The Unreusability of Diversified Search Test Collections

Tetsuya Sakai
Microsoft Research Asia, P.R.C.
tetsuyasakai@acm.org

ABSTRACT

Traditional "ad hoc" test collections, typically built based on depth-100 pools, are often used a posteriori by *non-contributors*, i.e., research groups that did not contribute the pools. The *Leave One Out* (LOO) test is useful for testing whether the test collections are actually *reusable*: that is, whether the non-contributors can be evaluated fairly relative to the *contributors*' official performances. In contrast, at the recent *web search result diversification* tasks of TREC and NTCIR, diversity test collections have been built using shallow pools: the pool depths lie between 20 and 40. Thus it is unlikely that these diversity test collections are reusable: in fact, the organisers of these diversity tasks never claimed that they are. Nevertheless, these collections are also used a posteriori by non-contributors. In light of this, Sakai *et al.* [21] demonstrated by means of LOO tests that the NTCIR-9 INTENT-1 Chinese diversity test collection is not reusable, and also showed that *condensed-list* evaluation metrics generally provide better estimates of the non-contributors' true performances than raw evaluation metrics. This paper generalises and strengthens their findings through LOO tests with the latest TREC 2012 diversity test collection.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

diversity, evaluation, leave-one-out test, reusability, test collection

1. INTRODUCTION

Traditional "adhoc" test collections, typically built based on depth-100 pools, are often used a posteriori by *non-contributors*, i.e., research groups that did not contribute the pools. The *Leave One Out* (LOO) test [33] is useful for testing whether the test collections are actually *reusable*: that is, whether the non-contributors can be evaluated fairly relative to the official performances of the *contributors*. i.e. participating teams at an evaluation task. Let n be the number of contributors, and let $C_k (k = 1, \ldots, n)$ denote the set of documents contributed by the k-th team for a particular topic, regardless of relevance. (Each team may submit multiple runs.) Then the pool for this topic is defined as $P = \bigcup_k C_k$. Moreover, the *unique contributions* from this team is defined as $U_k = C_k - \bigcup_{k' \neq k} C_{k'}$, that is, the set of documents that only this team managed to retrieve (above the pool depth) and contribute. Then the LOO data for this team is defined as $LOO_k = \bigcup_{k' \neq k} C_{k'} = P - U_k$. Thus, we pretend that the i-th team was never a contributor, and evaluate the runs from this team using its LOO data[1]. This mimics a situation where a *real* non-contributor is evaluated with the original test collection.

In contrast to the practices with traditional ad hoc test collections, at the recent *web search result diversification* tasks of TREC and NTCIR, diversity test collections were built using shallow pools: the pool depths lie between 20 and 40. Moreover, at NTCIR, the number of participating teams n was small (See Section 2.1). Thus it is unlikely that these diversity test collections are reusable: in fact, the organisers of these diversity tasks never claimed that they are. Nevertheless, these collections are also used a posteriori by non-contributors. In light of this, Sakai *et al.* [21] demonstrated by means of LOO tests that the NTCIR-9 INTENT-1 Chinese diversity test collection is not reusable, and also showed that *condensed-list* evaluation metrics [15] generally provide better estimates of the non-contributors' true performances than standard evaluation metrics. While standard evaluation metrics assume that both *judged nonrelevant* documents (i.e. pooled documents that have been labelled as nonrelevant) and *unjudged* documents (i.e. documents that were outside the pool and therefore could be either relevant or nonrelevant) are "nonrelevant," condensed-list metrics completely disregard unjudged documents in the ranked list that is being evaluated, so that judged documents are "promoted" from their original ranks, as we shall illustrate later in this paper.

This paper extends the work by Sakai *et al.* [21], and demonstrates that the latest TREC 2012 diversity test collection is not reusable, but that condensed-list metrics provide more accurate results for non-contributors than standard metrics do. More specifically, while standard evaluation metrics heavily underestimate the performances of non-contributors and condensed-metrics tend to overestimate them, the errors are considerably smaller with the condensed-list metrics. We thus recommend the use of condensed-list metrics for researchers who want to evaluate their new diversified search systems with existing test collections to which they did not contribute, although condensed lists are by no means a perfect solution to the unreusability problem.

2. PRIOR ART

2.1 Diversity Test Collections

There are two main venues for evaluating web search result diversification: the English *Diversity Tasks* of the TREC Web Track [7, 10, 8, 11] and the Chinese and Japanese *Document Ranking subtasks* of the NTCIR INTENT Task [27, 22]. Some properties of the diversity test collections constructed at these evaluation venues

[1]The original LOO method [33] removed the contributions from a *run* from the pool, but removing the contributions from all runs for a particular team, as we do in this study, is more realistic [28].

Table 1: TREC Web Track and NTCIR INTENT diversity test collections and runs. The number of intents (i.e. subtopics) excludes those that lack relevant documents.

	(a) topics	(b) topic type	(c) intents /topic	(d) pool depth	(e) runs (teams)	(f) relevance levels	(g) intent probabilities	(h) intent type
TREC 2009	50	ambiguous/faceted	3.98	20	47 (18)	0-1	no	nav/inf
TREC 2010	48	ambiguous/faceted	4.17	20	32 (12)	0-1	no	nav/inf
TREC 2011	50	ambiguous/faceted	3.28	25	25 (9)	0-3	no	nav/inf
					62 (16) incl. adhoc			
TREC 2012	50	ambiguous/faceted	3.74	20/30	20 (8)	0-4	no	nav/inf
				(25 topics each)	48 (12) incl. adhoc		no	nav/inf
NTCIR-9 INTENT-1 Chinese	100	-	8.60	20	24 (7)	0-4	yes	-
NTCIR-9 INTENT-1 Japanese	100	-	10.16	20	18 (4)	0-4	yes	-
NTCIR-10 INTENT-2 Chinese	97	amb/fac/nav	5.57	40	12 (3)	0-4	yes	nav/inf
NTCIR-10 INTENT-2 Japanese	95	amb/fac/nav	5.85	40	8 (2)	0-4	yes	nav/inf

are shown in Table 1. Note that we refer to TREC "subtopics" as "intents" to be consistent with the NTCIR parlance. The properties most relevant to the present study are Columns (d) and (e), which directly affect the reusability of test collections.

