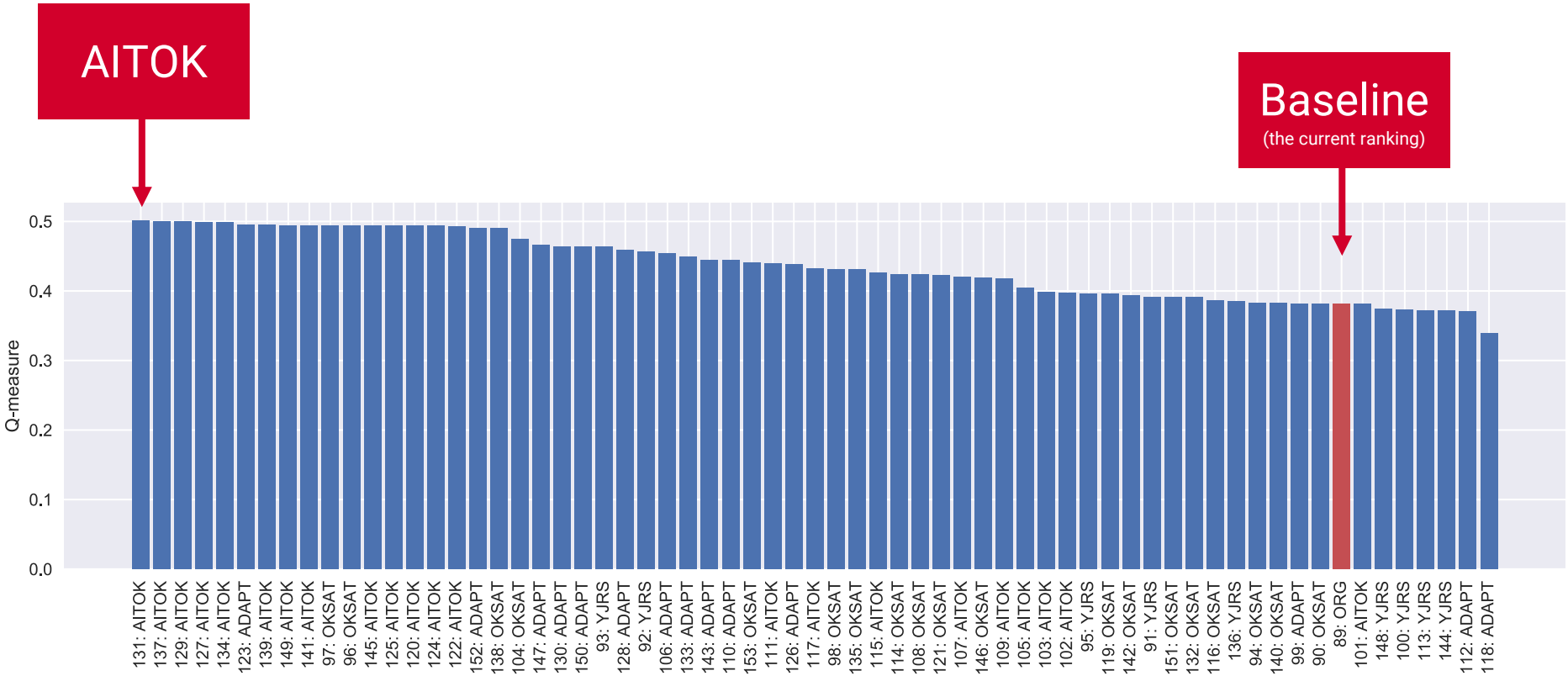
ID	Team Name	Description	Submission Time	Q
153	OKSAT	run-N7	2018-09-15 23:41:30 UTC	0.44076
152	ADAPT	Final run, normalized best features	2018-09-15 22:32:15 UTC	0.49051
151	OKSAT	run-S4	2018-09-14 23:36:29 UTC	0.39083
150	ADAPT	MixedModel TitleAnswerSnippet	2018-09-14 20:49:40 UTC	0.46404
149	AITOK	view count + answers x snippet cos word2vec double-weighted by norm query	2018-09-14 16:20:29 UTC	0.49437
148	YJRS	GBDT 77 features (tuned)	2018-09-14 12:53:22 UTC	0.37429

