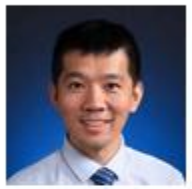




Overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) Task

Tetsuya Sakai, Sijie Tao, Zhumin Chu, Maria Maistro, Yujing Li, Nuo Chen, Nicola Ferro, Junjie Wang, Ian Soboroff, Yiqun Liu





June 15@NTCIR-16 (Virtual Event)

TALK OUTLINE

1. History and the importance of the task
2. The WWW-4 test collection
3. Participants
4. Results: progress
5. Results: reproducibility
6. Conclusions and future work

History

NTCIR-13 (Dec 2017)

NTCIR-14 (Jun 2019)

NTCIR-15 (Dec 2020)

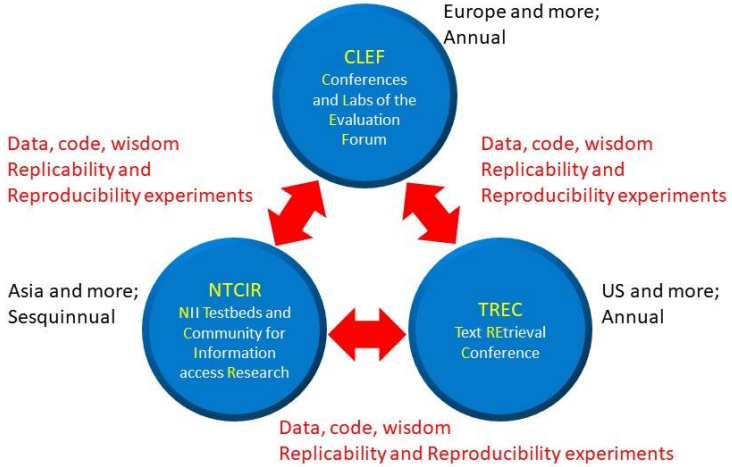
NTCIR-16 (Jun 2022)

TREC 2014

We Want Web 1 [Luo+17]

We Want Web 2 [Mao+19]

CENTRE = CLEF/NTCIR/TREC
(replicability and) reproducibility



CENTRE@TREC2018
CENTRE@CLEF2018, 2019

CENTRE [Sakai+19CENTRE]

We Want Web 3 [Sakai+20WWW3]

We Want Web 4

Why we're doing it

- Progress

- Web search isn't a solved problem! Neural approaches raise new research questions!

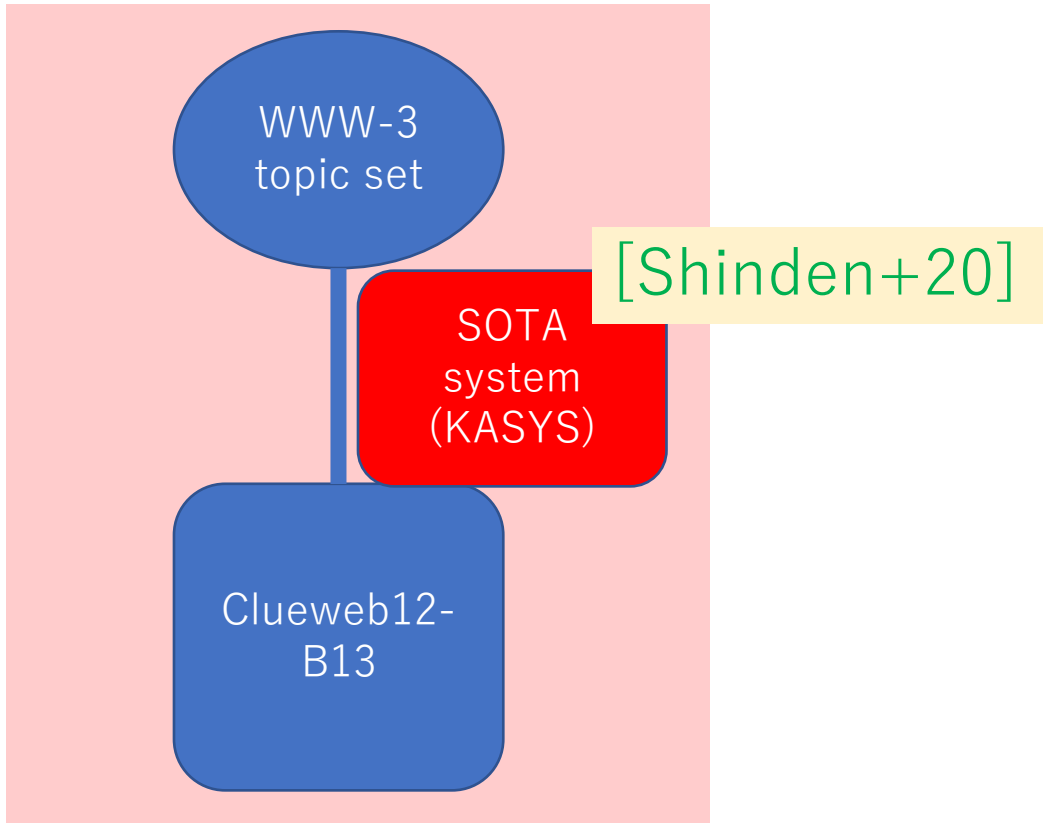
- We need a post-clueweb12 test collection to keep advancing the SOTA!

- Reproducibility

- If you can't even reproduce other people's work, you won't be able to advance the SOTA!

