Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents?

Tetsuya Sakai[1,a)] Sijie Tao[1,b)] Zhaohao Zeng[1,c)]

Abstract: In the context of depth-*k* pooling for constructing web search test collections, we compare two strategies for ordering pooled documents for relevance assessors: the prioritisation strategy (PRI) used widely at NTCIR, and simple randomisation (RND). More specifically, we utilise our WWW3E8 data set which contains eight independent relevance labels for 32,375 topic-document pairs, i.e., a total of 259,000 labels, to compare PRI and RND in terms of inter-assessor agreement and robustness to new systems that did not contribute to the pools. Four of the eight relevance labels were obtained from PRI-based pools; the other four were obtained from RND-based pools. We also utilise an assessor activity log we obtained as a byproduct of WWW3E8 to compare the two strategies in terms of assessment efficiency. Our main results are: (a) the presentation order has no substantial impact on assessment efficiency; (b) the difference between the inter-assessor agreement under the PRI condition and that under the RND condition is of no practical significance; (c) different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition; and (d) PRI-based relevance assessment files (qrels) are substantially and statistically significantly more robust to new systems than RND-based ones. This result suggests that PRI helps the assessors identify relevant documents that affect the evaluation of many systems, including those that did not contribute to the pools. Hence, in this respect, the PRI strategy does have an advantage over RND.

Keywords: information retrieval, pooling, relevance assessments, test collections, web search

1. Introduction

Over half a decade after the Cranfield II experiments of Cleverdon [4], [5], offline information retrieval system evaluation using pooling-based test collections stills remains important for providing researchers with insight into why some methods work while others do not, and for helping them advance the state-of-the-art by utilising that knowledge. In the context of *depth-k pooling* [8], [19] for constructing web search test collections, we compare two strategies for ordering pooled documents for relevance assessors. The first is the prioritisation strategy used widely at NTCIR, which we call PRI: using the `NTCIRPOOL` tool [19],[*1] the pooled documents are sorted by "pseudorelevance," where the first sort key is the number of runs containing the document at or above the pool depth *k* (the larger the better), and the second sort key is the sum of ranks of that document within those runs (the smaller the better). The second strategy, which we call RND, is simply to randomise the pooled documents [6], [23]. We utilise a large-scale data set that we have constructed called WWW3E8 [21][*2] as well as the assessor activity logs that we have obtained as a byproduct of WWW3E8 to address the following research questions.

RQ1 Which strategy enables more efficient relevance assessments?

RQ2 Which strategy enables higher inter-assessor agreements?

RQ3 Which strategy enables more stable system rankings across different versions of qrels files?

RQ4 Which strategy is more robust to the evaluation of systems that did not contribute to the pools?

WWW3E8 contains eight independent 3-point graded relevance labels for 32,375 topic-document pairs (which we call *topicdocs* for brevity), i.e., a total of 259,000 labels. The topicdocs represent the depth-15 pools for the 160 English topics of the NTCIR-15 We Want Web with CENTRE (WWW-3) task [22]. Each topicdoc in WWW3E8 has four relevance labels based on the PRI-based pools, and another four based on the RND-based pools. Hence eight different qrels files for the 160 topics are available, which we shall refer to as PRI1, ..., PRI4, RND1, ..., RND4. Note that, for example, RND1 does not represent the view of a single assessor: 24 assessors were hired to contruct the data, and they were were randomly assigned to topics and pool types.

Due to lack of space, we refer the reader to our paper on WWW3E8 [21] for the background of our research and discussions of related work (e.g., [1], [6], [7], [9], [10], [12], [25]).

2. RQ1: Efficiency

While constructing WWW3E8, we obtained assessor activity logs from our web browser-based relevance assessment interface called PLY [19]. Following Sakai and Xiao [23], we collected the following efficiency statistics for each topic-assessor pair (i.e., for

[1] Department of Computer Science and Engineering, Waseda University, Shinjuku-ku, Tokyo 169-8555, Japan
[a)] tetsuyasakai@acm.org
[b)] tsjmailbox@ruri.waseda.jp
[c)] zhaohao@fuji.waseda.jp

[*1] `http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html`

[*2] `https://waseda.app.box.com/WWW3E8dataset`

Table 1 Efficiency comparison for the eight qrels files (PRI1 through RND4) for the 160 topics. For each efficiency criterion, a paired Tukey HSD test at the 5% significance level was conducted to compare every pair of means. All statistically significant differences are indicated in the table with the p-value and the effect size (standardised mean difference). V_{E2} denotes the two-way ANOVA residual variance for computing the effect sizes [18].

