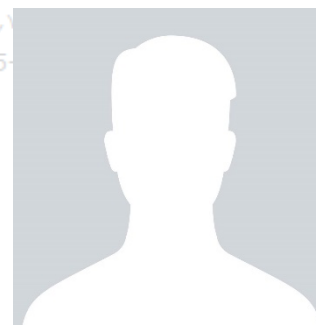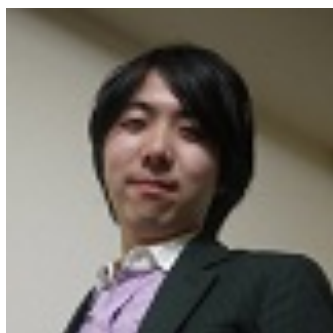# Overview of the NTCIR-13 OpenLiveQ Task

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

- **Task Design** **(3 slides)**

- **Data** **(5 slides)**

- **Evaluation Methodology** **(12 slides)**

- **Evaluation Results** **(6 slides)**

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)**

Effective for Fever   INPUT   ×   Q&A   Q 検索

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

OUTPUT

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.
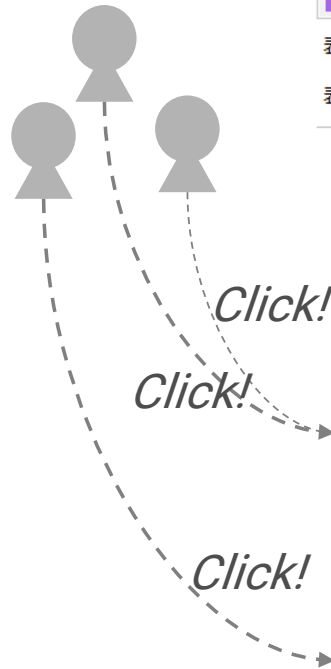
2 Answers   Posted on Jan 3, 2010

4

Real users

倍以上男子 | Q&A | Q 検索 ＋条件指定 **YAHOO! JAPAN** 知恵袋

ゲスト さん [ログイン] My知恵袋

約1,652件 1～10件目 Q&A

Q&A (1,652) | 知恵ノート (0)

表示： すべて | 回答受付中 (2) | 投票受付中 (2) | 解決済み (1,648)

表示順序： 関連度順 | RSS

20歳の**男子**学生が、年齢が**倍以上**行っている鈴木保奈美のファンだということ...
ファンだということはおかしいですか？
解決済み - 更新日時:2013/03/04 18:35:29 - 回答数：1 - 閲覧数：12
エンターテインメントと趣味 ＞ 芸能人 ＞ 俳優、女優

Insert → Team A

東大生は**男子**が女子の2倍のいるそうですが、どうしてですか？
2倍どころか、それ以上だと思いますよ。女性比率はもっと低いです。 理由としては、社会において男女平等が実現しつつある状況にはあっても、まだまだ女子側に、**男子**と同様の高等教育への意欲、社会における成功...
解決済み - 更新日時:2007/08/26 00:50:52 - 回答数：4 - 閲覧数：890
子育てと学校 ＞ 大学、短大、大学院 ＞ 大学

Insert → Team B

高1の不等式の問題なんですが、不等式が苦手でイマイチ分からなかったので解...
解説していただきたいと思い質問されていただきました。 次の関係を不等式で表しなさい。 Q.あるクラスの生徒数 は35人で、**男子**の人数xの3倍は、女子の人数の2**倍以上**である。 A.3x≧2(35-x)なんですがなぜカッコ...
解決済み - 更新日時:2013/07/10 18:57:48 - 回答数：3 - 閲覧数：25
教養と学問、サイエンス ＞ 数学、サイエンス ＞ 数学

Insert → Team C

Click!
Click!
Click!

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

5

Data

	Training	Testing
Queries	1,000	1,000
Documents (or questions)	984,576	982,698
Clickthrough data (with user demographics*)	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)
(basic features also available, i.e. language-independent)**

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

7

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	5701	0	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	4640	0	生き方と恋愛、人間関係の悩み > …	旦那との今後の夫婦生活、…	

8

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

9

Baselines

- **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
    - D. Metzler and W.B. Croft. Linear feature-based models for information retrieval. Information Retrieval, 10(3): 257-274, 2007.