How the WWW task quantifies progress and reproducibility

NTCIR-15
WWW-3

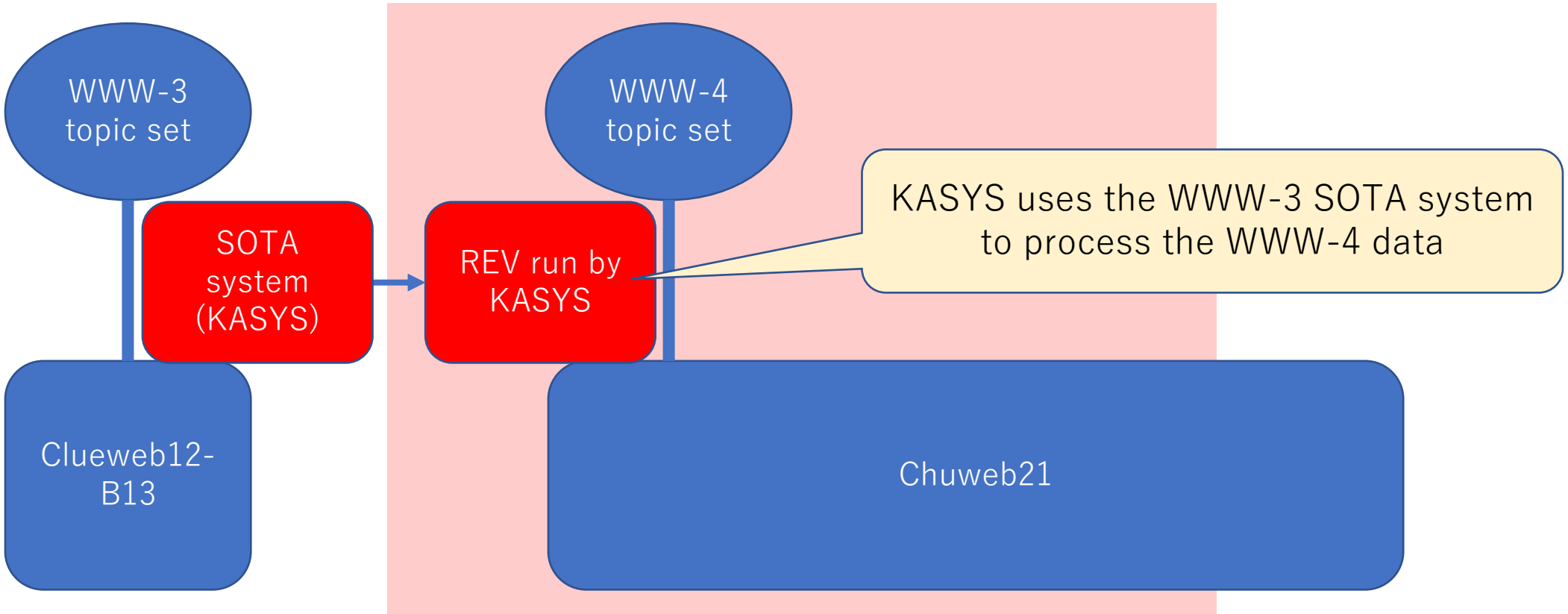


[Shinden+20]

