The Reusability of a Diversified Search Test Collection

Tetsuya Sakai[1], Zhicheng Dou[1], Ruihua Song[1], and Noriko Kando[2]

[1] Microsoft Research Asia, P.R. China
tetsuyasakai@acm.org, {zhichdou,Song.Ruihua}@microsoft.com
[2] National Institute of Informatics, Japan
kando@nii.ac.jp

Abstract. Traditional ad hoc IR test collections were built using a relatively large pool depth (e.g. 100), and are usually assumed to be reusable. Moreover, when they are reused to compare a new system with another or with systems that contributed to the pools ("contributors"), an even larger measurement depth (e.g. 1,000) is often used for computing evaluation metrics. In contrast, the web diversity test collections that have been created in the past few years at TREC and NTCIR use a much smaller pool depth (e.g. 20). The measurement depth is also small (e.g. 10-30), as search result diversification is primarily intended for the first result page. In this study, we examine the reusability of a typical web diversity test collection, namely, one from the NTCIR-9 INTENT-1 Chinese Document Ranking task, which used a pool depth of 20 and official measurement depths of 10, 20 and 30. First, we conducted additional relevance assessments to expand the official INTENT-1 collection to achieve a pool depth of 40. Using the expanded relevance assessments, we show that run rankings at the measurement depth of 30 are too unreliable, given that the pool depth is 20. Second, we conduct a leave-one-out experiment for every participating team of the INTENT-1 Chinese task, to examine how (un)fairly new runs are evaluated with the INTENT-1 collection. We show that, for the purpose of comparing new systems with the contributors of the test collection being used, condensed-list versions of existing diversity evaluation metrics are more reliable than the raw metrics. However, even the condensed-list metrics may be unreliable if the new systems are not competitive compared to the contributors.

1 Introduction

Traditional ad hoc IR test collections were built using a large *pool depth*: typically, top 100 documents were collected from every run that was submitted to an evaluation task, and these pooled documents were assessed for relevance (pool depth $pd = 100$). Although the target document collection is usually much larger than the pooled document sets, these IR test collections are often assumed to be *reusable*: they are used for comparing a new system with another and with systems that contributed to the pools ("contributors"). Moreover, in ad hoc IR, a *measurement depth* of 1,000 is often used: that is, top $l = 1000$ documents

Y. Hou et al. (Eds.): AIRS 2012, LNCS 7675, pp. 26–38, 2012.

returned by the system are used for computing evaluation metrics (e.g. [22]). While more intricate techniques for efficiently and reliably obtaining relevance assessments exist (e.g. [7]), the traditional method of using a static pool depth is still widely used, due to its simplicity, its convenience for assessment cost estimation, and its independence to the choice of evaluation metrics.

In contrast to the above practices in ad hoc (particularly non-web) IR, the web *diversity* test collections that have been created in the past few years at TREC and NTCIR use a much smaller pool depth, typically $pd = 20$ [8,17] or $pd = 25$ [9]. These collections are used specifically for evaluating *search result diversification*, which aims to produce a single Search Engine Result Page (SERP) that satisfies different users or user intents that share the same search query (e.g. [8,15]). In web diversity evaluation, the measurement depth is also very small (e.g. $l = 10, 30$), as the target of diversification is typically the *first* SERP (i.e. URLs ranked at the very top). Given the small pool depth, what is the appropriate measurement depth for diversity evaluation? Are existing web diversity test collections reusable to any degree? If they are, what are the appropriate ways to reuse them?

To address the above questions, we examine the reusability of a typical web diversity test collection, namely, one from the NTCIR-9 INTENT-1 Chinese Document Ranking task, which used $pd = 20$ and official measurement depths of $l = 10, 20, 30$ for ranking the submitted runs [17]. First, we conducted additional relevance assessments to expand the official relevance assessments of the INTENT-1 collection to achieve $pd = 40$. Using the expanded data, we show that run rankings at $l = 30$ are too unreliable, given that $pd = 20$. Second, we conduct a *Leave-One-Out* experiment (e.g. [3,12,22]) for every participating team of the INTENT-1 Chinese task, to examine how (un)fairly new runs are evaluated with the INTENT-1 collection. In addition to a set of state-of-the-art diversity evaluation metrics, we experiment with *condensed-list* versions of these metrics, which remove all *unjudged* documents from runs prior to computation [10]. We show that, for the purpose of comparing new systems with the contributors of the test collection being used, condensed-list diversity metrics are more reliable than the raw metrics. However, even the condensed-list metrics may be unreliable if the new systems are not competitive compared to the contributors.

