

Boiling Down Information Retrieval Test Collections

Tetsuya Sakai
Microsoft Research Asia, China
tetsuyasakai@acm.org

Teruko Mitamura
Carnegie Mellon University, USA
teruko@cs.cmu.edu

ABSTRACT

Constructing large-scale test collections is costly and time-consuming, and a few relevance assessment methods have been proposed for constructing “minimal” information retrieval test collections that may still provide reliable experimental results. In contrast to building up such test collections, we take existing test collections constructed through the traditional pooling approach and empirically investigate whether they can be “boiled down.” More specifically, we report on experiments with test collections from both NTCIR and TREC to investigate the effect of reducing both the topic set size and the pool depth on the outcome of a statistical significance test between two systems, starting with (approximately) 100 topics and depth-100 pools. We define *cost* (of manual relevance assessment) as the pool depth multiplied by the topic set size, and *error* as a system pair whose outcome of statistical significance testing differs from the original result based on the full test collection. Our main findings are: (a) Cost and the number of errors are negatively correlated, and any attempt at substantially reducing cost introduces some errors; (b) The NTCIR-7 IR4QA and the TREC 2004 robust track test collections all yield a comparable and considerable number of errors in response to cost reduction, and this is true despite the fact that the TREC relevance assessments relied on more than twice as many runs as the NTCIR ones; (c) Using 100 topics with depth-30 pools generally yields fewer errors than using 30 topics with depth-100 pools; and (d) Even with depth-100 pools, using fewer than 100 topics results in *false alarms*, i.e. two systems are declared significantly different even though the full topic set would declare otherwise.

Categories and Subject Descriptors

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

General Terms

Experimentation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RIA0'10, 2010, Paris, France.

Copyright CID

Keywords

test collection, relevance assessment, NTCIR, TREC.

1. INTRODUCTION

Information Retrieval (IR) evaluation with standard test collections constructed through collaborative efforts such as the Text Retrieval Conference (TREC), Cross-Language Evaluation Forum (CLEF) and NII Test Collection for IR systems (NTCIR) has been central to the progress of modern IR research. At these forums, large-scale test collections (or rather, test collections with a large target document collection) are usually constructed through *pooling* (e.g. [11]): since it is not feasible to judge the relevance of every document in the collection for every topic, top-ranked documents from participating *runs* (i.e. system output) are collected and only these documents are manually judged¹. Although this methodology has introduced new problems to IR evaluation such as *incompleteness* (i.e. not all relevant documents in the document collection have been identified) [1, 3, 14] and *biases* (i.e. relevance assessments favour some particular classes of systems or retrieved documents) [16, 27], the IR research community still relies heavily on test collections built through pooling.

Even with pooling, however, the relevance assessment process for test collection construction is costly and time-consuming. For example, a depth-100 pool, i.e. a set of documents obtained by taking the union of top 100 documents from each run for a particular topic, usually contains several hundred documents, depending on the topic itself as well as the number and the diversity of the runs. If you have 50 topics, the total number of documents that need to be judged may be well over 10,000. Hence the present study examines the possibility of reducing manual assessment cost for pooling-based test collections, with a particular focus on the three existing NTCIR-7 ACLIA IR4QA test collections [15, 17]². Although a few methods for allocating different amount of manual effort to different topics or to different systems were proposed more than a decade ago [6, 29], they have not been adopted at venues such as TREC due to bias con-

¹Manual interactive searches can be used to augment the pools, in the hope of improving the coverage of relevant documents [9].

²ACLIA stands for Advanced Cross-lingual Information Access; IR4QA stands for Information Retrieval for Question Answering. But the present study treats the IR4QA test collections as regular document retrieval test collections just like the “ad hoc” test collections from TREC, and the question answering tasks are outside its scope.

cerns [23]. While the TREC million query track has recently reported on successfully obtaining relevance assessments for over 1,800 queries by means of statistical techniques for selecting which documents to judge [4], the traditional pooling approach still offers several advantages, including its simplicity, its independence to any particular evaluation metric, and its “off-line” nature (i.e. document sets to be judged are determined in advance) which facilitates assessment cost estimation, collaboration across geographically-distributed task organisers (such as those for IR4QA, which covers multiple languages), and later analyses.

In contrast to trying to build up “minimal” test collections [4], we take the existing NTCIR-7 IR4QA and the TREC 2004 robust track test collections and empirically investigate whether they can be “boiled down.” More specifically, we investigate the effect of reducing both the topic set size and the pool depth on the outcome of a statistical significance test between two systems, starting with (approximately) 100 topics and depth-100 pools. We define *cost* (of manual relevance assessment) as the pool depth multiplied by the topic set size, and *error* as a system pair whose outcome of statistical significance testing differs from the original result based on the full test collection. Our main findings are: (a) Cost and the number of errors are negatively correlated, and any attempt at substantially reducing cost introduces some errors; (b) The NTCIR-7 IR4QA and the TREC 2004 robust track test collections all yield a comparable and considerable number of errors in response to cost reduction, and this is true despite the fact that the TREC relevance assessments relied on more than twice as many runs as the NTCIR ones; (c) Using 100 topics with depth-30 pools generally yields fewer errors than using 30 topics with depth-100 pools; and (d) Even with depth-100 pools, using fewer than 100 topics results in *false alarms*, i.e. two systems are declared significantly different even though the full topic set would declare otherwise. Hence, researchers should be careful about reporting experimental results based on a single test collection with 50 or even 70 topics – multiple test collections with a large topic set and multiple evaluation metrics should be used to minimise the risk of jumping to wrong conclusions.

