社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (1)
2008－ＦＩ－ 91 (1)
2008／6／19

Comparing Metrics across TREC and NTCIR: The Robustness to System Bias

Tetsuya Sakai

NewsWatch, Inc.
tetsuyasakai@acm.org,sakai@newswatch.co.jp

Abstract

Test collections are growing larger, and relevance data constructed through pooling are suspected of becoming more and more incomplete and biased. Several studies have used evaluation metrics specifically designed to handle this problem, but most of them have only examined the metrics under incomplete but unbiased conditions, using random samples of the original relevance data. This paper examines nine metrics in more realistic settings, by reducing the number of pooled systems. Even though previous work has shown that metrics based on a condensed list, obtained by removing all unjudged documents from the original ranked list, are effective for handling very incomplete but unbiased relevance data, we show that they are not necessarily superior to traditional metrics in the presence of system bias. Using data from both TREC and NTCIR, we first show that condensed-list metrics overestimate new systems while traditional metrics underestimate them, and that the overestimation tends to be larger than the underestimation. We then show that, when relevance data is heavily biased towards a single team or a few teams, the condensed-list versions of Average Precision (AP), Q-measure (Q) and normalised Discounted Cumulative Gain (nDCG), which we call AP′, Q′ and nDCG′, are not necessarily superior to the original metrics in terms of discriminative power, i.e., the overall ability to detect pairwise statistical significance. Nevertheless, AP′ and Q′ are generally more discriminative than bpref and the condensed-list version of Rank-Biased Precision (RBP), which we call RBP′.

1 Introduction

Test collections are growing larger, and relevance data constructed through pooling are suspected of becoming more and more *incomplete* and *biased* [6, 7, 9]. Relevance data are incomplete if there exist some relevant documents among the *unjudged* documents in the test collection. Furthermore, incomplete relevance data are biased if they represent some limited aspects of the complete set of relevant documents. For example, if the number of pooled systems is small, the resultant test collection may overestimate these systems and underestimate systems that did not contribute to the pool, since these new systems are likely to retrieve relevant documents that are outside the set of known relevant documents. We will refer to this phenomenon as *system bias*. Bias may also be caused by *shallow pools*: If only documents at the very top of submitted ranked lists are judged, the resultant relevance data may contain relevant documents that are very easy to retrieve, but not those that are difficult to retrieve. For example, Buckley *et al.* [7] report that the TREC 2005 HARD/Robust test collection is biased towards documents that contain topic title words due to shallow pools. We will refer to this phenomenon as *pool depth bias*.

The objective of this paper is to examine the robustness of retrieval effectiveness metrics in the presence of system bias, with an emphasis on those that can handle graded relevance. Several researchers have proposed evaluation metrics specifically for handling the incompleteness of relevance data, but most of them have only examined the metrics under *incomplete but unbiased* conditions, using random samples of the original relevance data [1, 4, 6, 15, 17, 25]. While random sampling may mimic a situation where the number of judged documents is extremely small compared to the entire document collection, it does not address the problems due to system bias and pool depth bias. Therefore, this paper examines metrics in more realistic settings, by reducing the number of pooled systems. We have also examined the effect of pool depth bias, but will report on the results elsewhere [18].

The main contributions of this paper are as follows. First, we examine as many as nine metrics for handling system bias in test collections. The metrics examined are: *Average Precision* (AP), *Q-measure* (Q) [13], *normalised Discounted Cumulative Gain* (nDCG) [10], *Rank-Biased Precision* (RBP) [12], *binary preference* (bpref) [6], AP′, Q′, nDCG′ and RBP′. The latter four metrics are AP, Q, nDCG and RBP applied to a *condensed list* [15], obtained by removing all unjudged documents from the original ranked list. Thus, just like bpref, these four metrics assume that retrieved unjudged documents are *nonexistent*, while traditional metrics assume that the unjudged documents are *nonrelevant*. Even though previous work has shown that condensed-list metrics are effective for handling very incomplete but unbiased relevance data, we show that they are not necessarily superior to traditional metrics in the presence of system bias. This discrepancy suggests that the results reported in previous studies that used random sampling should be interpreted with caution. Second, our extensive experiments cover two independent evaluation efforts, TREC and NTCIR, and utilise their graded relevance data. This is in contrast to most existing studies that are limited to TREC data and binary-relevance metrics [1, 6, 25]. Since our results are consistent across all of our data sets, we believe that our findings are general. Our main findings are:

1. Condensed-list metrics overestimate systems that did not contribute to the pool while traditional metrics underestimate them, and the overestimation is larger than the underestimation.
2. When runs from a single team or a few teams is used for forming the relevance data, AP′, Q′, nDCG′ are not necessarily superior to AP, Q and nDCG in terms of *discriminative power*, i.e., the overall ability to detect pairwise statistical significance [16]. Nevertheless, AP′ and Q′ are generally more discriminative than bpref and RBP′.

The first observation above substantially generalises a finding by Büttcher *et al.* [9], who analysed a TREC Terabyte data set and observed that "Where AP underestimates the performance of a [new] system, bpref overestimates it."

Section 2 discusses previous work, and Section 3 formally defines the nine metrics considered in this study. Section 4 describes the graded-relevance data and runs from TREC and NTCIR which we use for comparing the metrics. Section 5 reports on our leave-one-team-out experiments for examining how our metrics handle new systems. Section 6 reports on our take-just-one-team and take-just-three-teams experiments for examining the robustness of our metrics to heavy system bias. Finally, Section 7 concludes this paper.

2 Related Work

A decade ago, Zobel [26] examined the effect of pool depth and that of leaving out one run for forming the TREC relevance data. As TREC test collections at that time, i.e., TRECs 3-5, were based on binary relevance, he used binary-relevance metrics such as 11-point average precision. Subsequently, TREC adopted his leave-one-out methodology for validating their test collections, but chose to leave out one participating *team* at a time since each team usually contributes multiple runs to a pool [7, 21]. The present study also includes leave-one-*team*-out experiments as well as "take-just-one-team" experiments which relies on runs from a single team to form the relevance data. Sanderson and Joho [19] have examined a "take-just-one-*run*" approach, but they considered AP only, using data from TRECs 5-8. The present study compares nine metrics, and our analysis covers recent TREC and NTCIR data.

Büttcher *et al.* [9] and Aslam and Yilmaz [2] have proposed methods for "expanding" the original incomplete relevance data. Neither of these studies used graded relevance. In contrast, the focus of the present study is on the choice of metrics *given* a set of relevance assessments which may or may not be incomplete and biased.

