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

Tetsuya Sakai, Sijie Tao
Waseda University, Japan
tetsuyasakai@acm.org
tsjmailbox@ruri.waseda.jp

Zhumin Chu
Tsinghua University, P.R.C.
chuzm19@mails.tsinghua.edu.cn

Maria Maistro
University of Copenhagen, Denmark
mm@di.ku.dk

Yujing Li, Nuo Chen
Waseda University, Japan
liyujing@ruri.waseda.jp
pleviumtan@toki.waseda.jp

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Junjie Wang
Waseda University, Japan
wjj1020181822@toki.waseda.jp

Ian Soboroff
NIST, USA
ian.soboroff@nist.gov

Yiqun Liu
Tsinghua University, P.R.C.
yiqunliu@tsinghua.edu.cn

ABSTRACT

This is an overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) task, the fourth round of an evaluation series that aims to quantify the progress and reproducibility of web search algorithms in offline ad hoc retrieval settings. For WWW-4, we introduced a new English web corpus, which we named Chuweb21. Moreover, in addition to bronze relevance assessments (i.e., those given by assessors who are neither topic creators nor topic experts), we collected gold relevance assessments (i.e., those given by topic creators). We received 18 runs from 4 teams, including two runs from the organiser team. We describe the task, data, evaluation measures, and report on the official evaluation results.

1 INTRODUCTION

This paper presents an overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) Task,[1] the fourth round of an evaluation series that aims to quantify the progress and reproducibility of web search algorithms in offline ad hoc retrieval settings.

The We Want Web task (WWW-1) was launched at NTCIR-13 in 2017 [9], in response to the termination of the web track at TREC 2014.[2] WWW-1 received a total of 32 runs from five teams for the Chinese and English subtasks.

The NTCIR-14 WWW-2 task was held in 2019 [10], and received 31 runs from five teams for the same subtasks. Also, NTCIR-14 hosted the first CENTRE (CLEF/NTCIR/TREC Reproducibility) task [15], which was a "metatask" that spanned CLEF, TREC, and NTCIR [5–7, 21].

As NTCIR-14 CENTRE attracted only one participating team, for NTCIR-15 the WWW and CENTRE joined forces to organise the We Want Web with CENTRE (WWW-3) task. WWW-3 received a total of 48 runs from nine teams for the Chinese and English subtasks [17].

As NTCIR-15 WWW-3 attracted only two participating teams for the Chinese subtask, we focused on English web search at WWW-4.

[1] http://sakailab.com/www4/
[2] https://twitter.com/djoerd/status/536128465276530688

Table 1: WWW-4 timeline (time zone: UTC+9).

November 1, 2021	Topics released
November 4, 2021	BM25 Baseline run released
December 15, 2021	Run submissions due
December 21-January 20, 2021	Relevance assessments
February 1, 2021	Evaluation results released

Table 2: WWW-3 run statistics. Besides these 16 runs, we have two organisers runs: `ORG-TOPICDEV` and `baseline`.

Team	NEW	REV	REP	total
KASYS	5	1	N/A	6
SLWWW	4	N/A	1	5
THUIR	5	N/A	0	5
total	14	1	1	16

Moreover, instead of using Clueweb12-B13[3] again, we constructed a new English web corpus for the task, which we call Chuweb21. Another new feature of WWW-4 is that, in addition to *bronze* relevance assessments (i.e., those given by assessors who are neither topic creators nor topic experts), we collected *gold* relevance assessments (i.e., those given by topic creators) [1].

Table 1 shows the timeline of the WWW-4 task. Table 2 shows names of the participating teams and the number of runs submitted to the task. It is unfortunate that there was no participation besides University of Tsukuba (KASYS) whose runs served as the target of reproducibility experiments and two teams from the organisers' affiliations: SLWWW (Waseda University) and THUIR (Tsinghua University). As the table shows, we had three *run types*: NEW, REV, and REP, as we shall discuss in Section 2.

The remainder of this paper is organised as follows. Section 2 describes the task, and Section 3 describes the WWW-4 data. Section 4 describes the evaluation measures we use for quantifying retrieval effectiveness and reproducibility. Sections 5 and 6 report

[3] http://lemurproject.org/clueweb12/

on the retrieval effectiveness and reproducibility results, respectively. Finally, Section 7 concludes this paper.

2 TASK

The WWW-4 task is an ad hoc English web search task. Three types of runs were allowed.

REV (revived) run This is a run kindly provided by KASYS (University of Tsukuba). At the NTCIR-15 WWW-3 English subtask [17], their run `KASYS-E-CO-NEW-1` [20] was the top performer and therefore we treat this run as the SOTA (state-of-the-art) from WWW-3. It uses a BERT-based approach proposed in Yilmaz *et al.* [27]. We asked KASYS to use the exact algorithm used at WWW-3 to process the new WWW-4 topics to from a revived run. The resultant WWW-4 run is called **KASYS-CO-REV-6** [23].

NEW runs These runs are the regular adhoc runs designed to advance the SOTA. If a NEW run substantially outperforms the above REV run on the WWW-4 test collection, that suggests that we have a new SOTA. In this way, we can examine if we are seeing real technological progress.

REP runs These runs aim at reproducing what KASYS did at WWW-3 to generate the WWW-3 run `KASYS-E-CO-NEW-1`. Since the same algorithm was used to generate the WWW-4 run **KASYS-CO-REV-6**, we can discuss the reproducibility of the KASYS method by simply comparing the REP runs with `KASYS-CO-REV-6` on the new WWW-4 test collection. Unfortunately, however, only one REP run was submitted to the WWW-4 task [22].

Each team was allowed to submit up to five NEW/REP runs. KASYS submitted their REV runs as their sixth run.

Compared to the previous WWW tasks, WWW-4 is different in two ways. First, we use a new target English web corpus, Chuweb21, which we constructed as described in Section 3.1. Second, this time we hired *gold assessors* [1], that is, we collected the relevance assessments from the *topic creators*. In fact, the first seven authors of this paper served as the gold assessors! To maintain consistency with the previous WWW tasks, we also hired *bronze assessors* (i.e., those who are neither topic creators nor topic experts) [1] to construct an alternative version of qrels. The details are given in Sections 3.

3 DATA

3.1 The Chuweb21 Corpus

This corpus was named after one of the task organisers, Zhumin *Chu*[4]. Chuweb21 was generated based on the April 2021 block of Common Crawl dataset[5]. As the complete data block is too large to conduct the downstream data cleaning and indexing work, we sampled the web pages crawled between 2021-04-10 10:58:31 and 2021-04-11 11:56:10. This part of the data (denoted as subdata in the following content) contains 3,402,457 different domains and 858,616,203 different web pages, which occupies the space of 5.66TiB.