The remainder of this paper is organised as follows. Section 2 discusses previous work on reducing manual assessment cost to clarify the contributions of the present study. Section 3 describes the data from the NTCIR-7 ACLIA IR4QA task and the TREC 2004 robust track which we used in our experiments, as well as how we “boiled down” these collections. Section 4 describes how we measure IR effectiveness, statistical significance and the effect of reducing assessment cost. Section 5 presents our experimental results with the NTCIR-7 IR4QA and TREC robust track data. Finally, Section 6 summarises the findings of this paper and discusses future research directions.

2. RELATED WORK

This section discusses previous work on reducing manual assessment cost for test collection construction to clarify the contributions of the present study.

Over a decade ago, some methods for better managing the relevance assessment process at TREC were proposed. Zobel [29] suggested focussing the assessment effort on topics for which many relevant documents have been found so far, based on his experiments with data from TRECs 3-5.

Cormack, Palmer and Clarke [6] suggested focussing the assessment effort on runs that have found many relevant documents so far, based on their experiments with the TREC-6 data. Voorhees [22] expressed concerns for these approaches from the viewpoint of judgment biases, and neither of these methods was adopted at TREC, as mentioned earlier.

Using the TRECs 3-8 ad hoc track and the TRECs 9-10 Web track data, Voorhees and Buckley [23] examined the relationship between the topic set size and the *swap rate*, i.e. the probability of two experiments disagreeing with each other as to which of a given two systems is better. Since each of the TREC test collections they used contained only 50 topics, and since the swap rate method required splitting the full topic set into two, their swap rate measurement could not be performed beyond 25 topics, so they discussed the case of 50 topics based on extrapolation. Sanderson and Zobel [19] re-examined this method with data from TRECs 2-11, but after filtering out system pairs without a statistically significant difference. Again, their direct swap rate measurement was limited to 25 topics and fewer. Recently, Voorhees [25] has applied the swap rate method to the TREC 2004 robust track data with 100 topics from TRECs 7-8. Hence the swap rate for only 50 topics were considered in their experiments.

Sakai [12] proposed an alternative to the swap rate method, which is based on the bootstrap hypothesis test, and showed that the method can provide results that are similar to those with the swap rate method. Since Sakai’s method has a stronger theoretical background than the swap method and can directly examine the errors for the full topic set size, the present study adopts a variant of this method. In contrast to previous studies which examined the case of 25 topics [19, 23, 29] or 50 topics [25], our experiments directly examine the case of (approximately) 100 topics and fewer with each of our four test collections, including the aforementioned TREC data used by Voorhees [25].

One of the findings from the present study is that, using 100 topics with depth-30 pools generally yields fewer errors (in terms of statistical significance) than using 30 topics with depth-100 pools. This is in agreement with several previous studies, all of which are however limited to the cases with TREC data, as discussed below.

The aforementioned study by Sanderson and Zobel [19] suggests that “*it is better to have larger number of topics (perhaps 400) and shallower pools (perhaps depth 10).*” This recommendation was based on plotting the swap rates for both Precision at document cut-off 10 and Average Precision against *assessor effort*, which is exactly what we call *cost* in this paper, and also on the argument that examining a large number of topics with shallow pools will probably yield a large number of relevant documents (and therefore more data points) than examining a small number of topics with deep pools. Carterette and Smucker [3] used all (249) topics from the TREC 2004 robust track and demonstrated that, from the viewpoint of the power of the sign test, using 192 topics with fewer than 5 judgments for each is as good as using 25 topics with 166 judgments for each for their particular data set. However, these 249 topics originate from different TRECs, and used different runs, pool depths, and even relevance scales [24]. Hence we follow [25] and use only topics 351-450 with the TREC 2004 robust data. Webber, Moffat and Zobel [26] used the TREC 2004 robust data with topics 301-450 (along with other TREC data), and argued,

Table 1: Test collection statistics. *L3*-, *L2*- and *L1*-relevant mean “highly relevant,” “relevant” and “partially relevant,” respectively.

	IR4QA-CS	IR4QA-CT	IR4QA-JA	ROBUST04OLD	ROBUST04NEW
#topics	97	95	98	100 (351-400 from TREC-7, 401-450 from TREC-8) approx. 528,000	49 (651-700)
#documents	545,162	1,150,954	419,759		
pool depth	100	100	100	100	100
document language	simplified Chinese	traditional Chinese	Japanese	English	English
# <i>L3</i> -relevant/topic	-	-	-	-	12.5
# <i>L2</i> -relevant/topic	108.2	37.0	54.1	-	-
# <i>L1</i> -relevant/topic	61.6	53.7	56.0	94.0	28.8
#total relevant/topic	169.8	90.7	110.1	94.0	
#judged/topic	597.7	648.1	762.4	1671.8	710.0
#participating runs	40	26	25	110 (evaluated) 103 (used for judging 351-400) 116 (used for judging 401-450)	110
#participating teams	9	5	5	14 (evaluated) 49 (used for judging 351-400) 35 (used for judging 401-450)	14

from the viewpoint of the power of the *t*-test, that “*shallow evaluation of many topics is preferable to deep evaluation of a few.*” The analysis of the TREC million query track data by Carterette *et al.* [4] also supports these suggestions.

For reducing assessment cost, it is probably useful to consider not only *how many* topics to use, but exactly which topics or which combination of topics to use for IR evaluation [7, 28]. However, it is not yet clear how best to “boil down” an existing topic set in this way. Therefore, in our present study, we reduce the topic set size by first sorting the topics with the performance variance across systems, and gradually remove topics with the largest variances. This should represent a worst case scenario for each topic set size. Another related line of research is on the *reusability* of test collections, often examined by removing one system’s contribution to the pool at a time (e.g. [16, 29]). But reusability is beyond the scope of the present study.

There are also approaches to assessment cost reduction that go outside of the basic methodology of pooling followed by relevance assessments: For example, there are attempts at ranking systems without relevance assessments by forming “pseudo-qrels” from the pools (e.g. [17, 21]), and those at ranking systems based on a single system, thereby avoiding pooling altogether [18].