Participants

- **AITOK** Tokushima University
- **YJRS** Yahoo Japan Corporation
- **OKSAT** Osaka Kyoiku University
- **DCU-ADAPT** Dublin City University

- **ORG** Organizers

Offline Evaluation Results



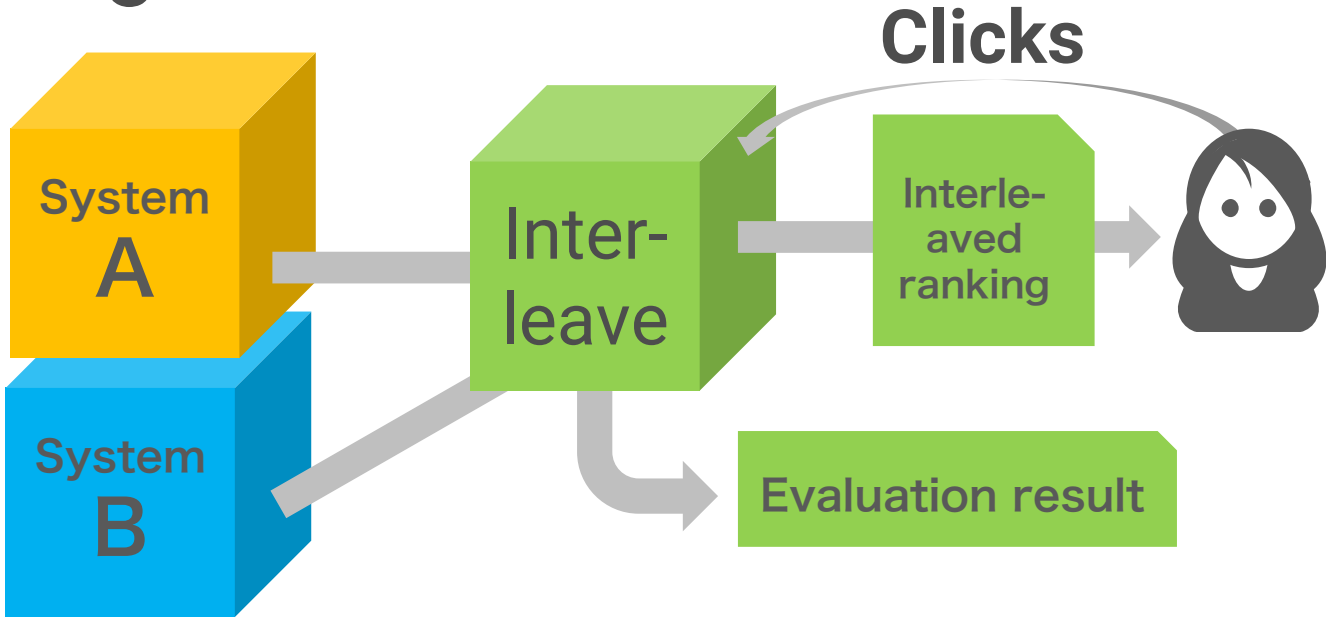
- AITOK achieved the best performances among five teams
- A concern about overfitting on test queries