- **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

11

- **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？

ブドウ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

Don't want to click

ブ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錠ほど舐めていました。
昨日、調剤薬局に行ったのでブ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

Want to click

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...

クリック

13

- ## 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)

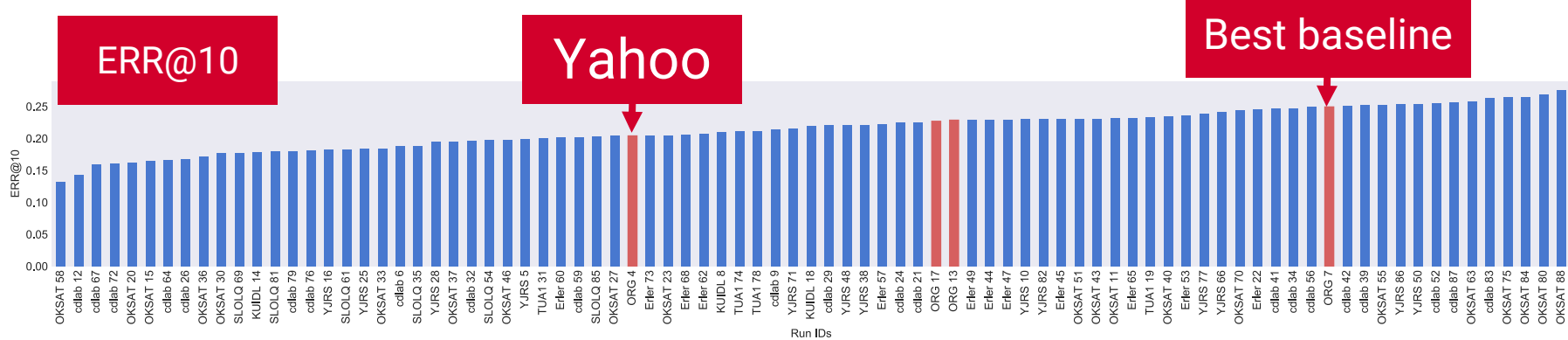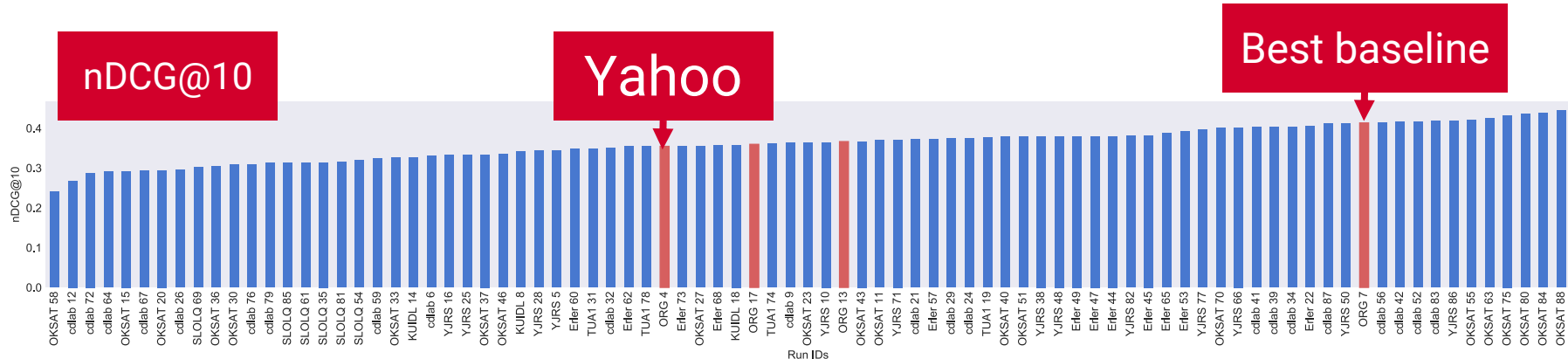 - **85** submissions from 7 teams

Leader Board

ID	Team Name	Description	Submission Time	nDCG@10
88	OKSAT	run20	2017-04-21 23:40:37 UTC	0.44471
87	cdlab	Last	2017-04-21 16:10:07 UTC	0.41131
86	YJRS	Baseline + multiple BM25F features + nDCG@10.	2017-04-21 01:06:33 UTC	0.41894
85	SLOLQ	test #05	2017-04-20 21:45:33 UTC	0.31329
84	OKSAT	run19	2017-04-20 16:19:10 UTC	0.43767
83	cdlab	#21	2017-04-20 15:41:28 UTC	0.41800
82	YJRS	8foldCV_LambdaMART	2017-04-20 00:36:57 UTC	0.38087
81	SLOLQ	test #04	2017-04-19 21:11:38 UTC	0.31516