It should be noted that there is very little work reported in the literature that uses data from both TREC and NTCIR and compares across these two forums for the purpose of IR evaluation. Exceptions include [7, 12, 14, 16]. The present study has a strength in that we use three data sets from NTCIR and one from TREC, each with (approximately) 100 topics, and also in that we use metrics designed for graded relevance. Note also that the test collection construction methods employed in the aforementioned TREC million query track [4] were designed specifically for evaluation with Average Precision – a binary relevance metric.

Reducing the pool size and/or the topic set size has also been tried outside TREC and NTCIR, for example, for the purpose of comparing the stability of different binary-relevance metrics for INEX [10].

3. DATA

Table 1 shows some statistics of the data sets that we used for our experiments in “boiling down” test collections, i.e. reducing either the topic set size, the pool depth, or both. “IR4QA-CS,” “IR4QA-CT” and “IR4QA-JA” from the NTCIR-7 ACLIA IR4QA task [15] and “ROBUST04OLD”

from the TREC 2004 robust track [24], each with (approximately) 100 topics, are the four data sets that we mainly use. According to Voorhees [25], the 100 topics of ROBUST04OLD “*were developed using the same methodology, using mostly the same set of assessors, and judged using the same pooling protocol with roughly equal numbers of runs contributing to the pools.*”

Unfortunately, while ROBUST04OLD enables comparisons across NTCIR and TREC, it is not an ideal data set for our purpose: Firstly, its relevance assessments are binary, even though we are primarily interested in evaluation with graded relevance. (*L3*-, *L2*- and *L1*-relevant mean “highly relevant,” “relevant” and “partially relevant,” respectively [15].) However, this is not a serious problem since graded relevance metrics are applicable to binary relevance test collections. A potentially more serious problem with this data set is that the relevance assessments come from TRECs 7-8, and were not constructed using the 110 runs submitted at TREC 2004. In other words, from the viewpoint of the ROBUST04OLD relevance assessments, all of these TREC 2004 runs are “new” systems. This makes a fair comparison across NTCIR and TREC a little difficult.

To remedy these issues with ROBUST04OLD, we conduct an additional set of experiments that involves a fifth data set, which we call “ROBUST04NEW” as shown in Table 1. This data set shares the runs with ROBUST04OLD, but the relevance assessments for these 49 topics were done by actually creating pools based on these runs. Moreover, these new TREC 2004 topics come with graded relevance assessments. The downside is that we only have 49 topics, but we can make comparisons across the five data sets by halving the other topic sets containing 100 topics. Details will follow in the Section 5.2.

Note that the NTCIR relevance assessments rely on only 25-40 runs (5-9 teams), while the TREC relevance assessments rely on over 100 runs (14-35 teams). It can be observed that the average number of documents judged per topic for ROBUST04OLD is more than twice as large as that for the other data sets.

In our main experiments, we “boiled down” test collections, by reducing either the topic set size starting with (approximately) 100 topics, or the pool depth starting with depth-100, or both. The original pool depth was 100 for each test collection. As was mentioned earlier, the original topic set was sorted by the variance across systems according to a graded relevance metric (See Section 4), and the

high-variance topics (i.e. those that are probably the most useful for system discrimination) were gradually removed. In this way, subsets containing 70/50/30 topics were created. (An even smaller topic set size would not be appropriate for statistical significance testing.) On the other hand, shallow-pool “qrels” (i.e. relevance assessments) were created as follows: Let $RET_{r,d}$ be the set of documents returned at or above rank d by run r for a topic. Then $P_d = \bigcup_r RET_{r,d}$ is the *depth- d* pool for this topic. For every document in P_{100} , i.e. depth-100 pool, we already have the relevance assessment in the original qrels. Then, for each topic, we create $P_{90}, P_{70}, P_{50}, P_{30}$ and filter the original qrels with the documents in these reduced pools. Since we have 4 different topic set sizes, and five different pool depths, we have 20 different versions of “test subcollection” for each test collection, including the original collection.

In our additional experiments, we start with (approximately) 50 topics so we only have two different topic set sizes (approximately 50 and 30). Hence we have 10 different versions of “test subcollection” for each test collection in these experiments, as we shall see later.

4. METRICS

4.1 Measuring Retrieval Effectiveness

For evaluating the retrieval performance of each run based on each “test subcollection,” we use Q-measure (Q) and nDCG, which are the official graded-relevance metrics of the NTCIR IR4QA task [15]. We omit Average Precision (AP), but it is known that AP and Q are very highly correlated.

Let \mathcal{L} be a relevance level, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an \mathcal{L} -relevant document [8]. For our graded-relevance data shown in Table 1, we let $gain(L1) = 1$, $gain(L2) = 2$ and $gain(L3) = 3$. Let $I(r) = 1$ if the document retrieved at rank r for a particular topic in a given run is \mathcal{L} -relevant (for some \mathcal{L}) and let $I(r) = 0$ otherwise. Let $C(r) = \sum_{i=1}^r I(i)$, i.e. the number of relevant documents seen so far at rank r . Furthermore, let $R(\mathcal{L})$ denote the number of known \mathcal{L} -relevant documents for a topic, and let $R = \sum_{\mathcal{L}} R(\mathcal{L})$, i.e. the total number of known relevant documents for that topic.

Let $g(r) = gain(\mathcal{L})$ if the document at rank r is \mathcal{L} -relevant and let $g(r) = 0$ otherwise. In particular, let $g^*(r)$ denote the gain at rank r of an *ideal ranked output*, obtained by listing up the R relevant documents in decreasing order of the relevance level. Let $cg(r) = \sum_{i=1}^r g(i)$ and $cg^*(r) = \sum_{i=1}^r g^*(i)$.