2 Related Work

It is well-known that IR test collections built through pooling are *incomplete* and may be *biased* [2,3]. Their relevance assessments are said to be incomplete if some relevant documents exist among the *unjudged* documents in the collection. Furthermore, the relevance assessments are said to be biased if they represent some limited aspects of the complete set of relevant documents. For example, *shallow pooling* (i.e. using a small *pd*) may cause a *pool depth bias*: the relevant document sets thus obtained may contain only documents that are very easy to retrieve, for example, by keyword matching [1]. Moreover, if the systems that participate in the pooling process all use similar search strategies, this will

cause a *system bias*: the relevant documents thus obtained are the ones that are retrievable by that particular class of systems. Hence relevant documents that can be retrieved by a "novel" system (which we really want to build) may be outside the relevance assessments.

In the context of ad hoc IR, several studies addressed the problem of test collection incompleteness by randomly sampling documents from the original relevance assessments (e.g. [2,10,14,21]). But random sampling does not directly address the bias problem. Zobel's seminal work [22] examined the effect of pool depth bias and system bias; in particular, his Leave-One-Out (LOO) methodology for studying system bias was later adopted by TREC for validating their test collections. This method removes *unique contributions* of a particular team from the original relevance assessments, where a unique contribution is a document contributed to the pool by that particular team only[1]. Thus, the team that has been left out can be seen as a "new" team that did not contribute to the pool. If the outcome of the evaluation for the "new" team based on the LOO relevance assessments is similar to that based on the original relevance assessments, then that suggests that the test collection may be reusable: many of the relevant documents retrieved by the "new" team are already covered by the test collection, even if this team did not contribute to the pool. More recent system bias and pool depth bias studies for ad hoc IR include the work by Büttcher *et al.* [3] and that by Sakai [11,12].

Some evaluation metrics have been designed specifically for the purpose of coping with incompleteness and bias [2,10,14,21]: among them, Sakai's simple approach of using *condensed lists* obtained by removing all unjudged documents from the runs is applicable to *any* existing evaluation metrics, including those that handle graded relevance. However, in his subsequnt study on handling system bias, Sakai [12] reported that "*condensed-list metrics overestimate new systems while traditional metrics underestimate them.*" When a new run is evaluated with a condensed-list metric, many relevant documents go up the ranks as the unjudged documents in the run are removed. The above work of Sakai generalises an earlier finding by Büttcher *et al.* who focussed on binary relevance metrics such as Average Precision (AP) and Binary Preference (bpref): "*Where AP underestimates the performance of a [new] system, bpref overestimates it*" [3]. However, these studies were about *traditional* IR, where typically $pd = 100$ and $l = 1000$.

More intricate approaches to handling incompleteness and bias exist. For example, Webber and Park [19] have proposed to adjust the evaluation metric values computed for new systems, but this methodology requires some new relevance assessments for the new systems. Carterette *et al.* [5] propose to quantify the reusability of test collections, but this requires several kinds of computation, such as estimating the relevance probability of each unjudged document using several features. Carterette *et al.* [6] propose an approach to conducting

[1] Zobel's original method removed documents contributed by a particular *run*, but it is now common practice to conduct a more stringent test by removing documents contributed by a particular *team*, as one team typically submits multiple runs [18].

relevance assessments while monitoring reusability. While the approach is interesting, the focus of the present study is to examine the reliability of an existing web diversity test collection that was constructed in a traditional manner.

3 NTCIR-9 INTENT-1 Task

3.1 Task and Data

NTCIR (NII Testbeds and Community for Information access Research) is a sesquiannual series of international evaluation workshops hosted by National Institute of Informatics (NII), Japan[2]. The first INTENT task (INTENT-1), introduced at the ninth NTCIR workshop (NTCIR-9), had two subtasks: *Subtopic Mining* (SM) and *Document Ranking* (DR). Both subtasks covered two languages: Chinese and Japanese [17].

The SM subtask was defined as: given a query, return a ranked list of *subtopic strings*, which represent diverse *intents* behind the query. By pooling subtopics submitted by the SM participants and manually clustering them, the INTENT-1 task organisers identified a set of *intents* for each query[3]. The organisers also estimated the probability of each intent given the query based on assessor voting.

The DR subtask is similar to the TREC web track diversity task [8,9]: given a query, each participating system returns a diversified ranked list of web pages. The main differences between the evaluation practices at TREC and INTENT-1 are: (a) While TREC treats each intent for a given topic as equally likely, INTENT-1 leveraged the intent probabilities, to prioritise documents relevant to popular intents; (b) While TREC evaluates runs based on per-intent *binary* relevance, INTENT-1 leveraged per-intent *graded* relevance[4]. Also, the evaluation metrics used in these two forums are different, as we shall discuss in Section 3.2.