Most existing studies that compared metrics for evaluation with incomplete data used the aforementioned random sampling [1, 4, 6, 15, 17, 25]. For example, Yilmaz and Aslam [25] used this approach to evaluate their proposed metrics, including *Induced AP* which is exactly what we call AP′, and *Inferred AP* which aims to estimate the true value of AP. An exception is the aforementioned work by Büttcher *et al.* [9] which included leave-one-team-out experiments to address the system bias issue. Their experiments covered a condensed-list version of precision at document cut-off 20 and *RankEff*[1]. However, precision is an unreliable metric [5, 13], and RankEff is in fact as unreliable as bpref by definition, as we shall clarify in Section 3.3.

Among the studies that used random sampling, Sakai [15] compared condensed-list metrics such as AP′, Q′, nDCG′ and bpref along with traditional metrics, using data sets from NTCIR. Sakai and Kando [17] repeated the experiments using graded-relevance data from TREC and NTCIR, and added RBP to their candidate metrics; they did not examine RBP′. The study showed that, under very incomplete but unbiased conditions, AP′, Q′, nDCG′ are superior to AP, Q, nDCG, bpref and RBP. In contrast, the present study shows that AP′, Q′, nDCG′ are not necessarily superior to AP, Q, nDCG in the presence of system bias.

Traditional metrics assume that retrieved unjudged documents are nonrelevant, while condensed-list metrics assume that retrieved unjudged documents are nonexistent. As a third approach, Baillie, Azzopardi and Ruthven [3] have proposed to quantify the uncertainty in system comparisons by reporting the proportion of unjudged documents within ranked lists, along with metric values such as AP.

3 Formal Definitions of Metrics

This section formally defines the nine metrics examined in this paper. It also explains why RankEff [9] is not included in our experiments.

3.1 AP, Q, nDCG and RBP

Let R denote the number of judged relevant documents. For any given ranked list of documents, let $isrel(r)$ be 1 if the document at rank r is relevant and 0 otherwise. Let $count(r) = \sum_{i \leq r} isrel(i)$. Clearly, precision at rank r is given by $P(r) = count(r)/r$. Hence AP is defined as:

$$AP = \frac{1}{R} \sum_r isrel(r) P(r) \,. \tag{1}$$

Let $\mathcal{L}$ denote a relevance level, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving a judged $\mathcal{L}$-relevant document. Without loss of generality, we follow the NTCIR tradition and let $\mathcal{L} \in \{S, A, B\}$ [11]. As for the TREC graded relevance data, we treat "highly relevant" documents as S-relevant and "relevant" documents as B-relevant. We chose not to treat the latter as A-relevant as it is known that binary relevance data created at TREC contain a considerable amount of partially or marginally relevant documents [20]. Moreover, we let $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$ hereafter as Q and nDCG are robust to the choice of gain values [13].

Let $g(r) = gain(\mathcal{L})$ if the document at rank r is $\mathcal{L}$-relevant and $g(r) = 0$ otherwise, i.e., if the document at rank r is either judged nonrelevant or unjudged. The *cumulative gain* at rank r is given by $cg(r) = \sum_{1 \leq i \leq r} g(i)$. Consider an *ideal* ranked list of documents, which satisfies $g(r) > 0$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$. For NTCIR, listing up all S-, A- and B-relevant documents in this order produces an ideal ranked output. Let $cg_I(r)$ denote the cumulative gain of the ideal list. Q is defined as:

$$Q\text{-}measure = \frac{1}{R} \sum_r isrel(r) BR(r) \tag{2}$$

$$BR(r) = \frac{\beta cg(r) + count(r)}{\beta cg_I(r) + r} \tag{3}$$

where β is a parameter for reflecting the persistence of the user [14]. Clearly, $\beta = 0$ reduces Q to AP; we let $\beta = 1$ throughout this paper.

For a given logarithm base a, let the *discounted* gain at Rank r be $dg(r) = g(r)/\log_a(r)$ for $r > a$ and $dg(r) = g(r)$ for $r \leq a$. Similarly, let $dg_I(r)$ denote the discounted gain for an ideal ranked list. nDCG at document cut-off l is defined as:

$$nDCG_l = \sum_{1 \leq r \leq l} dg(r) / \sum_{1 \leq r \leq l} dg_I(r) \,. \tag{4}$$

Throughout this paper, we let $l = 1000$ as it is known that small document cut-offs hurt the stability of nDCG [13]. This original definition of nDCG is "buggy" in that a relevant document retrieved at rank 1 and one retrieved at rank a receive the same credit. We adhere to the original nDCG but let $a = 2$ to alleviate the effect of the bug[1].

Let $\mathcal{H}$ denote the highest relevance level across all topics. In all of our experiments, $\mathcal{H} = S$. Let p be the persistence parameter that represents the fixed probability that the user moves from a document at rank r to rank $(r+1)$. RBP is defined as:

$$RBP = \frac{1-p}{gain(\mathcal{H})} \sum_r g(r) p^{r-1} . \quad (5)$$

Moffat and Zobel [12] explored $p = 0.5, 0.8, 0.95$, and Sakai and Kando [17] showed that $p = 0.95$ is the best choice among these three values in terms of system ranking stability and discriminative power. Hence we use $p = 0.95$ thoughout this paper. RBP is different from the other metrics considered in this paper in that it totally disregards recall. Sakai and Kando [17] have pointed out some weaknesses of this metric, including the fact that it does not average well and that it has low discriminative power.

3.2 Bpref and Other Condensed-List Metrics

Sakai [15] showed that a family of metrics, which are existing metrics applied to a *condensed list* of documents obtained by removing all unjudged documents from the original list, are simpler and better solutions than bpref. Bpref itself can be expressed as a metric based on a condensed list. Let r' denote the rank of a judged document in a condensed list, whose rank in the original list was $r(\geq r')$. Let N denote the number of judged nonrelevant documents. For any topic such that $R \leq N$, bpref reduces to *bpref_R*:

$$bpref\_R = \frac{1}{R} \sum_{r'} isrel(r')(1 - \frac{\min(R, r' - count(r'))}{R}) . \quad (6)$$

In fact, $R \leq N$ holds for every topic used in our experiments, and therefore *bpref is always bpref_R*. Whereas, for any topic such that $R \geq N$, bpref reduces to *bpref_N*:

$$bpref\_N = \frac{1}{R} \sum_{r'} isrel(r')(1 - \frac{r' - count(r')}{N}) . \quad (7)$$

Note that bpref_N does not require a minimum operator [15].