Let β be a positive constant. Q is defined as:

$$Q\text{-measure} = \frac{1}{R} \sum_r I(r) \frac{C(r) + \beta cg(r)}{r + \beta cg^*(r)}. \quad (1)$$

Letting $\beta = 0$ reduces Q to AP, and using a large β makes Q more forgiving to relevant documents found near the bottom of the ranked list [13]. We let $\beta = 1$ in this paper.

Let l be a document cut-off value. The version of nDCG we use is defined as [2]:

$$nDCG = \frac{\sum_{r=1}^l g(r) / \log(r+1)}{\sum_{r=1}^l g^*(r) / \log(r+1)}. \quad (2)$$

We let $l = 1000$ in this paper.

4.2 Measuring Statistical Significance

There are several ways to test statistical significance given paired data: but the t -test relies on the normality assumption; the paired Wilcoxon test relies on the symmetry assumption; and the sign test has low power. In contrast, the bootstrap test is distribution-free, and yet behaves similarly to the t -test in practice. It can also be used to replace the swap rate method (See Section 2) for estimating the absolute mean performance difference required to achieve statistical significance [12]. Hence we use the bootstrap test in our experiments³.

Let \mathcal{Q} be the set of topics provided in the test collection, and let $n = |\mathcal{Q}|$. Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ denote the per-topic effectiveness values of systems X and Y as measured by some metric (e.g. nDCG). We want to know whether the population means for X and Y , which we denote by μ_X and μ_Y , are any different. We thus set up the following hypotheses for $\mu = \mu_X - \mu_Y$:

$$H_0 : \mu = 0 \quad vs \quad H_1 : \mu \neq 0.$$

Let $\mathbf{z} = (z_1, \dots, z_n)$ where $z_i = x_i - y_i$. As with standard significance tests, we assume that \mathbf{z} is an independent and identically distributed sample drawn from an unknown distribution. Let $\bar{z} = \sum_i z_i / n$, and let $\mathbf{w} = (w_1, \dots, w_n)$ where $w_i = z_i - \bar{z}$, in order to create *bootstrap samples* \mathbf{w}^{*b} ($b = 1, \dots, B$) of per-topic performance differences that obey H_0 . Figure 1 shows the algorithm for obtaining B bootstrap samples of topics (\mathcal{Q}^{*b}) and the corresponding values for \mathbf{w}^{*b} . We let $B = 1000$ throughout this paper.

```

for  $b = 1$  to  $B$ 
  create topic set  $\mathcal{Q}^{*b}$  of size  $n = |\mathcal{Q}|$  by
  randomly sampling with replacement from  $\mathcal{Q}$ ;
  for  $i = 1$  to  $n$ 
     $q = i$ -th topic from  $\mathcal{Q}^{*b}$ ;
     $w_i^{*b} =$  observed value in  $\mathbf{w}$  for topic  $q$ ;

```

Figure 1: Algorithm for creating bootstrap samples \mathcal{Q}^{*b} and $\mathbf{w}^{*b} = (w_1^{*b}, \dots, w_n^{*b})$ for the paired bootstrap test.

For each trial b , let \bar{w}^{*b} and $\bar{\sigma}^{*b}$ denote the mean and the standard deviation of \mathbf{w}^{*b} . Consider a test statistic:

$$t(\mathbf{z}) = \frac{\bar{z}}{\bar{\sigma} / \sqrt{n}}$$

where $\bar{\sigma}$ is the standard deviation of \mathbf{z} , given by

$$\bar{\sigma} = \left(\sum_i (z_i - \bar{z})^2 / (n - 1) \right)^{\frac{1}{2}}.$$

Figure 2 shows how to compute the Achieved Significance Level (ASL) using \mathbf{w}^{*b} . In essence, we examine how *rare* the observed difference would be under H_0 . If $ASL < 0.05$, we reject H_0 : we have enough evidence to state that μ_X and μ_Y are probably different.

³A recent study [20] recommends the randomization test, which is also distribution-free. But it appears that the IR community has yet to agree on which significance tests are more reliable than others.

Table 2: The effect of cost reduction on errors (misses/false alarms). Significance test results based on the full test collection are taken to be the ground truth.