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

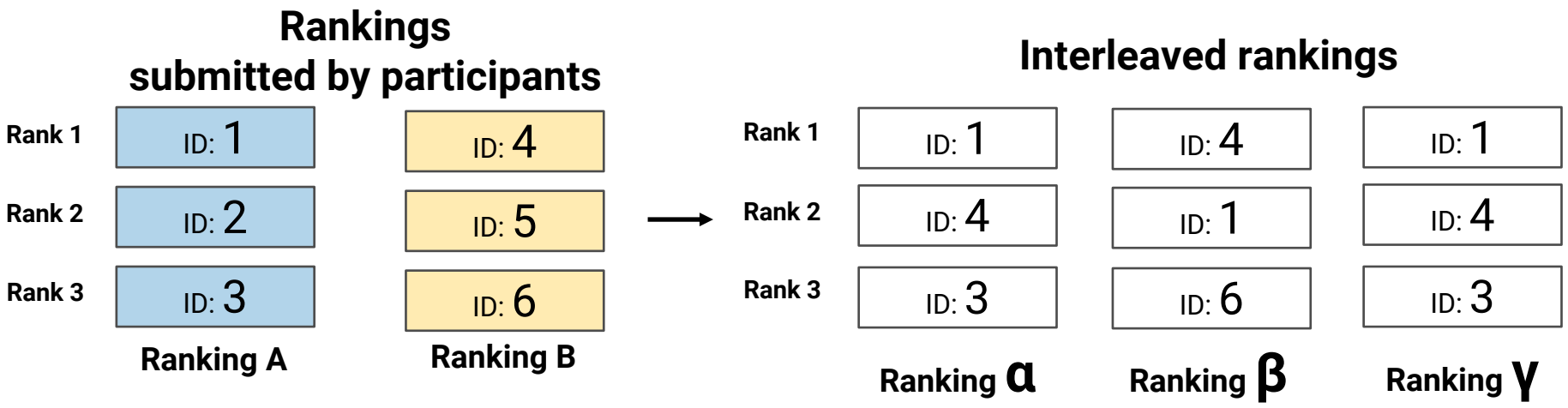
Interleaving: an alternative to A/B testing

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



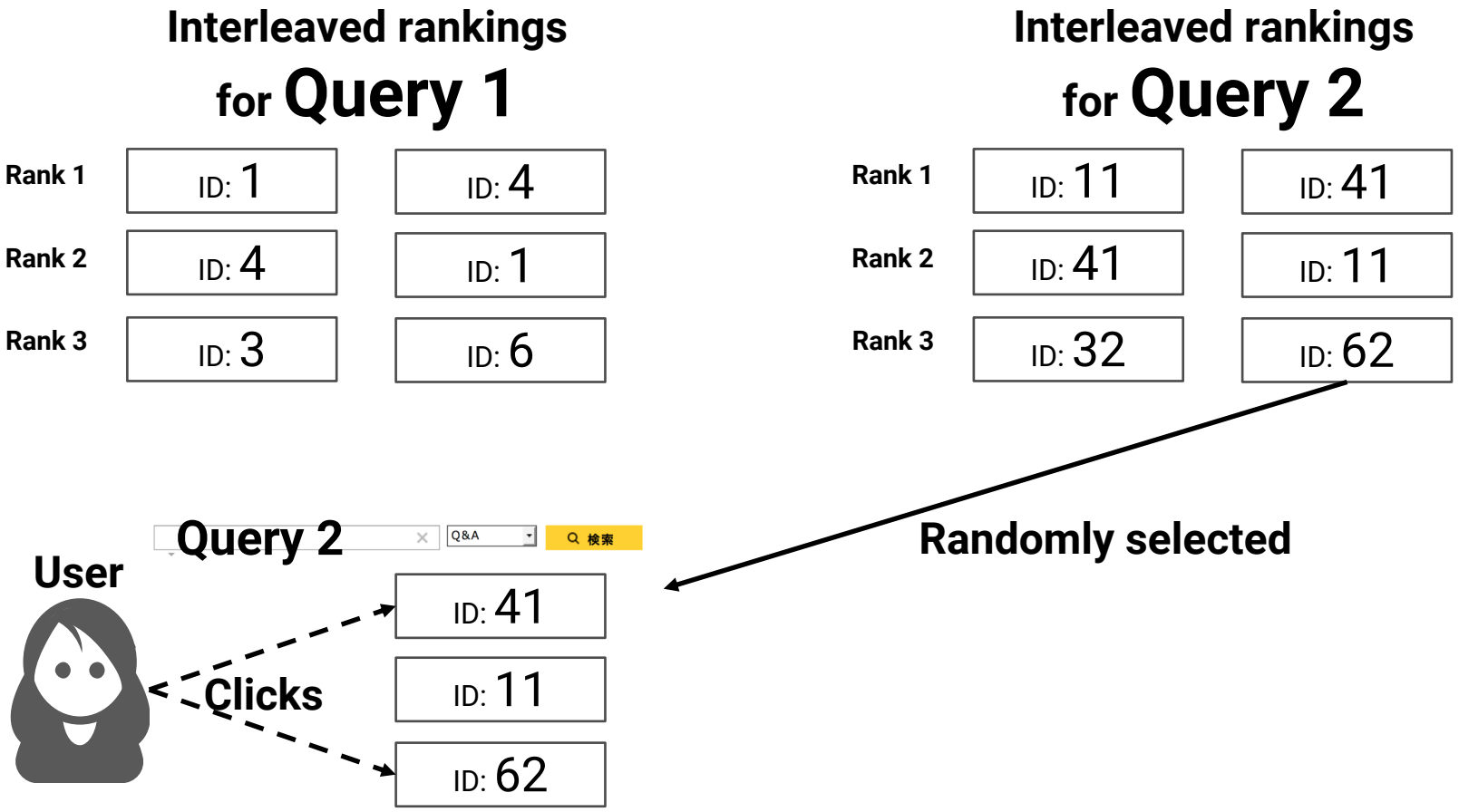
- **10-100 times as efficient as A/B testing**
- **Multileaving** = Interleaving for 3≧ rankings