Criterion	n	RND1	RND2	RND3	RND4	RND average	PRI1	PRI2	PRI3	PRI4	PRI average	V_{E2}
TJ1D (secs)	44	34.3	41.2	39.7	40.1	38.8	34.3	35.7	37.1	28.1	33.8	1200.5
TF1RH (secs)	72	25.1	28.9	28.0	26.9	27.2	31.3	31.5	29.9	29.0	30.4	920.3
TF1H (secs)	59	13.3	23.4	24.2	13.6	18.6	26.6	22.4	23.9	27.4	25.1	704.9
ATBJ (secs)	160	13.1♡ (p = 0.0281) (ES = 0.361)	15.8♡	14.5	14.8	14.6	14.5	14.3	14.7	15.52	14.7	56.53
NREJ (times)	160	8.41♠⋄	6.51	6.11	3.99⋄ (p = 0.00360) (ES = 0.427)	6.25	6.79	7.31	5.99	3.82♠ (p = 0.00199) (ES = 0.443)	5.98	107.4

each topic-qrels pair) to address **RQ1**, our assessment efficiency question.

TJ1D Time to judge the first document.

TF1RH Time to find the first relevant or highly relevant document.

TF1H Time to find the first highly relevant document.

ATBJ Average time between judging two documents.

NREJ Number of times the label of a judged document is corrected to another label.

For **TJ1D**, **TF1RH** and **TF1H**, times longer than three minutes were considered outliers and were replaced with an "NA," as we cannot tell from the log whether the assessors were actually reading a document or doing something else. Similarly, for computing **ATBJ**, times longer than three minutes were excluded when computing the average. Note that **ATBJ** is the most direct measure of assessor efficiency.

Table 1 shows, for each qrels file (PRI1 through RND4), our five efficiency criteria averaged across the topics. Note that the sample sizes are much smaller than 160 for **TJ1D**, **TF1RH**, and **TF1H** because we removed every topic that resulted in an "NA" for at least one version of the qrels. Scores averaged over all four PRI (RND) qrels versions are also shown. For each efficienty criterion, as we have eight mean scores to compare, we conducted a paired Tukey HSD test [18] at the 5% significance level. For **TJ1D**, **TF1RH**, and **TF1H**, none of the pairwise differences are statistically significant. It can also be observed that the effect sizes (i.e., standardised mean differences) [18] are also small. For example, for **TJ1D** in Table 1, the difference between the largest and smallest means is RND2 − PRI4 = 41.2 − 28.1 = 13.1; if we convert this to a standardised mean difference using the residual variance shown in the table, we obtain $13.1/\sqrt{1200.5} = 0.377$. That is, even the largest observed difference is less than half a standard deviation apart. Similarly, the largest effect sizes for **TF1RH** and **TF1H** are 0.211 (PRI2 − RND1) and 0.532 (PRI4 − RND4), respectively.

Table 1 also shows a few statistically significant differences for **ATBJ** and **NREJ**. However, the effect sizes are small. The statistically significant difference for **ATBJ** is between two versions of RND-based qrels (namely, RND1 and RND2) and is not interesting. As for **NREJ**, although the assessors involved in PRI4 corrected their labels statistically significantly less frequently compared to those involved in RND1, the effect size is only 0.443.

In summary, since none of the above differences in our efficiency criteria are of practical significance, our answer to **RQ1** is: *the choice of document ordering strategy (RND or PRI) has no substantial impact on assessor efficiency.*

Table 2 Overall inter-assessor agreement in terms of Krippendorff's α (for ordinal data) based on the 24 × 32,375 label matrices for the 160 topics. The original matrix contains 8 labels per topicdoc (4 based on RND, 4 based on PRI); the RND and PRI matrices each contain 4 labels per topicdoc. All other cells are stuffed with "NA."

All	RND	PRI
0.288	0.433	0.423

3. RQ2: Inter-Assessor Agreement

We now utilise the WWW3E8 data set to address **RQ2** (Which document ordering strategy enables higher inter-assessor agreements?). We quantify the inter-assessor agreements under RND and PRI conditions using Krippendorff's α for ordinal classes [11], [19]. For example, to quantify the inter-assessor agreement under the RND condition, all labels in the original WWW3E8 matrix that were obtained under the PRI condition can be replaced with NA's and then the α can be recomputed, so that each topicdoc has only four labels instead of eight. Table 2 shows the results. It can be observed that while the α scores for RND and PRI are very similar, they are much higher than the α score for the original matrix (0.288). That is, while the labels *within* each document ordering strategy are similar to each other, the labels *across* the two strategies differ substantially. It is clear that *the document ordering strategy substantially affects which documents are judged (highly) relevant.* We shall provide an explanation for this in Section 4.2.