		Q					nDCG				
		d100	d90	d70	d50	d30	d100	d90	d70	d50	d30
IR4QA-CS		#significantly different pairs=7/39					#significantly different pairs=7/39				
	97t	-	1 (1/0)	1 (1/0)	4 (2/2)	7 (2/5)	-	1 (1/0)	2 (2/0)	4 (3/1)	5 (3/2)
	70t	3 (3/0)	4 (4/0)	5 (3/2)	6 (3/3)	8 (3/5)	2 (2/0)	2 (2/0)	3 (3/0)	3 (3/0)	3 (3/0)
	50t	4 (3/1)	4 (3/1)	4 (3/1)	4 (3/1)	7 (3/4)	4 (3/1)	3 (3/0)	3 (3/0)	3 (3/0)	4 (3/1)
	30t	5 (3/2)	5 (3/2)	6 (4/2)	6 (3/3)	8 (5/3)	6 (5/1)	6 (5/1)	7 (5/2)	9 (6/3)	11 (6/5)
IR4QA-CT		#significantly different pairs=5/25					#significantly different pairs=4/25				
	95t	-	0 (0/0)	1 (1/0)	1 (1/0)	1 (1/0)	-	0 (0/0)	0 (0/0)	1 (0/1)	1 (0/1)
	70t	3 (1/2)	3 (1/2)	3 (1/2)	3 (1/2)	5 (1/4)	5 (1/4)	5 (1/4)	5 (1/4)	4 (0/4)	4 (0/4)
	50t	6 (1/5)	6 (1/5)	7 (1/6)	5 (1/4)	7 (2/5)	6 (1/5)	6 (1/5)	6 (0/6)	7 (1/6)	8 (1/7)
	30t	10 (2/8)	10 (2/8)	11 (2/9)	7 (2/5)	6 (2/4)	8 (2/6)	9 (2/7)	9 (1/8)	8 (1/7)	9 (2/7)
IR4QA-JA		#significantly different pairs=6/24					#significantly different pairs=5/24				
	98t	-	0 (0/0)	0 (0/0)	0 (0/0)	0 (0/0)	-	0 (0/0)	1 (1/0)	1 (1/0)	1 (1/0)
	70t	3 (1/2)	3 (1/2)	3 (1/2)	3 (1/2)	4 (2/2)	2 (1/1)	2 (1/1)	3 (2/1)	4 (2/2)	5 (3/2)
	50t	7 (4/3)	6 (3/3)	4 (2/2)	5 (3/2)	7 (4/3)	2 (1/1)	3 (2/1)	4 (2/2)	5 (3/2)	6 (4/2)
	30t	5 (4/1)	5 (4/1)	5 (4/1)	5 (4/1)	5 (4/1)	4 (4/0)	5 (5/0)	4 (4/0)	4 (4/0)	4 (4/0)
ROBUST04OLD		#significantly different pairs=3/109					#significantly different pairs=5/109				
	100t	-	1 (0/1)	1 (0/1)	1 (0/1)	1 (0/1)	-	2 (1/1)	2 (1/1)	2 (1/1)	3 (1/2)
	70t	4 (2/2)	6 (1/5)	6 (1/5)	5 (1/4)	5 (1/4)	4 (3/1)	6 (2/4)	8 (3/5)	9 (3/6)	14 (3/11)
	50t	6 (1/5)	8 (1/7)	8 (1/7)	7 (1/6)	8 (1/7)	9 (3/6)	10 (3/7)	11 (3/8)	11 (3/8)	13 (3/10)
	30t	6 (3/3)	8 (3/5)	8 (3/5)	8 (3/5)	9 (3/6)	8 (4/4)	10 (4/6)	12 (4/8)	9 (4/5)	12 (4/8)

```

count = 0;
for b = 1 to B
    t(w*b) = w*b / (σ*b / √n);
    if( |t(w*b)| ≥ |t(z)| ) then count++;
ASL = count/B;

```

Figure 2: Algorithm for estimating the Achieved Significance Level based on the paired bootstrap test.

4.3 Measuring the Effect of Reducing Assessment Cost

We measured the negative effect of boiling down test collections as follows. First, for each of the full test collection, the runs were sorted by mean Q or mean nDCG. Then, for every pair of adjacent runs in the sorted list, we conducted a two-sided paired bootstrap test as described above. Hence, for a run list of size n , we conducted $n-1$ significance tests⁴. Significantly different run pairs were recorded: this set is denoted by C_* . We then re-evaluated each run using a “test subcollection” and applied significance testing to the same sorted list (i.e. the same set of adjacent run pairs based on the full test collection). The set of significantly different run pairs thus obtained is denoted by C . Then we counted the number of *misses* given by $|C_* - C|$, *false alarms* given by $|C - C_*|$ and *errors* given as the sum of misses and false alarms⁵. Thus, if the full test collection detects a significance difference for a run pair and a test subcollection does not for the same pair, that is a miss. If a test subcollection detects a significance difference for a run pair, and the full test collection does not for the same pair, that is a false alarm. Note that we assume that the results with the full test collection are the ground truth: we have no other choice.

⁴Sakai [12] conducted a significant test for every run pair ($n(n-1)/2$ pairs) but we only consider adjacent runs to save computational cost. Note, however, that pairwise statistical significance is not transitive.

⁵In principle, another kind of error is possible, where runs X and Y are significantly different according to both test collections, but one says X outperforms Y while the other says Y outperforms X . We verified that such conflicts never occurred in our experiments.

In our experiments, we investigate the relationship between the number of errors with *cost*, which we define as the number of topics multiplied by the pool depth. We do not consider the cost of topic development [4], although we acknowledge that this in practice has a considerable impact on the overall cost of test collection construction.

5. RESULTS AND DISCUSSIONS

5.1 Main Results with 100 Topics

Table 2 summarises the results of our main experiments in boiling down test collections with (approximately) 100 topics. For each test collection, the “true” number of significantly different run pairs (i.e. $|C_*|$) are shown together with the total number of pairwise comparisons. For example, for IR4QA-CS, since we have 40 runs, comparing adjacent pairs in the sorted run list yields 39 comparisons. In the table, “30t” means a test subcollection with the topic set reduced to 30 topics, and “d30” means evaluation with depth-30 pools, and so on. For example, in our experiments with IR4QA-CT and Q-measure, it can be observed that, while 95 topics with depth-30 pools yield only one error (one miss and no false alarms), 30 topics with depth-100 pools yield as many as 10 errors (two misses and eight false alarms). With the exception of the IR4QA-CS results, the observable trend is that *using 100 topics with depth-30 pools generally yields fewer errors than using 30 topics with depth-100 pools*, as indicated in **bold**.

The “d100” columns in Table 2 show that, *even with depth-100 pools, using fewer than 100 topics yields false alarms*. This is alarming, given that many IR studies are based on evaluation with some 50 topics and finding a statistically significant difference between a proposed method and a baseline. The results demonstrate that the significance may well disappear when a larger topic set is used. Of course, significance testing is always associated with both the probability of Type I Error (concluding that systems X and Y are different although in truth they are not) and that of Type II Error (concluding that X and Y are the same although in truth they are not), so this is not altogether surprising. However, it should remind researchers to be careful about reporting experimental results based on a single test

collection – multiple test collections with a large topic set and multiple evaluation metrics should be used to minimise the risk of jumping to wrong conclusions.