Note that the TREC 2011 and 2012 data have two lines for Column (e). This is because, while the adhoc and diversity tasks formed pools separately at TREC 2009 and TREC 2010, a common pool was created for each topic at TREC 2011 and 2012. Thus, while TREC 2012 had only 20 runs (8 teams) for the diversity task, it had a total of 48 runs (12 teams), all of which contributed to the pools for diversity relevance assessments. Also, as Column (f) shows, TREC 2011 and TREC 2012 have graded relevance assessments, although they were not utilised in the official evaluations at TREC.

Note that the pool depth was 20 at TREC 2009 and TREC 2010. It was increased to 25 at TREC 2011. At TREC 2012, although the original plan was to further increase the pool depth to 30, in the end it was set to 30 for only 25 topics, and to 20 for the remaining 25 topics, due to time constraints [11].

In contrast to the TREC diversity test collections, it can be observed from Table 1 that the NTCIR collections have about 100 topics each; that they all have graded relevance assessments and intent probabilities. Sakai and Song [25] have discussed the benefit of utilising these features for diversified search evaluation. On the other hand, the number of runs (teams) is generally small: the NTCIR-9 Japanese and the two NTCIR-10 test collections should not be reused as they were built based on contributions from 2-4 teams, even though the NTCIR-10 collections used depth-40 pools. We shall discuss the NTCIR-9 Chinese collection in Section 2.3.

TREC has terminated the diversity task, and it is unlikely that NTCIR will continue the Document Ranking subtask given the few number of participating teams. Thus the reusability of diversity test collections is an important question for researchers who want to continue evaluating diversified search systems.

2.2 Diversity Evaluation Metrics

Unlike traditional IR evaluation where relevance is all that matters, diversity evaluation needs to evaluate the "right" balance between relevance and diversity. The goal is to satisfy different user intents behind a single query with a single search engine result page. Several researchers have proposed diversity evaluation metrics, and some of them are in use at TREC or NTCIR. As discussed below, all of them utilise *per-intent* relevance assessments. That is, each topic/query q has a set of intents $\{i\}$, and each document is judged for relevance with respect to i rather than the entire topic. Moreover, some metrics utilise $Pr(i|q)$, the likelihood of each intent given the query.

The *Intent-Aware* (IA) approach to diversity evaluation [1] first computes a traditional evaluation metric M_i for each intent i, and then combines them across intents: $M\text{-}IA = \sum_i Pr(i|q)M_i$. This simple approach has several drawbacks: the IA metric as defined above does not fully range between 0-1; in general it does not necessarily encourage diversity relative to relevance [9, 24]; and it has limited *discriminative power*, i.e. the ability to draw reliable conclusions from statistical tests in an experiment [24, 25].

Among the IA metrics, perhaps the most useful is *ERR-IA*, an IA version of *Expected Reciprocal Rank* (ERR) [6]. The "diversity extension of ERR" by Chapelle *et al.* [6] is equilavent to letting $M_i = ERR_i$ in the aforementioned definition of an IA metric, where ERR_i is the ERR computed for each intent i. (A more formal definition will be given in Section 4.) Clarke *et al.* [9] and Chapelle *et al.* [5] have independently described α-nDCG and ERR-IA in a single framework. What distinguishes α-nDCG and ERR-IA from other diversity metrics is their per-intent *diminishing return* property [5]: every time a document relevant to an intent is found, the value of the next document found that is relevant to the same intent is discounted. Thus these metrics penalise redundant information for each intent, and thereby encourages diversity across intents. Hence one of the aforementioned disadvantages of the IA approach, namely that it does not necessarily encourage diversity, may not apply to α-nDCG and ERR-IA.

In contrast to TREC where the intent probabilities are assumed to be uniform (i.e. $Pr(i|q) = 1/|\{i\}|$ for any i) and graded relevance assessments are not utilised, the NTCIR INTENT task utilises these types of information by leveraging the "D♯" evaluation framework of Sakai and Song [24]. More specifically, a diversity version of *normalised discounted cumulative gain* (nDCG) [13] called *D-nDCG* is computed, based on the *global gain* which consolidates per-intent graded relevance assessments and intent probabilities for each document. Unlike the IA approach to computing nDCG, D-nDCG defines a single ideal ranked list for a given topic. Moreover, as D-nDCG is an overall relevance metric and does not have the diminishing return property, Sakai and Song proposed to combine this with *intent recall* (a.k.a. *subtopic recall* [32]), to compute D♯-nDCG. At the NTCIR INTENT task, each participating run is plotted on a D-nDCG/I-rec graph, which can visualise which runs are relevance-oriented and which runs are diversity-oriented. (At TREC, precision and intent recall have been used for similar purposes.)

Clarke, Kolla and Vechtomova [12] proposed to utilise the ambiguous/faceted topic tags for diversity evaluation, but this has not been put into practice. Sakai [17] proposed to utilise the informational/navigational intent tags, to encourage systems to allocate more space to informational ones in the search engine result page. His metrics, *DIN-nDCG* and *P+Q* have been used as additional metrics in the NTCIR-10 INTENT-2 task [22]. However, these metrics are outside the scope of the present study.

In the present study, we use the D♯-nDCG framework as well as a version of ERR-IA to discuss the reusability of diversity test collec-

tions. We use the NTCIREVAL toolkit[2] to compute these metrics as it lets us easily compute *condensed-list* versions of evaluation metrics. In Section 4, we quantify exactly how NTCIREVAL's ERR-IA differs from the "official" ERR-IA of TREC, before examining the test collection reusability.