The above analysis computed a single α score for the entire matrix. In contrast, Table 3 compares the inter-assessor agreement under the RND and PRI conditions based on mean *per-topic* α scores, averaged over the 160 topics. According to a paired t-test, the difference between the RND and PRI conditions is not statistically significant. More importantly, the effect size (Glass's Δ, a form of standardised mean difference [18]) in terms of α is very small ($\Delta = 0.0859$), and power analysis[*3] asks for over 1,000 topics to achieve 70% statistical power for such a small effect size. From these results, we conclude that *even though the document ordering strategy substantially affects which documents are judged (highly) relevant, the difference between the inter-assessor agreement under the RND condition and that under the PRI condition is of no practical significance.*

[*3] Sakai's tool `future.sample.pairedt` [17] was used for the analysis.

Table 3 Mean per-topic Krippendorff's α (for ordinal data) averaged over the 160 topics. Each per-topic matrix contains 4 labels (either based on RND or PRI) per document. All other cells are stuffed with "NA." The mean α when all 8 labels are included in the matrix is 0.125. A paired t-test at the 5% significance level was conducted. Glass's Δ [18] is based on the standard deviation from the RND data.

n	RND	PRI	t statistic	p-value	Glass's Δ	Achieved power (n required for 70% power)
160	0.293	0.279	0.949	0.344	0.0859	15.7% ($n = 1,098$)

Table 4 System ranking agreement as measured by Kendall's τ between system ranking pairs ($n = 36$ runs) according to the four official measures of the NTCIR-15 WWW-3 task. Correlation strengths are visualised in color: $\tau > 0.8$, $0.5 < \tau \leq 0.8$, $\tau \leq 0.5$.

(a) nDCG	RND2	RND3	RND4	PRI1	PRI2	PRI3	PRI4
RND1	0.752	0.721	0.775	0.340	0.302	0.337	0.340
RND2	-	0.765	0.705	0.378	0.340	0.368	0.371
RND3	-	-	0.698	0.308	0.283	0.305	0.289
RND4	-	-	-	0.349	0.317	0.333	0.356
PRI1	-	-	-	-	0.905	0.927	0.924
PRI2	-	-	-	-	-	0.940	0.917
PRI3	-	-	-	-	-	-	0.940

(b) Q	RND2	RND3	RND4	PRI1	PRI2	PRI3	PRI4
RND1	0.705	0.724	0.775	0.375	0.333	0.387	0.410
RND2	-	0.689	0.657	0.333	0.311	0.352	0.356
RND3	-	-	0.714	0.314	0.286	0.308	0.311
RND4	-	-	-	0.422	0.349	0.410	0.451
PRI1	-	-	-	-	0.889	0.892	0.908
PRI2	-	-	-	-	-	0.883	0.873
PRI3	-	-	-	-	-	-	0.914

(c) nERR	RND2	RND3	RND4	PRI1	PRI2	PRI3	PRI4
RND1	0.587	0.603	0.641	0.330	0.302	0.314	0.324
RND2	-	0.597	0.597	0.337	0.327	0.359	0.337
RND3	-	-	0.632	0.251	0.241	0.229	0.238
RND4	-	-	-	0.213	0.171	0.229	0.225
PRI1	-	-	-	-	0.895	0.902	0.905
PRI2	-	-	-	-	-	0.905	0.908
PRI3	-	-	-	-	-	-	0.927

(d) iRBU	RND2	RND3	RND4	PRI1	PRI2	PRI3	PRI4
RND1	0.606	0.517	0.511	0.346	0.289	0.302	0.314
RND2	-	0.587	0.530	0.352	0.327	0.321	0.327
RND3	-	-	0.670	0.479	0.460	0.460	0.454
RND4	-	-	-	0.390	0.390	0.346	0.384
PRI1	-	-	-	-	0.854	0.892	0.905
PRI2	-	-	-	-	-	0.886	0.886
PRI3	-	-	-	-	-	-	0.892

Table 5 Comparison of mean system ranking τ's based on the τ's shown in Table 4. Results of Tukey HSD tests for unpaired data with sample sizes 6, 6, 16 are shown. The effect sizes are standardised mean differences based on the one-way ANOVA residual variance V_{E1} [18].

Measure	Mean τ (sample size)			Residual	p-value (effect size)		
	RND-RND ($n_1 = 6$)	PRI-PRI ($n_1 = 6$)	RND-PRI ($n_1 = 16$)	variance V_{E1}	RND-RND vs RND-PRI	PRI-PRI vs RND-PRI	PRI-PRI vs RND-RND
nDCG	0.736	0.926	0.332	0.000754	≈ 0 (14.7)	≈ 0 (21.6)	≈ 0 (6.90)
Q	0.711	0.893	0.357	0.00174	≈ 0 (8.48)	≈ 0 (12.9)	≈ 0 (4.37)
nERR	0.610	0.907	0.277	0.00210	≈ 0 (7.26)	≈ 0 (13.8)	≈ 0 (6.49)
iRBU	0.570	0.886	0.371	0.00316	≈ 0 (3.54)	≈ 0 (9.16)	≈ 0 (5.62)

4. RQ3: System Ranking Agreement

4.1 Kendall's τ Results

Using the eight qrels files (PRI1 through RND4) available in WWW3E8, we now address **RQ3** (Which strategy enables more stable system rankings across different versions of qrels files?). More specifically, using each qrels file, we rank the 36 runs submitted to the NTCIR-15 WWW-3 task [22] with the official measuers used in the task.[*4] We then quantify the system ranking similarity with Kendall's τ [16].