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

16

nDCG@10

ERR@10

Similar results. The top performers are OKSAT, cdlab, and YJRS

nDCG@10 bars: cdlab 87, YJRS 50, ORG 7, cdlab 56, cdlab 42, cdlab 52, cdlab 83, YJRS 86, OKSAT 55, OKSAT 63, OKSAT 75, OKSAT 80, OKSAT 84, OKSAT 88

ERR@10 bars: ORG 7, cdlab 42, cdlab 39, OKSAT 55, YJRS 86, YJRS 50, cdlab 52, cdlab 87, OKSAT 63, cdlab 83, OKSAT 75, OKSAT 84, OKSAT 80, OKSAT 88

Q-measure

Different results. The top performers are YJRS and Erler
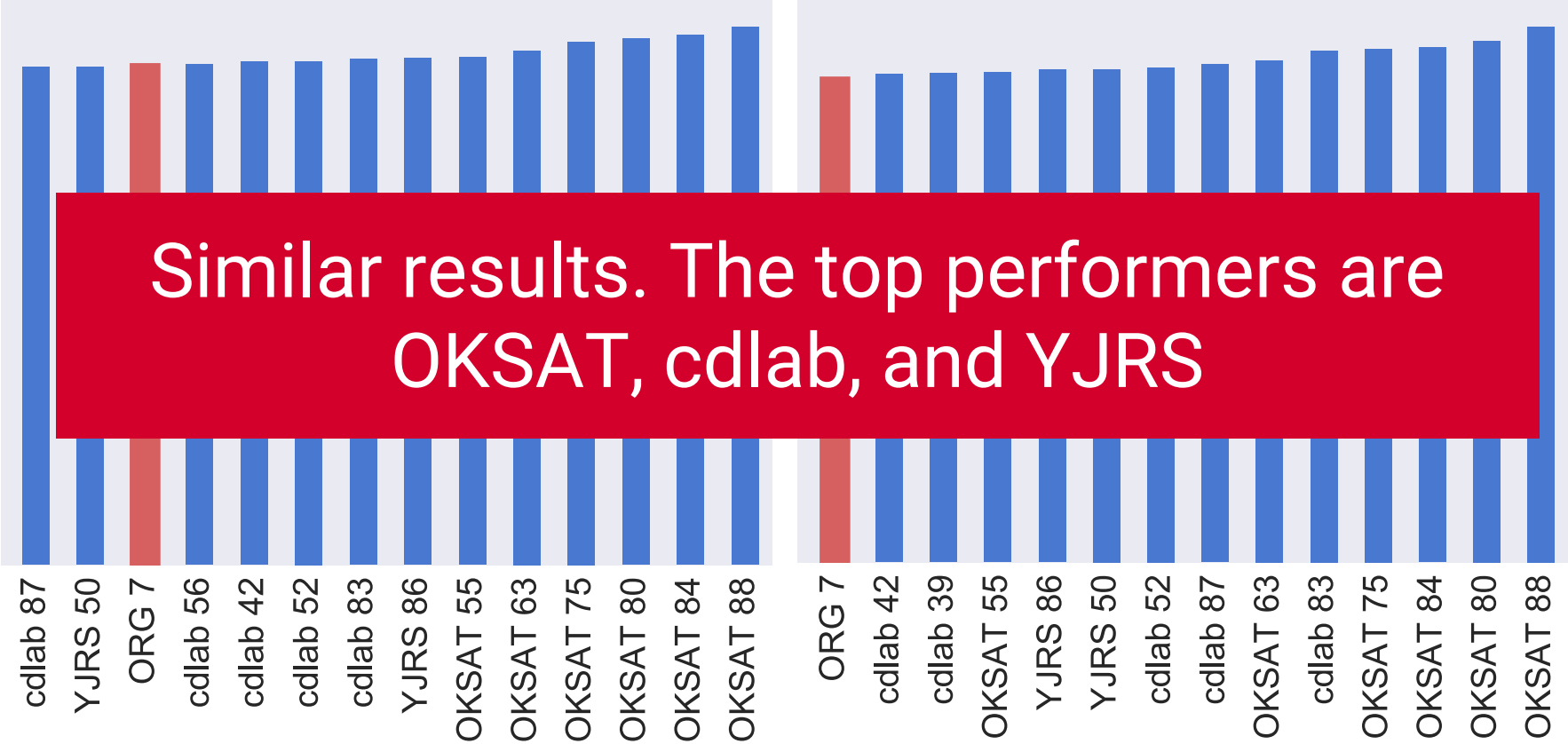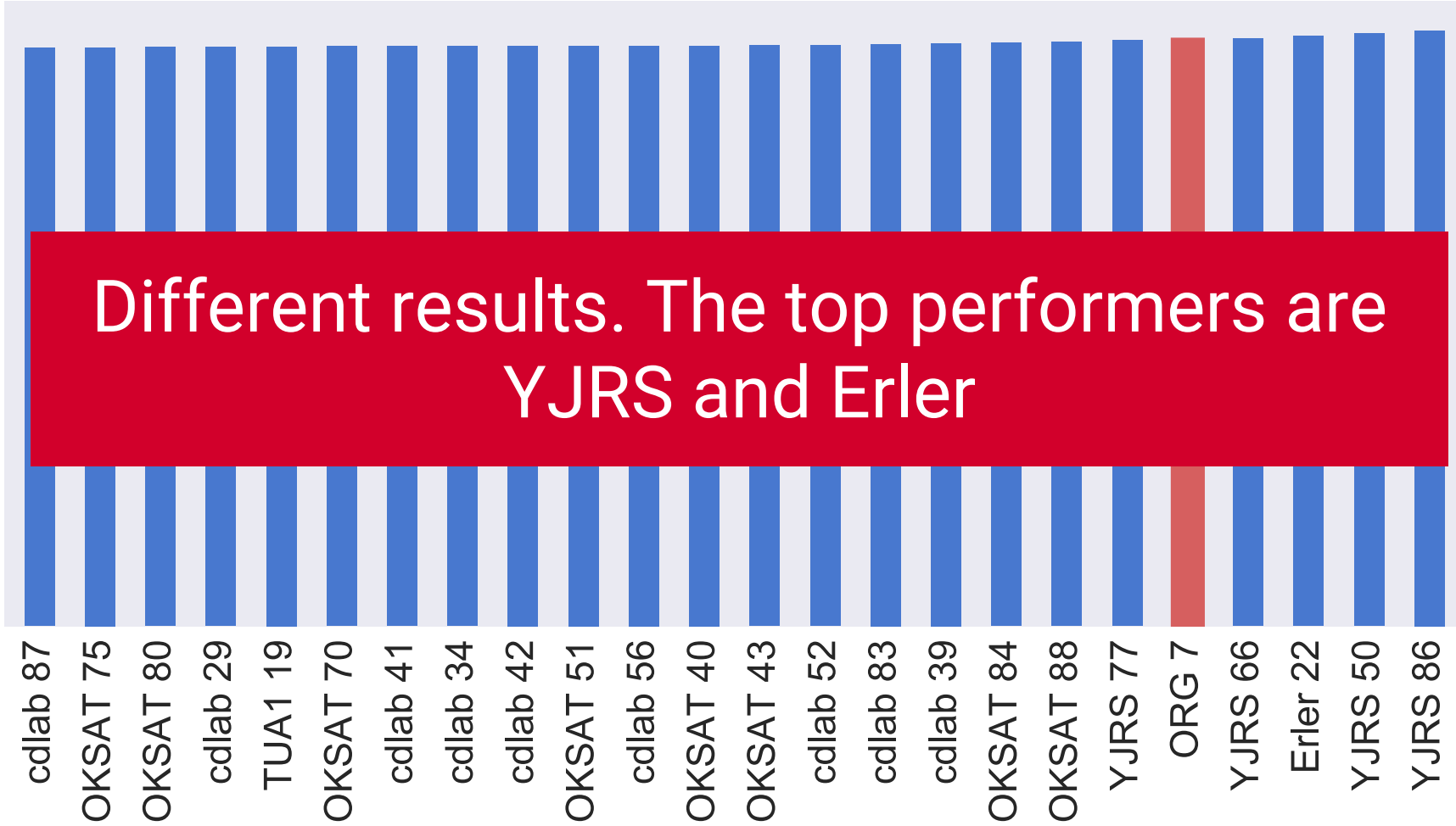
Turned out to be more consistent with the online evaluation

- **Multileaved comparison methods are used in the online evaluation**
  - Schuth, Sietsma, Whiteson, Lefortier, de Rijke: Multileaved comparisons for fast online evaluation, CIKM2014.