It has been shown, in several experimental settings using NTCIREVAL, that D(♯)-nDCG outperforms ERR-IA in terms of discriminative power and/or *concordance tests*, which reflect how often diversity metrics agree with simpler and more "intuitive" metrics such as precision and intent recall [24, 25, 17]. Sanderson *et al.* [26] have tried to study how diversity metrics agree with user preferences, but their study treated each intent (i.e. subtopic) as an independent topic. It is indeed difficult to assess diversity metrics from the user's point of view, as they try to accomodate different user intents and different users with a single ranked list.

More recently, Sakai and Dou [19] have proposed an evaluation framework which is very different from rank-based metrics such as α-nDCG, ERR-IA and D(♯)-nDCG: their U-measure discounts relevant pieces of information based on the *amount of text the user has read* instead of ranks, and can handle diversity evaluation. While the present study does not test their new metrics, their approach does not have any mechanism specifically designed to handle reusability. Sakai's subsequent *concordance test* experiments [18] suggest, however, that these new metrics may be more intuitive than α-nDCG and ERR-IA.

2.3 Handling Incompleteness

The main reason why researchers worry about the reusability of test collections is that they were built based on pooling, and therefore their relevance assessments are *incomplete*: only a small part of the entire corpus has been judged for relevance, and if a system retrieves many unjudged documents from the corpus, how should the system be evaluated?

In the context of traditional IR evaluation, Buckley and Voorhees [2] proposed a family of evaluation metrics collectively known as *bpref* for the purpose of robust evaluation with incomplete relevance assessments. Subsequently, Sakai [15] showed that computing traditional Average Precision (AP) after removing all unjudgd documents from the original ranked list (i.e. obtaining a condensed list) is actually more robust to incompleteness than bpref: the only difference between bpref and the condensed-list AP is that the former lacks the sensitiveness to changes near the top ranks (*top-heaviness*). Yilmaz and Aslam proposed *Inferred AP*, a version of AP specifically designed to handle incompleteness [31]. They also proposed *Induced AP*, which is exactly the condensed-list version of AP. One advantage of the condensed-list approach is that it can readily be applied to *any* existing evaluation metric for ranked retrieval. Sakai and Kando [23] further demonstrated the advantages of condensed-list metrics for handling incompleteness.

Also in the context of traditional IR evaluation, Büttcher *et al.* [3] advocated the use of a metric called *RankEff* for the purpose of robust evaluation with incomplete relevance assessments. However, Sakai [16] showed that RankEff is in fact a known variant of bpref known as *bpref_N* (and "bpref_allnonrel" in the trec_eval evaluation tool by Chris Buckley) and that it also lacks the top-heaviness property. He also conducted LOO tests with TREC and NTCIR adhoc test collections, and showed that *while traditional raw-list evaluation metrics underestimate systems from non-contributors, condensed-list metrics overestimate them, and that the error of the overestimation may be larger than that of the underestimation.* While earlier work randomly downsampled the original relevance assessments to study the effect of incompleteness, Büttcher *et al.* [3] and Sakai [16] directly addressed the reusability problem by means of the LOO test, which we believe is more realistic.

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

Table 2: TREC 2012 web track teams and runs. (a): team; (b): #unique contributions; (c): #adhoc runs; (d): #diversity runs; (e): official best run; (f): ERR-IA@20 by ndeval.

(a)	(b)	(c)	(d)	(e)	(f)
(I) Teams that participated in the diversity task					
uogTr	4924	3	3	uogTrA44xu	.5048
uottawa	3670	2	2	DFalah121D	.4315
utwente	4273	3	3	utw2012c1	.4046
srchvrs	1852	1	2	srchvrs12c00	.3860
ICTNET	6848	3	3	ICTNET12DVR1	.3257
udel	7834	3	3	autoSTA	.3253
LIA	3360	3	1	lcm4res	.3176
udel_fang	2963	-	3	UDInfoDivSt	.3001
(II) Teams that did not participated in the diversity task					
QUT_Para	-	2	-	-	-
IRRA	-	3	-	-	-
qutir12	-	2	-	-	-
BudapestAcad	-	3	-	-	-
total		28	20	-	-

Sakai *et al.* [21] followed Sakai's approach [16] but in the context of diversty evaluation. Their hypothesis was as follows. In adhoc IR evaluation, the *measurement depth* (i.e. the cutoff of the ranked list to be evaluated) is typicaly 1000. Therefore, if a condensed list is formed from such a large ranked list, many relevant documents may be promoted dramatically as a result of removing unjudged documents, and this should cause serious overestimation of systems from non-contributors. However, in diversity evaluation where typically a small measurement depth (e.g. 10-30) is used, we can expect that fewer documents will be promoted, and that the promotion for each document will be smaller. Therefore, the overestimation error of condensed-list metrics may not be as great as that in the case of traditional search.

Through LOO experiments with the NTCIR-9 INTENT Chinese test collection. Sakai *et al.* [21] indeed demonstrated that the test collection should not be regarded as reusable, but that condensed-list diversity metrics may be more reliable than raw-list metrics for evaluating runs from non-contributors. The objective of the present to study is to generalise the finding by examining the latest TREC diversity test collection. Note that we are questioning the reusability of *existing* test collections: thus, "on-line" methods such as one that monitors the reusability of a test collection while building it [4] is outside the scope of this study. Also, while a *score adjustment* approach [29] may be useful for evaluating non-contributors accurately, this is also beyond our scope as it requires some new relevance assessments for the non-contributors.

3. DATA

In this study, we test the reusability of the TREC 2012 diversity test collection by means of LOO tests. As Table 1 shows, this collection was built using depth-30 pools for half of the topic set, and all of the pools were created from 48 runs from 12 teams, including the runs for the adhoc task. Hence we regard this as representative of the TREC diversity test collections: if this collection is not reusable, it is highly likey that the others are also unreusable.