The only essential difference between bpref and AP applied to a condensed list, which we call AP′, is that bpref lacks the top-heaviness property of AP [15]. Note that, from Eq. 1, AP′ can be expressed as:

$$AP' = \frac{1}{R} \sum_{r'} isrel(r')(1 - \frac{r' - count(r')}{r'}) . \quad (8)$$

Thus, for each retrieved relevant document, AP′ uses r' while bpref uses a very large constant (either R or N) for scaling $r' - count(r')$ i.e., the number of judged nonrelevant documents ranked above the relevant one at rank r'. Scaling by a large constant is not good: For example, consider a condensed list that has a judged nonrelevant document at rank 1 and a relevant one at rank 2. For this relevant document, the "misplacement penalty" is $r' - count(r') = 2 - 1 = 1$ and and $P(r') = 1/2$. Thus, the existence of the judged nonrelevant document at rank 1 weighs heavily. In contrast, this nonrelevant document has very little impact on bpref, because the same misplacement penalty is divided by R, or N which is generally even larger than R. In addition to discussing these inherent properties of AP′ and bpref, Sakai [15] demonstrated experimentally that AP′ is in fact superior to bpref in terms of system ranking stability and discriminative power, given incomplete but unbiased relevance data.

Condensed-list versions of Q, nDCG and RBP will be denoted by Q′, nDCG′ and RBP′. Thus this paper considers four metrics (AP, Q, nDCG and RBP) plus five condensed-list metrics (AP′, Q′, nDCG′, RBP and bpref). Among these, AP, AP′ and bpref cannot handle graded relevance.

3.3 A Note on RankEff, a.k.a. Bpref_N

Ahlgren and Grönqvist [1] have claimed that a binary-relevance metric called *RankEff* provides more stable system ranking than *bpref-10* [6] and AP when the relevance data is reduced at random. Let d be a judged relevant document, and let $I(d)$ denote the number of judged nonrelevant documents ranked *lower* than d. RankEff is defined as:

$$RankEff = \frac{1}{R} \sum_d \frac{I(d)}{N} . \quad (9)$$

However, let us rewrite RankEff using a condensed list. Recall that the number of judged nonrelevant documents ranked *above* a relevant one at rank r' is given by $r' - count(r')$. Hence the number of judged nonrelevant documents ranked *below* r', including those not retrieved at all, is given by $N - (r' - count(r'))$. Whereas, for each relevant document not retrieved, $I(d) = 0$ by definition. That is, the summation in Eq. 9 is essentially over *retrieved* relevant documents rather than all judged relevant documents. Hence,

$$RankEff = \frac{1}{R} \sum_{r'} isrel(r') \frac{N - (r' - count(r'))}{N} . \quad (10)$$

It is clear that *RankEff is none other than bpref_N* (Eq. 7) which has been known as `bpref_allnonrel` in trec_eval.

As discussed earlier, both theory and experiments have shown that bpref_N (RankEff) is not a desirable metric, in that it is very insensitive to change in the top ranked documents [15]. Ahlgren and Grönqvist [1] themselves report that the metric correlates poorly with AP: In their study, a system ranked at number 42 by AP was ranked at number 7 by RankEff.

4 Data

Table 1 provides some statistics of the TREC and NTCIR data we used for evaluating the nine metrics in the presence of system bias. The "TREC03" and "TREC04" data are from the TREC 2003 and 2004 robust track [22, 23], and the "NTCIR-6J" (Japanese) and "NTCIR-6C" (Chinese) data are from the

[1] The Microsoft version of nDCG uses $dg(r) = g(r)/\log_a(r+1)$ for all r [8], but this cancels out a, depriving nDCG of a persistence parameter.

Table 1. TREC and NTCIR data used.

	TREC03	TREC04	NTCIR-6J	NTCIR-6C
#topics	50	49	50	50
#docs	approx. 528,000		858,400	901,446
pool depth	125	100	100	100
average N	925.5	654.6	1157.9	999.4
range N	[292, 2050]	[132, 1371]	[480, 2732]	[414, 1907]
average R	33.2	41.2	95.3	88.1
range R	[4, 115]	[3, 161]	[4, 311]	[15, 400]
S-relevant	8.1	12.5	2.5	21.6
A-relevant	-	-	61.1	30.4
B-relevant	25.0	28.8	31.7	36.1
#all runs	78	110	74	46
#teams	16	14	10(12)	10(11)

NTCIR-6 CLIR task [11]. For forming system-biased relevance data from the NTCIR data, we considered teams that submitted at least one monolingual run. For example, we excluded two teams from the NTCIR-6J data as they submitted cross-lingual runs only.

Consider a particular topic. Let t denote a participating team, and let D_t denote the set of documents contributed to the pool by this team. For TREC03, for example, D_t is the union of the top 125 documents of each run submitted by t. The set of *unique contributions* by t is defined as $U_t = D_t - \cup_{t' \neq t} D_{t'}$. Similarly, let $D_t^{rel} (\subseteq D_t)$ denote the set of judged relevant documents obtained from t. The set of *unique relevant documents* from t is defined as $U_t^{rel} = D_t^{rel} - \cup_{t' \neq t} D_{t'}^{rel}$. Table 2 shows the participating teams that we used, along with some statistics on U_t^{rel}. For example, Table 2(c) shows that a team called NICT contributed as many as 229.7 unique relevant documents per topic on average, and this was achieved by submitting 20 runs (including cross-lingual runs). For one topic, this team contributed 721 unique relevant documents.

Let J denote the complete set of judged documents for a topic. Our leave-one-team-out experiments reported in Section 5 replace J with $J - U_t$ for each t. That is, unique contributions from t are removed from the original relevance data, so that t can be treated as a "new" team. In Section 6, we go to the other extreme and replace J with D_t. That is, runs from a single team is used for forming the relevance data. In these "take-just-one-team" experiments, the teams labelled with a "†" in Table 2 failed to contribute a relevant document (i.e., $D_t^{rel} = \phi$) for at least one topic, and were therefore excluded from our analysis. In addition, we chose three teams from each data set to conduct "take-just-three-teams" experiments, by replacing J with $\cup_{t \in T} D_t$, where T is the set of chosen teams. As indicated by "*"'s in Table 2, we chose three "ordinary" teams: Ones with the smallest number of unique relevant documents.