The official measures used in the WWW-3 task are nDCG (normalised Discounted Cumulative Gain), Q-measure, nERR (normalised Expected Reciprocal Rank) [16], and iRBU (intentwise Rank-Biased Utility) [24]. These were computed using `NTCIREVAL`.[*5] with an exponential gain value setting: that is, $2^2 - 1 = 3$ for **highly relevant** and $2^1 - 1 = 1$ for **relevant**.

Table 4 shows the results of comparing all pairs of qrels versions in terms of τ. Correlation strengths are visualised in color ($\tau > 0.8$, $0.5 < \tau \leq 0.8$, $\tau \leq 0.5$). The trends are similar across all four evaluation measures, and are very clear. More specifically:

- The four PRI-based qrels files produce very similar system rankings ($\tau > 0.8$);
- The four RND-based qrels files produce moderately similar system rankings ($0.5 < \tau \leq 0.8$);
- The RND-based rankings and the PRI-based ones are substantially different ($\tau \leq 0.5$).

The above three levels of system ranking agreement can be examined more closely as follows. From Table 4, we can compute, for each evaluation measure, a mean τ that represents the agreement within the RND condition by averaging the six values shown in red. Similarly, we can obtain a mean τ within the PRI condition by averaging the six values shown in black. Finally, we can obtain a mean τ across the two conditions by averaging the $4 * 4 = 16$ values shown in blue. To discuss the differences in means for these three cases, we can apply a Tukey HSD test for unpaired data at the 5% significance level [18].

Table 5 shows the results of the unpaired Tukey HSD test for each evaluation measure. Again, the results are similar for all four measures: the "Mean τ" columns show that system rankings within the PRI condition are very similar, those within the RND condition are less so, and that those across the two conditions are substantially different. As the "p-value" columns show, all of these differences in means are statistically highly significant. The "V_{E1}" column shows the residual variance from one-way ANOVA for computing the effect sizes, since we are dealing with unpaired data here [18]. For example, the difference in mean nDCG between the within-RND condition (RND-RND) and the within-PRI condition (PRI-PRI) is $0.926 - 0.736 = 0.190$; therefore the effect size can be computed as $0.190/\sqrt{0.000754} = 6.90$. That is, the two means are about seven standard deviations apart. We conclude that *different system rankings under the PRI condition are substantially more similar to one another than those under*

[*4] The task actually received 37 runs but, as recommended in the WWW3E8 paper [21], we exclude one run file as this was not generated using a single system.

[*5] `http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html` (version 200626)

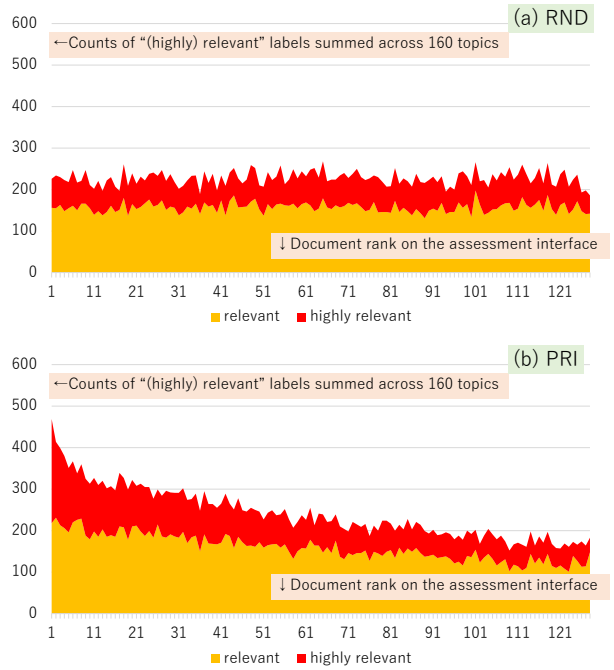

Fig. 1 Document presentation order vs. counts of relevance labels (relevant, highly relevant) based on the main experiment with the 160 topics.

the RND condition. Moreover, as we have observed in Table 4, *PRI-based rankings and RND-based rankings are substantially different from each other.*

4.2 Why Does PRI Produce Similar Rankings?

This section discusses why different PRI-based qrels files produce similar rankings. To examine this phenomenon closely, Figure 1 visualises the relationship between the total number of **highly relevant** and **relevant** labels obtained and the document presentation order as seen by the assessors when WWW3E8 was constructed. The x-axis represents the document ranks shown on the PLY assessment interface. As the *minimum* pool size across the 160 topics was 128 (i.e., every topic had at least 128 pooled documents), we count the assessors' labels (**highly relevant** or **relevant**) across the 160 topics for ranks 1-128. As each topic was judged by four assessors for each document ordering strategy, the maximum possible value for the y-axis is $4 * 160 = 640$: this would happen if, at a particular rank, all four assessors gave a **highly relevant** or **relevant** label for all 160 topics.

