A Further Note on Alternatives to Bpref

Tetsuya Sakai† Noriko Kando‡

† NewsWatch, Inc. ‡National Institute of Informatics
sakai@newswatch.co.jp, kando@nii.ac.jp

Abstract

This paper compares the robustness of information retrieval (IR) metrics to incomplete relevance assessments, using four different sets of graded-relevance test collections with submitted runs – two from TREC and two from NTCIR. We investigate the effect of reducing the original relevance data on discriminative power (i.e., how often statistical significance can be detected given the probability of Type I Error) and on Kendall's rank correlation between two system rankings. According to these experiments, Q′, nDCG′ and AP′ proposed by Sakai are superior to bpref proposed by Buckley and Voorhees and to Rank-Biased Precision proposed by Moffat and Zobel. We also clarify some properties of these metrics that immediately follow from their definitions.

1 Introduction

Information Retrieval (IR) evaluation using *incomplete* relevance assessments is beginning to receive attention. Large-scale test collections constructed through *pooling* such as the TREC, CLEF and NTCIR collections are all incomplete to some degree, in that only a small sample of the document collection has been judged for relevance for each topic. While the collection sizes tend to grow monotonically in order to mimic real-world data such as the Web, the available manpower for relevance assessments often remain more or less constant, and therefore IR researchers are expected to live with the incompleteness issue as long they adhere to the *Cranfield paradigm* [3].

At ACM SIGIR '04, Buckley and Voorhees [3] proposed an IR evaluation metric called *bpref* (binary preference) which is highly correlated with Average Precision (AP) when full relevance assessments are available and is yet more robust when the relevance assessments are reduced. Recent TREC tracks have started using this metric along with AP. Bpref penalises a system if it ranks a *judged nonrelevant* document above a judged relevant one, and is indepedendent of how the *unjudged* documents are retrieved.

At SIGIR '07, Moffat, Webber and Zobel [8] introduced an IR evaluation metric called Rank-Biased Precision (RBP) which they claimed is suitable for evaluation with incomplete relevance data. RBP assumes that the probability that the user moves from a document at Rank r to Rank $(r+1)$ is a constant p, regardless of the relevance (level) of the document at Rank r. As it does not have a recall component, adding more relevant documents to the "qrels" (the relevance data file) always increases the RBP score.

Also at SIGIR'07, Sakai [14] reported that applying Q-measure (Q), AP and normalised Discounted Cumulative Gain (nDCG) to a *condensed list*, i.e., a ranked list of documents obtained by removing all unjudged documents from the original list, is a simpler and a better solution than bpref for handling relevance data incompleteness. The metrics applied to condensed lists will hereafter be referred to as Q′, AP′ and nDCG′, respectively.

This paper compares the robustness of Q(′), AP(′), nDCG(′), bpref and RBP to incomplete relevance assessments, using four different sets of graded-relevance test collections with submitted runs – two from TREC and two from NTCIR. We investigate the effect of reducing the original relevance data on *discriminative power* [11], or how often statistical significance can be detected given the probability of Type I Error, and on *Kendall's rank correlation* between two system rankings [11]. According to these experiments, Q′, nDCG′ and AP′ are superior to bpref and RBP. As these results hold across two different evaluation efforts, namely TREC and NTCIR, we believe that these findings are very general.

This paper generalises Sakai's work [14], in that (a) While he used the NTCIR-3 and NTCIR-5 Japanese/Chinese data, we use TREC 2003 and TREC 2004 robust track data plus the NTCIR-6 Japanese/Chinese data to obtain more general and substantial conclusions; (b) We compare RBP with the other metrics, after discussing some properties of the metrics that immediately follow from their definitions.

2 Related Work

There are at least two approaches to tackling the relevance data incompleteness problem: One is to try to construct a better test collection more efficiently, and another is to devise or choose reliable IR metrics, *given* a test collection. This paper takes the latter approach, of choosing reliable IR metrics for handling relevance data incompleteness. Along this line, Aslam [1] at SIGIR '06 and Yilmaz and Aslam [17] at CIKM '06 proposed *Induced AP*, *Subcollection AP* and *Inferred AP*. Induced AP is exactly what we call AP′ (or AveP′ [14]). We do not consider Subcollection AP and Inferred AP in our present study, because (a) While the goal of Yilmaz and Aslam was to estimate the true AP values, ours is not: We prefer to explore different metrics, especially those that can handle graded relevance; (b) Both Subcollection AP and Inferred AP require knowledge of the *pooled but unjudged* documents, which limits their applicability; (Subcollection AP requires even more knowledge, namely, how small the subcollection with relevance assessments is compared to the entire document collection.) (c) According to Bompada *et al.* [2], Inferred AP is not as robust as (the original) nDCG for evaluation with incomplete relevance data.