5 Leave one team out

To compare the robustness of our metrics to runs that did not contribute to the pool, we formed leave-one-team-out relevance data $J - U_t$ for each team t, as explained earlier. Then, for each t, we randomly selected one monolingual run from t and evaluated this run using $J - U_t$. Recall that *all* runs submitted by t have been left out to form $J - U_t$.

Table 3 shows, for each t from NTCIR-6J, how a selected monolingual run from t is affected when the original relevance data J is replaced by $J - U_t$. For example, when a run from "BRKLY" is evaluated using nDCG with this team' s leave-one-team-out relevance data, the run's absolute score goes down by .0016, and its rank among the 10 selected runs goes down from rank 6 to rank 8. In contrast, a run from "HUM" goes up from rank 6 to rank 5 according to nDCG′ and this team's leave-one-team-out relevance data. It can be observed that, according to condensed-list metrics, i.e., AP′, Q′, nDCG′, RBP′ and bpref, the scores and the ranks tend to go up with the use of each leave-one-team-out relevance data, while, according to traditional metrics, i.e., AP, Q, nDCG and RBP, the scores and the ranks tend to go down. Moreover, the "average absolute performance change" row of Table 3 shows that the average score changes are higher for the condensed-list metrics than for the traditional ones. For example, for AP, this is computed as $(.0021 + .0002 + .0005 + .0000 + .0013 + .0050 + .0011 + .0000 + .0003 + .0003)/10 = .0011$. Whereas, the average for AP′ is .0062. The trends are similar for TREC03, TREC04 and NTCIR-6C, but the results are omitted due to lack of space. Hence, our first observation is that *condensed-list metrics overestimate new systems while traditional metrics underestimate them, and that the overestimation tends to be larger than the underestimation.* A new run contains many unjudged documents. Therefore, condensing its ranked list may move up the ranks of retrieved relevant documents dramatically. This is why condensed-list metrics, including bpref, overestimate new systems.

Our main criterion for comparing metrics is Sakai's *discriminative power* [16]. Let C be the set of all pairs of runs that are being considered. For a given significance level α, let $C_* (\subseteq C)$ be the set of pairs of runs with a statistically significant performance difference in terms of a given metric according to a two-sided, *paired bootstrap hypothesis test*. Then discriminative power is defined as C_*/C: It means how often a metric manages to detect a statistically significant difference for a fixed probability of Type I Error. Although C_*/C can also be defined using a significance test other than the bootstrap test, one of the advantages of Sakai's method is that it can also estimate the minimum performance difference required to achieve statistical significance.

Suppose that C_* was obtained using a given metric and the original relevance data. Now, let C'_* denote the set of pairs of runs with a statistically significant difference in terms of the same metric but with a *different* relevance data set. Assuming that the results based on the original relevance data are the ground truth, we can quantify the discrepancy between C_* and C'_* by reporting the number of *misses* $|C_* - C'_*|$ and that of *false alarms* $|C'_* - C_*|$. Bompada *et al.* [4] have also examined misses and false alarms for comparing bpref, nDCG and Inferred AP, but they used random sampling and did not consider system bias.

Table 4 summarises the results of our discriminative power experiments using $\alpha = 0.05$ with the leave-one-team-out relevance data. For example, Table 4(a) shows that Q manages to detect a statistically significant difference for 80 run pairs out of 120 (66.7%) using the original relevance data, and this is the highest discriminative power achieved across all metrics for TREC03, as indicated in bold. Moreover, given the 50 topics of TREC03, the performance difference required to achieve a significance level of $\alpha = 0.05$ is around 0.07 in Q. On the other hand, Table 4(a) also shows the corresponding results averaged over the 16 leave-one-team-out relevance data. For example, it can be oberved that, by replacing the original relevance data of TREC03 with a leave-one-team-out relevance data, the discriminative power of Q′ goes up from 64.2% to 64.9%, but this is due to false alarms, which occur 0.81 times

Table 2. Participating teams, #runs and #unique relevant documents per topic (mean and range).

(a) TREC03			(b) TREC04			(c) NTCIR-6J			(d) NTCIR-6C		
MU03rob	5	47.1 [5,145]	Juru	10	15.3 [0,161]	BRKLY	8	64.8 [4,239]	BRKLY	8	166.8 [11,355]
NLPR03*†	5	5.1 [0,38]	NLPR04*	11	6.2 [0,91]	HUM	5	120.6 [95,199]	CCNU*	2	12.9 [0,87]
SABIR03	3	24.1 [1,171]	SABIR04	6	16.2 [0,62]	JSCCL*	4	12.8 [0,99]	HUM	5	130.9 [95,225]
Sel	5	16.9 [0,73]	apl04rs	5	15.0 [0,57]	KLE	3	28.8 [0,152]	I2R*	4	22.3 [0,107]
THUIRr030*	5	9.0 [0,51]	fub04	10	8.4 [0,80]	NCUTW†	5	54.4 [1,232]	ISQUT†	3	82.8 [6,183]
UAmsT03R	5	31.0 [1,82]	humR04	10	23.2 [0,95]	NICT	20	229.7 [5,721]	NCUTW	5	25.4 [0,90]
UIUC03R*	5	11.1 [0,129]	icl04pos	9	42.9 [6,123]	OKSAT	5	65.9 [2,216]	NTNU†	4	32.9 [0,115]
VT	5	26.6 [1,201]	mpi04r†	10	62.7 [4,176]	TSB†	12	37.5 [0,270]	UniNE*	5	13.4 [0,76]
aplrob03	5	15.0 [0,114]	pircRB04*	10	6.4 [0,27]	UniNE*	5	14.3 [0,130]	WTG	4	66.6 [0,186]
fub03l†	5	12.9 [0,94]	polyu	6	26.0 [1,94]	YLMS*	3	7.9 [0,73]	pircs	4	59.6 [0,177]
humR03	5	18.0 [0,82]	uic0401†	1	12.6 [0,68]						
oce03	5	39.5 [1,135]	uogRob*	10	6.2 [0,101]						
pircRB	5	30.6 [1,200]	vtum	8	9.5 [0,55]						
rutcor03†	5	103.6 [7,262]	wdo	4	16.7 [0,91]						
uic030†	5	22.6 [1,81]									
uwmtCR	5	17.4 [0,72]									

†Not used for take-just-one-team experiments; *Used for take-just-three-teams experiments (See Section 6).