Figure 3 visualises Table 2, by plotting the misses, false alarms and errors against *cost*. For a full test collection with 100 topics, each with a depth-100 pool, the cost is 10,000. A test subcollection with 30 topics and depth-70 pools, and one with 70 topics and depth-30 pools, both correspond to the cost of 2,100. It is clear that *cost and the number of errors are negatively correlated, and any attempt at substantially reducing cost introduces some errors*.

Figure 3 also shows that *the NTCIR-7 IR4QA and the TREC 2004 robust track test collections all yield a comparable and considerable number of errors in response to cost reduction, and this is true despite the fact that the TREC relevance assessments relied on more than twice as many runs as the NTCIR ones*: over 100 vs. 25-40 runs (See Table 1). Recall also that the number of judged documents for ROBUST04OLD is also at least twice as high as those for IR4QA, which indicates the diversity of the TREC runs. Hence, collecting many runs does not seem to have helped here. But are we underestimating the robustness of the TREC data to cost reduction just because the 110 evaluated runs are “new” from the viewpoint of the ROBUST04OLD relevance assessments? We shall discuss this question in the next section.

5.2 Additional Results with 50 Topics

To remedy the fact that the runs that were evaluated with the ROBUST04OLD topics were not used for the relevance assessments of these topics, we conducted additional sets of experiments with ROBUST04NEW which has 49 topics with relevance assessments that are actually based on the 110 runs. As was explained in Section 3, we also used the 50-topics versions of the three IR4QA and the ROBUST04OLD data sets, which we obtained in our main experiments through variance-based topic set reduction. The only difference between our main experiments and our additional experiments with the IR4QA and ROBUST04OLD data is that, in the additional experiments, the reduced topic set containing 50 topics was treated as the ground truth. That is, we act as if these four test collections had only 50 topics from the beginning. This makes comparisons across the five data sets easier.

Table 3 summarises the results of boiling down the five test collections, starting with (approximately) 50 topics. Unfortunately, for ROBUST04NEW, there was no significantly different run pair from the beginning (i.e. $|C_*| = 0$), and therefore, by definition, misses do not happen. However, it can be observed that even ROBUST04NEW is quite vulnerable to topic set reduction: for example, with 30 topics and depth-30 pools, it yields as many as 13 false alarms in terms of Q-measure, which is the largest number of errors observed across the five data sets. Hence, regardless of whether the relevance assessments originate from the evaluated runs (ROBUST04NEW) or not (ROBUST04OLD), the TREC data appear to be as vulnerable to any attempt at cost reduction as the NTCIR data. (With the full set of 49 topics, ROBUST04NEW appears to be relatively robust to shallow pool depths. However, this may be an artifact of the fact that $|C_*| = 0$.)

6. CONCLUSIONS AND FUTURE WORK

Through extensive experiments using data from both NTCIR and TREC, each with (approximately) 100 topics, we explored the possibilities of boiling down existing test collections to reduce manual assessment cost. Our main findings are: (a) Cost and the number of errors are negatively correlated, and any attempt at substantially reducing cost introduces some errors; (b) The NTCIR-7 IR4QA and the TREC 2004 robust track test collections all yield a comparable and considerable number of errors in response to cost reduction, and this is true despite the fact that the TREC relevance assessments relied on more than twice as many runs as the NTCIR ones; (c) Using 100 topics with depth-30 pools generally yields fewer errors than using 30 topics with depth-100 pools; and (d) Even with depth-100 pools, using fewer than 100 topics results in *false alarms*, i.e. two systems are declared significantly different even though the full topic set would declare otherwise.

Finding (a) suggests that these test collections are already “minimal” in a way, and that the effort and cost spent on the relevance assessments were basically worthwhile. Finding (b) suggests that using many participating runs/teams for constructing a test collection does not guarantee a higher reliability of experiments using that collection. Finding (c) is in agreement with previous studies [3, 4, 19, 26], but we have generalised this beyond TREC data. Finding (d) should remind researchers to be careful about reporting experimental results based on a single test collection with 50 or even 70 topics. This also generalises a recent finding [25] beyond TREC data.

Clearly, there are limitations to this study. Our basic assumption is that the results based on the full test collection is the ground truth, but obviously even these results could be “wrong.” Just as significant differences based on 50 topics can disappear when we add 50 more topics (as we have demonstrated), significant differences based on 100 topics may well disappear when we add (say) 100 more topics. This we do not know because we do not have 200 topics. Of course, there will always be Type I Errors and Type II Errors – researchers should use multiple test collections with a large topic set and multiple evaluation metrics to minimise the risk of jumping to wrong conclusions.

One important direction for future research is selecting a minimal topic set for reliable IR evaluation [7, 28]. Which combination of topics to include is probably more important than just how many. Another direction is selecting which documents to judge for relevance without introducing any biases towards run or documents, and without relying on a particular evaluation metric such as Average Precision. Finally, a selective usage of submitted runs for forming a reliable and reusable relevance assessments is also an interesting research question, since the present study has witnessed that a test collection built based on many runs does not necessarily yield more reliable experiments than one built based on much fewer runs does. Intuitively, it would make sense to select runs *per topic*, rather than to use a frozen subset of runs for all topics.