Grönqvist's *RankEff* metric [5], a simple variant of bpref, was subsequently used by Büttcher *et al.* [4] at SIGIR '07. However, as we shall discuss in Section 3, its weakness is clear from its definition. Büttcher *et al.* [4] also uses Precision at l *judged* documents, which relies on condensed lists just like Q′, AP′ and nDCG′. However, it is known that Precision is very unstable and insensitive, and does not average well [12].

3 Metrics

3.1 Q('), AP('), nDCG(') and bpref

Let $\mathcal{L}$ denote a *relevance level*, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an $\mathcal{L}$-relevant document. Without loss of generality, this paper assumes that we have S-relevant (highly relevant), A-relevant (relevant) and B-relevant (partially relevant) documents as in NTCIR [7] in addition to judged nonrelevant documents. Moreover, we let $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$ hereafter as it is known that metrics such as Q and nDCG are robust to the choice of gain values [12]. As for the TREC data, which only have "highly relevant" and "relevant" documents, we treat the former as S-relevant and the latter as B-relevant (rather than A-relevant). This is because it is known that typically one-half of the relevant documents in the TREC qrels are only partially or marginally relevant [15].

Let $R(\mathcal{L})$ denote the number of $\mathcal{L}$-relevant documents, and let $R = \sum_{\mathcal{L}} R(\mathcal{L})$. Let $cg(r) = \sum_{1 \leq i \leq r} g(i)$ denote the *cumulative gain* at Rank r of the system output, where $g(i) = gain(\mathcal{L})$ if the document at Rank i is $\mathcal{L}$-relevant and $g(i) = 0$ otherwise (i.e., if the document at Rank i is either judged nonrelevant or unjudged). Let $g_I(r)$ and $cg_I(r)$ denote the (cumulative) gain of an *ideal* ranked output, where an ideal ranked output is one that satisfies $g(r) > 0$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$. For NTCIR, for example, listing up all S-relevant documents, followed by all A-relevant documents, followed by all B-relevant documents produces an ideal ranked output. Let $isrel(r)$ be one if the document at Rank r is relevant and zero otherwise, and let $count(r) = \sum_{1 \leq i \leq r} isrel(i)$. Clearly, precision at Rank r is given by $P(r) = count(r)/r$.

Q-measure is defined as follows:

$$Q\text{-}measure = \frac{1}{R} \sum_{r} isrel(r) BR(r) \quad (1)$$

$$BR(r) = \frac{\beta cg(r) + count(r)}{\beta cg_I(r) + r} \quad (2)$$

where $BR(r)$ is called the *blended ratio* and β is a *persistence* parameter. Because $BR(r)$ has an r in the denominator (just like $P(r)$), Q-measure is guaranteed to become smaller as a relevant document goes down the ranked list. A large β (e.g., $\beta = 100$) alleviates this effect, and makes Q-measure more forgiving for relevant documents near the bottom of the ranked list. Conversely, a small β (e.g., $\beta = 1$) imposes more penalty. Sakai [13] showed empirically that $\beta = 1, 10$ are good choices, so we take $\beta = 1$ throughout this paper. Note also that $\beta = 0$ reduces Q-measure to AP:

$$AP = \frac{1}{R} \sum_{r} isrel(r) \frac{count(r)}{r} = \sum_{r} isrel(r) P(r) \,. \quad (3)$$

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:

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

Throughout this paper, we let $l = 1000$ as it is known that small document cut-offs hurt the stability of nDCG [12]. Moreover, we let $a = 2$ because it is known that using a large logarithm base makes nDCG counterintuitive and insensitive [13], despite the fact that this parameter was designed to reflect persistence just like Q-measure's β. We shall come back to this issue in Section 3.2.

At SIGIR 2007, Sakai [14] reported that Q', AP' and nDCG' (the application of Q, AP an nDCG to *condensed lists*; See Section 1) are simpler and better solutions to the problem of evaluating IR systems with incomplete relevance data than *bpref* [3].

Let r' denote the rank of a document in a condensed list, whose original rank was $r (\geq r')$. Let N denote the number of *judged nonrelevant* documents. Then *bpref* can be expressed as follows [14]:

$$bpref = \frac{1}{R} \sum_{r'} isrel(r') \left(1 - \frac{\min(R, r' - count(r'))}{\min(R, N)}\right) \quad (5)$$

where $r' - count(r')$ is the number of judged nonrelevant documents ranked above the relevant one at Rank r', or the *misplacement penalty* with respect to this particular relevant document. Clearly, for any topic such that $R \leq N$, bpref reduces to:

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

In fact, $R \leq N$ in holds for all of our TREC and NTCIR topics (See also Table 2), so *bpref is always bpref_R* in our study.