How the WWW task quantifies progress and reproducibility

NTCIR-15
WWW-3

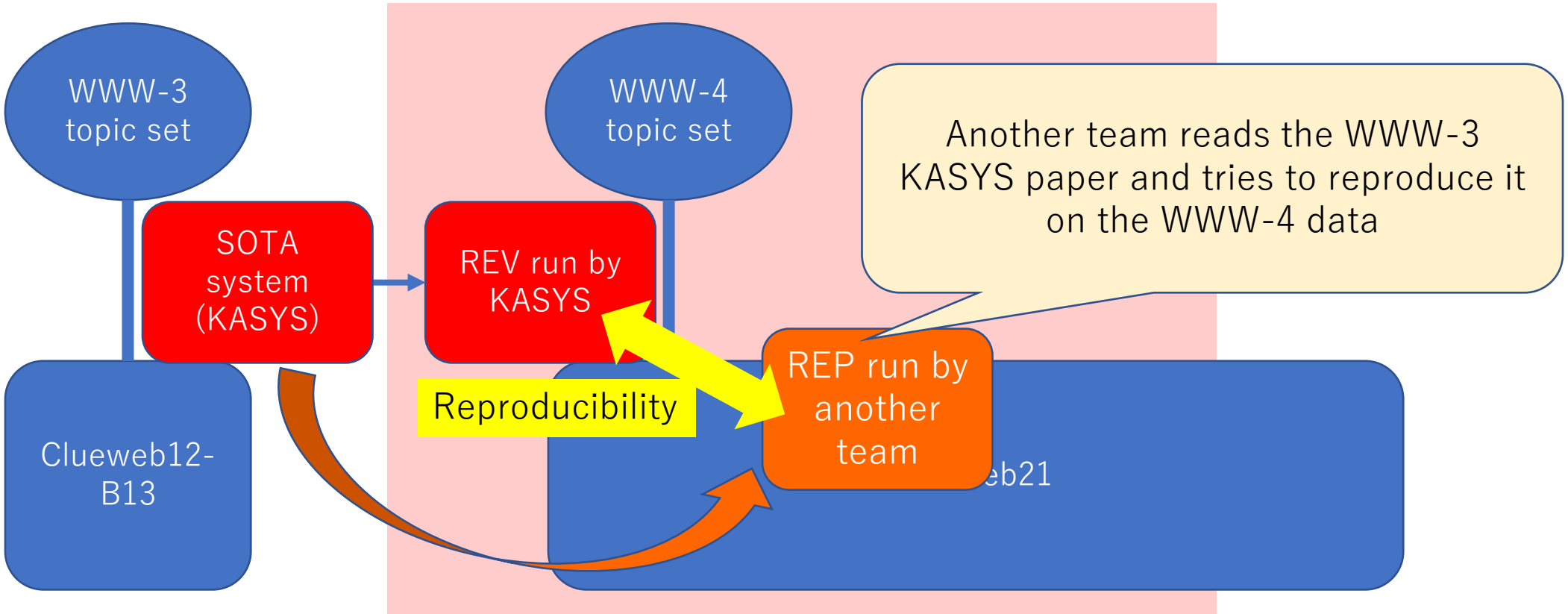
NTCIR-16
WWW-4



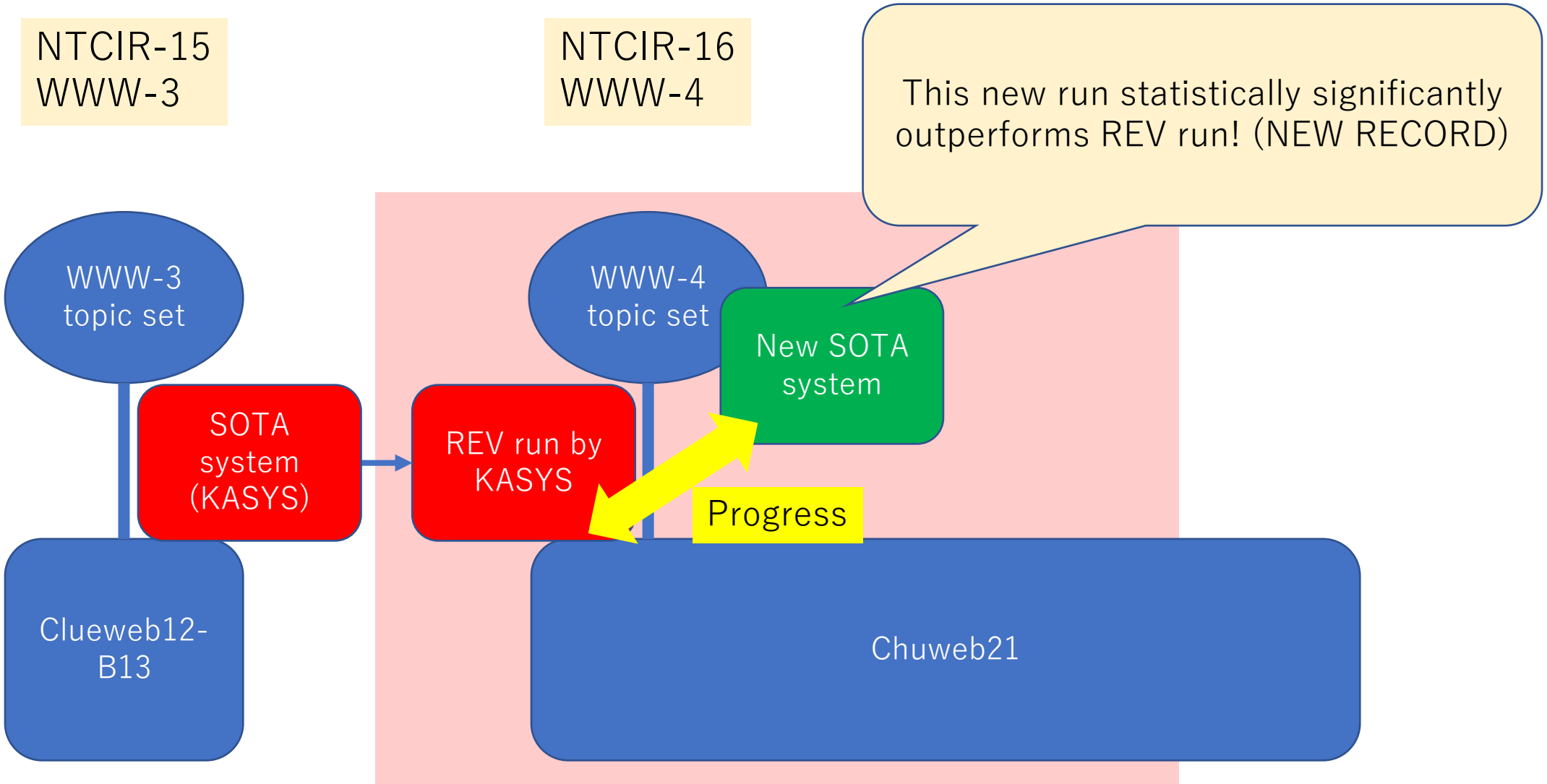
How the WWW task quantifies progress and reproducibility