In this study, we examine the reusability of the INTENT-1 *Chinese DR* test collection, because we were able to hire Chinese assessors and obtain additional relevance assessments for this collection, *and* because the INTENT-1 *Japanese* DR collection is highly unlikely to be reusable: it only involved three participating teams. Table 1 shows some statistics of the Chinese DR test collection.

The additional relevance assessments were done in exactly the same way as the official relevance assessments of $pd = 20$. The same assessor interface was used, which let assessors to view each pooled document and to select a relevance grade for each intent: "highly relevant", "relevant" and "nonrelevant." Two assessors were assigned to each topic, and the relevance grades were aggregated to form a five-point relevance scale, from $L0$ (judged nonrelevant) to $L4$ (highest relevance level) [17][5]. The only difference is how the document pools were obtained: at

[2] `http://research.nii.ac.jp/ntcir/`

[3] In the TREC web diversity tasks, the intents are referred to as "subtopics."

[4] The TREC 2011 web diversity test collection actually contains graded relevance assessments, but they were treated as binary relevance assessments in the evaluation.

[5] We assume that the disagreements between the old and the new assessors are negligible. Ideally, this assumption should be verified by letting the new assessors re-judge some of the "old" documents and computing kappa statistics.

Table 1. Statistics of the INTENT-1 Chinese DR test collection.

#web pages	138 million (SogouT collection)
#topics	100
#intents	917 across 100 topics (max.16; min 4)
#teams (#runs)	7 (24)
relevance levels	5-point ($L0$: judged nonrelevant – $L4$: highest level)
pool depth	20
#relevant docs	12,144 across 100 topics (max.182; min 9)
#relevant intents/doc	mean across 12,144: 1.94 (max. 10; min 1)
#$L0$ (judged nonrelevant)	6,335

Table 2. Statistics of the expanded relevance assessments

pool depth	40
#relevant docs	21,596 across 100 topics (max. 343; min 13)
#relevant intents/doc	mean across 21,596: 1.89 (max. 10; min 1)
#$L0$ (judged nonrelevant)	15,176

Table 3. Statistics of each team at the INTENT-1 Chinese DR subtask ($pd = 20$)

Team	#runs	Unique contributions per topic	Unique relevant per topic	Best run	Official Mean D♯-nDCG @10
THUIR	5	29.92	16.18	THUIR-D-C-5	.5717
uogTr	5	19.51	15.28	uogTr-D-C-5	.5499
MSINT	5	18.10	7.87	MSINT-D-C-1	.5461
HIT2jointNLPlab	2	23.04	16.30	HIT2jointNLPLab-D-C-2	.4749
NTU	1	10.90	7.70	NTU-D-C-1	.4747
SJTUBCMI	5	34.19	22.70	SJTUBCMI-D-C-2	.4663
III_CYUT_NTHU	1	17.10	8.59	III_CYUT_NTHU-D-C-1	.3335

INTENT-1, the top 20 documents from every run was included in the pool; in our study, we first obtained the top 40 documents from every run, and then removed the aforementioned depth-20 documents[6]. Table 2 shows the statistics of the expanded relevance assessments thus obtained.

The left half of Table 3 shows some statistics for each of the seven teams that participated in INTENT-1. The *unique contributions* of a team are documents that were contributed to the pool by *this team only*. The *unique relevant* documents of a team are the relevant documents among its unique contributions. We will discuss the right half of Table 3 in Section 4.1.

3.2 Evaluation Metrics

In this study, we use the evaluation metrics that were used officially for ranking the INTENT-1 runs, namely, *I-rec* (intent recall), *D-nDCG* and *D♯-nDCG* [16]. I-rec is simply the proportion of intents covered by a ranked list; D-nDCG is a ranked retrieval metric for diversity evaluation that takes into account the popularity of intents and per-intent graded relevance. Intuitively, it encourages systems to retrieve documents that are highly relevant to many popular intents

[6] In both cases, the pooled documents are sorted by "popularity" prior to assessments [13]. This practice is not used at TREC.

before those that are marginally relevant to a few minor intents. D♯-nDCG is simply a linear combination of I-rec and D-nDCG, and has been shown to have several advantages over other diversity metrics [15,16].