7. ACKNOWLEDGMENTS

We thank Hideki Shima for his technical support, as well as all NTCIR and TREC organisers/participants for their efforts and contributions to test collection building and eval-

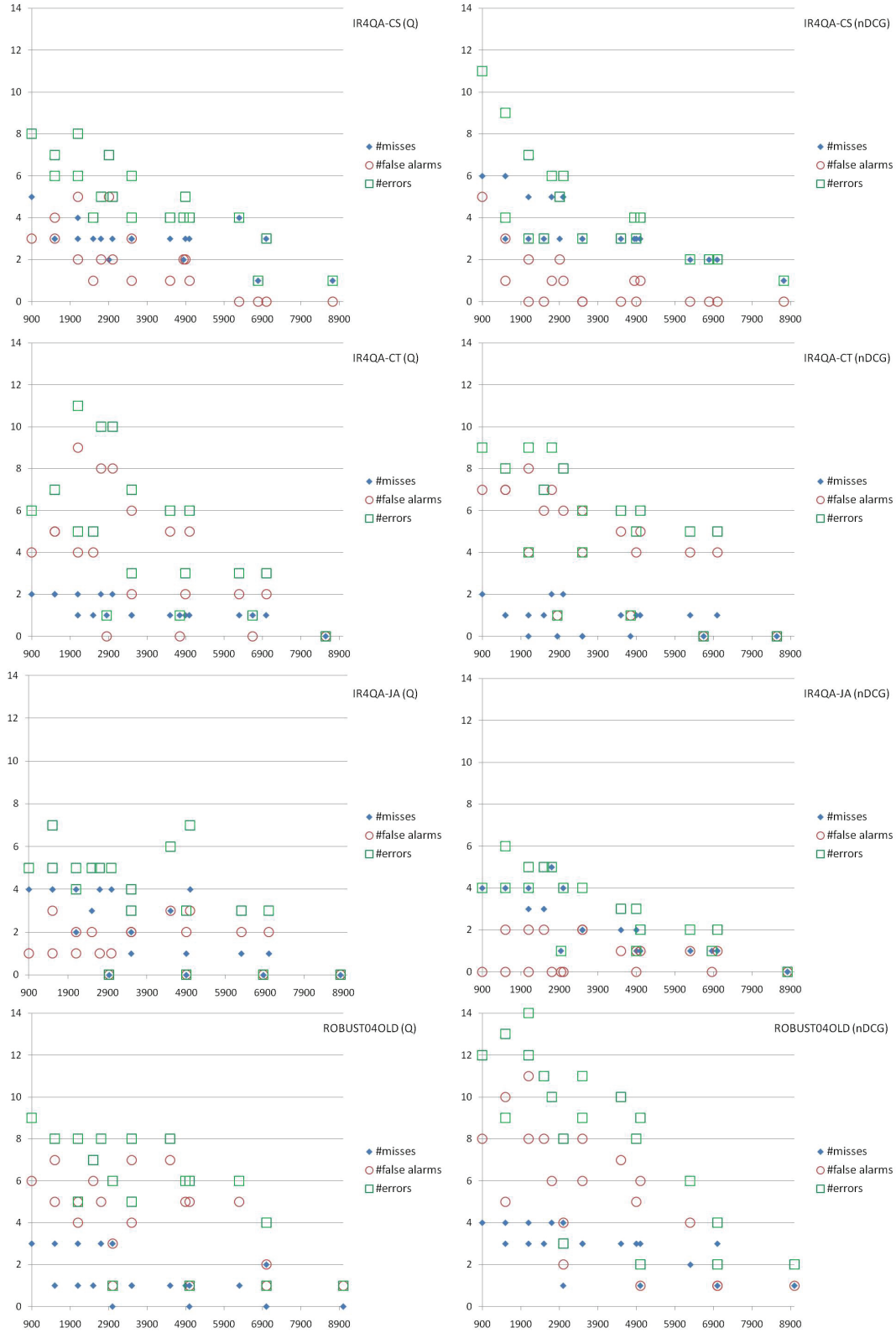


Figure 3: The relationship between cost and errors/false alarms/misses.

Table 3: Additional results on the effect of cost reduction on errors (misses/false alarms). For ROBUST04NEW, significance test results based on the full test collection (49 topics with depth-100 pools) are taken to be the ground truth. For the other data sets, the results based on the size-50 reduced topic sets with depth-100 pools are taken to be the ground truth.

		Q					nDCG				
		d100	d90	d70	d50	d30	d100	d90	d70	d50	d30
IR4QA-CS		#significantly different pairs=5/39					#significantly different pairs=5/39				
	50t	-	0 (0/0)	2 (1/1)	2 (1/1)	5 (1/4)	-	4 (2/2)	2 (1/1)	2 (1/1)	5 (1/4)
	30t	1 (0/1)	1 (0/1)	2 (1/1)	2 (0/2)	4 (2/2)	6 (4/2)	6 (4/2)	7 (4/3)	7 (4/3)	9 (4/5)
IR4QA-CT		#significantly different pairs=9/25					#significantly different pairs=8/25				
	50t	-	0 (0/0)	1 (0/1)	3 (2/1)	5 (3/2)	-	0 (0/0)	2 (0/2)	1 (0/1)	2 (0/2)
	30t	6 (2/4)	6 (2/4)	7 (2/5)	7 (4/3)	6 (4/2)	4 (2/2)	3 (1/2)	5 (1/4)	4 (1/3)	3 (1/2)
IR4QA-JA		#significantly different pairs=5/24					#significantly different pairs=5/24				
	50t	-	1 (0/1)	3 (1/2)	2 (1/1)	2 (1/1)	-	1 (1/0)	2 (1/1)	3 (2/1)	4 (3/1)
	30t	4 (3/1)	4 (3/1)	4 (3/1)	4 (3/1)	4 (3/1)	4 (4/0)	5 (5/0)	4 (4/0)	4 (4/0)	4 (4/0)
ROBUST04OLD		#significantly different pairs=7/109					#significantly different pairs=8/109				
	50t	-	4 (1/3)	4 (1/3)	3 (1/2)	4 (1/3)	-	1 (0/1)	2 (0/2)	4 (0/4)	4 (0/4)
	30t	8 (6/2)	10 (6/4)	10 (6/4)	10 (6/4)	11 (6/5)	9 (6/3)	5 (3/2)	5 (2/3)	6 (4/2)	5 (2/3)
ROBUST04NEW		#significantly different pairs=0/109					#significantly different pairs=2/109				
	49t	-	0 (0/0)	0 (0/0)	1 (0/1)	1 (0/1)	-	0 (0/0)	0 (0/0)	0 (0/0)	0 (0/0)
	30t	11 (0/11)	11 (0/11)	11 (0/11)	11 (0/11)	13 (0/13)	5 (2/3)	5 (2/3)	5 (2/3)	5 (1/4)	5 (1/4)

uation. We also thank the RIAO reviewers for their advice.