NTCIR-15
WWW-3

NTCIR-16
WWW-4



How the WWW task quantifies progress and reproducibility

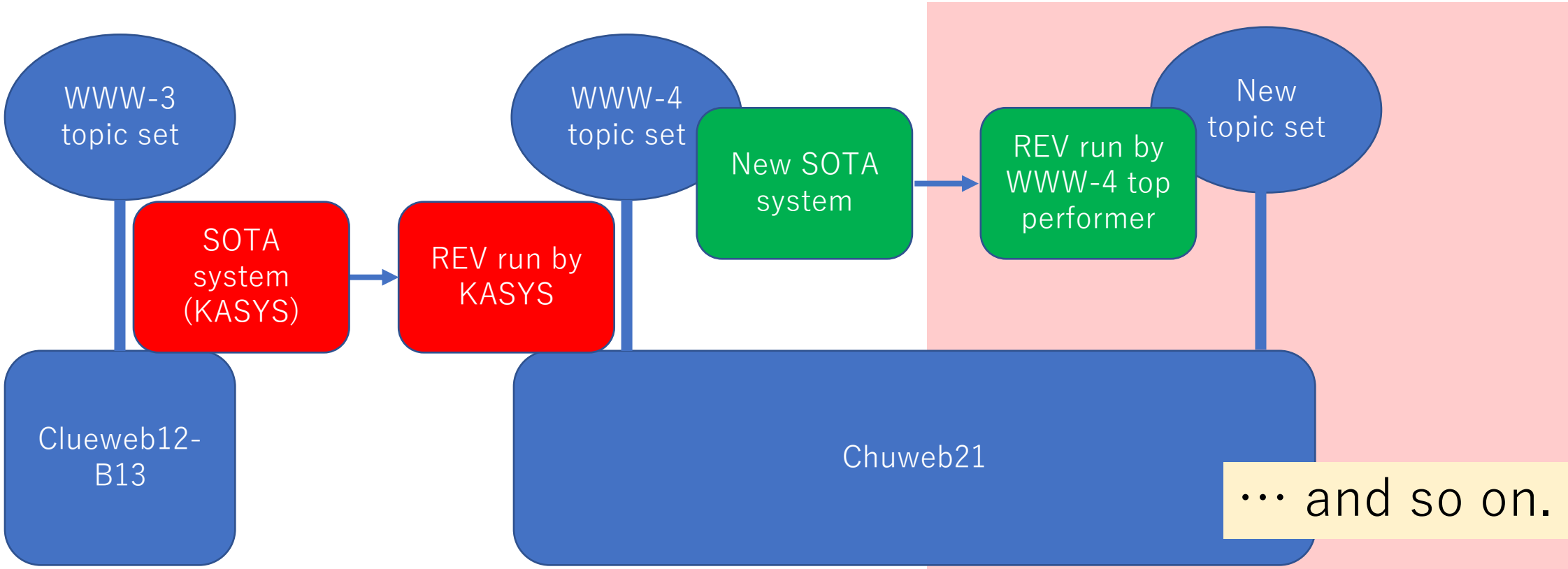


How the WWW task quantifies progress and reproducibility

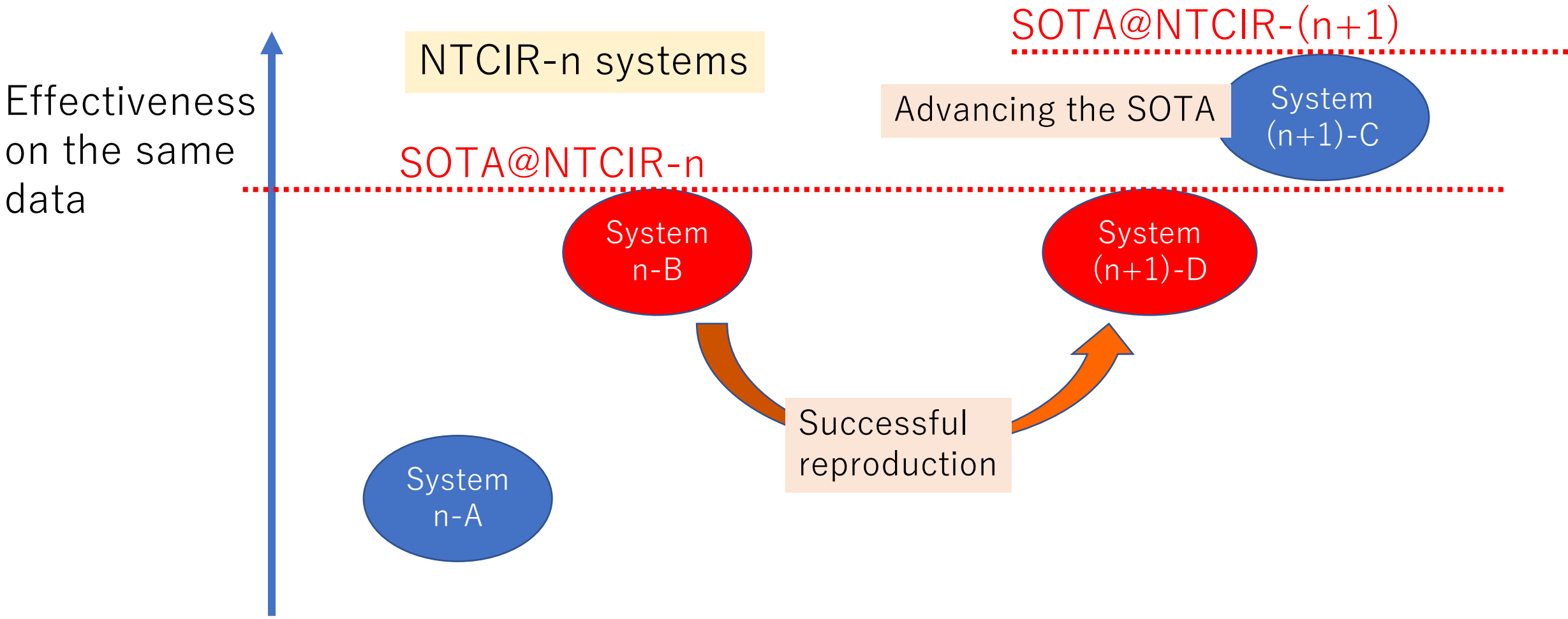
NTCIR-15
WWW-3

NTCIR-16
WWW-4

NTCIR-17
New task?