Sakai [14] pointed out that the only essential difference between AP' and bpref is that, while the former uses r' for scaling each misplacement penalty $r' - count(r')$, the latter uses a constant (e.g., R). Compare Eq. 6 with

$$AP' = \frac{1}{R} \sum_{r'} isrel(r') \frac{count(r')}{r'} \quad (7)$$

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

Scaling by a constant is generally not good, especially if the constant is large, because this means that the misplacement penalties with respect to the top ranked relevant documents are virtually ignored [14]. In other words, bpref lacks the "top heaviness" of AP', which is one of the main strengths of the original AP. It is clear that bpref_N [14] and RankEff [5] both suffer from this problem, as they use N and $R * N$ for scaling the misplacement penalty, respectively. Sakai [14] showed experimentally that bpref_N indeed performs very poorly.

3.2 RBP and Persistence

We now formally define RBP [8, 9]. Let $\mathcal{H}$ denote the highest relevance level across all topics. RBP can be expressed as follows:

$$RBP = \frac{1-p}{gain(\mathcal{H})} \sum_{r} g(r) p^{r-1} \quad (9)$$

where $p (\leq 1)$ is a persistence parameter. A high value of p represents a persistent user; a low value represents an impatient one. As Moffat and Zobel [9] explored $p = 0.5, 0.8, 0.95$, we start our own experiments with the same values, denoting each version of RBP by RBP.5, RBP.8 and RBP.95. In all of our experiments, we let $gain(\mathcal{H}) = gain(S) = 3$. Recall that our NTCIR data have S-, A- and B-relevant documents, but our TREC data have S- and B-relevant documents only.

The assumption behind RBP is that the user, after examining the document at Rank r, will examine the document at Rank $(r+1)$ with probability p or stop scanning the ranked list with probability $1-p$. Thus the model assumes that the transition probability is independent of the relevance of the document at Rank r, which is not necessarily realistic. On the other hand, this assumption makes RBP easy to interpret and to compute. Moreover, Moffat, Webber and Zobel [8] argue that RBP is suitable for evaluation with incomplete relevance data as it is guaranteed to increase as more relevance judgments are added (since it does not have a recall component) and the error due to unjudged documents can be quantified.

However, we can discuss RBP's possible weaknesses. Firstly, RBP may give a very low score even to an ideal ranked output: In fact, the fact that it does not rely on recall implies that it denies the very existence of an "ideal" ranked output. From Eq. 9, it is clear that the RBP for an ideal ranked list in a binary relevance environment equals $(1-p)\sum_{r=1}^{R} p^{r-1}$. Table 1 shows the RBP value for an ideal ranked output for $p = 0.5, 0.8, 0.95$ and $R = 1, 10, 100, 1000$. When $p = 0.95$, for example, an ideal ranked output for a topic with $R = 10$ relevant documents receives an RBP of .4013, while one for a topic with $R = 100$ relevant documents receives .9941. Whether it is good to average such a measurement across topics can be debated, but it is at least a fact that topics with many relevant documents can have a far larger impact on Mean RBP than those with few relevant ones. Moreover, Table 1 shows the extreme cases of when $R = 1$: It can be observed that the RBP of an ideal ranked output (i.e., one that has the only one relevant document at Rank 1) can range from 0.05 ($p = 0.95$) and 0.5 ($p = 0.5$), since RBP in this case equals $1-p$. It is not clear why the user's persistence (the probability of moving from a document from Rank r to that at Rank $(r+1)$) should influence the effectiveness value of the same ranked output so drastically, even though only the document at Rank 1 is being examined.

We further note that depending on recall is not necessarily bad. The real user may have some idea of the number of relevant documents, due to his background knowledge, or if not, by looking at the total number of hits shown in the IR interface. Moreover, even if this is not the case, a good IR performance metric is not necessarily one that closely mimics "user satisfaction". For example, a user may be very satisfied with the ranked output, having found a decent document, but he may have missed ten other documents that are in fact more relevant than the one he has found. That is, *the user may be happy, just because he is ignorant.* From a conscientious system developer's point of view, if he *knows* that there are ten relevant documents that should be retrieved, then he would design a system that can retrieve as many of them as possible rather than a system that makes the user "happy" by showing just one relevant document and hiding the other relevant ones completely. Hence Q and AP depend directly on R, the number of judged relevant documents, and even nDCG depends on it indirectly, as it relies on an ideal ranked output.

Figure 1 compares the "top-heaviness" of RBP, AP, Q and nDCG, by considering a ranked output that contains exactly one relevant document, and making it move from Rank 1 to Rank 20. The graph at the top shows the situation when $R = 10$, and the one at the bottom shows the situation when $R = 100$, both under a binary relevance environment. Note that the three RBP curves are not affectd by the value of R.

Table 1. Values of RBP for an ideal ranked output.