In addition to the three official metrics, our LOO experiments consider their condensed-list versions, which we call I-rec′, D-nDCG′ and D♯-nDCG′. While standard metrics treat both judged nonrelevant documents ($L0$ documents) and *unjudged* documents (i.e. those that were never included in the pool) as nonrelevant, condensed-list metrics *remove* all unjudged documents from the ranked list before computation. Sakai [10] and Sakai and Kando [14] showed that some condensed-list metrics are more robust to the *incompleteness* of test collections than others such as bpref. As was mentioned in Section 2, however, Sakai [12] showed in his study on test collection *bias* that "*condensed-list metrics overestimate new systems while traditional metrics underestimate them*" and that "*the overestimation [by the condensed-list metrics] tends to be larger than the underestimation [by the raw metrics].*" The overestimation occurs because, when a new run is evaluated with a condensed-list metric, many relevant documents go up the ranks as the unjudged documents in the run are removed. However, while his study was about traditional IR where the measurement depth was $l = 1000$, our present study concerns diversified search where the measurement depth is very small, e.g. $l = 10$. We thought it possible that the overestimation effect of condensed-list metrics may be small in our case, as the number of unjuded documents that are removed will be small.

4 Experiments

4.1 Pool Bias: Pool Depth and Measurement Depth

The INTENT-1 task used $pd = 20$ and officially reported Mean I-rec, D-nDCG and D♯-nDCG values at the measurement depths of $l = 10, 20, 30$. In this section, we examine the effect on the evaluation outcome at $l = 10, 30$ when relevance assessments based on $pd = 40$ are used instead[7]. Figure 1 shows how run rankings change if $pd = 40$ is used instead of $pd = 20$, for all three official metrics. Kendall's τ rank correlation and *symmetric* τ_{ap} values are also shown [20]. τ_{ap} is similar to τ but is more sensitive to the rank changes near the top. The left half of the figure shows that expanding the relevance assessments has very little effect on the system ranking for $l = 10$; while the right half of the figure shows that the effect is not negligible for $l = 30$. For example, the graphs for Mean D♯-nDCG@30 show that top runs actually change if we add more relevance assessments. This shows that the system rankings with $l = 30$, given the pool depth of $pd = 20$, should not be trusted.

We also examined the effect of adding more relevance assessments to statistical significance testing, in particular, significant differences across different teams, as this is important at evaluation forums like NTCIR and TREC. From each of the

[7] Note that this section does not discuss condensed-list metrics, as there are no unjudged documents in the top 20 of any of the runs.

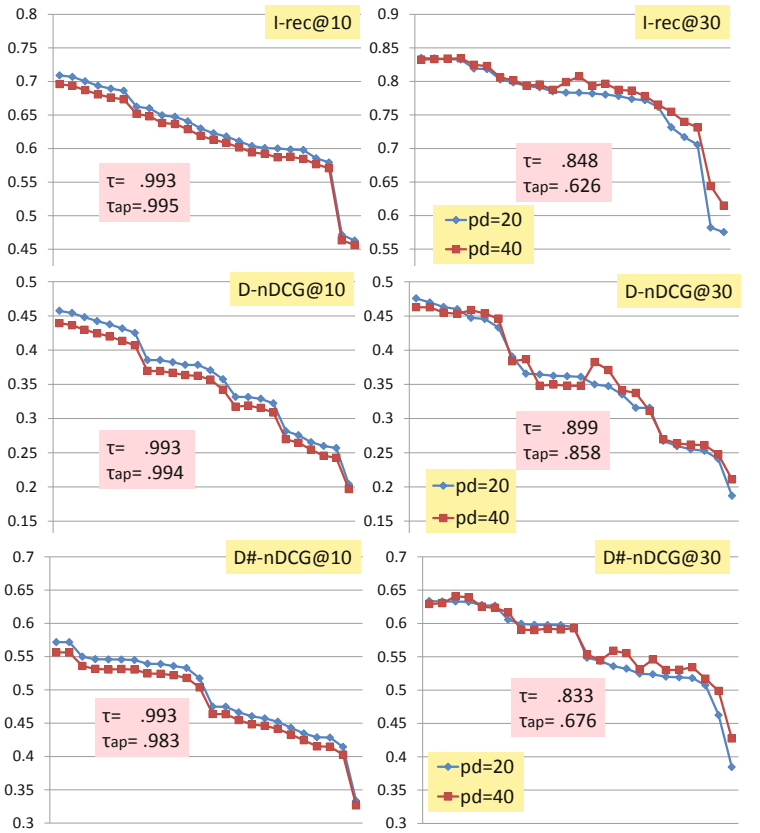

Fig. 1. How run rankings change if $pd = 40$ is used instead of $pd = 20$. The x-axis represents runs sorted by the performance based on relevance assessments with $pd = 20$. The y-axis represents the performance values.