Progress and reproducibility



TALK OUTLINE

1. History and the importance of the task
2. The WWW-4 test collection
3. Participants
4. Results: progress
5. Results: reproducibility
6. Conclusions and future work

Brand new corpus: chuweb21

Zhumin Chu (*)

Tsinghua University, China

- Statistics: 82,451,337 HTMLs, 1.69TiB compressed content
- Generated based on the April 2021 block of Common Crawl dataset[1]
- Filtering constraints

More recent than ClueWeb12!

- root domain: .com, .org or .net
- WARC-Type is “response”
- HTML content: more than 1,000
- Document language: P(Eng.) > 0.99

[1] https://commoncrawl.org/the-data/

Topic set size design [Sakai16IRJ][Sakai18book]

http://www.f.waseda.jp/tetsuya/samplesizeANOVA2.xlsx

Residual variances from WWW-3 results:

Smallest: 0.00716 (iRBU); Largest: 0.0284 (nERR)

Under Cohen's five-eighty convention (α=0.05, β=0.20,)

- We need 44 topics for a minimum detectable diff of 0.1 in nERR;
- We need 45 topics for a minimum detectable diff of 0.05 in iRBU.

Let's go with 50 topics to satisfy the above statistical power requirements.

Realistic topics – organisers' info needs







https://waseda.box.com/www4topicsxml

v1 **www4topics.xml**

16WWW4 · Updated Oct 29, 2021 by 酒井　哲也

```
<queries>
    <query>
        <qid>0201</qid>
        <content>Timnit Gebru Google</content>
        <description>I want to know the details regarding Google's firing of Dr. Timnit Gebru.</description>
    </query>
```

Guess which organisers came up with which topics!

Gold and bronze assessments

See [Bailey+08]

5 students, 10 topics each

5 employees, 10 topics each

Waseda and Tsinghua merged

relevance level	Gold (1 assessor/topic)	Bronze-Waseda (1 assessor/topic)	Bronze-Tsinghua (1 assessor/topic)	Bronze-All (2 assessors/topic)
L0	7,154	5,584	6,571	4,900
L1	1,806	3,158	1,986	1,881
L2	1,373	1,591	1,776	1,485
L3	N/A	N/A	N/A	1,241
L4	N/A	N/A	N/A	826
total	10,333	10,333	10,333	10,333

Bronze assessors tend to be more liberal (esp. Bronze-W)

Low agreements between different qrels versions…

Gold and bronze assessors disagree substantially!

Mean per-topic inter-assessor agreement in terms of quadratic weighted Cohen's κ ($n = 50$ topics).

qrels version	mean κ
Gold-Waseda	0.242
Gold-Tsinghua	0.280
Waseda-Tsinghua	0.458

Waseda-Tsinghua mean kappa statistically significantly higher than the Gold-Bronze mean kappas!

Comparison of the mean κ's with a randomised Tukey HSD test ($B = 5,000$ trials). The effect sizes are based on the two-way ANOVA residual variance $V_{E2} = 0.0345$ [11].

[Sakai18book]

Gold-Waseda vs. Gold-Tsinghua	$p = 0.679, ES_{E2} = 0.202$
Gold-Waseda vs. Waseda-Tsinghua	$p \approx 0, ES_{E2} = 0.958$
Gold-Tsinghua vs. Waseda-Tsinghua	$p \approx 0, ES_{E2} = 1.160$

The low disagreements are due to the way documents were ordered for the assessors
[Details to be reported elsewhere]

PRI: order by pseudorelevance RND: randomize [Sakai+22TOIS]

Table 4. Inter-assessor agreement in terms of mean quadratic weighted κ for each pair of assessment conditions. A Tukey HSD test for unpaired data shows that every difference between a within-PRI agreement (rows a, b, and c) and a RND-PRI agreement (rows d and e) is statistically highly significant ($p \approx 0.0000000$), while none of the other differences are statistically significant ($p > 0.238$), as indicated in the rightmost column. Residual variance for computing effect sizes: $V_{E1} = 0.0225$ [31].

Gold assessors used PRI for some topics and RND for others
Bronze assessors used PRI for every topic

Assessment condition pair	#topics	Mean κ	Statistically significantly outperforms ($\alpha = 0.05$)
a. PRI-Gold vs. PRI-BronzeT	25	0.5324	d,e
b. PRI-BronzeW vs PRI-BronzeT	50	0.4575	d,e
c. PRI-Gold vs. PRI-BronzeW	25	0.4445	d,e
d. RND-Gold vs. PRI-BronzeW	25	0.0395	
e. RND-Gold vs. PRI-BronzeT	25	0.0268	

TALK OUTLINE

1. History and the importance of the task
2. The WWW-4 test collection
3. Participants
4. Results: progress
5. Results: reproducibility
6. Conclusions and future work

Participating groups – only three ☹

WWW-3 run statistics. Besides these 16 runs, we have two organisers runs: ORG-TOPICDEV and baseline.