Table 2 provides details of the teams and runs that were involved in the TREC 2012 web track. Part (I) shows the eight teams that participated in the diversity task, together with the best run from each team according to the *official* ERR-IA at the measurement depth of 20, as computed by the ndeval program[3]. The teams have

[3] http://trec.nist.gov/data/web/12/ndeval.c

been sorted by the official score. Column (b) shows the unique contributions U_k from each team, based on the actual pool depths that were used at TREC 2012 (i.e. 20 or 30 depending on the topic) [11]. It can be observed, for example, that Team udel is the most "novel" in that it contributed more unique documents than any other team. This is possibly because one of its runs is a manual run.

The LOO tests are conducted as follows. First, all of the 20 diversity runs (See Column (d)) are evaluated using the original relevance assessments. Then, we create eight different LOO relevance assessment data sets, $LOO_k = P - U_k$ for each topic (See Section 1). Note that we remove *all* contributions from the i-th team, including those from the adhoc runs (See Table 2(II)), since we are assuming that this team did not participate in the TREC web track at all. With each LOO set, we re-evaluate the 20 runs, and then examine what happened to the best run from the team that has been left out in absolute and relative terms. By absolute terms, we mean what happens to its absolute evaluation scores: are they overestimated or underestimated relative to the *true* values obtained based on the full relevance data? By relative terms, we mean what happens to its ranking among the 20 runs. For example, if we evaluate the top run uogTrA44xu using the LOO data for uogTr, will the run still be the top run or will it go down several ranks?

4. DEFINITIONS OF METRICS

This section provides formal definitions of the diversity metrics considered in our experiments. All of them are implemented in NTCIREVAL. Following the recent practices at the TREC web track [7, 10, 8], we report on each metric computed at the measurement depth l of 20. Recall that the objective of search result diversification is to diversify the *first* search engine result page, and that considering deeper ranks is probably not practical.

Intent recall (I-rec), also known as *subtopic recall* [32], is defined as:

$$I\text{-}rec = |\{i'\}|/|\{i\}| \tag{1}$$

where $\{i\}$ is the entire set of known intents for a topic and $\{i'\}(\subseteq \{i\})$ is the set of intents actually covered by the ranked list being evaluated.

Let $g_i(r)$ denote the "local" gain value for the document at rank r with respect to intent i: throughout this study, we let the local gain value be 1, 2, 3 and 4 for per-intent relevance levels 1, 2, 3 and 4, respectively, for all of our metrics. In the $D\sharp$ framework, the *global gain* for the document at rank r is defined as $GG(r) = \sum_i Pr(i|q)g_i(r)$. Based on GG, we can easily define *D-measures*, such as *D-nDCG*:

$$D\text{-}nDCG = \frac{\sum_{r=1}^{l} GG(r)/\log(r+1)}{\sum_{r=1}^{l} GG^*(r)/\log(r+1)} \tag{2}$$

where $GG^*(r)$ is the global gain at rank r in a "globally ideal" ranked list, defined by sorting all documents in descending order of the global gain.

In the $D\sharp$ framework, D-nDCG values (representing the overall relevance of a SERP) are plotted againt I-rec (representing the diversity of a SERP). In addition, a simple single-value metric that combines the two axes can be computed, e.g.:

$$D\sharp\text{-}nDCG = \gamma I\text{-}rec + (1-\gamma)D\text{-}nDCG \tag{3}$$

where γ is a parameter, set to 0.5 throughout this study.

Next, we define ERR-IA as implemented in NTCIREVAL, which utilises graded relevance unlike the TREC version. Let $Pr_i(r)$ denote the relevance probability of a document at rank r with respect to intent i: in accordance with the aforementioned linear local gain value setting, we let the probabilities be 1/5, 2/5, 3/5 and 4/5 for per-intent relevance levels 1, 2, 3 and 4, respectively[4]. Then ERR for intent i is given by:

$$ERR_i = \sum_{r=1}^{l} \prod_{k=1}^{r-1} (1 - Pr_i(k)) Pr_i(r) \frac{1}{r} . \tag{4}$$

Finally, ERR-IA is defined as:

$$ERR\text{-}IA = \sum_i Pr(i|q) ERR_i . \tag{5}$$

In addition to I-rec, D-nDCG, D♯-nDCG and ERR-IA computed with the raw ranked lists from the runs, we also compute their condensed-list versions. Following Sakai [15], the condensed-list version of Metric M is denoted by M'. With NTCIREVAL, it is very easy to compute condensed-list metrics: while traditional metrics can be computed as follows,

```
% cat [rankedlist] | ntcir_eval glabel -I [ideallist] |
ntcir_eval gcompute -cutoffs 20 -I [ideallist]
```

the corresponding condensed-list metrics can be obtained by just adding the `-j` option to the `glabel` command,

```
% cat [rankedlist] | ntcir_eval glabel -j -I [ideallist] |
ntcir_eval gcompute -cutoffs 20 -I [ideallist]
```

as this option removes all unjudged documents while labelling each judged document in the ranked list with its global gain value. More details are available in the README files of NTCIREVAL.

Figure 1 compares the TREC 2012 diversity run rankings according to the above metrics computed by NTCIREVAL to the official ranking according to ndeval's ERR-IA at 20, using the original relevance assessments in both cases. It can be observed that the ranking by NTCIREVAL's ERR-IA is similar to that by the ndeval version,

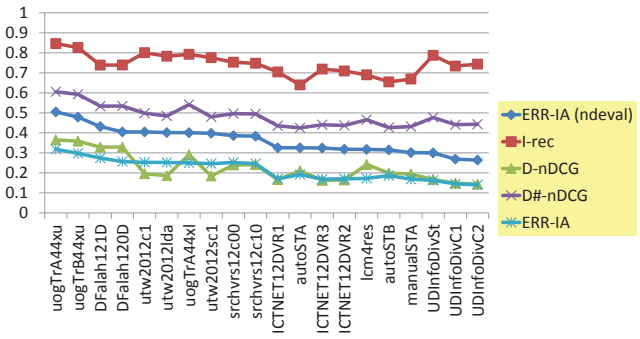

Figure 1: Comparison of the TREC 2012 diversity run rankings by I-rec, D-nDCG, D♯-nDCG and ERR-IA as computed by NTCIREVAL with the official ranking by ERR-IA as computed by ndeval ($l = 20$). The x axis represents runs sorted by ERR-IA with ndeval.