- **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**

19

A Comparative Live Evaluation of Multileaving Methods on a Commercial cQA Search

20

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

Clicks

System A		Inter-leave		Interle-aved ranking	

System B

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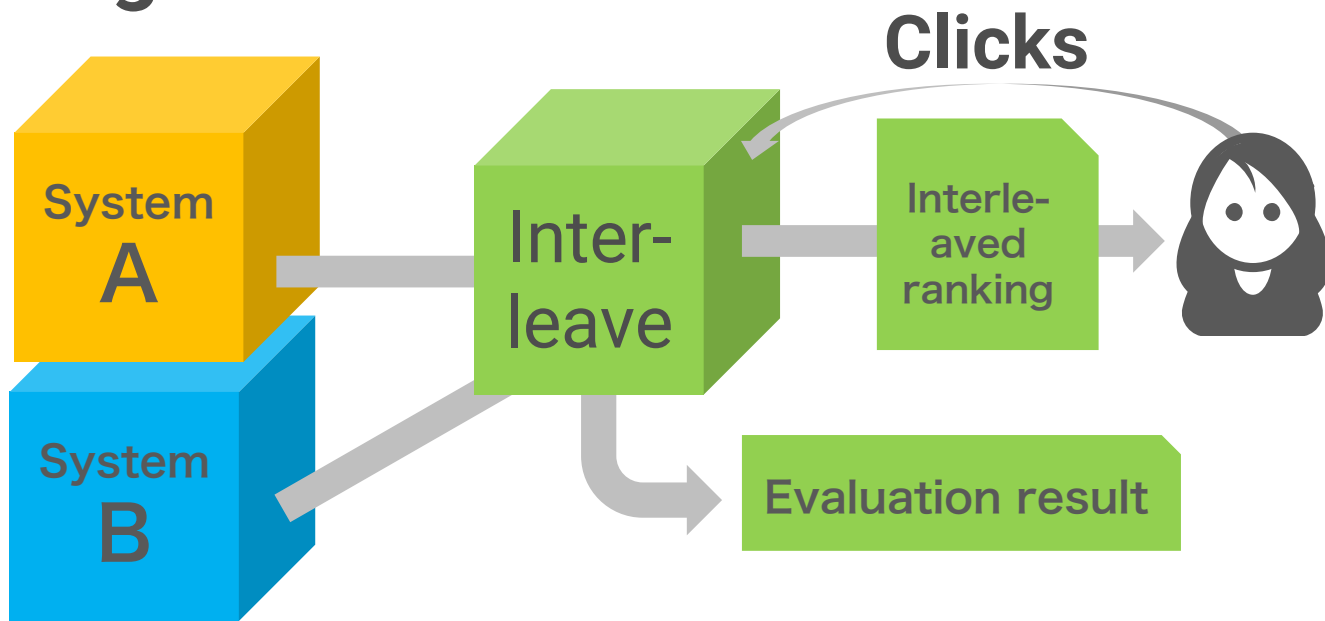
- **10-100 times as efficient as A/B testing**
- **Multileaving = Interleaving for 3≧ rankings**

21

Intuitive Explanation of Optimized Multileaving (OM)