University of Tsukuba (top team from WWW-3)

Waseda University

Tsinghua University

Team	NEW	REV	REP	total	
KASYS	5	1	N/A	6	[Usuha+22]
SLWWW	4	N/A	1	5	[Ubukata+22]
THUIR	5	N/A	0	5	[Yang+22]
total	14	1	1	16	

Only Waseda worked on reproducibility ☹

SYSDESC of the top run from each team (in terms of Mean nDCG, gold qrels)

SLWWW-CO-REP-1 0.3686
Rep run of the KASYS system

KASYS-CO-REV-6 0.3682
Revival of KASYS-E-CO-NEW-1 at NTCIR-15 WWW-3

SOTA from WWW-3

THUIR-CO-NEW-2 0.3670
We first use BM25 to retrieve the top-100 documents of each query, and then use PROP to rerank the top-100 documents. In the training, we use all 280 queries in www1-3 dataset as the train set to fine-tune PROP and no validation set.

pre-training with representative words prediction

No real progress at WWW-4 (though the THUIR run is the only run that outperformed 5 other runs in Mean nDCG with bronze qrels)

TALK OUTLINE

1. History and the importance of the task
2. The WWW-4 test collection
3. Participants
4. Results: progress
5. Results: reproducibility
6. Conclusions and future work

Gold-based results (top 6 runs)

Run name	(a) Mean nDCG
SLWWW-CO-REP-1	0.3686
KASYS-CO-REV-6	0.3682
THUIR-CO-NEW-2	0.3670
SLWWW-CO-NEW-4	0.3650
THUIR-CO-NEW-1	0.3596
THUIR-CO-NEW-5	0.3405

Run name	(b) Mean Q
THUIR-CO-NEW-2	0.2944
THUIR-CO-NEW-1	0.2931
SLWWW-CO-NEW-4	0.2891
KASYS-CO-REV-6	0.2890
SLWWW-CO-REP-1	0.2886
SLWWW-CO-NEW-2	0.2718

Run name	(c) Mean nERR
THUIR-CO-NEW-2	0.5289
SLWWW-CO-NEW-3	0.5248
SLWWW-CO-NEW-2	0.5129
THUIR-CO-NEW-1	0.5102
SLWWW-CO-REP-1	0.5098
KASYS-CO-REV-6	0.5098

Run name	(d) Mean iRBU
SLWWW-CO-NEW-4	0.7986
SLWWW-CO-REP-1	0.7840
KASYS-CO-REV-6	0.7811
THUIR-CO-NEW-5	0.7545
THUIR-CO-NEW-2	0.7544
THUIR-CO-NEW-4	0.7510

None of the differences are statistically significant: no real progress

Gold-based results (top 6 runs)

Run name	(a) Mean nDCG
SLWWW-CO-REP-1	0.3686
KASYS-CO-REV-6	0.3682
THUIR-CO-NEW-2	0.3670
SLWWW-CO-NEW-4	0.3650
THUIR-CO-NEW-1	0.3596
THUIR-CO-NEW-5	0.3405

Run name	(b) Mean Q
THUIR-CO-NEW-2	0.2944
THUIR-CO-NEW-1	0.2931
SLWWW-CO-NEW-4	0.2891
KASYS-CO-REV-6	0.2890
SLWWW-CO-REP-1	0.2886
SLWWW-CO-NEW-2	0.2718

Run name	(c) Mean nERR
THUIR-CO-NEW-2	0.5289
SLWWW-CO-NEW-3	0.5248
SLWWW-CO-NEW-2	0.5129
THUIR-CO-NEW-1	0.5102
SLWWW-CO-REP-1	0.5098
KASYS-CO-REV-6	0.5098

Run name	(d) Mean iRBU
SLWWW-CO-NEW-4	0.7986
SLWWW-CO-REP-1	0.7840
KASYS-CO-REV-6	0.7811
THUIR-CO-NEW-5	0.7545
THUIR-CO-NEW-2	0.7544
THUIR-CO-NEW-4	0.7510

Reproduction looks successful at least in terms of mean effectiveness scores!

Bronze-All-based results (top 6 runs)

Successful in terms of nDCG – the only run that statistically significantly outperformed 5 other runs
(but delta over KASYS not statistically significant)

Run name	(a) Mean nDCG
THUIR-CO-NEW-2	0.6249
THUIR-CO-NEW-1	0.6111
KASYS-CO-REV-6	0.5931
SLWWW-CO-REP-1	0.5846
SLWWW-CO-NEW-4	0.5750
SLWWW-CO-NEW-2	0.5600

Run name	(b) Mean Q
THUIR-CO-NEW-2	0.5857
KASYS-CO-REV-6	0.5743
THUIR-CO-NEW-1	0.5691
SLWWW-CO-REP-1	0.5629
SLWWW-CO-NEW-4	0.5397
SLWWW-CO-NEW-2	0.5316

Run name	(c) Mean nERR
THUIR-CO-NEW-2	0.7967
THUIR-CO-NEW-1	0.7962
KASYS-CO-REV-6	0.7634
SLWWW-CO-REP-1	0.7537
SLWWW-CO-NEW-2	0.7330
SLWWW-CO-NEW-3	0.7242