We grouped the subdata by root domain and found that the top-10 frequent root domains include .com (43.7%), .org (5.7%), .ru (5.2%), .de (4.2%), .net (3.7%), .uk (2.3%), .jp (1.9%), .fr (1.8%), .it (1.8%) and .nl (1.6%). To avoid the inclusion of many non-English contents in the corpus, we only retained web pages under the .com, .org and .net domains. Specifically, we adopted the following constraints to filter for useful html pages:

- The root domain must be one of the .com, .org, .net domains;
- The WARC-Type must be "response" (actually the web pages as we need);
- The character length of HTML content must be larger than 1,000;
- The probability that the document content belongs to English is greater than 0.99.

We used 4 servers (32 processes) running for about two weeks to complete the aforementioned data cleaning jobs. The final Chuweb21 corpus contains 82,451,337 (9.6% of the subdata) HTMLs or 1.69 TiB of compressed content. Similar to the ClueWeb12-B13 data, Chuweb21 has been reorganized with the "warc.tar" format to facilitate the users. The dataset is already accessible to the public for academic usage.[6]

3.2 Topics

3.2.1 Topic Set Size. As with the previous WWW rounds, we decided on the number of topics to create using Sakai's topic set size design tool for comparing $m = 2$ systems with a t-test (or ANOVA) [13].[7]

From the 160×36 topic-by-run score matrices from the WWW-3 English subtask, we obtained the variance estimates of the four evaluation measures [17] as residual variances from two-way ANOVA (V_{E2}) [13, p. 120]: nERR (normalised Expected Reciprocal Rank) [12] had the largest variance (0.0284) and iRBU (intentwise Rank-Biased Utility) [18] had the smallest variance (0.00716). Under Cohen's five-eighty convention ($\alpha = 0.05, \beta = 0.20$),[8] the topic set size design tool tells us the following:

- We need 44 topics for a *minimum detectable difference* [13] (for 80% statistical power) of 0.1 in terms of nERR;
- We need 45 topics for a minimum detectable difference (for 80% statistical power) of 0.05 in terms of iRBU.

Based on the above estimates, we decided to construct 50 topics for the WWW-4 task.

3.2.2 Topic Creation. There was an indication that the WWW-3 bronze English relevance assessments contained some noise [11]. Moreover, although we created multiple versions of qrels files at WWW-3 using two different document ordering strategies for the relevance assessors (RND and PRI [16]), it was not possible to say which versions were *correct* as all of the assessors involved were bronze assessors. We therefore decided to construct a gold qrels file along with a bronze one. Gold-relevant documents are what the topic creators want, and therefore can be treated as the right answers, although they are not immune to human errors.

In light of the above situation, the first seven authors of this paper volunteered to serve as the topic creators *and* the gold relevance

[4]This is not a misspelling of Clueweb.
[5]https://commoncrawl.org/2021/04/april-2021-crawl-archive-now-available/
[6]https://drive.google.com/drive/folders/11hi_R6cSIHEZx3QwyG5KQjgRVmxXhWta?usp=sharing
[7]http://www.f.waseda.jp/tetsuya/samplesizeANOVA2.xlsx
[8]This means we want the Type I and Type II error probabilities to be 5% and 20%, respectively.

```
<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>

<query>
<qid>0202</qid>
<content>New Orleans restaurants</content>
<description>Tell me about good restaurants in New Orleans.</description>
</query>
```

Figure 1: The top part of the WWW-4 topic file.

assessors. Each of the seven organisers tried to create realistic topics, i.e., those based on their actual information needs. The first author of this paper was responsible for creating eight topics and providing gold assessments for them; similarly, the other six authors each handled seven topics.

To develop the topics, the seven organisers used a browser-based topic development tool to conduct some pilot searches on the Chuweb21 corpus to ensure that there is at least one relevant document. They were allowed to formulate and reformulate their own queries.

The WWW-4 test topic file is publicly available.[9] Figure 1 shows the top part of this file.

3.2.3 Topic Set File. The `content` field represents the Assessor query that the topic creator is likely to enter, and the `description` field concisely describes the topic creator's information need. If a run file name contains "CO," that means only the `content` field was used as the input to the system; if it contains "CD," that means both `content` and `description` fields were utilised.

3.3 Organisers' Runs

During topic development, the topic creators (i.e., seven organisers) identified at least one relevant document for each topic. We created a "manual" run from these documents, which we call `ORG-TOPICDEV`. This run contains only 97 topic-document pairs for the 50 test topics, and the documents for each topic are not necessarily sorted by perceived relevance.

We also provided an Anserini-based [25] vanilla BM25 baseline run together with the contents of the retrieved documents to participants, so that the participants can optionally rerank the baseline to produce their own runs. This run is a CO (content-only) run, and is simply called `baseline`.

3.4 Runs and Pool Files

Table 3 shows the run names and the system descriptions of the 16 runs submitted by KASYS [23], SLWWW [22], and THUIR [26]. Note that `SLWWW-CO-REP-1` is the only REP run. That is, this is the only run that tackled the reproducibility problem.

From the 18 runs (16 participant runs plus the 2 organiser runs), we formed a depth-60 pool for each topic, and obtained a total of 10,333 topic-document pairs to judge (206.7 docs per topic on average). We created the following two types of pool file for each topic:

- **RND** The pooled documents are randomly ordered;
- **PRI** The pooled documents are ordered by pseudorelevance, based on the number of runs that returned that document and the ranks of that document in those runs, using the `NTCIRPOOL` script [16].[10]

Elsewhere, we plan to report on a study that compares RND-based and PRI-based relevance assessments from gold assessors, to follow up on the work of Sakai, Tao, and Zeng [16] that compared RND and PRI under the bronze setting.

3.5 Relevance Assessments

For evaluating the WWW-4 runs, we constructed 2 versions of qrels (relevance assessment) files: the Gold version and the Bronze-All version.

The Gold file is based on the relevance assessments given by the topic creators, i.e., the first seven authors of this paper. By definition, each topic was judged by exactly one assessor. For each topic, each gold assessor processed either the RND pool or the PRI pool assigned at random.

The Bronze-All file is actually the result of merging two different versions of relevance assessments given by *bronze assessors* (i.e., those who are neither topic creators nor topic experts [1]). At Waseda University, Japan, five English-course computer science students were hired as bronze assessors, as in previous English WWW subtasks. At Tsinghua University, China, five more bronze assessors were hired through a Chinese company. All 10 bronze assessors used the PRI pool files, not the RND pool files; every topic was assessed by one Waseda assessor and one Tsinghua assessor; the topics were assigned at random so that each bronze assessor handled exactly 10 topics.