Table 3: τ/τ_{ap} rank correlation between different metrics ($l = 20$).

	I-rec	D-nDCG	D♯-nDCG	ERR-IA
ERR-IA (ndeval)	.453/.522	.600/.637	.579/.635	.874/.897
I-rec	-	.179/.319	.705/.661	.347/.439
D-nDCG	-	-	.474/.580	.663/.662
D♯-nDCG	-	-	-	.558/.622

[4]Chapelle *et al.* [6] used an exponential gain value setting with ERR so that the gain value of a level-4 document is set to $15/16 = 0.94$. This means that the chance of the user examining beyond a 4-relevant document found is only about 6%. With our linear gain value setting, the corresponding probability is 20%, which we believe is more suitable at least for informational intents.

Table 4: Leaving out Team srchvrs for Topic 187 with two intents. Columns (b) shows the gain values for the global ideal list: leaving out Team srchvrs does not affect the top 20 global gain values in this particular example. Columns (c)-(e) show the raw ranked list from Run srchvrs12c00 with its per-intent relevance levels before and after leaving out srchvrs. Columns (f)-(h) show the condensed list after leaving out srchvrs, with its per-intent relevance levels and the original ranks in the raw list.

(a) Rank	(b) Global gains of ideal list (for both original and LOO)	Raw list			Condensed list obtained after leaving out srchvrs		
		(c) docno	(d) Original global (local) gains	(e) LOO global (local) gains	(f) docno	(g) LOO global (local) gains	(h) Rank in raw list
1	4	enwp03-23-13642	1(1/1)	1(1/1)	enwp03-23-13642	1(1/1)	1
2	2.5	enwp01-89-16198	0	0	enwp01-89-16198	0	2
3	2.5	enwp03-38-13766	0	unj	enwp01-69-08702	0	4
4	2.5	enwp01-69-08702	0	0	enwp00-57-17228	0.5(0/1)	5
5	2.5	enwp00-57-17228	0.5(0/1)	0.5(0/1)	enwp01-57-05548	0	13
6	2	enwp01-93-15529	0	unj	enwp01-00-00742	0	14
7	2	enwp01-75-07749	0	unj	enwp02-01-15551	0	15
8	2	enwp02-04-15336	0	unj	enwp02-07-02124	1(1/1)	16
9	2	enwp01-54-21784	0	unj	enwp00-69-18058	0	18
10	2	enwp02-14-17076	0	unj	en0009-42-14770	0.5(0/1)	22
11	2	enwp01-02-21542	0.5(0/1)	**unj**	enwp03-35-11503	0	24
12	2	enwp00-78-04939	0	unj	enwp00-48-22702	0.5(0/1)	27
13	2	enwp01-57-05548	0	0	enwp01-89-16199	0	29
14	2	enwp01-00-00742	0	0	enwp00-55-13168	0	38
15	2	enwp02-01-15551	0	0	enwp00-40-03741	0	41
16	1	enwp02-07-02124	1(1/1)	1(1/1)	enwp00-91-15335	0.5(0/1)	48
17	1	enwp01-60-12924	0	unj	en0003-91-26385	0	52
18	1	enwp00-69-18058	0	0	enwp01-80-16643	1(1/1)	58
19	1	enwp01-60-12971	0	unj	enwp00-14-12323	0.5(0/1)	65
20	1	enwp01-97-15488	0	unj	en0009-26-18188	1(1/1)	70

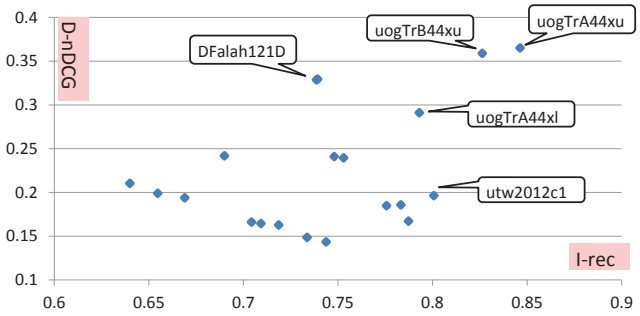

Figure 2: I-rec/D-nDCG graph for the TREC 2012 diversity runs ($l = 20$).

even though our version utilises per-intent graded relevance unlike ndeval. The rankings by the other metrics are more different. These differences are quantified in Table 3, in terms of Kendall's τ as well as the symmetric τ_{ap} [30] which is a top-heavy version of τ. Figure 2 plots the same set of runs in the "NTCIR style" (i.e. on an I-rec/D-nDCG graph): it can be observed, for example, that the TREC official top performer uogTrA44xu is also the top performer in terms of I-rec (pure diversity), D-nDCG (overall relevance) and D♯-nDCG (both combined).

Here we illustrate how D♯-nDCG and ERR-IA are computed with the full and LOO relevance assessments using an example shown in Table 4, Columns (a)-(e). This topic has only two intents, and the global gain values of the ideal list are shown in Column (b): in this particular example, leaving out Team srchvrs does not affect the top 20 global gain values. That is, this ideal list applies to both the original and the LOO relevance assessments. Column (c) is the actual ranked list from Run srchvrs12c00, and Column (d) shows the global and local gain values for each document if it is relevant. For example, the document at rank 1 is 1-relevant to both Intent 1 and Intent 2, and therefore its global gain value is $0.5 * 1 + 0.5 * 1 = 1$ under the uniform intent distribution assumption. Column (e) shows what happens when we leave out the unique contributions from Team srchvrs: many documents are now treated as unjudged, but in this particular example we lose only one relevant document, namely, the one at rank 11. Using Eq. 2 with the global gain values from Columns (b) and (e), the D-nDCG with the LOO data is computed to be .0906; since the ranked list covers both of the intents, I-rec is 1; and therefore $D\sharp\text{-}nDCG = (1 + .0906)/2 = .5453$. Using Eq. 4 with the local gain values from Column (e), the ERRs for the two intents are .2100 and .2400, respectively; and therefore $ERR\text{-}IA = .2250$.