It is clear from Figure 1 that while the counts of **highly relevant** and **relevant** labels across topics are not correlated under the RND condition, we obtain more and more **highly relevant** and **relevant** labels as we approach the top of the PRI-based document ranks. There are two possible (mutually nonexclusive) explanations for this phenomenon: (I) the pseudorelevance as computed by `NTCIRPOOL` is accurate to some degree, and often manages to present truly relevant documents before nonrelevant ones; (II) under the PRI condition, the assessors tend to *overrate* the documents that they encounter early. Recall that, for each topic, each assessor receives either a RND pool or a PRI pool at random; they are not even aware that there are two kinds of document ordering strategies. The sharp contrast shown in Figure 1 despite this blind nature of the experiment suggests that the PRI strategy tends to prioritise documents that clearly *look* relevant (e.g., documents that contain query terms in the title field). That is, it is possible that because they look relevant, assessors tend to label them as so. Again, note that this does not rule out Explanation (I): the documents that *look* relevant may often be truly relevant.

Recall that the document sort keys for the PRI strategy are (a) the number of runs that returned the document; and (b) the sum of the ranks of that document in each of the above runs. In essence, the PRI strategy orders documents based on majority votes of the participating runs, and the assessors tend to agree with the majority votes. That is, these "popular" documents tend to be rated highly regardless of who the assessor is. Put another way, *different PRI-based qrels files produce similar system rankings probably because they are all "similarly biased" towards popular documents.*

Table 6 Number of topicdocs in leave-one-team-out qrels files. The original qrels size is 32,375.

team left out	#runs	unique contributions	#topicdocs in LOTO qrels
Group 1	3	1,763	30,612
Group 2	5	1,508	30,867
Group 3	5	1,510	30,865
Group 4	5	6,366	26,009
Group 5	5	4,031	28,344
Group 6	5	4,191	28,184
Group 7	5	1,040	31,335
Group 8	1	136	32,239
Group 9	3	530	31,845

5. RQ4: Robustness to New Systems

We now know that RND-based and PRI-based labels substantially differ from each other, and that the system ranking similarities under the PRI condition are higher than those under the RND condition. However, a more practically important question is **RQ4**: which strategy is more robust to the evaluation of systems that did not contribute to the pool? It is known that relevance assessments of test collections (especially those based on a small pool depth) are *incomplete*, and that new systems tend to be underrated if evaluated with such collections [14], [20], because the new systems may return relevant documents that are outside the pools. While researchers should be aware of this, we still would not want test collections to fail catastrophically when evaluating new systems.

The robustness to new systems can be quantified using *Leave-One-Team-Out* (LOTO) tests [15], [16], [20], [25]. That is, for each of the eight versions of qrels and for each team (G) that participated in the NTCIR-15 WWW-3 task [22], we remove G's *unique contributions* from the original qrels to form a "leave-out-G" qrels file. Here, a unique contribution is a topicdoc that was originally contributed to the pool by team G and by no other team. The WWW-3 task received runs from nine teams, and therefore we created $8 * 9 = 72$ LOTO qrels files. Table 6 shows the relevant statistics of our LOTO experiments. For example, by removing the 1,763 unique contributions of Group 1 from the original qrels file that contained 32,375 topicdocs, we create a "leave-out-Group-1" qrels, with which we can simulate a situation where "new" runs from Group 1 are evaluated using an existing test collection. We then compare the system ranking based on the

Table 7 Mean system ranking τ over nine leave-one-team out experiments. For example, the RND1 column compares the original RND1 qrels with nine leave-one-team-out versions of the qrels. For each evaluation measure, a paired Tukey HSD test at the 5% significance level was conducted. V_{E2} denotes the two-way ANOVA residual variance for computing effect sizes.