Pairwise Preference Multileaving (PPM) 1/3



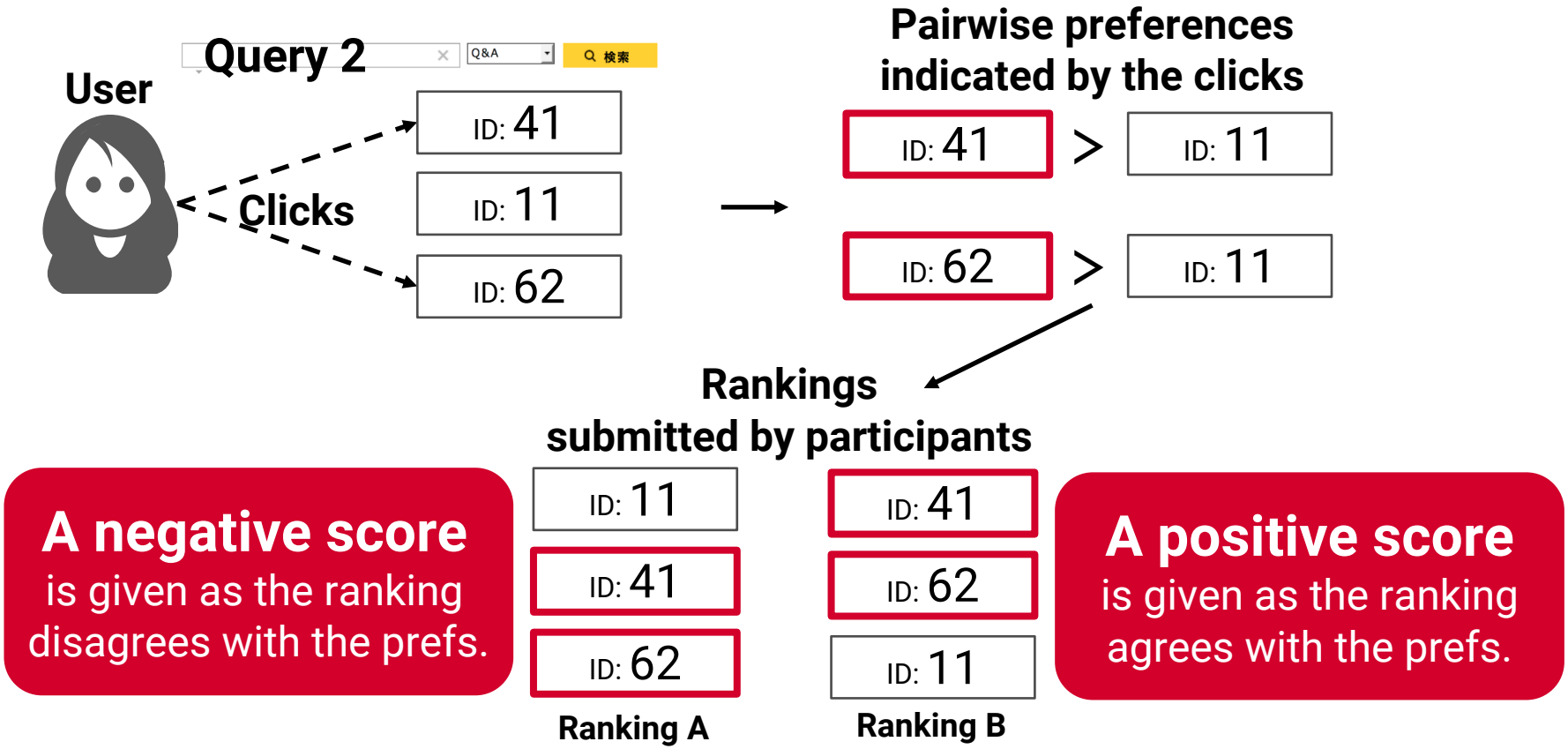
- Given multiple rankings $\mathcal{R}$, PPM generates **interleaved rankings** such that
  - A document at k-th rank is selected from documents at $1, \ldots, k$-th rank in $\mathcal{R}$
  - A document can be selected only once
- Example of **Ranking α**
  - Rank 1: 1 ~ {1, 4}, Rank 2: 4 ~ {2, 4, 5}, Rank 3: 3 ~ {2, 3, 5, 6}