All gold and bronze assessors used the PLY interface [14, Figure 4] for the relevance assessments, and each document was labeled either highly relevant, relevant, nonrelevant, or *error* [17]. These

[9] https://waseda.box.com/www4topicsxml

[10] https://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html

Table 3: 16 participant runs and their system description (SYSDESC) fields.

KASYS-CD-NEW-1	On the topic that include proper noun, We rerank top 10 from documents retrieved from bm25. Other then that, We use the ranking from bm25 directly. For the re-ranking method, we use BERT fine-tuned on the SQuAD 2.0 dataset. For question generation, we translate www4 topics to question manually.
KASYS-CD-NEW-3	We rerank top 10 from documents retrieved from bm25. For the re-ranking method, we use BERT fine-tuned on the SQuAD 2.0 dataset. For question generation, we translate www4 topics to question manually.
KASYS-CD-NEW-5	We rerank top 100 from documents retrieved from bm25 on the topic that include proper noun. For the re-ranking method, we use BERT fine-tuned on the SQuAD 2.0 dataset. For question generation, we translate www4 topics to question manually.
KASYS-CO-NEW-2	On the topic that include proper noun, We rerank top 10 from documents retrieved from bm25. For the re-ranking method, we use BERT fine-tuned on the SQuAD 2.0 dataset. For the question generation, we use Encoder-Decoder neural machine translation model.
KASYS-CO-NEW-4	We rerank top 10 from documents retrieved from bm25. For the re-ranking method, we use BERT fine-tuned on the SQuAD 2.0 dataset. For the question generation, we use Encoder-Decoder neural machine translation model.
KASYS-CO-REV-6	Revival of KASYS-E-CO-NEW-1 at NTCIR-15 WWW-3
SLWWW-CO-NEW-2	COIL, contextualized exact lexical match. Split into chunks. Smaller corpus
SLWWW-CO-NEW-3	COIL, contextualized exact lexical match. Split into chunks. Bigger corpus
SLWWW-CO-NEW-4	COIL, contextualized exact lexical match. No splitting. Smaller corpus
SLWWW-CO-NEW-5	Reproduction of PARADE full transformer based model
SLWWW-CO-REP-1	Rep run of the KASYS system
THUIR-CO-NEW-1	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 206 queries in www1-3 dataset (280 queries totally) as train set to fine-tune PROP, and the remaining 52 queries as validation set.
THUIR-CO-NEW-2	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.
THUIR-CO-NEW-3	We first use BM25 to retrieve the top-100 documents of each query, and then use BERT-Prompt to rerank the top-100 documents. BERT-Prompt is trained by cloze prompt method based on BERT on 206 queries in www1-3 dataset (280 queries totally) as train set, and the remaining 52 queries as validation set.
THUIR-CO-NEW-4	The LambdaMART model implemented by Ranklib. We adopt MQ2007&2008, www1-3 as our training and development datasets. The features contain TF, IDF, TF*IDF, DL, BM25, LM.ABS, LM.DIR, LM.JM in the fields of content, title, url, anchor text.
THUIR-CO-NEW-5	The Coordinate Ascent model implemented by Ranklib. We adopt MQ2007&2008, www1-3 as our training and development datasets. The features contain TF, IDF, TF*IDF, DL, BM25, LM.ABS, LM.DIR, LM.JM in the fields of content, title, url, anchor text.

labels were mapped to scores of 2, 1, 0, and 0, respectively. Thus the Gold qrels file contains 3-point (0,1,2) relevance levels. Similarly, from Waseda's and Tsinghua's bronze assessments, we obtained a 3-point relevance level file, respectively, which we refer to as Bronze-Waseda and Bronze-Tsinghua. Finally, for computing the official evaluation scores for the WWW-4 task, we created a "Bronze-All" file, by adding the relevance scores from Bronze-Waseda and Bronze-Tsinghua and forming 4-point relevance levels.

We report on retrieval effectiveness based on the Gold file and that based on the Bronze-All file separately: the former represents results based on "correct" relevance assessments as defined by the topic creators; the latter is based on views of multiple bronze assessors and are more similar to previous WWW evaluation settings.

Table 4 shows the distribution of relevance labels for each version of the qrels. Table 5 shows the mean inter-assessor agreement for each pair of qrels files. Note that each qrels file consists of labels from multiple assessors. Table 6 compares the mean κ's shown in Table 5 in terms of statistical significance. These two tables show that the Gold-Bronze agreements are substantially and statistically highly significantly lower than the Bronze-Bronze (i.e., Waseda-Tsinghua) agreements.

Tables 7 and 8 examine the per-topic κ's at the individual assessor level. For example, the "Gold01" row of Table 7 compares the labels of Gold01 (the first author of this paper) with those of Waseda and Tsinghua assessors, and shows the mean κ over the eight topics that he was in charge of. It can be observed that the Gold-Tsinghua agreements are higher than the Gold-Waseda agreements for Gold01, Gold 02, Gold03, and Gold07 on average, but not for Gold04, Gold05, and Gold06. On the other hand, in Table8, the Waseda-Tsinghua (i.e., Bronze-Bronze) agreements are substantially higher than the Bronze-Gold agreements for every Bronze assessor.

In summary, according to the WWW4 data, different versions of bronze relevance assessments are relatively similar to each other, but they are substantially different from gold relevance assessments.

Table 4: Distribution of pooled documents over the relevance levels in the gold and bronze qrels files.

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

Table 5: 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

Table 6: 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 $V_{E2} = 0.0345$ [13].

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	$p \approx 0, ES_{E2} = 1.160$

This suggests that system evaluations based on gold and bronze relevance assessments may also be substantially different. Section 5 discusses the actual evaluation results based on Gold and Bronze-All files.

A further investigation showed that the low Gold-Bronze agreements were largely due to the differences in the document presentation order for the assessors: that is, the agreements were extremely low when the Gold assessors used the RND pool files while the Bronze assessors used the PRI pool files. (Recall that all Bronze assessments are based on PRI bools.) Put another way, the disagreements reflect the differences between RND and PRI document ordering strategies rather than the differences between Gold and Bronze assessors. Details will be reported elsewhere.