	RND1	RND2	RND3	RND4	Average RND	PRI1	PRI2	PRI3	PRI4	Average PRI	V_{E2}
nDCG	0.911	0.884	0.895	0.902	0.898	0.955	0.955	0.957	0.963	0.958	0.00105
Q	0.899	0.871	0.882	0.901	0.888	0.942	0.944	0.940	0.957	0.946	0.00137
nERR	0.927	0.898	0.910	0.917	0.913	0.973	0.971	0.967	0.970	0.970	0.000704
iRBU	0.926	0.900	0.896	0.895	0.904	0.966	0.963	0.958	0.965	0.963	0.001054

original qrels with the new ranking based on each LOTO qrels in terms of Kendall's τ. If the τ is low, that means that the LOTO qrels substantially underrate the "new" runs, which by extension suggests that the original qrels file is also not robust to real new runs that did not contribute to the pools.

Table 7 shows, for each of the eight qrels files (PRI1 through RND4) and for each evaluation measure, the mean τ scores averaged over the nine LOTO trials. Table 8 shows the accompanying results of the paired Tukey HSD tests. They can be summarised as follows.

- None of the differences *within* each document ordering strategy are statistically significant.
- All statistically significant differences are cases where a PRI-based qrels file outperforms a RND-based qrels file. The largest effect size observed is over 2.0 for every evaluation measure (e.g., 2.17 for iRBU, 2,81 for nERR).

Hence the answer to **RQ4** is clear: *the PRI strategy substantially outperforms the RND strategy in terms of robustness to new systems*. The result suggests that the PRI strategy often helps the assessors identify relevant documents *that affect the evaluation of many systems, regardless of whether they contributed to the pools or not.* Put another way, because the RND strategy ignores the "popularity" of documents, it is liable to miss relevant documents that are useful for evaluating many systems fairly.

To examine the above result more closely, Figure 2 visualises the LOTO results with RND2 and PRI4 for nDCG, whose mean τ's are the lowest and the highest among the eight versions of qrels (0.884 and 0.963 as shown in Table 7, respectively). The y-axis represents the mean nDCG scores, while the x-axis represents the runs from all nine groups sorted according to the original qrels file. For example, "Group1-1" means Run 1 from Group 1. Runs that are heavily underrated by a LOTO qrels file can be identified as a "V" in the curves. For example, in Figure 2(b), it is easy to observe from the red curve that if Group 4 is left out, this group's runs (e.g., Group4-5, Group4-3) are heavily underrated. Leaving out this particular group disrupts the ranking this much because this group had as many as 6,366 unique contributions to the pools (See Table 6). If we compare Figure 2(a) and (b), it can be observed that:

(i) Compared to the PRI-based qrels file, the RND-based qrels file gives similar scores to all runs, suggesting that PRI is indeed biased towards "popular" documents;

(ii) For both RND-based and PRI-based qrels files, the top half of the runs do not suffer much even when they are left out from the pools; it is the bottom half of the runs that are heavily underrated when treated as new runs. Moreover, the LOTO qrels files under the RND condition suffer from this more often (i.e., there are more large "V"'s).

Table 8 All statistically significantly different pairs of qrels versions in terms of robustness to new systems (mean τ over $n = 9$ leave-one-team-out experiments), based on a paired Tukey HSD test at the 5% significance level. Effect sizes are based on the two-way ANOVA residual variances shown in Table 7.

qrels pairs	p-value	effect size	qrels pairs	p-value	effect size
(a) mean nDCG			(b) mean Q		
PRI4-RND2	0.0000850	2.44	PRI4-RND2	0.000195	2.33
PRI3-RND2	0.000372	2.24	PRI4-RND3	0.00176	2.02
PRI1-RND2	0.000540	2.19	PRI2-RND2	0.00217	1.99
PRI2-RND2	0.000581	2.18	PRI1-RND2	0.00313	1.94
PRI4-RND3	0.00104	2.10	PRI3-RND2	0.00457	1.88
PRI3-RND3	0.00408	1.90	PRI2-RND3	0.0158	1.68
PRI4-RND4	0.00436	1.89	PRI1-RND3	0.0219	1.63
PRI1-RND3	0.00572	1.85	PRI3-RND3	0.0305	1.57
PRI2-RND3	0.006114	1.84	PRI4-RND1	0.0316	1.57
PRI3-RND4	0.0155	1.69	PRI4-RND4	0.0443	1.51
PRI1-RND4	0.0211	1.64			
PRI2-RND4	0.0224	1.63			
PRI4-RND1	0.0229	1.63			
(c) mean nERR			(d) mean iRBU		
PRI1-RND2	0.0000049	2.81	PRI1-RND4	0.000594	2.17
PRI2-RND2	0.0000075	2.75	PRI4-RND4	0.000688	2.15
PRI4-RND2	0.0000097	2.71	PRI1-RND3	0.000741	2.14
PRI3-RND2	0.0000249	2.60	PRI4-RND3	0.000858	2.12
PRI1-RND3	0.000153	2.36	PRI2-RND4	0.00115	2.08
PRI2-RND3	0.000229	2.30	PRI2-RND3	0.00142	2.05
PRI4-RND3	0.000293	2.27	PRI1-RND2	0.00172	2.02
PRI3-RND3	0.000708	2.15	PRI4-RND2	0.00198	2.00
PRI1-RND4	0.00111	2.09	PRI2-RND2	0.00323	1.93
PRI2-RND4	0.00162	2.03	PRI3-RND4	0.00323	1.93
PRI4-RND4	0.00204	2.00	PRI3-RND3	0.00398	1.90
PRI3-RND4	0.00463	1.88	PRI3-RND2	0.00868	1.78
PRI1-RND1	0.0142	1.70			
PRI2-RND1	0.0198	1.65			
PRI4-RND1	0.0241	1.61			
PRI3-RND1	0.0481	1.49			