8. REFERENCES

- [1] Buckley, C and Voorhees, E. M. (2004). Retrieval Evaluation with Incomplete Information, In *Proceedings of ACM SIGIR 2004*, pp. 25-32.
- [2] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G. (2005). Learning to Rank using Gradient Descent, In *Proceedings of ACM ICML 2005*, pp. 89-96.
- [3] Carterette, B. and Smucker, M. (2007). Hypothesis Testing with Incomplete Relevance Judgments. In *Proceedings of ACM CIKM 2007*, pp. 643-652.
- [4] Carterette, B., Pavlu, V., Kanoulas, E., Aslam, J. A. and Allan, J. (2008). Evaluation over Thousands of Queries. In *Proceedings of ACM SIGIR 2008*, pp. 651-658.
- [5] Carterette, B. (2009). On Rank Correlation and the Distance between Rankings. In *Proceedings of ACM SIGIR 2009*, pp. 436-443.
- [6] Cormack, G. V., Palmer, C. R. and Clarke, L. A. (1998). Efficient Construction of Large Test Collections, In *Proceedings of ACM SIGIR '98*, pp. 282-289.
- [7] Guiver, J., Mizzaro, S. and Robertson, S. (2009). A Few Good Topics: Experiments in Topic Set Reduction for Retrieval Evaluation. *ACM Transactions on Information Systems*, to appear.
- [8] Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20(4), pp. 422-446.
- [9] Kuriyama, K., Yoshioka, M. and Kando, N. (2001). The Effect of Cross-Lingual Pooling on Evaluation, In *Proceedings of NTCIR-2*, pp. 297-310.
- [10] Pal, S., Mitra, M and Chakraborty, A. (2008). Stability of INEX 2007 Evaluation Measures, In *Proceedings of EVIA 2008*, pp. 23-29.
- [11] Robertson, S. (2008). On the History of Evaluation in IR. *Journal of Information Science*, 34(4), pp. 439-456.
- [12] Sakai, T. (2007). Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests. *IPSJ Transactions on Databases*, Vol.48, No.SIG 9 (TOD35), pp.11-28.
- [13] Sakai, T. (2007). On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pp.32-43.
- [14] Sakai, T. and Kando, N. (2008). On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments. *Information Retrieval*, 11(5), pp. 447-470.
- [15] Sakai, T., Kando, N., Lin, C.-J., Mitamura, T., Shima, H., Ji, D., Chen, K.-H. and Nyberg, E. (2008). Overview of the NTCIR-7 ACLIA IR4QA Task. In *Proceedings of NTCIR-7*, pp. 77-114.
- [16] Sakai, T. (2008). On the Robustness of Information Retrieval Metrics to Biased Relevance Assessments, *Journal of Information Processing*, pp. 156-166.
- [17] Sakai, T., Kando, N., Shima, H., Lin, C.-J., Song, R., and Mitamura, T. (2009). Ranking the NTCIR ACLIA IR4QA Systems without Relevance Assessments. *DBSJ Journal*, 8(2), pp.1-6.
- [18] Sanderson, M. and Joho, H. (2004). Forming Test Collections with No System Pooling. In *Proceedings of ACM SIGIR 2004*, pp. 33-40.
- [19] Sanderson, M. and Zobel, J. (2005). Information Retrieval Evaluation: Effort, Sensitivity, and Reliability. In *Proceedings of ACM SIGIR 2005*, pp. 162-169.
- [20] Smucker, M. D., Allan, J. and Carterette, B.: Agreement Among Statistical Significance Tests for Information Retrieval Evaluation at Varying Sample Sizes, In *Proceedings of ACM SIGIR 2009*, pp. 630-631.
- [21] Soboroff, I., Nicholas, C. and Cahan, P. (2001). Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of ACM SIGIR 2001*, pp. 66-73.
- [22] Voorhees, E. M. (2002). The Philosophy of Information Retrieval Evaluation. In *Revised Papers from the Second Workshop of CLEF (Lecture Notes in Computer Science 2406)*, pp. 355-370.
- [23] Voorhees, E. M. and Buckley, C. (2002). The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of ACM SIGIR 2002*, pp. 316-323.
- [24] Voorhees, E. M. (2005). Overview of the TREC 2004 Robust Retrieval Track. In *Proceedings of TREC 2004*.
- [25] Voorhees, E. M. (2009). Topic Set Size Redux. In *Proceedings of ACM SIGIR 2009*, pp. 806-807.
- [26] Webber, W., Moffat, A. and Zobel, J.: Statistical Power in Retrieval Experimentation. In *Proceedings of ACM CIKM 2008*, pp. 571-580.
- [27] Webber, W. and Park, L. A.: Score Adjustment for Correction of Pooling Bias. In *Proceedings of ACM SIGIR 2009*, pp. 444-451.
- [28] Zhu, J., Wang, J., Cox, I. and Vinay, V. (2009). Topic (Query) Selection for IR Evaluation. In *Proceedings of ACM SIGIR 2009*, pp. 802-803.
- [29] Zobel, J. (1998). How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In *Proceedings of ACM SIGIR '98*, pp. 307-314.