Columns (f)-(h) in Table 4 illustrate how condensed-list metrics are computed for the same example, by removing all unjudged documents shown in Column (e). As shown in Column (h), many judged documents get promoted: for example, it can be observed that the relevant document which was at rank 70 in the raw list is now at rank 20. From Table 4, it is easy to see the inherent properties of raw-list and condensed-list evaluation metrics: raw-list metrics tend to underestimate runs from non-contributors because some of the retrieved unjudgd documents are actually relevant (See Column (e)); condensed-list metrics overestimate them because relevant documents get promoted (See Column (h)). For this example, $D\sharp\text{-}nDCG' = .5791$ and $ERR\text{-}IA' = .2581$ with the LOO data; whereas the true values are: $D\sharp\text{-}nDCG = .5497$; $ERR\text{-}IA = .2300$.

5. RESULTS AND DISCUSSIONS

Recall that we have nine different relevance assessment data sets: the original TREC relevance assessments plus a LOO data set for each of the eight teams. We rank the 20 diversity runs with each data set, using eight evaluation metrics: I-rec($'$), D-nDCG($'$), D♯-nDCG($'$), ERR-IA($'$). Since the pool depth is either 20 or 30 depending on the topic (See Table 1) and our measurement depth l is 20, note that when the full relevance assessments are used, condensing a ranked list has no effect on the top 20 documents. Thus, for example, I-rec$'$ equals I-rec with the full relevance assessments.

For each of the eight participating teams, we compare the performance of its best run (See Table 2) before and after leaving out the unique contributions from that team. Let M^* be the absolute evaluation metric value for a run when the full relevance assessments are used, and let r^* be its rank among the 20 runs. Let M^{LOO} and r^{LOO} denote the corresponding values for the LOO

Table 5: TREC 2012 diversity task Leave-One-Out results ($l = 20$). Each team's official best run (according to ndeval's ERR-IA@20) is re-evaluated with that team's LOO relevance assessments. 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 when all 20 runs are ranked. In Part (b), results that are effective in absolute/relative terms are shown in bold.

	(a) Raw-list metrics				(b) Condensed-list metrics			
	I-rec	D-nDCG	D♯-nDCG	ERR-IA	I-rec′	D-nDCG′	D♯-nDCG′	ERR-IA′
uogTr	−.1220	−.0942	−.1081	−.0649	**−.0306**	**−.0022**	**−.0165**	**+.0035**
	1↓12	1↓4	1↓7	1↓5	**1→1**	**1↓2**	**1→1**	**1→1**
uottawa	−.1374	−.1097	−.1235	−.0683	**+.0310**	**+.0195**	**+.0253**	**+.0432**
	12↓20	4↓7	5↓20	3↓9	**12↑9**	**4↑3**	**5↑3**	**3↑2**
utwente	−.0684	−.0397	−.0540	−.0273	**+.0173**	+.0595	**+.0384**	+.0341
	3↓11	11↓16	6↓11	5↓9	**3↓5**	**11↑6**	**6↓8**	**5↑4**
srchvrs	−.0770	−.0423	−.0597	−.0247	**+.0383**	**+.0325**	**+.0354**	**+.0094**
	8↓16	8↓10	7↓15	7↓9	**8↑6**	**8↑7**	**7→7**	7↑4
ICTNET	−.1616	−.0768	−.1192	−.0704	**+.0540**	**+.0563**	**+.0552**	**+.0279**
	16↓18	16↓18	17↓18	15↓20	16↑10	16↑10	17↑10	**15↑13**
udel	−.0510	−.0322	−.0417	−.0133	**−.0187**	**+.0164**	**−.0012**	+.0219
	20↑19	9↓12	20↑19	11→11	20↑19	**9→9**	20↑19	11→11
LIA	−.0507	−.0585	−.0546	−.0261	**+.0473**	**+.0298**	**+.0386**	**+.0224**
	17↓20	6↓14	12↓20	13↓18	17↑13	**6→6**	**12↑6**	**13↑11**
udel_fang	−.0816	−.0273	−.0544	−.0089	**+.0167**	**+.0227**	**+.0197**	+.0134
	5↓13	15↓18	11↓18	18→18	**5↑3**	**15↑13**	**11↑7**	18↑13

Table 6: NTCIR-9 INTENT-1 Chinese Document Ranking Leave-One-Out results ($l = 10$), copied from Sakai *et al.* [21] Table 5. Each team's official best run (according to D♯-nDCG) is re-evaluated with that team's LOO relevance assessments. 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 when all 24 runs are ranked. In Part (b), results that are effective in absolute/relative terms are shown in bold.

	(a) Raw-list metrics			(b) Condensed-list metrics		
	I-rec	D-nDCG	D♯-nDCG	I-rec′	D-nDCG′	D♯-nDCG′
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

case. We define the *absolute error* as $|M^{LOO} - M^*|$ (i.e. the score delta), and the *relative error* as $|r^{LOO} - r^*|$ (i.e. the rank change). Furthermore, we compare the errors of a condensed-list metric with those of a corresponding raw-list metric. We say that a condensed-list metric is *effective in absolute terms* if its absolute error is smaller than that of the corresponding raw-list metric; we say that a condensed-list metric is *effective in relative terms* if its relative error is smaller than that of the corresponding raw-list metric. That is, if a condensed-list metric is effective in absolute/relative terms, that means it provides a more accurate estimate of the true performance of a "new" system than the raw-list metric does.