Pairwise Preference Multileaving (PPM) 2/3



- 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

Pairwise Preference Multileaving (PPM) 3/3



- A ranking receives a positive score **if it agrees with pairwise prefs. indicated by the clicks**

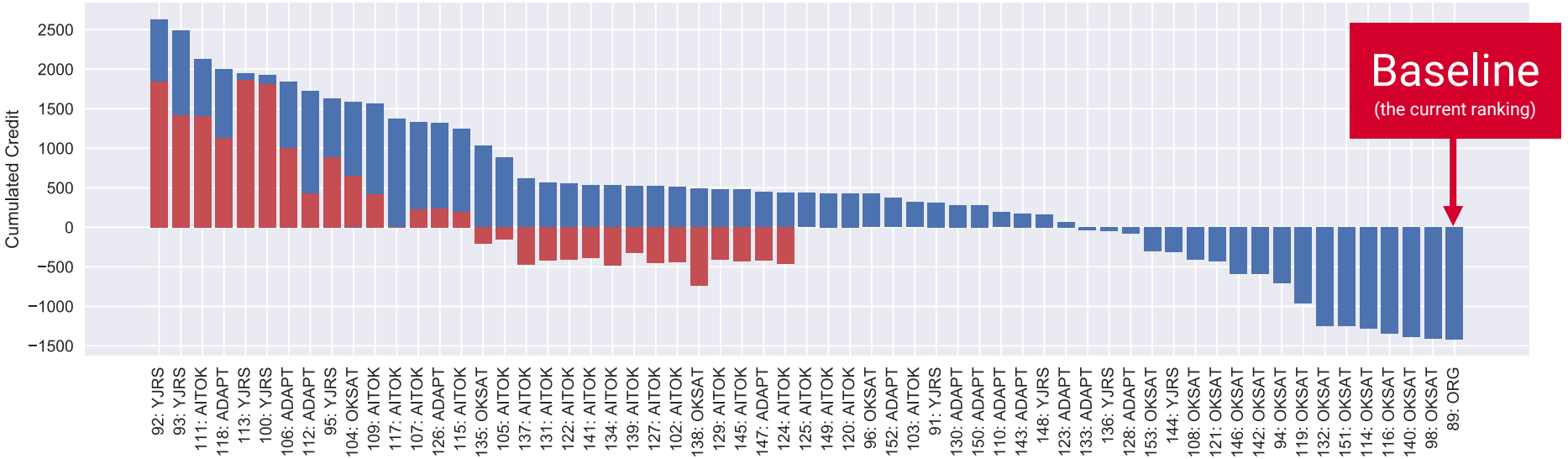
Two-phase Strategy for Large-scale Interleaving