Table 7: Mean per-topic inter-assessor agreement for each gold assessor in terms of quadratic weighted Cohen's κ ($n = 8$ topics for Gold01; $n = 7$ topics for the others). For example, the labels of Gold01 are compared with those given by the Waseda and Tsinghua bronze assessors.

sassessor	mean κ (with Waseda)	mean κ (with Tsinghua)
Gold01	0.218	0.306
Gold02	0.258	0.343
Gold03	0.226	0.283
Gold04	0.326	0.305
Gold05	0.258	0.221
Gold06	0.154	0.145
Gold07	0.258	0.350

Table 8: Mean per-topic inter-assessor agreement for each bronze assessor in terms of quadratic weighted Cohen's κ ($n = 10$ topics). For example, the labels of Waseda01 are compared with those given by the Gold and Tsinghua assessors.

assessor	mean κ (with Gold)	mean κ (with Tsinghua)
Waseda01	0.214	0.450
Waseda02	0.226	0.459
Waseda03	0.247	0.444
Waseda04	0.166	0.428
Waseda05	0.358	0.507
assessor	mean κ (with Gold)	mean κ (with Waseda)
Tsinghua06	0.352	0.476
Tsinghua07	0.303	0.485
Tsinghua08	0.241	0.395
Tsinghua09	0.201	0.500
Tsinghua10	0.301	0.432

4 EVALUATION MEASURES

4.1 Effectiveness Measures

Following the NTCIR-15 WWW-3 task, we use nDCG@10 (MSnDCG@10), Q@10, and nERR@10 [12], and iRBU@10 (with $p = 0.99$) [18] to evaluate the runs in terms of retrieval effectiveness, using the `NTCIREVAL` tool with a linear gain value setting.[11] According to the experiments reported by Sakai and Zeng [19], nDCG and iRBU outperformed other measures in terms of agreement with users' SERP preferences.

4.2 CENTRE Evaluation Measures

In this round of CENTRE, we quantify reproducibility as suggested in Breuer et al. [2][12]. We cannot quantify replicability, since this would require runs generated by the same system on different test collections. Details on reproducibility measures are presented in the following.

First, we evaluate whether the reproduced run can retrieve the same exact ranking of documents retrieved by the original run.

[11] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html

[12] Note that reproducibility terminology has changed since 2020. In this paper we adopt the updated terminology: https://www.acm.org/publications/policies/artifact-review-badging

We compute Kendall's τ union (KTU) [6, 7], which compares the relative order of documents by computing Kendall's τ with respect to the union of the original and replicated rankings. This is necessary since Kendall's τ is defined for permutations of items from the same list [8], while reproduced runs can rank documents that were not retrieved by the original run.

Let r be the original run and r' the reproduced run, r_j denotes the ranked list of document ids for topic j for the original run and similarly r'_j is the ranked list of documents for the reproduced run. KTU is computed as follows:

(1) consider the union of r_j and r'_j by removing duplicate entries;
(2) consider the rank position of documents from the union in r_j and r'_j;
(3) compute Kendall's τ between these two lists of rank positions.

Kendall's τ at step 3 is computed as follows:

$$\mathrm{KTU}_j(r, r') = \tau_j(l, l') = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}} \tag{1}$$

where l and l' are the list of rank positions obtained at step 2, P is the total number of concordant pairs, Q is the total number of discordant pairs, U and V are the number of ties, in l and l' respectively.

As reported in previous work [2, 6, 7], Kendall's τ can be too strict when comparing 2 lists of documents and is not top heavy. Therefore, in addition to Kendall's τ we also compute Rank-Biased Overlap (RBO) [24]. RBO for the j-th topic is computed as follows:

$$\mathrm{RBO}_j(r, r') = (1 - \phi) \sum_{i=1}^{\infty} \phi^{i-1} \cdot O_i \tag{2}$$

where $\phi \in [0, 1]$ is a parameter to adjust the measure top-heaviness: the smaller ϕ, the more top-weighted the measure; and O_i is the proportion of overlap up to rank i, which is defined as the cardinality of the intersection between r_j and r'_j up to i divided by i. Therefore, RBO accounts for the overlap of two rankings and discounts the overlap while moving towards the end of the ranking, since it is more likely for two rankings to have a greater overlap when many rank positions are considered.

In addition to differences in how runs rank documents, we consider the per topic differences in effectiveness scores. Let $M_j(r)$ denote the effectiveness score for topic j of the original run. Similarly, let $M_j(r')$ denote the corresponding score of the reproduced run. Following CENTRE@CLEF [6, 7], the Root Mean Square Error (RMSE) for replicating absolute per-topic differences is computed as follows.

$$RMSE = \sqrt{\frac{1}{T} \sum_{j=1}^{T} (M_j(r') - M_j(r))^2}\,, \tag{3}$$

where T is the number of topics. Note that *RMSE* focuses on the per-topic measure scores rather than the actual documents retrieved.

Finally, we compare the original and reproduced runs from a statistical point of view [2]. We run a two tailed paired t-test between $M_j(r)$ and $M_j(r')$, for $j \in \{1, \ldots, T\}$. The p-value returned by the t-test informs on the success of the reproducibility experiment: the smaller the p-value, the stronger the evidence that r and r' are statistically significantly different.

5 RETRIEVAL EFFECTIVENESS RESULTS

Table 9 shows the official effectiveness results based on the Gold relevance assessments. Table 10 shows the results of statistical significance tests for Table 9: note that none of the differences in terms of Mean nERR are statistically significant. From the statistical significance point of view, all of the WWW-4 runs except `ORG-TOPICDEV` are tied even in terms of the other three measures, as none of the run pairs are statistically significantly different except those that involve `ORG-TOPICDEV`. In other words, from the Gold-based results, all we know for certain is that `ORG-TOPICDEV` substantially underperform the others. This is not surprising, since this run contains only a few documents per topic (See Section 3.3).

The following are probably worth noting from Table 9, however.

- None of the runs substantially outperform the REV run, which suggests that we are not seeing any substantial technological advance in this round of the task;
- The only REP run (`SLWWW-CO-REP-1`) performs very similarly to the REV run in terms of all four evaluation measures, which suggests that the reproducibility effort may be successful to some degree (see Section 6).

Table 11 shows the official effectiveness results based on the Bronze-All relevance assessments. Tables 12 and 13 show the results of statistical significance tests for Table 11. It can be observed that `THUIR-CO-NEW-2` is quite successful: this is the only run that statistically significantly outperform five other runs in terms of Mean nDCG (Table 12(a)). However, we cannot say with confidence that this run is now the new SOTA since the difference between `THUIR-CO-NEW-2` and **KASYS-CO-REV-6** is not statistically significant. Note also that `THUIR-CO-NEW-2` is ranked first in the Gold Mean Q ranking (Table 9(b)).

Regarding reproducibility, the trend is similar to the Gold-based results in that `SLWWW-CO-REP-1` performs very similarly to **KASYS-CO-REV-6** in terms of all four evaluation measures. Section 6 discusses reproducibility in more detail.