Table 3. Performance change and rank change when leaving out one team and evaluating a run from that team (NTCIR-6J). A "+" indicates that a run is overestimated; a "−" indicates that it is underestimated. Rank changes are indicated in bold: For example, "6↑5" means going up from rank 6 to rank 5.

	AP'	Q'	nDCG'	RBP'	bpref	AP	Q	nDCG	RBP
BRKLY	+.0071 4→4	+.0059 4→4	+.0025 5→5	+.0030 4→4	+.0083 4→4	−.0021 7→7	−.0019 7→7	−.0016 **6↓8**	−.0009 4→4
HUM	+.0069 8→8	+.0065 8→8	+.0035 **6↑5**	+.0016 8→8	+.0074 8→8	−.0002 9→9	+.0001 8→8	+.0003 5→5	−.0004 8→8
JSCCL	+.0031 6→6	+.0024 6→6	+.0013 4→4	+.0013 5→5	+.0044 **6↑4**	−.0005 5→5	−.0004 5→5	−.0005 4→4	−.0001 5→5
KLE	+.0037 7→7	+.0036 **7↑6**	+.0022 7→7	+.0010 6→6	+.0038 7→7	.0000 6→6	.0000 **6↓7**	−.0002 **7↓8**	−.0011 6→6
NCUTW	+.0055 **9↑8**	+.0048 9→9	+.0029 9→9	+.0043 9→9	+.0067 **9↑8**	−.0013 **8↓9**	−.0013 9→9	−.0013 9→9	−.0009 9→9
NICT	+.0138 **5↑4**	+.0133 **5↑4**	+.0079 8→8	+.0027 7→7	+.0125 **5↑4**	+.0050 **4↓5**	+.0053 **4↓5**	+.0038 8→8	−.0002 7→7
OKSAT	+.0145 10→10	+.0120 10→10	+.0080 10→10	+.0086 10→10	+.0156 10→10	−.0011 10→10	−.0007 10→10	−.0003 10→10	−.0010 10→10
TSB	+.0033 1→1	+.0028 1→1	+.0018 1→1	+.0011 1→1	+.0029 1→1	.0000 1→1	−.0002 1→1	+.0001 1→1	−.0002 1→1
UniNE	+.0040 3→3	+.0033 3→3	+.0021 2→2	+.0021 3→3	+.0052 3→3	+.0003 3→3	+.0004 3→3	+.0001 2→2	−.0001 3→3
YLMS	+.0005 2→2	+.0003 2→2	.0000 3→3	+.0002 2→2	+.0007 2→2	−.0003 2→2	−.0003 2→2	−.0005 3→3	.0000 2→2
average abs. performance change	.0062	.0055	.0032	.0026	.0065	.0011	.0010	.0009	.0005

on average.

The results with the original relevance data in Table 4 confirm those by Sakai and Kando [17], in that AP, Q, nDCG and their condensed-list versions are more discriminative than bpref and RBP. In addition, they are also more discriminative than RBP′, which has been examined for the first time. It can also be observed that our discriminative power results using leave-one-team-out relevance data are similar to the ones using the original relevance data. This is because the only difference between the two relevance data sets is U_t, the unique contributions from one team.

6 Take Just one team

The leave-one-team-out experiments replaced J with $J - U_t$. We now discuss a more extreme case of system bias, by replacing J with D_t, the contributions from a single team. As we have explained in Section 4, we also form relevance data using contributions from three teams with the smallest number of unique relevant documents.

Table 5 summarises our take-just-one-team and take-just-three-teams results for NTCIR-6J in a way similar to Table 3. Thus, for each team t, the table shows how a particular monolingual run from t (which is the same as the one we used for the leave-one-team-out experiments) is affected when the original relevance data J is replaced by D_t. For example, when a run from "BRKLY" is evaluated using Q′ with this team's contributions only, the run goes down from rank 4 to rank 8. In contrast, when the same run is evaluated using Q with this team's contributions only, it goes up from rank 7 to rank 4. As we have explained in Section 4, two teams with a "†" in Table 2 are excluded here. It can be observed that, *if a single team t is used for forming the relevance data, the run score for t goes up for all metrics (except for RBP and RBP′); however, while traditional metrics overestimate the rank of a run from t, condensed-list metrics understimate it.* Condensed-list metrics underestimate the rank of a run from t because all the other runs from $t' (\neq t)$ are substantially *overestimated*: These other runs are "new" to the take-just-one-team relevance data of t, and we have already observed in Section 5 that condensed-list metrics overestimate new runs. As for RBP and RBP′, replacing J with D_t does not substantially affect the run score for t, because this merely turns some relevant documents below the pool depth within that run, i.e., those that belong to $J - D_t$, into nonrelevant documents. The stability of scores for RBP and RBP′ reflects the fact that they totally disregard recall, and not necessarily that they are superior: Note that the ranks ac-

Table 4. Discriminative power at $\alpha = 0.05$: leaving one team out. For each experimental condition, the highest discriminative power is indicated in bold.