- 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

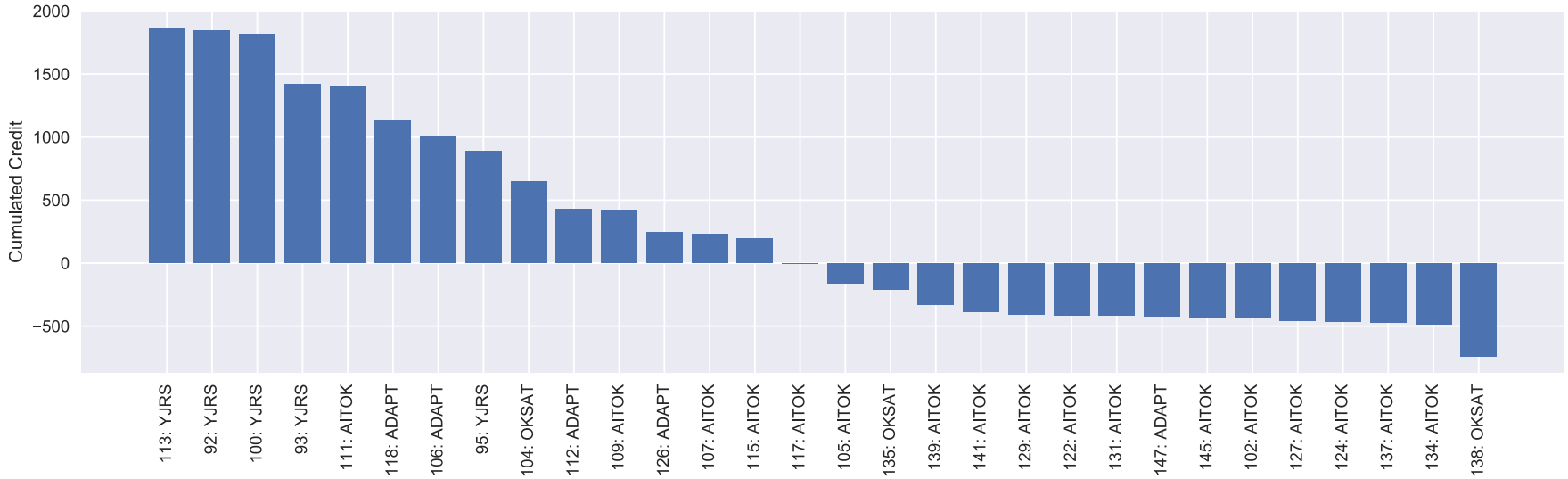
***Kato et al. Challenges of Multileaved Comparison in Practice: Lessons from NTCIR-13 OpenLiveQ Task, CIKM 2018.**