seven teams, we selected the official "best" run in terms of Mean D♯-nDCG@10, as shown in the right half of Table 3. Then we applied a randomised version of two-sided Tukey's Honestly Significant Differences (HSD) test as described by Carterette [4] for the three evaluation metrics with different settings. Table 4 shows the results. For example, from Table 4(I), we can see that even though the difference between THUIR and MSINT was officially significant in terms of D-nDCG@10 (but not in terms of I-rec@10 or D♯-nDCG@10), the difference is *not* statistically significant when less incomplete (i.e. $pd = 40$) relevance assessments are used. Thus, assuming that less incomplete assessments provide more reliable conclusions, the significant difference in terms of D-nDCG@10 is a *false alarm*. Such discrepancies between the $pd = 20$ and $pd = 40$ results are indicated by $\star$'s in Table 4, and it can be observed by comparing Parts (I) and (II) of the table that the false alarms occur more frequently when $l = 30$. These significance test results also show that results with $l = 30$ are not to be trusted when

Table 4. Statistically significant differences across teams according to a randomised Tukey's HSD test at $\alpha = 0.05$. Significant differences by I-rec, D-nDCG and D♯-nDCG are indicated by I, D and D♯, respectively. If none of the differences is significant, this is indicated by a "no." Discrepancies across the pool depths are indicated by ⋆'s..

(I) Measurement depth $l = 10$							
		(b)	(c)	(d)	(e)	(f)	(g)
(a) THUIR	$pd = 20$	no	D⋆	I, D, D♯	D, D♯	I, D, D♯	I, D, D♯
	$pd = 40$	no	no	I, D, D♯	D, D♯	I, D, D♯	I, D, D♯
(b) uogTr	$pd = 20$	–	no	I, D♯	D, D♯	D, D♯	I, D, D♯
	$pd = 40$	–	no	I, D♯	D, D♯	D, D♯	I, D, D♯
(c) MSINT	$pd = 20$	–	–	I, D♯	I, D♯	I, D♯	I, D, D♯
	$pd = 40$	–	–	I, D♯	I, D♯	I, D♯	I, D, D♯
(d) HIT2jointNLPLab	$pd = 20$	–	–	–	no	no	I, D, D♯
	$pd = 40$	–	–	–	no	no	I, D, D♯
(e) NTU	$pd = 20$	–	–	–	–	no	I, D, D♯
	$pd = 40$	–	–	–	–	no	I, D, D♯
(f) SJTUBCMI	$pd = 20$	–	–	–	–	–	I, D, D♯
	$pd = 40$	–	–	–	–	–	I, D, D♯
(g) III_CYUT_NTHU	$pd = 20$	–	–	–	–	–	–
	$pd = 40$	–	–	–	–	–	–
(II) Measurement depth $l = 30$							
		(b)	(c)	(d)	(e)	(f)	(g)
(a) THUIR	$pd = 20$	no	D	I, D, D♯	I, D, D♯	D, D♯	I, D, D♯
	$pd = 40$	no	D	I, D, D♯	I, D, D♯	D, D♯	I, D, D♯
(b) uogTr	$pd = 20$	–	D	I⋆, D, D♯	I⋆, D, D♯	D, D♯	I, D, D♯
	$pd = 40$	–	D	D, D♯	D, D♯	D, D♯	I, D, D♯
(c) MSINT	$pd = 20$	–	–	I, D♯⋆	I, D♯⋆	no	I, D, D♯
	$pd = 40$	–	–	I	I	no	I, D, D♯
(d) HIT2jointNLPLab	$pd = 20$	–	–	–	no	no	I, D, D♯
	$pd = 40$	–	–	–	no	no	I, D, D♯
(e) NTU	$pd = 20$	–	–	–	–	no	I, D, D♯
	$pd = 40$	–	–	–	–	no	I, D, D♯
(f) SJTUBCMI	$pd = 20$	–	–	–	–	–	I, D, D♯
	$pd = 40$	–	–	–	–	–	I, D, D♯
(g) III_CYUT_NTHU	$pd = 20$	–	–	–	–	–	–
	$pd = 40$	–	–	–	–	–	–

$pd = 20$. Thus we recommend that the INTENT task organisers focus on $l = 10$ measurements when officially announcing the participants's performances in the future. In Section 4.2 where we discuss the effect of system biases by means of LOO tests, we will focus on $l = 10$ measurements only.