6. Conclusions

The present study addressed a few questions that remained open for the past two decades or so regarding two document ordering strategies for relevance assessors: PRI (practiced at NTCIR) and RND (recommended elsewhere). Our experiments, which involved eight independent relevance assessments for 32,375 topic-document pairs (i.e., a total of 259,000 labels), provide some answers to them. Our conclusions are as follows.

RQ1 *Which strategy enables more efficient relevance assessments?* The choice of document ordering strategy has negligible impact on assessor efficiency statistics such as average time between judging two documents.

RQ2 *Which strategy enables higher inter-assessor agrements?* While the choice of the strategy substantially affects which documents are judged (highly) relevant, the difference between the inter-assessor agreement under the PRI condition and that under the RND condition is of no practical significance.

RQ3 *Which strategy enables more stable system rankings*

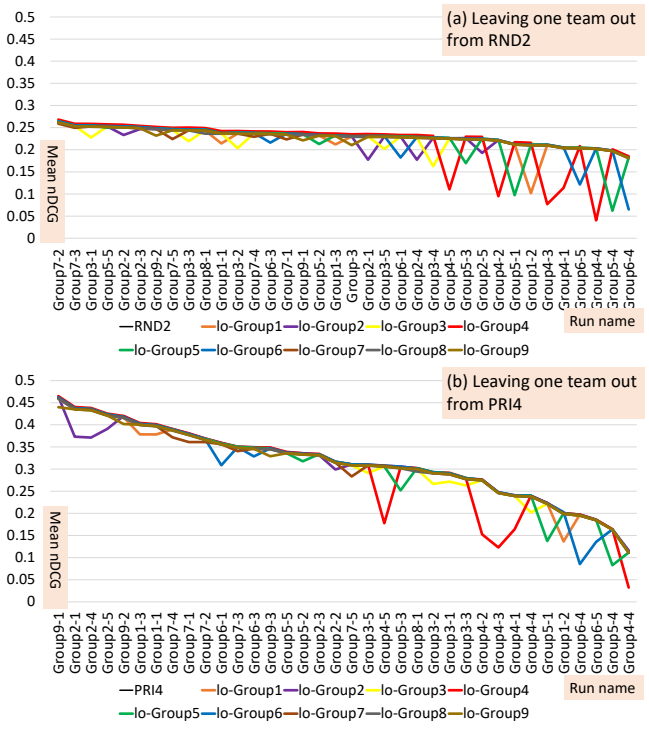

Fig. 2 Effect of leaving one team out on Mean nDCG for (a) RND2 qrels and (b) PRI4 qrels.

across different versions of qrels files? Different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition. Moreover, PRI-based rankings and RND-based rankings are substantially different from each other.

RQ4 *Which strategy is more robust to systems that did not contribute to the pools?* Our leave-one-team-out results show that the PRI-based relevance assessments are substantially more robust. Our interpretation of this finding is that, while the PRI-based qrels files are probably biased towards popular documents, PRI often helps the assessors identify relevant documents *that affect the evaluation of many systems, including those that did not contribute to the pools.* Hence, we conclude that, in this respect, the PRI strategy does have an advantage over RND.

One substantial limitation of the present study is that WWW3E8 is a large collection of "bronze assessor" labels: the relevance assessors were students, not topic originators. There is some indication that not all of the WWW3E8 relevance labels are "correct" [13], although we argue that this issue is orthogonal to our comparison of PRI and RND. As future work, we would like to examine the effect of document ordering on "gold assessor" labels, i.e., those obtained from the topic originators [2] (also called "query owners" [3]). However, such a study will also have its own limitation: by definition, each topic can have only one set of gold assessor labels, based on either a PRI-based or RND-based pool. On the other hand, as the gold assessor labels can be treated as the ground truth, we will be able to assess the *accuracy* of bronze labels, which was not possible in the present study.

References

[1] Allan, J., Harman, D., Kanoulas, E., Li, D., Van Gysel, C. and Voorhees, E.: TREC Common Core Track Overview, *Proceedings of TREC 2017* (2018).