Rankings submitted by participants

Rank 1	ID: 1	ID: 4
Rank 2	ID: 2	ID: 5
Rank 3	ID: 3	ID: 6
	Ranking A	Ranking B

$\longrightarrow$

Interleaved rankings

Rank 1	ID: 1	ID: 4	ID: 1
Rank 2	ID: 4	ID: 1	ID: 4
Rank 3	ID: 3	ID: 6	ID: 3
	Ranking α	Ranking β	Ranking γ
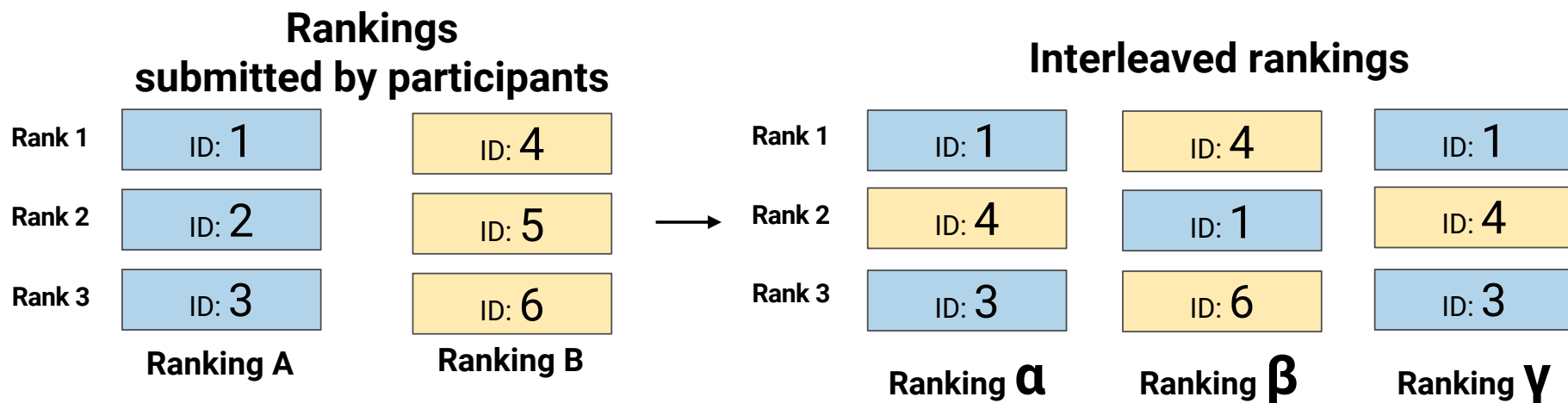
- Interleaved rankings are shown to users with prob. $P(\alpha)$, $P(\beta)$, and $P(\gamma)$, respectively

- Give a credit $1/k$ to each ranking if its document at rank k is clicked
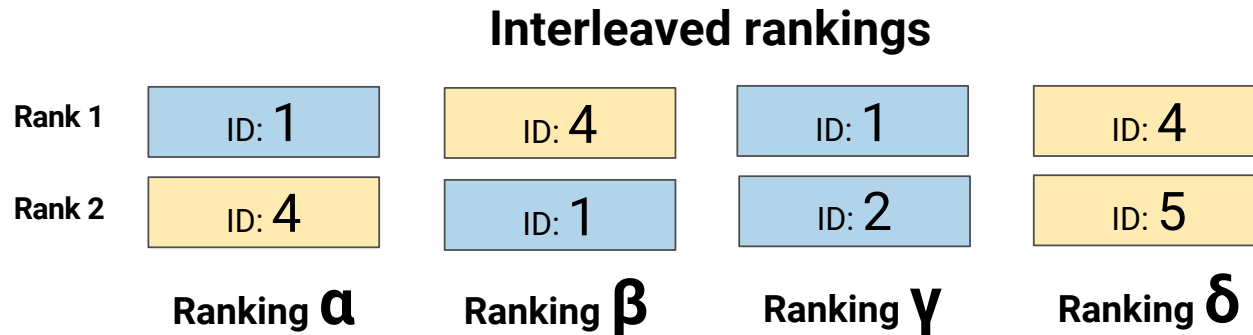
- Evaluate rankers by the cumulated credits

22

Interleaved rankings

	Ranking α	Ranking β	Ranking γ
Rank 1	ID: 1	ID: 4	ID: 1
Rank 2	ID: 4	ID: 1	ID: 4
Rank 3	ID: 3	ID: 6	ID: 3

- 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

23

Forcing Comparison of Rankings

Interleaved rankings

	Ranking α	Ranking β	Ranking γ	Ranking δ
Rank 1	ID: 1	ID: 4	ID: 1	ID: 4
Rank 2	ID: 4	ID: 1	ID: 2	ID: 5

- (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
  $\rightarrow$ **Less chances to know the difference**

- **OM optimizes the presentation probability to maximize the chance of comparison**