Table 14 compares the pairs of run rankings in terms of Kendall's τ. Part (a) compares the Gold-based rankings with different evaluation measures; Part (b) compares the Bronze-All-based rankings with different evaluation measures; and Part (c) compares the Gold-based and Bronze-All-based rankings for each evaluation measure. Parts (a) and (b) shows that nDCG and Q are very highly correlated, while the correlation between nERR and iRBU is relatively low. Part (c) shows that while the Q-based run ranking is the most robust when Gold assessments are replaced with Bronze-All ones, nERR's run ranking changes completely when this is done: the 95%CI shows that the correlation between the Gold nERR-based ranking and the Bronze-All nERR-based ranking is not statistically significant. Assuming that the Gold-based results are correct, the above result suggests that (n)ERR-based evaluation with bronze assessors should be interpreted with caution.

As we have mentioned earlier, it appears that the Gold-Bronze disagreements are largely due to the use of different document ordering strategies (PRI and RND). We shall report on further findings elsewhere.

Table 9: Official results based on the Gold file (mean over the 50 WWW-4 test topics).

Run name	(a) Mean nDCG	Run name	(b) Mean Q
SLWWW-CO-REP-1	0.3686	THUIR-CO-NEW-2	0.2944
KASYS-CO-REV-6	0.3682	THUIR-CO-NEW-1	0.2931
THUIR-CO-NEW-2	0.3670	SLWWW-CO-NEW-4	0.2891
SLWWW-CO-NEW-4	0.3650	**KASYS-CO-REV-6**	0.2890
THUIR-CO-NEW-1	0.3596	SLWWW-CO-REP-1	0.2886
THUIR-CO-NEW-5	0.3405	SLWWW-CO-NEW-2	0.2718
SLWWW-CO-NEW-2	0.3398	SLWWW-CO-NEW-3	0.2670
SLWWW-CO-NEW-3	0.3388	THUIR-CO-NEW-5	0.2667
KASYS-CO-NEW-4	0.3312	KASYS-CO-NEW-4	0.2566
KASYS-CD-NEW-1	0.3294	KASYS-CD-NEW-1	0.2548
KASYS-CD-NEW-3	0.3280	KASYS-CO-NEW-2	0.2539
KASYS-CO-NEW-2	0.3273	SLWWW-CO-NEW-5	0.2538
THUIR-CO-NEW-3	0.3222	KASYS-CD-NEW-3	0.2538
baseline	0.3205	THUIR-CO-NEW-3	0.2494
SLWWW-CO-NEW-5	0.3193	baseline	0.2473
THUIR-CO-NEW-4	0.3094	THUIR-CO-NEW-4	0.2288
KASYS-CD-NEW-5	0.2879	KASYS-CD-NEW-5	0.2086
ORG-TOPICDEV	0.1626	ORG-TOPICDEV	0.0857
Run name	(c) Mean nERR	Run name	(d) Mean iRBU
THUIR-CO-NEW-2	0.5289	SLWWW-CO-NEW-4	0.7986
SLWWW-CO-NEW-3	0.5248	SLWWW-CO-REP-1	0.7840
SLWWW-CO-NEW-2	0.5129	**KASYS-CO-REV-6**	0.7811
THUIR-CO-NEW-1	0.5102	THUIR-CO-NEW-5	0.7545
SLWWW-CO-REP-1	0.5098	THUIR-CO-NEW-2	0.7544
KASYS-CO-REV-6	0.5098	THUIR-CO-NEW-4	0.7510
SLWWW-CO-NEW-4	0.5052	THUIR-CO-NEW-1	0.7449
KASYS-CO-NEW-4	0.4971	SLWWW-CO-NEW-3	0.7368
THUIR-CO-NEW-5	0.4783	SLWWW-CO-NEW-2	0.7358
KASYS-CD-NEW-1	0.4769	KASYS-CD-NEW-1	0.7351
KASYS-CO-NEW-2	0.4747	KASYS-CD-NEW-3	0.7348
KASYS-CD-NEW-3	0.4733	KASYS-CO-NEW-4	0.7346
THUIR-CO-NEW-4	0.4672	KASYS-CO-NEW-2	0.7343
KASYS-CD-NEW-5	0.4580	baseline	0.7327
baseline	0.4541	KASYS-CD-NEW-5	0.7206
ORG-TOPICDEV	0.4510	THUIR-CO-NEW-3	0.7166
SLWWW-CO-NEW-5	0.4288	SLWWW-CO-NEW-5	0.7133
THUIR-CO-NEW-3	0.4281	ORG-TOPICDEV	0.4526

6 CENTRE: REPRODUCIBILITY RESULTS

Reproducibility is evaluated by comparing `KASYS-CO-REV-6` as the original run and `SLWWW-CO-REP-1` as the reproduced run. KTU and RBO are computed at varying cut-offs thresholds. RBO is computed with $\phi = 0.9$, which roughly corresponds to greater weight on the top 10 rank positions (the smaller ϕ, the more top-heavy the measure) [24]. RMSE and the t-test are instantiated with the same effectiveness measures used for performance evaluation: nDCG@10, Q@10, nERR@10 and iRBU@10, computed with the Gold relevance assessments. All reproducibility measures are computed with the `repro_eval`[13] library [3].

Reproducibility results are reported in Figure 2. The reproducibility run `SLWWW-CO-REP-1` achieves high scores with respect to all the reproducibility measures, thus it represents a successful reproducibility attempt. This is aligned with the results in Table 9, showing that `SLWWW-CO-REP-1` performs similarly to the original run `KASYS-CO-REV-6`.

Figure 2a reports KTU at varying cut-offs. The average KTU across topics is KTU = 0.1477 with respect to the entire run, i.e., 1000 rank positions. Even if this value is quite low, it is higher than those reported in other reproducibility attempts [2, 17]. Recall that KTU is the strictest measures because it requires the same document at each rank position for the reproduced and original runs.

Figure 2b reports RBO at varying cut-offs. RBO for the entire run averaged across topics is close to one: RBO = 0.9686. Note that KTU and RBO have different trends: KTU decreases at higher cut-offs while RBO increases (compare Figures 2a and 2b). This

[13] https://github.com/irgroup/repro_eval

Table 10: Randomised Tukey HSD test results ($B = 5,000$ trials) for the Gold-based results in Table 9. The runs in the left column are statistically significantly better than those in the right column at the 5% significance level. Note that this table omits Section (c) as none of the differences in terms of Mean nERR are statistically significant.