[2] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P. and Yilmaz, E.: Relevance Assessment: Are Judges Exchangeable and Does It Matter?, *Proceedings of ACM SIGIR 2008*, pp. 667–674 (2008).

[3] Chouldechova, A. and Mease, D.: Differences in Search Engine Evaluations Between Query Owners and Non-Owners, *Proceedings of ACM WSDM 2013*, pp. 103–112 (2013).

[4] Cleverdon, C. and Keen, M.: Factors Determining the Performance of Indexing Systems; Volume 2, Technical report, College of Aeronautics, Cranfield, UK (1966).

[5] Cleverdon, C., Mills, J. and Keen, M.: Factors Determining the Performance of Indexing Systems; Volume 1: Design, Technical report, College of Aeronautics, Cranfield, UK (1966).

[6] Damessie, T. T., Culpepper, J. S., Kim, J. and Scholer, F.: Presentation Ordering Effects on Assessor Agreement, *Proceedings of ACM CIKM 2018*, pp. 723–732 (2018).

[7] Eisenberg, M. and Barry, C.: Order Effects: A Study of the Possible Influence of Presentation Order on User Judgments of Document Relevance, *Journal of the American Society for Information Science*, Vol. 39, No. 5, pp. 293–300 (1988).

[8] Harman, D. K.: The TREC Test Collections, *TREC: Experiment and Evaluation in Information Retrieval* (Voorhees, E. M. and Harman, D. K., eds.), The MIT Press, chapter 2 (2005).

[9] Huang, M.-H. and Wang, H.-Y.: The Influence of Document Presentation Order and Number of Documents Judged on Users' Judgments of Relevance, *Journal of the American Society for Information Science*, Vol. 55, No. 11, pp. 970–979 (2004).

[10] Kando, N.: Evaluation of Information Access Technologies at the NTCIR Workshop, *Proceedings of CLEF 2003 (LNCS 3237)*, pp. 29–43 (2004).

[11] Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology (Fourth Edition)*, SAGE Publications (2018).

[12] Losada, D. E., Parapar, J. and Álvaro Barreiro: When to Stop Making Relevance Judgments? A Study of Stopping Methods for Building Information Retrieval Test Collections, *Journal of the Association for Information Science and Technology* (2018).

[13] Muraoka, M., Zeng, Z. and Sakai, T.: SLWWW at the NTCIR-15 WWW-3 Task, *Proceedings of NTCIR-15*, pp. 243–246 (2020).

[14] Sakai, T.: Alternatives to Bpref, *Proceedings of ACM SIGIR 2007*, pp. 71–78 (2007).

[15] Sakai, T.: The Unreusability of Diversified Search Test Collections, *Proceedings of EVIA 2013*, pp. 1–8 (online), available from ⟨http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/EVIA/01-EVIA2013-SakaiT.pdf⟩ (2013).

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

[17] Sakai, T.: Statistical Significance, Power, and Sample Sizes: A Systematic Review of SIGIR and TOIS, 2006-2015, *Proceedings of ACM SIGIR 2016*, pp. 5–14 (2016).

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

[19] Sakai, T.: How to Run an Evaluation Task, *Information Retrieval Evaluation in a Changing World* (Ferro, N. and Peters, C., eds.), Springer (2019).

[20] Sakai, T., Dou, Z., Song, R. and Kando, N.: The Reusabilty of a Diversified Search Test Collection, *Proceedings of AIRS 2012 (LNCS 7675)*, pp. 26–38 (2012).

[21] Sakai, T., Tao, S. and Zeng, Z.: WWW3E8: 259,000 Relevance Labels for Studying the Effect of Document Presentation Order for Relevance Assessors, *in preparation* (2021).

[22] Sakai, T., Tao, S., Zeng, Z., Zheng, Y., Mao, J., Chu, Z., Liu, Y., Dou, Z., Ferro, N., Maistro, M. and Soboroff, I.: Overview of the NTCIR-15 We Want Web with CENTRE (WWW-3) Task, *Proceedings of NTCIR-15*, pp. 219–234 (2020).

[23] Sakai, T. and Xiao, P.: Randomised vs. Prioritised Pools for Relevance Assessments: Sample Size Considerations, *Proceedings of AIRS 2019 (LNCS 12004)*, pp. 94–105 (2019).

[24] Sakai, T. and Zeng, Z.: Retrieval Evaluation Measures that Agree with Users' SERP Preferences: Traditional, Preference-based, and Diversity Measures, *ACM TOIS*, Vol. 39, No. 2 (2020).

[25] Voorhees, E. M.: The Philosophy of Information Retrieval Evaluation, *Proceedings of CLEF 2001 (LNCS 2406)*, pp. 355–370 (2002).