4.2 System Bias: Evaluating New Systems

To examine how the INTENT-1 Chinese DR test collection evaluates a "new" system that did not contribute to the pool, we conducted a Leave-One-Out (LOO) experiment for each of the seven participating teams at the INTENT-1 Chinese DR task. For example, as Table 1 shows, Team THUIR, the official top performer of the task in terms of Mean D♯-nDCG, had a total of 2,992 unique contributions across the 100 topics when $pd = 20$. The LOO relevance assessment set for this team at $pd = 20$ ("pd20loo-THUIR") was constructed by removing all of these unique contributions from the original relevance assessments with $pd = 20$. Then, all of the 24 runs were evaluated using pd20loo-THUIR. Note that, when the runs from THUIR are evaluated using pd20loo-THUIR, the evaluation relies entirely on contributions from teams *other than* THUIR. The above

Table 5. Leave-One-Out results for each participating team at the INTENT-1 Chinese Document Ranking subtask. For each team, the first row shows the difference between the performance with the original relevance assessments and the performance with that teams' LOO relevance assessments; the second row show the ranks before and after leaving out that team. In Part (b), the results that are more *effective* than those in Part (a) are shown in **bold**.

	(a) Raw-list metrics			(b) Condensed-list metrics		
	I-rec	D-nDCG	D♯-nDCG	I-rec′	D-nDCG′	D♯-nDCG′
(I) $pd = 20$ ($l = 10$)						
THUIR	−.0720	−.1015	−.0867	**+.0183**	**+.0279**	**+.0231**
	6↓12	1↓12	1↓10	**6↓7**	**1→1**	**1↓2**
uogTr	−.0708	−.1123	−.0915	**+.0212**	**+.0233**	**.0223**
	7↓16	5↓16	3↓14	**7↓8**	**5→5**	**3↓6**
MSINT	−.0774	−.0843	−.0809	**+.0318**	**+.0695**	**+.0507**
	2↓9	8↓14	4↓12	**2↓3**	**8↑5**	**4↑2**
HIT2jointNLPlab	−.1221	−.1396	−.1308	**+.0418**	**+.0515**	**+.0466**
	22↓23	13↓22	13↓22	**22↑14**	**13↑9**	**13→13**
NTU	−.1527	−.1260	−.1394	**−.0090**	**+.0353**	**+.0131**
	14↓23	16↓23	14↓23	**14↓15**	**16↑14**	**14↑13**
SJTUBCMI	−.2081	−.1718	−.1900	**−.0390**	**+.0071**	**−.0160**
	17↓20	15↓20	15↓20	**17↓18**	**15→15**	**15↓16**
III_CYUT_NTHU	−.2123	−.1398	−.1760	**+.0928**	**+.0648**	**+.0788**
	24→24	24→24	24→24	24↑23	24↑21	24→24
(II) $pd = 40$ ($l = 10$)						
THUIR	−.0558	−.0671	−.0615	**+.0045**	**+.0170**	**+.0107**
	6↓10	1↓8	1↓10	**6↓8**	**1→1**	**1→1**
uogTr	−.0442	−.0670	−.0556	**+.0058**	**+.0146**	**+.0102**
	7↓10	5↓11	3↓9	**7→7**	**5↓6**	**3↓4**
MSINT	−.0461	−.0511	−.0485	**+.0339**	+.0636	+.0488
	2↓5	8↓14	4↓9	**2↑1**	**8↑4**	**4↑1**
HIT2jointNLPlab	−.0869	−.0929	−.0899	**+.0329**	**+.0327**	**+.0328**
	22→22	13↓19	13↓22	22↑16	**13↑9**	**13→13**
NTU	−.0692	−.0712	−.0702	**+.0058**	**+.0360**	**+.0209**
	14↓22	15↓21	14↓23	**14↑13**	**15↑14**	**14↑13**
SJTUBCMI	−.1423	−.1206	−.1315	**−.0011**	**+.0359**	**+.0174**
	17↓20	16↓20	15↓20	**17→17**	**16↑14**	**15↑13**
III_CYUT_NTHU	−.1618	−.1040	−.1329	**+.1205**	**+.0926**	**+.1066**
	24→24	24→24	24→24	24↑22	24↑19	24↑20

process was repeated for all of the seven teams, and also for $pd = 40$. Thus, we constructed 14 different sets of LOO relevance assessments and evaluated all 24 runs with each of them.

Table 5 summarises our LOO experimental results for $l = 10$. Part (a) of this table shows the results for Mean I-rec, D-nDCG and D♯-nDCG, and the takeaway from this is that the INTENT-1 Chinese DR collection is indeed *not* reusable when these raw (as opposed to condensed-list) metrics are used to evaluate a "new" run. For example, from Table 5(a)(I), we can observe that when the best run from THUIR is evaluated using pd20loo-THUIR, its absolute Mean D♯-nDCG@10 value is smaller than the original one by 0.0867, and more importantly, it is ranked at 10 among the 24 runs even though it is in fact the top performer. That is, the team that has been left out is *heavily underestimated*[8].

[8] Note that the rank of III_CYUT_NTHU does not change when its run is evaluated using this team's LOO assessments as the run was ranked at 24 even before leaving out the team.