(a) TREC03 (16 runs)		AP'	Q'	nDCG'	RBP'	bpref'	AP	Q	nDCG	RBP
original relevance data	disc. power	77/120 =64.2%	77/120 =64.2%	71/120 =59.2%	55/120 =45.8%	69/120 =57.5%	77/120 =64.2%	80/120 **=66.7%**	71/120 =59.2%	55/120 =45.8%
	diff. required	0.09	0.07	0.08	0.04	0.08	0.07	0.07	0.08	0.04
average over 16 leave-one-team-out	disc. power	64.3%	64.9%	59.1%	46.0%	57.1%	64.2%	**66.7%**	59.2%	46.0%
	#misses	0.00	0.00	0.13	0.69	0.00	0.00	0.00	0.00	0.00
	#false alarms	0.06	0.81	0.06	0.38	0.19	0.00	0.00	0.00	0.19
(b) TREC04 (14 runs)		AP'	Q'	nDCG'	RBP'	bpref'	AP	Q	nDCG	RBP
original relevance data	disc. power	61/91 =67.0%	62/91 =68.1%	58/91 =63.7%	46/91 =50.5%	57/9 =62.6%	61/91 =67.0%	63/91 **=69.2%**	58/91 =63.7%	45/91 =49.5%
	diff. required	0.07	0.08	0.09	0.05	0.09	0.07	0.08	0.08	0.05
average over 14 leave-one-team-out	disc. power	67.0%	68.1%	63.1%	50.1%	62.3%	67.1%	**69.0%**	62.8%	49.4%
	#misses	0.14	0.07	0.57	0.36	0.43	0.00	0.29	0.93	0.07
	#false alarms	0.14	0.07	0.00	0.00	0.14	0.07	0.07	0.07	0.00
(c) NTCIR-6J (10 runs)		AP'	Q'	nDCG'	RBP'	bpref'	AP	Q	nDCG	RBP
original relevance data	disc. power	25/45 =55.6%	28/45 =62.2%	33/45 **=73.3%**	26/45 =57.8%	23/45 =51.1%	26/45 =57.8%	28/45 =62.2%	33/45 **=73.3%**	26/45 =57.8%
	diff. required	0.07	0.09	0.08	0.04	0.08	0.08	0.07	0.08	0.05
average over 10 leave-one-team-out	disc. power	55.6%	62.4%	**73.3%**	57.6%	52.0%	57.6%	62.4%	**73.3%**	57.8%
	#misses	0.40	0.00	0.00	0.20	0.10	0.10	0.20	0.00	0.00
	#false alarms	0.40	0.10	0.00	0.10	0.50	0.00	0.30	0.00	0.00
(d) NTCIR-6C (10 runs)		AP'	Q'	nDCG'	RBP'	bpref'	AP	Q	nDCG	RBP
original relevance data	disc. power	36/45 =80.0%	34/45 =75.6%	34/45 =75.6%	32/45 =71.1%	34/45 =75.6%	37/45 **=82.2%**	36/45 =80.0%	34/45 =75.6%	32/45 =71.1%
	diff. required	0.07	0.06	0.07	0.07	0.07	0.08	0.07	0.08	0.06
average over 10 leave-one-team-out	disc. power	79.8%	75.8%	75.4%	71.5%	75.4%	**82.2%**	79.6%	75.6%	71.1%
	#misses	0.10	0.10	0.10	0.00	0.20	0.00	0.20	0.00	0.00
	#false alarms	0.00	0.20	0.00	0.20	0.10	0.00	0.00	0.00	0.00

Table 5. Performance change and rank change when taking one team and evaluating a run from that team (NTCIR-6J). A "+" indicates that a run is overestimated; a "−" indicates that it is underestimated. Rank changes are indicated in bold.

	AP'	Q'	nDCG'	RBP'	bpref'	AP	Q	nDCG	RBP
BRKLY	+.0770 **4↓8**	+.0790 **4↓8**	+.0399 **5↓7**	.0000 **4↓8**	+.0646 **4↓8**	+.0741 **7↑4**	+.0724 **7↑4**	+.0377 **6↑4**	−.0003 4→4
HUM	+.1002 8→8	+.1065 8→8	+.0643 **6↑5**	−.0001 8→8	+.0785 **8↓9**	+.1020 **9↑4**	+.1054 **8↑4**	+.0647 **5↑4**	−.0003 **8↑5**
JSCCL	+.1038 **6↓7**	+.1043 **6↓7**	+.0622 4→4	.0000 **5↓7**	+.0877 **6↓7**	+.0980 **5↑2**	+.0963 **5↑3**	+.0598 4→4	−.0003 **5↑4**
KLE	+.0951 **7↓8**	+.1011 **7↓8**	+.0580 7→7	.0000 **6↓8**	+.0726 **7↓8**	+.0879 **6↑3**	+.0908 **6↑4**	+.0548 **7↑4**	−.0004 **6↑4**
NICT	+.0333 **5↓8**	+.0354 **5↓8**	+.0193 8→8	+.0001 **7↓8**	+.0261 **5↓8**	+.0244 4→4	+.0256 4→4	+.0156 **8↑4**	.0000 **7↑6**
OKSAT	+.1083 10→10	+.1172 10→10	+.1017 10→10	.0000 10→10	+.0871 10→10	+.1049 **10↑7**	+.1099 **10↑8**	+.1006 **10↑9**	−.0001 **10↑6**
UniNE	+.0978 3→3	+.0957 3→3	+.0462 2→2	.0000 3→3	+.0829 3→3	+.0934 **3↑2**	+.0897 **3↑2**	+.0445 2→2	−.0003 **3↑2**
YLMS	+.1174 2→2	+.1148 2→2	+.0494 3→3	−.0002 2→2	+.1024 2→2	+.1213 **2↑1**	+.1176 **2↑1**	+.0511 **3↑2**	−.0005 **2↑1**
average abs. performance change	.0916	.0943	.0551	.0001	.0752	.0882	.0885	.0536	.0003

cording to RBP and RBP′ are altered just like the other metrics. Similar results for TREC03, TREC04 and NTCIR-6C are omitted due to lack of space.

Table 6 compares, for each data set and metric, the ranking of the aforementioned selected runs based on the original relevance data and that based on a take-just-one-team or a take-just-three-teams relevance data. The similarity between two rankings is quantified using Kendall's tau rank correlation, which would be 1 if the two rankings are identical and -1 if the two rankings are the exact inverse of each other. The rank correlation values for the take-just-one-team relevance data have been averaged across teams. It can be observed that the correlation values are generally very high. That is, it is possible to replace the original relevance data with one that is based on a single team (or three teams) and still maintain a similar system ranking. As mentioned in Section 2, this generalises a finding by Sanderson and Joho [19] who considered only AP and binary-relevance TREC data. However, obtaining a system ranking that is similar to the full relevance data is not sufficient for sound evaluation: We later show that strong system bias can introduce much noise in statistical significance tests.

For the two NTCIR data sets, the take-just-one-team rankings with AP′, Q′, nDCG′ appear to be more consistent with the original rankings than those with AP, Q and nDCG. However, we refrain from making a claim based on these results because the NTCIR rankings contain only eight teams (See Table 2) and the trend is not clear for the two TREC data sets.

Table 7 summarises the results of our discriminative power experiments using $\alpha = 0.05$ with the take-just-one-team and take-just-three-teams relevance data, in a way similar to Table 4. The "original relevance data" rows have been copied from Table 4. For example, Table 7(a) shows that replacing the original relevance data with a take-just-three-teams relevance data superficially raises the discriminative power of Q from 66.7% to 68.3%, but this is due to four false alarms with two misses. False alarms in particular are not welcome in retrieval experiments: the take-just-three-teams relevance data declared

Table 6. Kendall's rank correlation: the original ranking vs. that by taking one team / three teams.