Run name	(d) Mean iRBU
KASYS-CO-REV-6	0.9424
SLWWW-CO-REP-1	0.9397
SLWWW-CO-NEW-2	0.9244
SLWWW-CO-NEW-4	0.9213
SLWWW-CO-NEW-3	0.9192
THUIR-CO-NEW-1	0.9106

Bronze-All-based results (top 6 runs)

Run name	(a) Mean nDCG
THUIR-CO-NEW-2	0.6249
THUIR-CO-NEW-1	0.6111
KASYS-CO-REV-6	0.5931
SLWWW-CO-REP-1	0.5846
SLWWW-CO-NEW-4	0.5750
SLWWW-CO-NEW-2	0.5600

Run name	(b) Mean Q
THUIR-CO-NEW-2	0.5857
KASYS-CO-REV-6	0.5743
THUIR-CO-NEW-1	0.5691
SLWWW-CO-REP-1	0.5629
SLWWW-CO-NEW-4	0.5397
SLWWW-CO-NEW-2	0.5316

Run name	(c) Mean nERR
THUIR-CO-NEW-2	0.7967
THUIR-CO-NEW-1	0.7962
KASYS-CO-REV-6	0.7634
SLWWW-CO-REP-1	0.7537
SLWWW-CO-NEW-2	0.7330
SLWWW-CO-NEW-3	0.7242

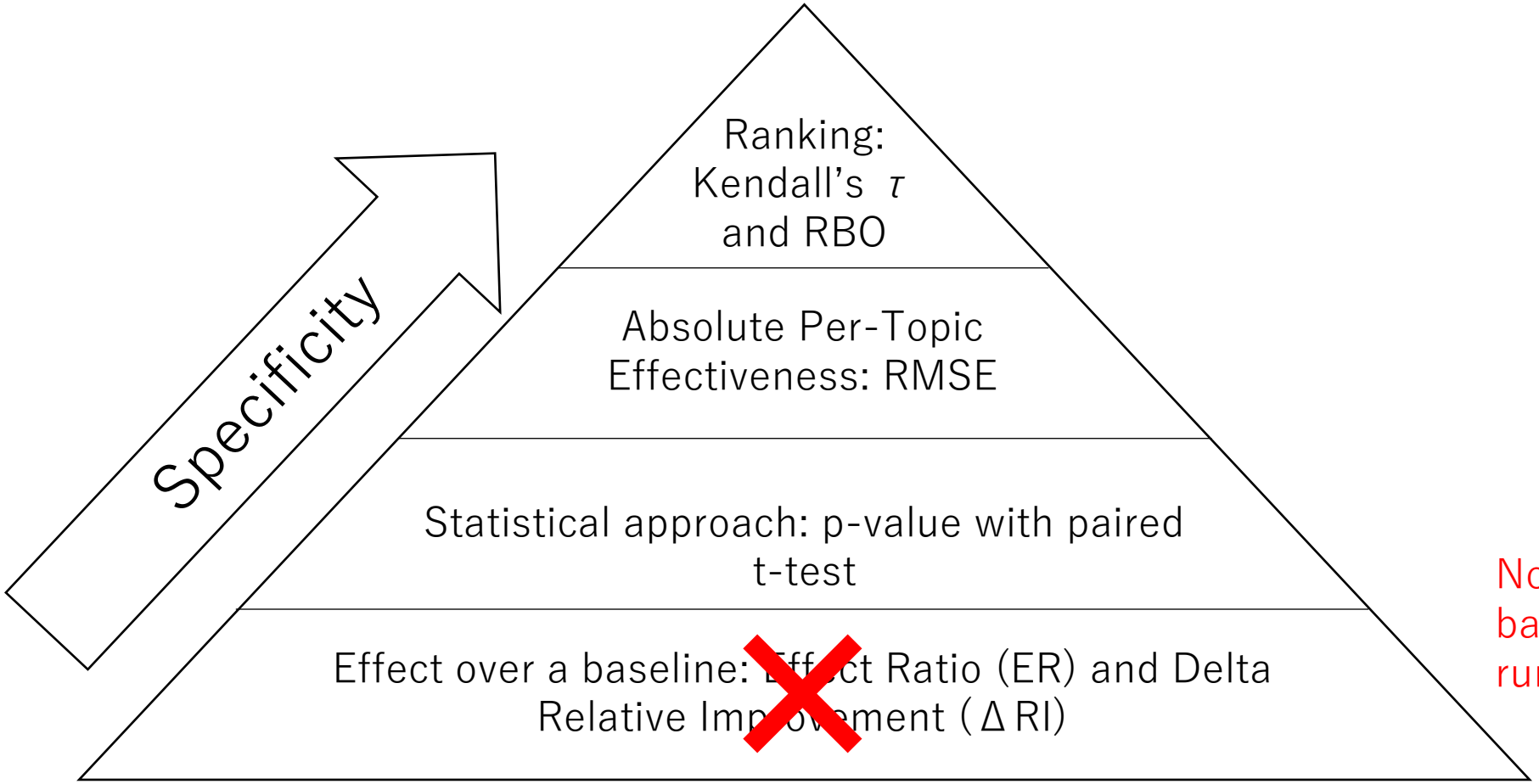
Run name	(d) Mean iRBU
KASYS-CO-REV-6	0.9424
SLWWW-CO-REP-1	0.9397
SLWWW-CO-NEW-2	0.9244
SLWWW-CO-NEW-4	0.9213
SLWWW-CO-NEW-3	0.9192
THUIR-CO-NEW-1	0.9106

Reproduction looks successful at least in terms of mean effectiveness scores!