While Table 5(a)(II) shows that the absolute errors are a little smaller when $pd = 40$ (e.g. the LOO performance in terms of Mean D♯-nDCG@10 for THUIR is smaller than the original performance by 0.0615), the dramatic rank changes do not seem to be alleviated even when $pd = 40$.

Part (b) of Table 5, on the other hand, gives us some hope. It shows the results for the three condensed-list metrics, I-rec′, D-nDCG′ and D♯-nDCG′. For example, when we evaluate the best run from THUIR using pd20loo-THUIR, the run's absolute Mean D♯-nDCG′@10 value is *higher* than the original one by 0.0231, and the run is ranked at 2 when it is in fact the top performer[9]. We say that a condensed-list metric based on a LOO set is *effective in absolute terms* if the absolute difference between the LOO performance and the original performance for the team that has been left out is smaller than the case with the raw metric. For example, in the aforementioned case with THUIR, the absolute error of Mean D♯-nDCG′@10 is 0.0231 while that of Mean D♯-nDCG@10 is 0.0867, so D♯-nDCG′@10 with pd20loo-THUIR is effective in absolute terms. Similarly, we say that a condensed-list metric based on a LOO set is *effective in relative terms* if the absolute rank change of the team that has been left out is smaller than the case with the raw metric. For example, in the aforementioned case with THUIR, the absolute rank change by Mean D♯-nDCG′@10 is $2 - 1 = 1$ while that by Mean D♯-nDCG@10 is $10 - 1 = 9$, so D♯-nDCG′@10 with pd20loo-THUIR is effective in relative terms as well. In Table 5(b), the effective cases are indicated in bold. The results suggest that *condensed-list diversity metrics may be more useful than raw diversity metrics for the purpose of evaluating new systems with an existing diversity test collection.*

The above finding may be true, however, only if the new system to be evaluated is competitive compared to the systems that contributed to the pools. Note that, in Table 5, the condensed-list results for Team III_CYUT_NTHU are not impressive: for example, in Part (II), it can be observed that the absolute error for III_CYUT_NTHU by Mean D♯-nDCG′@10 is as high as 0.1066 (whereas the corresponding error by Mean D♯-nDCG@10 is 0.1329), and that the metric overestimates III_CYUT_NTHU by ranking it at 20 rather than 24. Similar trends can be observed even for the low performers submitted by the top performing team THUIR: Figure 2 visualises the effect of leaving out THUIR with $pd = 20$ for the entire set of 24 runs, which shows that while D♯-nDCG′@10 is more accurate than D♯-nDCG@10 for evaluating the *high performing* runs from THUIR using pd20loo-THUIR, it is no more accurate than D♯-nDCG@10 for evaluating the *low performing* runs from the same team. (Compare the absolute errors of THUIR-D-C-5 and THUIR-D-C-1 with those of THUIR-D-C-3 and THUIR-D-C-4.) Thus, it appears that *condensed-list diversity metrics may overestimate new systems as much as raw diversity metrics underestimate them if the new systems are low performers.* This is probably because low performers contain relevant

[9] Recall that, for example, D♯-nDCG′@10 is the same as the raw D♯-nDCG@10 when the original relevance assessments are used, since there will be no unjudged documents involved.

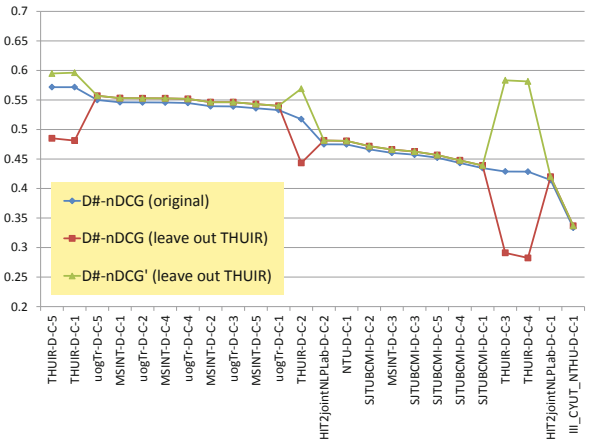

Fig. 2. Original Mean D♯-nDCG ranking vs. Mean D♯-nDCG ranking based on the "leave out THUIR" relevance data ($pd = 20$)

documents at low ranks, and therefore the effect of condensing the list tends to be greater: the relevant documents are "promoted" more dramatically.

Based on the above observation, we have also considered an evaluation method that serves as a compromise between the traditional and the condensed-list methods: the new method computes evaluation metrics based on the condensed-list, but discounts the gain value of each promoted document based on the number of promoted ranks when compared to the original list. The larger the promotion is, the more uncertain we are about the evaluation outcome. However, we have not obtained a promising method that achieves the desired effect and leave it for future work.