		AP'	Q'	nDCG'	RBP'	bpref	AP	Q	nDCG	RBP
(a) TREC03	take-three-teams	.950	.917	.900	.967	.933	.933	.933	.933	.967
	average over 12 take-one-team	.951	.918	.929	.958	.944	.932	.920	.935	.947
(b) TREC04	take-three-teams	.978	.956	.978	.934	.956	1	1	.956	1
	average over 12 take-one-team	.932	.936	.898	.903	.903	.906	.907	.860	.926
(c) NTCIR-6J	take-three-teams	.956	.911	.956	.867	.956	.822	.822	.956	.956
	average over 8 take-one-team	.876	.880	.925	.893	.894	.756	.782	.885	.849
(d) NTCIR-6C	take-three-teams	1	.956	1	.911	.956	1	1	1	1
	average over 8 take-one-team	.960	.929	.991	.880	.920	.853	.867	.898	.907

Table 7. Discriminative power at $\alpha = 0.05$: take one team / three teams. For each experimental condition, the highest discriminative power is indicated in bold.

(a) TREC03 (16 runs)		AP'	Q'	nDCG'	RBP'	bpref	AP	Q	nDCG	RBP
original relevance data	disc. power	77/120 =64.2%	77/120 =64.2%	71/120 =59.2%	55/120 =45.8%	69/120 =57.5%	77/120 =64.2%	80/120 **=66.7%**	71/120 =59.2%	55/120 =45.8%
	diff. required	0.09	0.07	0.08	0.04	0.08	0.07	0.07	0.08	0.04
take-three-teams relevance data	disc. power	61.7%	62.5%	55.0%	42.5%	55.8%	67.5%	**68.3%**	63.3%	49.2%
	#misses	5	6	8	8	3	2	2	2	1
	#false alarms	2	4	3	4	1	6	4	7	5
average over 12 take-one-team	disc. power	59.6%	59.0%	54.4%	43.1%	51.7%	66.4%	**67.6%**	61.6%	52.6%
	#misses	8.42	9.50	8.92	9.25	8.83	5.67	5.50	4.67	3.17
	#false alarms	2.92	3.33	3.17	6.00	1.83	8.33	6.58	7.58	11.25
(b) TREC04 (14 runs)		AP'	Q'	nDCG'	RBP'	bpref	AP	Q	nDCG	RBP
original relevance data	disc. power	61/91 =67.0%	62/91 =68.1%	58/91 =63.7%	46/91 =50.5%	57/9 =62.6%	61/91 =67.0%	63/91 **=69.2%**	58/91 =63.7%	45/91 =49.5%
	diff. required	0.07	0.08	0.09	0.05	0.09	0.07	0.08	0.08	0.05
take-three-teams relevance data	disc. power	63.7%	65.9%	56.0%	40.7%	54.9%	69.2%	**70.3%**	61.5%	48.4%
	#misses	3	2	7	10	8	0	0	2	1
	#false alarms	0	0	0	1	1	2	1	0	0
average over 12 take-one-team	disc. power	61.6%	62.3%	53.5%	45.9%	56.4%	64.4%	**67.0%**	59.7%	50.1%
	#misses	7.92	7.00	11.50	9.00	9.67	7.17	7.00	7.50	3.75
	#false alarms	3.00	1.75	2.25	4.75	4.00	4.83	5.00	3.83	4.33
(c) NTCIR-6J (10 runs)		AP'	Q'	nDCG'	RBP'	bpref	AP	Q	nDCG	RBP
original relevance data	disc. power	25/45 =55.6%	28/45 =62.2%	33/45 **=73.3%**	26/45 =57.8%	23/45 =51.1%	26/45 =57.8%	28/45 =62.2%	33/45 **=73.3%**	26/45 =57.8%
	diff. required	0.07	0.09	0.08	0.04	0.08	0.08	0.07	0.08	0.05
take-three-teams relevance data	disc. power	57.8%	64.4	**71.1%**	42.2%	44.4%	66.7%	68.9%	**71.1%**	62.2%
	#misses	1	1	1	7	3	2	1	1	0
	#false alarms	2	2	0	0	0	6	4	0	2
average over 8 take-one-team	disc. power	61.9%	66.7%	66.1%	49.2%	50.6%	66.4%	67.2%	**67.8%**	61.4%
	#misses	1.00	1.13	3.38	5.13	3.00	4.13	4.25	4.25	3.50
	#false alarms	3.88	3.13	0.13	1.25	2.75	8.00	6.50	1.75	5.13
(d) NTCIR-6C (10 runs)		AP'	Q'	nDCG'	RBP'	bpref	AP	Q	nDCG	RBP
original relevance data	disc. power	36/45 =80.0%	34/45 =75.6%	34/45 =75.6%	32/45 =71.1%	34/45 =75.6%	37/45 **=82.2%**	36/45 =80.0%	34/45 =75.6%	32/45 =71.1%
	diff. required	0.07	0.06	0.07	0.07	0.07	0.08	0.07	0.08	0.06
take-three-teams relevance data	disc. power	71.1%	73.3%	75.6%	66.7%	64.4%	**82.2%**	80.0%	77.8%	77.8%
	#misses	4	1	0	2	5	0	0	0	0
	#false alarms	0	0	0	0	0	0	0	1	3
average over 8 take-one-team relevance data	disc. power	73.9%	75.0%	71.1%	72.5%	70.3%	**80.3%**	79.7%	75.0%	75.0%
	#misses	3.00	1.50	2.13	2.25	4.25	3.63	3.38	2.63	2.00
	#false alarms	0.25	1.25	0.13	2.88	1.88	2.75	3.25	2.38	3.75

that the four run pairs are significantly different, even though they are not significantly different according to the original relevance data.

According to Table 7, take-just-one-team relevance data generally yield more misses and false alarms than take-three-teams relevance data. Hence we observe that, even though take-just-one-team relevance data may produce a system ranking that is very similar to that produced by the original relevance data, *pooling runs from several teams is better than pooling runs from a single team for obtaining reliable conclusions based on statistical significance tests.* The focus of this study, however, is on the comparison of different metrics under the same condition, and not on how many and what kind of teams are required to obtain reliable conclusions.