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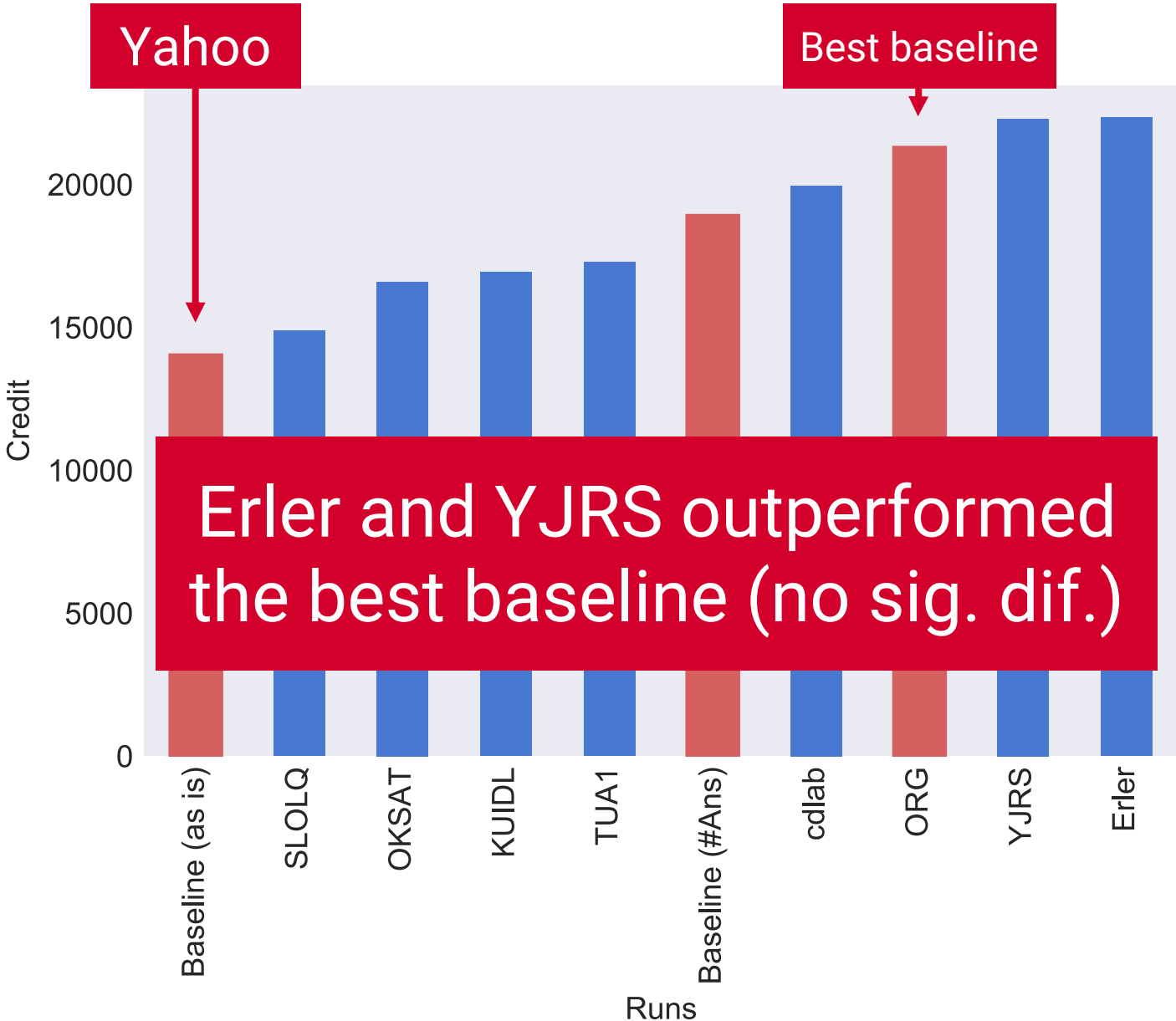= expected variance of cumulated
credits for each ranker