(a) Mean nDCG		(b) Mean Q		(d) Mean iRBU	
SLWWW-CO-REP-1	ORG-TOPICDEV	THUIR-CO-NEW-2	ORG-TOPICDEV	SLWWW-CO-NEW-4	ORG-TOPICDEV
KASYS-CO-REV-6	ORG-TOPICDEV	THUIR-CO-NEW-1	ORG-TOPICDEV	SLWWW-CO-REP-1	ORG-TOPICDEV
THUIR-CO-NEW-2	ORG-TOPICDEV	SLWWW-CO-NEW-4	ORG-TOPICDEV	**KASYS-CO-REV-6**	ORG-TOPICDEV
SLWWW-CO-NEW-4	ORG-TOPICDEV	**KASYS-CO-REV-6**	ORG-TOPICDEV	THUIR-CO-NEW-5	ORG-TOPICDEV
THUIR-CO-NEW-1	ORG-TOPICDEV	SLWWW-CO-REP-1	ORG-TOPICDEV	THUIR-CO-NEW-2	ORG-TOPICDEV
THUIR-CO-NEW-5	ORG-TOPICDEV	SLWWW-CO-NEW-2	ORG-TOPICDEV	THUIR-CO-NEW-4	ORG-TOPICDEV
SLWWW-CO-NEW-2	ORG-TOPICDEV	SLWWW-CO-NEW-3	ORG-TOPICDEV	THUIR-CO-NEW-1	ORG-TOPICDEV
SLWWW-CO-NEW-3	ORG-TOPICDEV	THUIR-CO-NEW-5	ORG-TOPICDEV	SLWWW-CO-NEW-3	ORG-TOPICDEV
KASYS-CO-NEW-4	ORG-TOPICDEV	KASYS-CO-NEW-4	ORG-TOPICDEV	SLWWW-CO-NEW-2	ORG-TOPICDEV
KASYS-CD-NEW-1	ORG-TOPICDEV	KASYS-CD-NEW-1	ORG-TOPICDEV	KASYS-CD-NEW-1	ORG-TOPICDEV
KASYS-CD-NEW-3	ORG-TOPICDEV	KASYS-CO-NEW-2	ORG-TOPICDEV	KASYS-CD-NEW-3	ORG-TOPICDEV
KASYS-CO-NEW-2	ORG-TOPICDEV	SLWWW-CO-NEW-5	ORG-TOPICDEV	KASYS-CO-NEW-4	ORG-TOPICDEV
THUIR-CO-NEW-3	ORG-TOPICDEV	KASYS-CD-NEW-3	ORG-TOPICDEV	KASYS-CO-NEW-2	ORG-TOPICDEV
baseline	ORG-TOPICDEV	THUIR-CO-NEW-3	ORG-TOPICDEV	baseline	ORG-TOPICDEV
SLWWW-CO-NEW-5	ORG-TOPICDEV	baseline	ORG-TOPICDEV	KASYS-CD-NEW-5	ORG-TOPICDEV
THUIR-CO-NEW-4	ORG-TOPICDEV	THUIR-CO-NEW-4	ORG-TOPICDEV	THUIR-CO-NEW-3	ORG-TOPICDEV
KASYS-CD-NEW-5	ORG-TOPICDEV	KASYS-CD-NEW-5	ORG-TOPICDEV	SLWWW-CO-NEW-5	ORG-TOPICDEV

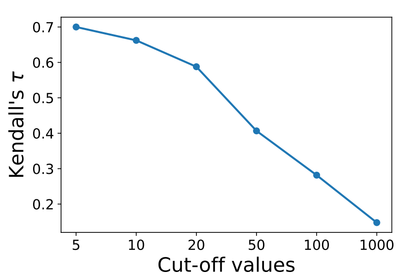
happens because as the cut-off increases also the overlap between the original and reproduced runs increases, consequently RBO score increases. Conversely, when KTU considers a higher cut-off the number of discordant pairs increases, so KTU score decreases.

Finally, Table 2c reports RMSE scores and p-values. As for ranking measures, these results are better than those reported in other reproducibility experiments [2, 17]. With respect to RMSE, the worst value is obtained with nERR. This might happen because nERR is one of the most top-heavy measures and even a small error at rank position 1 or 2 can affect the measure score to a great extent [4]. All p-values are much higher than 0.05, showing that the difference between the original and reproduced runs is not statistically significant (see also discussion in Section 5).

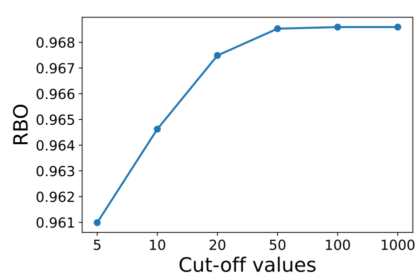
Table 11: Official results based on the Bronze-All file (mean over the 50 WWW-4 test topics).

Run name	(a) Mean nDCG	Run name	(b) Mean Q
THUIR-CO-NEW-2	0.6249	THUIR-CO-NEW-2	0.5857
THUIR-CO-NEW-1	0.6111	**KASYS-CO-REV-6**	0.5743
KASYS-CO-REV-6	0.5931	THUIR-CO-NEW-1	0.5691
SLWWW-CO-REP-1	0.5846	SLWWW-CO-REP-1	0.5629
SLWWW-CO-NEW-4	0.5750	SLWWW-CO-NEW-4	0.5397
SLWWW-CO-NEW-2	0.5600	SLWWW-CO-NEW-2	0.5316
SLWWW-CO-NEW-3	0.5464	SLWWW-CO-NEW-3	0.5137
SLWWW-CO-NEW-5	0.5410	SLWWW-CO-NEW-5	0.5113
THUIR-CO-NEW-3	0.5304	THUIR-CO-NEW-3	0.4853
baseline	0.5170	KASYS-CD-NEW-1	0.4842
KASYS-CD-NEW-1	0.5147	baseline	0.4806
KASYS-CD-NEW-3	0.5130	KASYS-CD-NEW-3	0.4799
KASYS-CO-NEW-2	0.5090	KASYS-CO-NEW-2	0.4733
THUIR-CO-NEW-5	0.5054	KASYS-CO-NEW-4	0.4658
KASYS-CO-NEW-4	0.5025	THUIR-CO-NEW-5	0.4629
THUIR-CO-NEW-4	0.4814	THUIR-CO-NEW-4	0.4402
KASYS-CD-NEW-5	0.4097	KASYS-CD-NEW-5	0.3739
ORG-TOPICDEV	0.2468	ORG-TOPICDEV	0.1384