Table 7 also shows that AP, Q and nDCG are generally more discriminative than AP′, Q′ and nDCG′, respectively, even with take-just-one-team or take-just-three-teams relevance data. For example, for TREC03, the discriminative power of Q averaged over 12 take-just-one-team relevance data is 67.6% while the corresponding value for Q′ is only 59.0%, even though the number of misses and that of false alarms are more or less comparable. Thus, *condensed-list metrics are not necessarily superior to traditional metrics when the relevance data is heavily biased towards one team or a few teams.* On the other hand, even with take-just-three-teams and take-just-one-team relevance data, AP′, Q′ and nDCG′ are generally more discriminative than bpref, RBP and RBP′, although bpref sometimes does as well as nDCG′.

7 Conclusions

Several recent studies [1, 4, 6, 15, 17, 25] discussed the effect of incomplete relevance data in retrieval evaluation using random samples of the original relevance data. Hence they discussed neither system bias nor pool depth bias.

This paper examined the effect of system bias. Even though Sakai [15] and Sakai and Kando [17] showed that AP′, Q′ and nDCG′ are effective for handling very incomplete but unbiased data, we showed that these results do not hold in the presence of system bias. Using data from both TREC and NTCIR, we first showed that condensed-list metrics overestimate new systems while traditional metrics underestimate them, and that the overestimation tends to be larger than the underestimation. We then showed that, when relevance data is heavily biased towards a sigle team or a few teams, AP′, Q′ and nDCG′ are not necessarily more discriminative than AP, Q and nDCG. Nevertheless, AP′ and Q′ *are* generally more discriminative than bpref and RBP′ in the presence of system bias.

Our separate study [18] shows that AP′, Q′ and nDCG′ are not necessarily superior to AP, Q and nDCG in the presence of *pool depth bias* either. Hence previous studies that used random sampling should be interpreted with caution. In reality, relevance data formed through pooling are never a random sample of the full relevance data.

Traditional metrics assume that retrieved unjudged documents are nonrelevant, while condensed-list metrics, including bpref, assume that they are nonexistent. The present study showed that the latter assumption is no better than the former. In our future work, we would like to couple efficient and reliable test construction methods with reliable graded-relevance metrics. We also plan to establish quantitative criteria for choosing good evaluation metrics: Although we believe that discriminative power is one important criterion, there are probably other aspects that need to be examined, including the ability to predict performance on new topics in terms of "user-oriented" metrics such as precision-at-ten [24].

References

[1] Ahlgren, P. and Grönqvist, L.: Evaluation of Retrieval Effectiveness with Incomplete Relevance Data: Theoretical and Experimental Comparison of Three Measures, *Information Processing and Management*, Volume 44, pp. 212-225, 2008.

[2] Aslam, J. A. and Yilmaz, E.: Inferring Document Relevance from Incomplete Information, *ACM CIKM 2007 Proceedings*, pp. 633-642, 2007.

[3] Baillie, M., Azzopardi, L. and Ruthven, I.: Evaluating Epistemic Uncertainty under Incomplete Assessments, *Information Processing and Management*, 44(2), pp. 811-837, 2008.

[4] Bompada, T. *et al.*: On the Robustness of Relevance Measures with Incomplete Judgments, *ACM SIGIR 2007 Proceedings*, pp. 359-366, 2007.

[5] Buckley, C. and Voorhees, E. M.: Evaluating Evaluation Measure Stability, *ACM SIGIR 2000 Proceedings*, pp. 33-40, 2000.

[6] Buckley, C. and Voorhees, E. M.: Retrieval Evaluation with Incomplete Information, *ACM SIGIR 2004 Proceedings*, pp. 25-32, 2004.

[7] Buckley, C. *et al.*: Bias and the Limits of Pooling for Large Collections, *Information Retrieval*, Vol. 10, Number 6, pp. 491-508, 2007.

[8] Burges, C. *et al.*: Learning to Rank using Gradient Descent, *ACM ICML 2005 Proceedings*, pp. 89-96, 2005.

[9] Büttcher *et al.*: Reliable Information Retrieval Evaluation with Incomplete and Biased Judgements, *ACM SIGIR 2007 Proceedings.*, pp. 63-70, 2007.

[10] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422-446, 2002.

[11] Kando, N.: Overview of the Sixth NTCIR Workshop, *NTCIR-6 Proceedings*, pp. i-ix, 2007.

[12] Moffat, A. and Zobel, J.: Rank-Biased Precision for Measurement of Retrieval Effectiveness, under review, 2008.

[13] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, 43(2), pp. 531-548, 2007.

[14] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First International Workshop on Evaluating Information Access (EVIA 2007)*, pp. 32-43, 2007.

[15] Sakai, T.: Alternatives to Bpref, *ACM SIGIR 2007 Proceedings*, pp. 71-78, 2007.

[16] Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, *IPSJ Transactions on Databases*, Vol.48, No.SIG 9 (TOD35), pp.11-28, 2007. `http://www.jstage.jst.go.jp/article/ipsjdc/3/0/625/_pdf`

[17] Sakai, T. and Kando, N.: On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments, *Information Retrieval*, 2008. `http://www.springerlink.com/content/k41j115214032614/fulltext.pdf`

[18] Sakai, T.: Comparing Metrics across TREC and NTCIR: The Robustness to Pool Depth Bias, *ACM SIGIR 2008*, to appear, 2008.

[19] Sanderson, M. and Joho, H.: Forming Test Collections with No System Pooling, *ACM SIGIR 2004 Proceedings*, pp. 33-40, 2004.

[20] Sormunen, E.: Liberal Relevance Criteria of TREC - Counting on Negligible Documents? *ACM SIGIR 2002 Proceedings*, pp. 324-330, 2002.

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

[22] Voorhees, E. M.: Overview of the TREC 2003 Robust Retrieval Track, *TREC 2003 Proceedings*, 2004.

[23] Voorhees, E. M.: Overview of the TREC 2004 Robust Retrieval Track, *TREC 2004 Proceedings*, 2005.

[24] Webber, W., Moffat, A., Zobel, J. and Sakai, T.: Precision-At-Ten Considered Redundant, *ACM SIGIR 2008 Proceedings*, to appear, 2008.

[25] Yilmaz, E. and Aslam, J. A.: Estimating Average Precision with Incomplete and Imperfect Judgments, *ACM CIKM 2006 Proceedings*, 2006.

[26] Zobel, J.: How Reliable are the Results of Large-Scale Information Retrieval Experiments? *ACM SIGIR '98 Proceedings*, pp. 307-314, 1998.