5 Conclusions and Future Work

To our knowledge, the present study is the first to have addressed the issue of reusability for diversified search test collections. Although we do not claim that our findings apply to every existing diversity test collection as our study is limited to the case of the NTCIR-9 INTENT-1 Chinese DR test collection, it should be noted that the TREC web diversity test collections were constructed in a similar manner, using a shallow pool depth of either 20 or 25 with a comparable number of participating teams. By conducting additional relevance assessments to achieve a pool depth of 40 for the INTENT-1 collection, we showed that run rankings at the measurement depth of 30 are too unreliable given the pool depth of 20. Thus we recommend that the future INTENT task use the measurement depth of 10 only. Moreover, through leave-one-out experiments for every participating team of the INTENT-1 Chinese task, we showed that condensed-list

versions of existing diversity evaluation metrics may be more reliable than the raw metrics for comparing new systems with the contributors of the test collection, However, it appears that condensed-list metrics can be successful only if the new systems to be evaluated are competitive relative to the contributors. These findings should be useful for diversity task organisers as well as researchers who want to reuse existing diversity test collections.

We plan to generalise our findings by examining other diversity test collections from TREC and NTCIR.

References

1. Buckley, C., Dimmick, D., Soboroff, I., Voorhees, E.M.: Bias and the limits of pooling for large collections. Information Retrieval 10(6), 491–508 (2007)
2. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Proceedings of ACM SIGIR 2004, pp. 25–32 (2004)
3. Büttcher, S., Clarke, C.L., Yeung, P.C., Soboroff, I.: Reliable information retrieval evaluation with incomplete and biased judgements. In: ACM SIGIR 2007 Proceedings, pp. 63–70 (2007)
4. Carterette, B.: Multiple testing in statistical analysis of systems-based information retrieval experiments. ACM TOIS 30(1) (2012)
5. Carterette, B., Gabrilovich, E., Josifovski, V., Metzler, D.: Measuring the reusability of test collections. In: Proceedings of WSDM 2012, pp. 231–240 (2010)
6. Carterette, B., Kanoulas, E., Pavlu, V., Fang, H.: Reusable test collections through experimental design. In: Proceedings of ACM SIGIR 2010, pp. 547–554 (2010)
7. Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J.A., Allan, J.: Evaluation over thousands of queries. In: Proceedings of ACM SIGIR 2008, pp. 651–658 (2008)
8. Clarke, C.L., Craswell, N., Soboroff, I.: Overview of the TREC 2009 web track. In: Proceedings of TREC 2009 (2009)
9. Clarke, C.L., Craswell, N., Soboroff, I., Voorhees, E.: Overview of the TREC 2011 web track. In: Proceedings of TREC 2011 (2012)
10. Sakai, T.: Alternatives to bpref. In: Proceedings of ACM SIGIR 2007, pp. 71–78 (2007)
11. Sakai, T.: Comparing metrics across TREC and NTCIR: The robustness to pool depth bias. In: Proceedings of ACM SIGIR 2008, pp. 691–692 (2008)
12. Sakai, T.: Comparing metrics across TREC and NTCIR: The robustness to system bias. In: Proceedings of ACM CIKM 2008, pp. 581–590 (2008)
13. Sakai, T., Kando, N.: Are popular documents more likely to be relevant? a dive into the ACLIA IR4QA pools. In: Proceedings of EVIA 2008, pp. 8–9 (2008)
14. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Information Retrieval 11, 447–470 (2008)
15. Sakai, T., Song, R.: Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. Information Retrieval (to appear)
16. Sakai, T., Song, R.: Evaluating diversified search results using per-intent graded relevance. In: Proceedings of ACM SIGIR 2011 (2011)
17. Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q., Orii, N.: Overview of the NTCIR-9 INTENT task. In: Proceedings of NTCIR-9, pp. 82–105 (2011)

18. Voorhees, E.M.: The Philosophy of Information Retrieval Evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) CLEF 2001. LNCS, vol. 2406, pp. 355–370. Springer, Heidelberg (2002)
19. Webber, W., Park, L.A.: Score adjustment for correction of pooling bias. In: Proceedings of ACM SIGIR 2009, pp. 444–451 (2009)
20. Yilmaz, E., Aslam, J., Robertson, S.: A new rank correlation coefficient for information retrieval. In: Proceedings of ACM SIGIR 2008, pp. 587–594 (2008)
21. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: ACM CIKM 2006 Proceedings, pp. 102–111 (2006)
22. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: Proceedings of ACM SIGIR 1998, pp. 307–314 (1998)