Run name	(c) Mean nERR	Run name	(d) Mean iRBU
THUIR-CO-NEW-2	0.7967	**KASYS-CO-REV-6**	0.9424
THUIR-CO-NEW-1	0.7962	SLWWW-CO-REP-1	0.9397
KASYS-CO-REV-6	0.7634	SLWWW-CO-NEW-2	0.9244
SLWWW-CO-REP-1	0.7537	SLWWW-CO-NEW-4	0.9213
SLWWW-CO-NEW-2	0.7330	SLWWW-CO-NEW-3	0.9192
SLWWW-CO-NEW-3	0.7242	THUIR-CO-NEW-1	0.9106
SLWWW-CO-NEW-4	0.7209	THUIR-CO-NEW-2	0.9028
THUIR-CO-NEW-3	0.7091	THUIR-CO-NEW-3	0.8979
ORG-TOPICDEV	0.6977	KASYS-CD-NEW-3	0.8922
SLWWW-CO-NEW-5	0.6939	baseline	0.8920
THUIR-CO-NEW-4	0.6783	KASYS-CO-NEW-4	0.8912
baseline	0.6711	KASYS-CD-NEW-1	0.8905
KASYS-CD-NEW-3	0.6629	KASYS-CO-NEW-2	0.8902
THUIR-CO-NEW-5	0.6557	SLWWW-CO-NEW-5	0.8888
KASYS-CD-NEW-1	0.6519	THUIR-CO-NEW-5	0.8793
KASYS-CO-NEW-2	0.6427	THUIR-CO-NEW-4	0.8781
KASYS-CO-NEW-4	0.6384	KASYS-CD-NEW-5	0.8399
KASYS-CD-NEW-5	0.5666	ORG-TOPICDEV	0.6998



(a) Kendall's τ Union (KTU)



(b) Rank Biased Overlap (RBO)

	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

Figure 2: Reproducibility results: ranking measures KTU and RBO with varying cut-offs (Figures 2a and 2b) and reproducibility effectiveness measures RMSE and p-values (Table 2c).

Table 12: Randomised Tukey HSD test results ($B = 5,000$ trials) for the Bronze-All-based results in Table 11(a) and (b). The runs in the left column are statistically significantly better than those in the right column at the 5% significance level.

(a) Mean nDCG	
THUIR-CO-NEW-2	THUIR-CO-NEW-5,KASYS-CO-NEW-4,THUIR-CO-NEW-4,KASYS-CD-NEW-5,ORG-TOPICDEV
THUIR-CO-NEW-1	THUIR-CO-NEW-4,KASYS-CD-NEW-5,ORG-TOPICDEV
KASYS-CO-REV-6	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-REP-1	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-4	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-2	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-3	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-5	KASYS-CD-NEW-5,ORG-TOPICDEV
THUIR-CO-NEW-3	KASYS-CD-NEW-5,ORG-TOPICDEV
baseline	ORG-TOPICDEV
KASYS-CD-NEW-1	ORG-TOPICDEV
KASYS-CD-NEW-3	ORG-TOPICDEV
KASYS-CO-NEW-2	ORG-TOPICDEV
THUIR-CO-NEW-5	ORG-TOPICDEV
KASYS-CO-NEW-4	ORG-TOPICDEV
THUIR-CO-NEW-4	ORG-TOPICDEV
KASYS-CD-NEW-5	ORG-TOPICDEV
(b) Mean Q	
THUIR-CO-NEW-2	THUIR-CO-NEW-4,KASYS-CD-NEW-5,ORG-TOPICDEV
KASYS-CO-REV-6	THUIR-CO-NEW-4,KASYS-CD-NEW-5,ORG-TOPICDEV
THUIR-CO-NEW-1	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-REP-1	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-4	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-2	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-3	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-5	KASYS-CD-NEW-5,ORG-TOPICDEV
THUIR-CO-NEW-3	ORG-TOPICDEV
KASYS-CD-NEW-1	ORG-TOPICDEV
baseline	ORG-TOPICDEV
KASYS-CD-NEW-3	ORG-TOPICDEV
KASYS-CO-NEW-2	ORG-TOPICDEV
KASYS-CO-NEW-4	ORG-TOPICDEV
THUIR-CO-NEW-5	ORG-TOPICDEV
THUIR-CO-NEW-4	ORG-TOPICDEV
KASYS-CD-NEW-5	ORG-TOPICDEV

Table 13: Randomised Tukey HSD test results ($B = 5,000$ trials) for the Bronze-All-based results in Table 11(c) and (d). The runs in the left column are statistically significantly better than those in the right column at the 5% significance level.

(c) Mean nERR	
THUIR-CO-NEW-2	KASYS-CO-NEW-2,KASYS-CO-NEW-4,KASYS-CD-NEW-5
THUIR-CO-NEW-1	KASYS-CO-NEW-2,KASYS-CO-NEW-4,KASYS-CD-NEW-5
KASYS-CO-REV-6	KASYS-CD-NEW-5
SLWWW-CO-REP-1	KASYS-CD-NEW-5
SLWWW-CO-NEW-2	KASYS-CD-NEW-5
SLWWW-CO-NEW-3	KASYS-CD-NEW-5
SLWWW-CO-NEW-4	KASYS-CD-NEW-5
(d) Mean iRBU	
KASYS-CO-REV-6	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-REP-1	KASYS-CD-NEW-5,ORG-TOPICDEV
SLWWW-CO-NEW-2	ORG-TOPICDEV
SLWWW-CO-NEW-4	ORG-TOPICDEV
SLWWW-CO-NEW-3	ORG-TOPICDEV
THUIR-CO-NEW-1	ORG-TOPICDEV
THUIR-CO-NEW-2	ORG-TOPICDEV
THUIR-CO-NEW-3	ORG-TOPICDEV
KASYS-CD-NEW-3	ORG-TOPICDEV
baseline	ORG-TOPICDEV
KASYS-CO-NEW-4	ORG-TOPICDEV
KASYS-CD-NEW-1	ORG-TOPICDEV
KASYS-CO-NEW-2	ORG-TOPICDEV
SLWWW-CO-NEW-5	ORG-TOPICDEV
THUIR-CO-NEW-5	ORG-TOPICDEV
THUIR-CO-NEW-4	ORG-TOPICDEV
KASYS-CD-NEW-5	ORG-TOPICDEV

Table 14: Run ranking correlations in terms of Kendall's τ with 95%CIs ($n = 18$ runs).

(a) Gold	Q	nERR	iRBU
nDCG	0.824 [0.677, 0.908]	0.627 [0.372, 0.794]	0.725 [0.517, 0.852]
Q	-	0.699 [0.477, 0.837]	0.601 [0.335, 0.778]
nERR	-	-	0.536 [0.247, 0.737]
(b) Bronze-All	Q	nERR	iRBU
nDCG	0.961 [0.924, 0.980]	0.725 [0.517, 0.852]	0.699 [0.477, 0.837]
Q	-	0.686 [0.457, 0.830]	0.712 [0.497, 0.845]
nERR	-	-	0.503 [0.204, 0.716]

(c) Gold vs. Bronze-All	
nDCG	0.595 [0.327, 0.775]
Q	0.680 [0.449, 0.826]
nERR	0.327 [−0.007, 0.595]
iRBU	0.438 [0.123, 0.673]

7 CONCLUSIONS

This paper provided an overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) task. Our conclusions are as follows:

- Our Gold and Bronze relevance assessments differ substantially. This is largely because while all Bronze assessments are based on the RND (randomised) pool files, the Gold assessments are based on PRI (prioritised) pool files for one half of the topic set. (Further details will be reported elsewhere.) Due to the disagreements, the Gold and Bronze-All system rankings in terms of Mean nERR are not even statistically significantly correlated.
- In both Gold and Bronze-All evaluations, none of the runs statistically significantly outperform the REV run (i.e., SOTA from NTCIR-15). Hence we are not seeing any substantial technological advance. However, THUIR-CO-NEW-2 [26] is quite successful in the Bronze-All evaluation in the sense that it is the only run that managed to outperform five other runs in terms of Mean nDCG.
- The only REP run, SLWWW-CO-REP-1 [22], is quite successful. Its effectiveness is very similar to **KASYS-CO-REV-6** [23], whose algorithm is identical to that of KASYS-E-CO-NEW-1 from the NTCIR-15 WWW-3 task [20]. Our suite of reproducibility measures also suggest that this run is more successful than previous reproducibility efforts.