Table 5 summarises our main results. Each row evaluates a team's best run using that team's LOO data set. In Column (b), boldface values indicate cases where the condensed-list metrics are effective in absolute/relative terms, i.e., they are more accurate than the raw-list metrics. It can be observed that condensed-list metrics are generally more accurate than raw-list metrics. For example, if the official top run uogTrA44xu (See Table 2) is evaluated with uogTr's LOO data set, it is ranked at 7 with D♯-nDCG and at 5 with ERR-IA, and its absolute scores are smaller than the true values by .1081 in D♯-nDCG and .0649 in ERR-IA (Column (a)). In contrast, the run is still ranked at 1 with both D♯-nDCG′ and ERR-IA′, and the absolute errors are as small as .0165 and .0035, respectively.

Table 6 has been duplicated from Sakai *et al.* [21], just to emphasise the fact that our LOO results with the TREC data are highly consistent with their NTCIR results. This table is not part of our new contribution.

Figures 3 and 4 visualise the above results of uogTr for D♯-nDCG(′) and ERR-IA(′). The x axis represents the original ranking of the 20 runs; those from uogTr are indicated by arrows. It is clear that while raw-list metrics heavily underestimate the "new" run, condensed-list metrics provide more accurate estimates of the true performances. Figures 5 and 6 provide similar graphs for Team udel: we chose this team because it contributed more unique documents than any other team (See Table 2 Column (b)). Thus this represents a case where a relatively novel non-contributor is evaluated with an existing test collection. Again, the condensed-list results look more accurate in general, except that in Figure 6, ERR-IA′ overestimates autoSTA and autoSTB at least as much as ERR-IA underestimates them in absolute terms (See also Table 5 Row "udel", Columns ERR-IA and ERR-IA′). The LOO results for udel remind us that the condensed-list metrics are not a perfect solution for handling new systems with existing diversity test collections, even though they are more accurate than raw-list metrics. Sakai *et al.* [21] have reported a similar observation based on the NTCIR-9 INTENT-1 Chinese data.

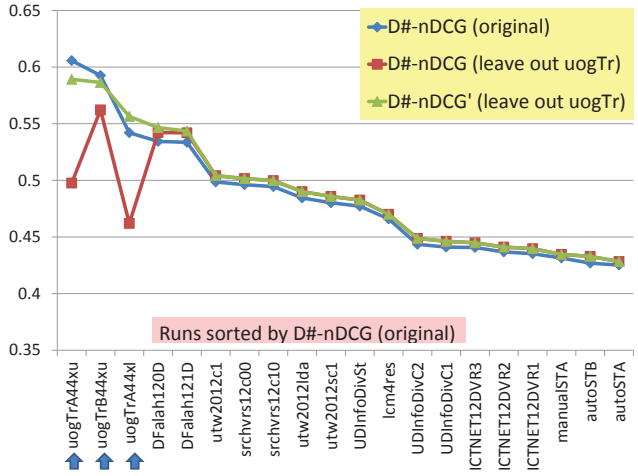

Figure 3: Comparison of the TREC 2012 diversity run rankings by D♯-nDCG with the original relevance assessments, D♯-nDCG with uogTr's leave one out data, and D♯-nDCG′ with uogTr's leave one out data. Runs from uogTr are indicated with arrows. ($l = 20$).

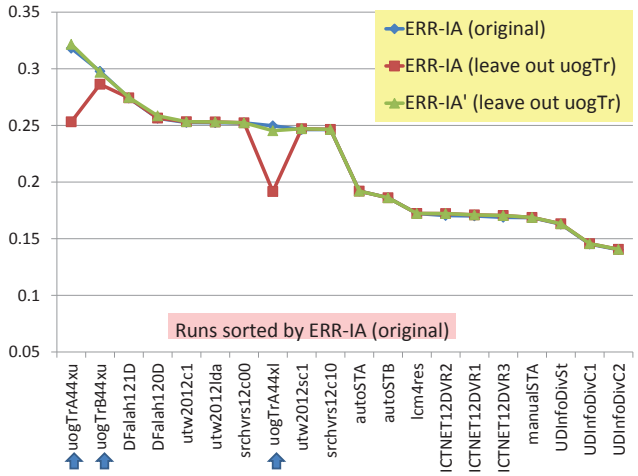

Figure 4: Comparison of the TREC 2012 diversity run rankings by ERR-IA with the original relevance assessments, ERR-IA with uogTr's leave one out data, and ERR-IA′ with uogTr's leave one out data. Runs from uogTr are indicated with arrows. ($l = 20$).

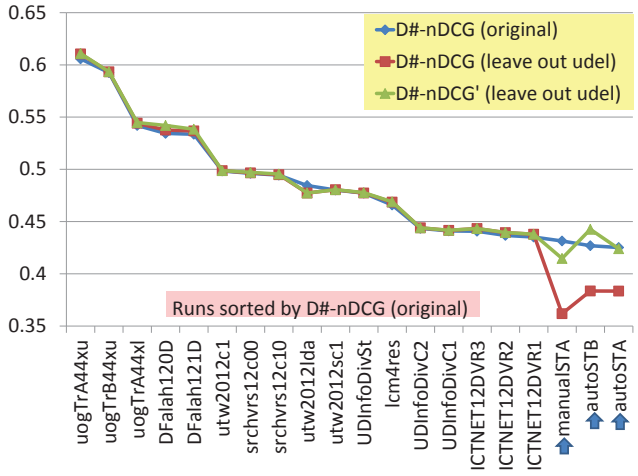

Figure 5: Comparison of the TREC 2012 diversity run rankings by D♯-nDCG with the original relevance assessments, D♯-nDCG with udel's leave one out data, and D♯-nDCG′ with udel's leave one out data. Runs from udel are indicated with arrows. ($l = 20$).

Figure 6: Comparison of the TREC 2012 diversity run rankings by ERR-IA with the original relevance assessments, ERR-IA with udel's leave one out data, and ERR-IA′ with udel's leave one out data. Runs from udel are indicated with arrows. ($l = 20$).