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

$$\forall r, \forall j, j' - \lambda_r \leq \mathbb{E}\big[g(I_j, r)\big] - \mathbb{E}\big[g(I_{j'}, r)\big] \leq \lambda_r$$
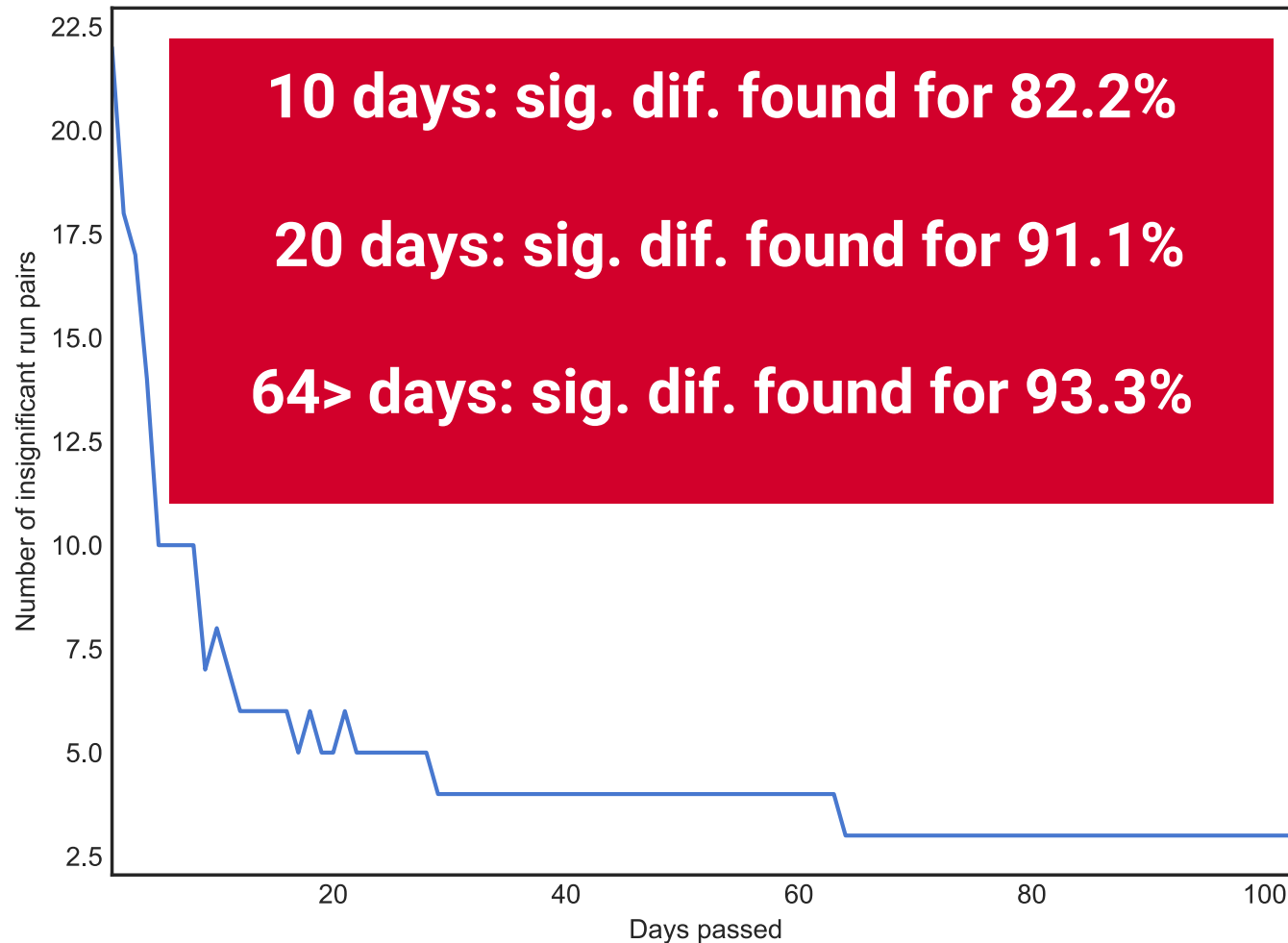
Bias = the difference of expected
cumulated credits of rankers

- Slightly different from Schuth et al.'s
  - Theirs sometimes fails due to "no solution"

Online Evaluation Result

Yahoo

Best baseline

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

Credit

20000

15000

10000

5000

0

Baseline (as is)
SLOLQ
OKSAT
KUIDL
TUA1
Baseline (#Ans)
cdlab
ORG
YJRS
Erler

Runs

26

- ***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%

27

1. **Some differences from the offline evaluation**
   - Offline: OKSAT > cdlab $\fallingdotseq$ YJRS > Erler
   - Online: Erler $\fallingdotseq$ 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)