Unfortunately, only the University of Tsukuba (KASYS), Waseda University (SLWWW), and Tsinghua University (THUIR) participated in WWW-4, so we will not continue the task in its current form. Our current plan is to propose a group-fair web search task for NTCIR-17 by leveraging the new Chuweb21 corpus.

ACKNOWLEDGEMENTS

The WWW-4 participants (KASYS, SLWWW, THUIR) have made our work possible. Thank you! In particular, we are indebted to KASYS for providing a REV run, which made our discussion of progress and reproducibility possible. We also thank the NTCIR chairs and staff for their constant support.

This paper was partially supported by the the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667.

DISCLAIMER

Certain companies and products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products or companies identified are necessarily the best available for the purpose.

REFERENCES

[1] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter?. In *Proceedings of ACM SIGIR 2008*. 667–674.

[2] Timo Breuer, Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, Philipp Shaer, and Ian Soboroff. 2020. How to Measure the Reproducibility of System-oriented IR Experiments. In *Proceedings of ACM SIGIR 2020*. 349–358.

[3] Timo Breuer, Nicola Ferro, Maria Maistro, and Philipp Schaer. 2021. repro_eval: A Python Interface to Reproducibility Measures of System-Oriented IR Experiments. In *Proceedings of ECIR 2021 Part II (LNCS 12657)*. 481–486.

[4] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2015. Towards a Formal Framework for Utility-oriented Measurements of Retrieval Effectiveness. In *The 2015 International Conference on The Theory of Information Retrieval*. 21–30.

[5] Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2019. CENTRE@CLEF 2019. In *Proceedings of ECIR 2019 Part II (LNCS 11438)*. 283–290.

[6] Nicola Ferro, Norbert Fuhr, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2019. CENTRE@CLEF: Sequel in the Systematic Reproducibility Realm. In *Proceedings of CLEF 2019 (LNCS 11696)*. 287–300.

[7] Nicola Ferro, Maria Maistro, Tetsuya Sakai, and Ian Soboroff. 2018. Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm. In *Proceedings of CLEF 2018 (LNCS 11018)*. 239–246.

[8] M. G. Kendall. 1948. *Rank correlation methods.* Griffin, Oxford, England.

[9] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the NTCIR-13 We Want Web Task. In *Proceedings of NTCIR-13*. 394–401. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/pdf/ntcir/01-NTCIR13-OV-WWW-LuoC.pdf

[10] Jiaxin Mao, Tetsuya Sakai, Cheng Luo, Peng Xiao, Yiqun Liu, and Zhicheng Dou. 2019. Overview of the NTCIR-14 We Want Web Task. In *Proceedings of NTCIR-14*. 455–467. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-WWW-MaoJ.pdf

[11] Masaki Muraoka, Zhaohao Zeng, and Tetsuya Sakai. 2020. SLWWW at the NTCIR-15 WWW-3 Task. In *Proceedings of NTCIR-15*. 243–246.

[12] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173)*. 116–163.

[13] Tetsuya Sakai. 2018. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power.* Springer. https://link.springer.com/book/10.1007/978-981-13-1199-4

[14] Tetsuya Sakai. 2019. How to Run an Evaluation Task: with a Primary Focus on Ad Hoc Information Retrieval. In *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, Nicola Ferro and Carol Peters (Eds.). Springer, 71–102.

[15] Tetsuya Sakai, Nicola Ferro, Ian Soboroff, Zhahao Zeng, Peng Xiao, and Maria Maistro. 2019. Overview of the NTCIR-14 CENTRE Task. In *Proceedings of NTCIR-14*. 494–509. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/ntcir/01-NTCIR14-OV-CENTRE-SakaiT.pdf

[16] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2022. Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents? *ACM TOIS* 40, 4 (2022), Article 76.

[17] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Maria Maistro, Zhicheng Dou, Nicola Ferro, and Ian Soboroff. 2020. Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task. In *Proceedings of NTCIR-15*. 219–234.

[18] Tetsuya Sakai and Zhaohao Zeng. 2019. Which Diversity Evaluation Measures are "Good"?. In *Proceedings of ACM SIGIR 2019*. 595–604.

[19] Tetsuya Sakai and Zhaohao Zeng. 2020. Retrieval Evaluation Measures that Agree with Users' SERP Preferences: Traditional, Preference-based, and Diversity Measures. *ACM TOIS* 39, 2 (2020), Article 14.

[20] Kohei Shinden, Atsuki Maruta, and Makoto P. Kato. 2020. KASYS at the NTCIR-15 WWW-3 Task. In *Proceedings of NTCIR-15*. 235–238. https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/02-NTCIR15-WWW-ShindenK.pdf

[21] Ian Soboroff, Nicola Ferro, Maria Maistro, and Tetsuya Sakai. 2020. Overview of the TREC 2018 CENTRE Track. In *Proceedings of TREC 2018*.

[22] Yuya Ubukata, Masaki Muraoka, Sijie Tao, and Tetsuya Sakai. 2022. SLWWW at the NTCIR-16 WWW-4 Task. In *Proceedings of NTCIR-16*. to appear.

[23] Kota Usuha, Kohei Shinden, and Makoto P. Kato. 2022. KASYS at the NTCIR-16 WWW-4 Task. In *Proceedings of NTCIR-16*. to appear.

[24] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM TOIS* 4, 28 (2010), Article 20.

[25] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *Journal of Data and Information Quality* 10, 4 (2018), Article 16.

[26] Shenghao Yang, Haitao Li, Zhumin Chu, Jingtao Zhan, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. THUIR at the NTCIR-16 WWW-4 Task. In *Proceedings of NTCIR-16*. to appear.

[27] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proceedings of EMNLP-IJCNLP 2019*. 3490–3496. https://aclanthology.org/D19-1352