6. CONCLUSIONS

This paper extended the work by Sakai *et al.* [21], and demonstrated that the latest TREC 2012 diversity test collection is not reusable, but that condensed-list metrics provide more accurate results for non-contributors than standard metrics do. More specifically, while standard evaluation metrics heavily underestimate the performances of non-contributors and condensed-metrics tend to overestimate them, the errors are considerably smaller with the condensed-list metrics. We thus recommend the use of condensed-list metrics for researchers who want to evaluate their new diversified search systems with existing test collections to which they did not contribute to, although condensed lists are by no means a perfect solution to the unreusability problem.

Constructing diversity test collections is costly, as it requires relevance assessments per intent, not per topic. It is a shame that such expensive collections are not reusable, although we have shown that condensed-list metrics may provide more accurate results for non-contributors than traditional metrics. Given this situation, one useful future direction for diversity evaluation would be to establish a methodology for efficient and economical construction of *disposable* diversity test collections: instead of explicitly defining a set of possible intents for each topic a priori[5], would it be possible to automatically extract implicit intents from a given set of systems and rank them by "relative diversity"? Would the relative diversity correlate well with the users' diversity preferences? As for relevance, would it be possible to (semi)automatically assign a set of (pseudo)relevant documents to each implicit intent using nugget-based techniques (e.g. [14])? Addressing these questions may in fact be more important than pursuing reusability.

Acknowledgements

The author thanks Ellen Voorhees for providing the TREC 2012 web track data. He also thanks Charlie Clarke for useful discussions and the reviewers for their insightful comments.

7. REFERENCES

[1] R. Agrawal, G. Sreenivas, A. Halverson, and S. Leong. Diversifying search results. In *Proceedings of ACM WSDM 2009*, pages 5–14, 2009.

[2] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of ACM SIGIR 2004*, pages 25–32, 2004.

[3] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *ACM SIGIR 2007 Proceedings*, pages 63–70, 2007.

[4] B. Carterette, E. Kanoulas, V. Pavlu, and H. Fang. Reusable test collections through experimental design. In *Proceedings of ACM SIGIR 2010*, pages 547–554, 2010.

[5] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S.-L. Wu. Intent-based diversification of web search results: Metrics and algorithms. *Information Retrieval*, 14(6):572–592, 2011.

[6] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of ACM CIKM 2009*, pages 621–630, 2009.

[7] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *Proceedings of TREC 2009*, 2009.

[8] C. L. Clarke, N. Craswell, I. Soboroff, and E. Voorhees. Overview of the TREC 2011 web track. In *Proceedings of TREC 2011*, 2012.

[9] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *Proceedings of ACM WSDM 2011*, pages 75–84, 2011.

[5]Recently, Sakai, Dou and Clarke [20] have discussed the effect of the choice of explicit intents on diversity system ranking by leveraging the TREC 2012 diversity topic set, which has "subtopics" and diversity runs from TREC *and* a separate set of intents from the NTCIR-10 INTENT-2 English Subtopic Mining subtask.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 web track. In *Proceedings of TREC 2010*, 2010.

[11] C. L. A. Clarke, N. Craswell, and E. Voorhees. Overview of the TREC 2012 web track. In *Proceedings of TREC 2012*, 2013.

[12] C. L. A. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *Advances in Information Retrieval Theory (ICTIR 2009), LNCS 5766*, pages 188–199, 2009.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[14] S. Rajput, M. Ekstrand-Abueg, V. Pavlu, and J. A. Aslam. Constructing test collections by inferring document relevance via extracted relevant information. In *Proceedings of ACM CIKM 2012*, pages 145–154, 2012.

[15] T. Sakai. Alternatives to bpref. In *Proceedings of ACM SIGIR 2007*, pages 71–78, 2007.

[16] T. Sakai. Comparing metrics across TREC and NTCIR: The robustness to system bias. In *Proceedings of ACM CIKM 2008*, pages 581–590, 2008.

[17] T. Sakai. Evaluation with informational and navigational intents. In *Proceedings of WWW 2012*, pages 499–508, 2012.

[18] T. Sakai. How intuitive are diversified search metrics? Concordance test results for the diversity U-measures. In *IPSJ SIG Technical Report 2013-IFAT-111*, 2013.

[19] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of ACM SIGIR 2013*, 2013.

[20] T. Sakai, Z. Dou, and C. L. A. Clarke. The impact of intent selection on diversified search evaluation. In *Proceedings of ACM SIGIR 2013*, 2013.

[21] T. Sakai, Z. Dou, R. Song, and N. Kando. The reusability of a diversified search test collection. In *Proceedings of AIRS 2012 (LNCS 7675)*, pages 26–38, 2012.

[22] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, R. Song, M. P. Kato, and M. Iwata. Overview of the NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.

[23] T. Sakai and N. Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11:447–470, 2008.

[24] T. Sakai and R. Song. Evaluating diversified search results using per-intent graded relevance. In *Proceedings of ACM SIGIR 2011*, 2011.

[25] T. Sakai and R. Song. Diversified search evaluation: Lessons from the NTCIR-9 INTENT task. *Information Retrieval*, 2013.

[26] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of ACM SIGIR 2010*, pages 555–562, 2010.

[27] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *Proceedings of NTCIR-9*, pages 82–105, 2011.

[28] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF 2001 (LNCS 2406)*, pages 355–370, 2002.

[29] W. Webber and L. A. F. Park. Score adjustment for correction of pooling bias. In *Proceedings of ACM SIGIR 2009*, pages 444–451, 2009.

[30] E. Yilmaz, J. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of ACM SIGIR 2008*, pages 587–594, 2008.

[31] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *ACM CIKM 2006 Proceedings*, pages 102–111, 2006.

[32] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of ACM SIGIR 2003*, pages 10–17, 2003.

[33] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of ACM SIGIR 1998*, pages 307–314, 1998.