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

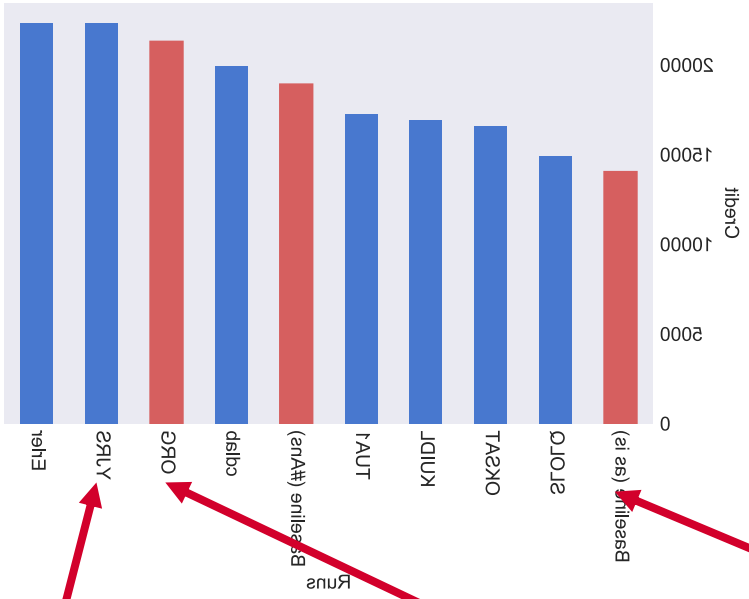
Online Evaluation Result at the 2nd Phase



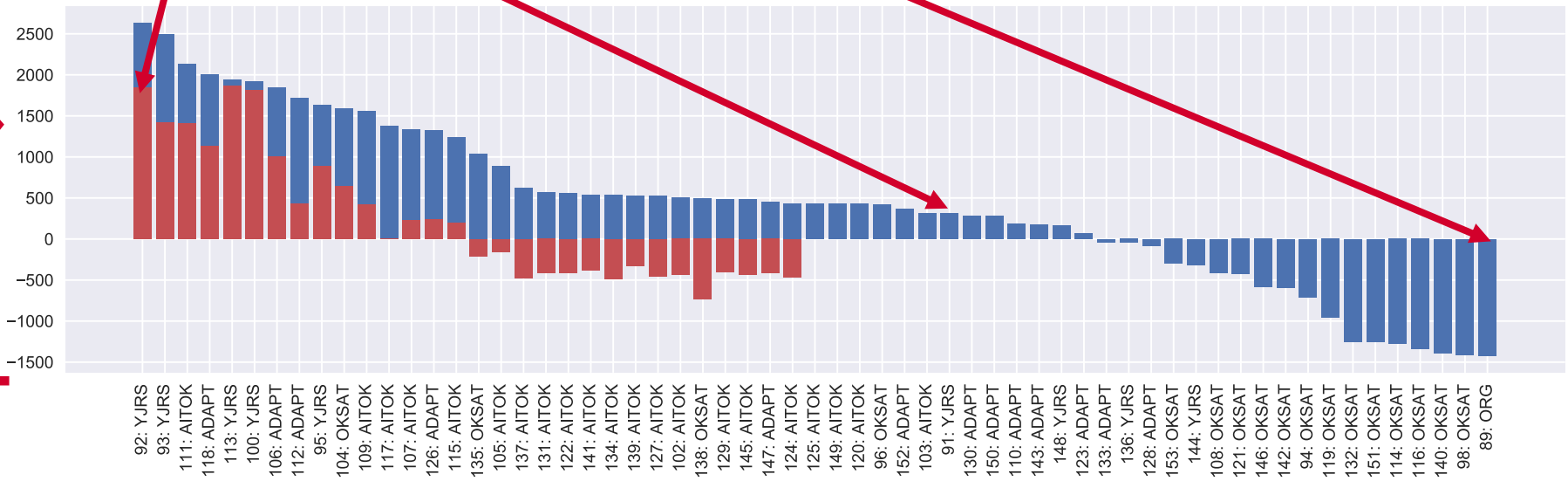
- **Quite different from the offline evaluation results**
  - Confirmed the importance of evaluating all the runs online
- **YJRS** achieved the best performance, while no sig. diff. from the top eight runs

Progress from OpenLiveQ-1

OpenLiveQ-1



OpenLiveQ-2



- **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 1st and 2nd 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