TALK OUTLINE

1. History and the importance of the task
2. The WWW-4 test collection
3. Participants
4. Results: progress
5. Results: reproducibility
6. Conclusions and future work

Reproducibility Measures [Breuer+20]

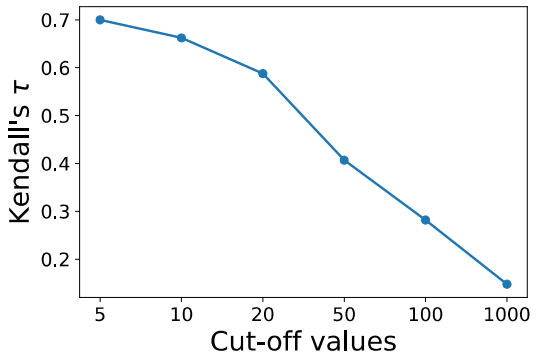


No reproduced baseline/advanced runs

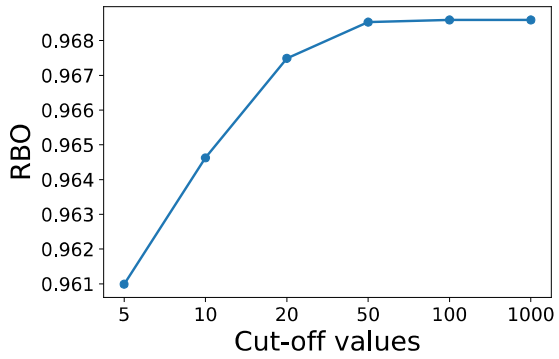
Reproducibility Results

- Original run: KASYS-CO-REV-6 and reproduced run: SLWWW-CO-REP-1;
- Successful reproducibility experiment!

Most specific measures, but higher scores than similar experiments



(a) Kendall's τ Union (KTU)



(b) Rank Biased Overlap (RBO)

Least specific measures, still better scores than similar experiments

	RMSE	p-values
nDCG	0.0253	0.9109
Q	0.0277	0.9098
nERR	0.0337	0.9944
iRBU	0.0271	0.4612

(c) Reproducibility effectiveness measures

TALK OUTLINE

1. History and the importance of the task
2. The WWW-4 test collection
3. Participants
4. Results: progress
5. Results: reproducibility
6. Conclusions and future work

Conclusions

- Some concerns

- Few participating teams (3, including two organiser institutions)
- Bronze assessments substantially different from Gold ones (assessments depend heavily on PRI vs RND) [Detailed to be reported elsewhere]

- Progress

No real progress: THUIR performs well but the delta over the KASYS REV run not statistically significant

- Reproducibility

SLWWW successful with respect to all measures.

For NTCIR-17

We plan to propose a new group-fair web search task using the Chuweb21 corpus and a new evaluation framework [Sakai+22arxiv-fair] .

Detail TBD.

References (1)

[Bailey+08] P. Bailey et al.: Relevance Assessment: Are Judges Exchangeable and Does It Matter? ACM SIGIR 2008.

[Breuer+20] T. Breuer et al.: How to Measure the Reproducibility of System-oriented IR Experiments, Proceedings of ACM SIGIR 2020.

http://waseda.box.com/sigir2020reproducibility (preprint)

[Luo+17] C. Luo et al.: Overview of the NTCIR-13 We Want Web Task, Proceedings of NTCIR-13, 2017. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-NTCIR13-OV-WWW-LuoC.pdf

[Mao+19] J. Mao et al.: Overview of the NTCIR-14 We Want Web Task, Proceedings of NTCIR-14, 2019.

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-WWW-MaoJ.pdf

[Sakai16IRJ] T. Sakai: Topic Set Size Design, Information Retrieval Journal, 19(3), 2016. http://link.springer.com/content/pdf/10.1007%2Fs10791-015-9273-z.pdf (open access)

[Sakai18book] T. Sakai: Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power, Springer, 2018.

References (2)

[Sakai+19CENTRE] T. Sakai et al.: Overview of the NTCIR-14 CENTRE Task, Proceedings of NTCIR-14, 2019.

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-CENTRE-SakaiT.pdf

[Sakai+20WWW3] T. Sakai et al.: Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task, Proceedings of NTCIR-15, 2020.

http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/01-NTCIR15-OV-WWW-SakaiT.pdf

[Sakai+22TOIS] T. Sakai et al.: Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents?, ACM TOIS 40(4) https://dl.acm.org/doi/pdf/10.1145/3494833 (open access)

[Sakai+22arxiv-fair] T. Sakai et al.: Versatile Framework for Evaluating Ranked Lists in terms of Group Fairness and Relevance, 2022. http://arxiv.org/abs/2204.00280

[Shinden+20] K. Shinden et al.: KASYS at the NTCIR-15 WWW-3 Task, Proceedings of NTCIR-15, 2020.

https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/02-NTCIR15-WWW-ShindenK.pdf

[Ubukata+22] Y. Ubukata et al.: SLWWW at the NTCIR-16 WWW-4 Task, Proceedings of NTCIR-16, 2022.

[Usuha+22] K. Usuha et al.: KASYS at the NTCIR-16 WWW-4 Task, Proceedings of NTCIR-16, 2022.

[Yang+22] S. Yang et al.: THUIR at the NTCIR-16 WWW-4 Task, Proceedings of NTCIR-16, 2022.