	RBP.5	RBP.8	RBP.95
$R = 1$	.5	.2	.05
$R = 10$	.9990	.8926	.4013
$R = 100$	1	1	.9941
$R = 1000$	1	1	1

From the figure, it can be observed that RBP.5 is probably too top-heavy: it basically ignores any relevant document retrieved below Rank 10. This makes evaluation very unstable, as we shall see in our experiments in Section 5. RBP.8 gives a reasonable "rank-bias": RBP.95 looks almost like a straight line, compared to other metrics such as Q-measure and nDCG.

In Figure 1, the top-heaviness curve of AP is almost completely hidden by that of Q-measure, because in a binary relevance environment, $Q\text{-}measure = AP$ holds if there is no relevant document below Rank R, while $Q\text{-}measure > AP$ holds if there is at least one relevant document below Rank R [10]. Thus the AP curve actually begins to deviate from the Q-measure one at Rank 11 in the graph at the top (where $R = 10$).

It can also be observed that the top-heaviness curves of nDCG have a minor problem: nDCG with a logarithm base of 2 cannot distinguish between a system that has a relevant document at Rank 1 and one that has a relevant document at Rank 2. This is because, according to the original definition of nDCG (which we stick to), *gain discounting* cannot be applied to ranks above $a(=2)$. This is precisely why using a large a with nDCG is no good [13]: it makes the top-heaviness curve even flatter. (One rather inelegant way to avoid this problem is, instead of letting $dg(r) = g(r)/\log_a(r)$ for $r > a$, to let $dg(r) = g(r)/\log_a(r+a-1)$ for all r so that gain discounting can be applied to every rank.) It should also be noted that the top-heaviness curve for *nCG* [6], the undiscounted version of nDCG, is a completely flat line. That is, to nCG, it does not matter at all at which rank the relevant document is found. This explains why nCG performs very poorly [12].

To sum up, the IR metrics we consider in this study all have a mechanism, each in its own way, of penalising relevant documents found near the bottom of the ranked list. But the graphs suggest that using $p = 0.5$ for RBP may not be good for reliable evaluation. This we will verify in our experiments described below.

4 Full and Reduced Data

Table 2 provides some statistics of the TREC and NTCIR data we used for evaluating the IR metrics for the purpose of evaluation with incomplete relevance assessements. We chose these data sets as we wanted "ad hoc" test collections with graded relevance data. The "TREC03" and "TREC04" data are from the TREC 2003 and 2004 robust track, and the "NTCIR-6J" and "NTCIR-6C" data are from the NTCIR-6 Crosslingual track. The TREC runs are English monolingual runs, and the NTCIR-6J (NTCIR-6C) runs include both monolingual and crosslingual runs for the Japanese (Chinese) document retrieval subtask.

For conducting our discriminative power experiments described in Section 5, we randomly selected one run from each

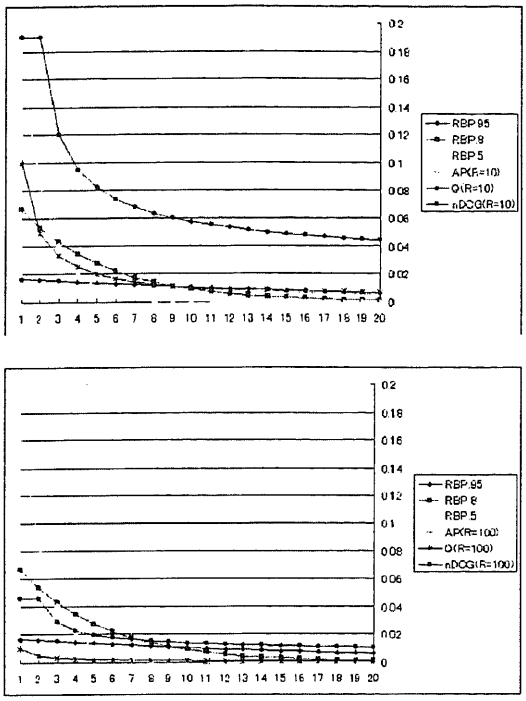

Figure 1. Comparison of "top-heaviness".

Table 2. TREC and NTCIR data used in our experiments.

	TREC03	TREC04	NTCIR-6J	NTCIR-6C
#topics	50	49	50	50
average N	925.5	654.6	1157.9	999.4
average R	33.2	41.2	95.3	88.1
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
#teams	16	14	12	11
#runs used (#all runs)	30(78)	30(110)	30(73)	30(45)

participating teams. Thus, with the TREC03 data, for example, we used 16 runs, which yields 16*15/2=120 combinations of teams for significance testing. For computing Kendall's rank correlation, we wanted more runs, so we randomly sampled 30 runs from each data set, disregarding which team each run comes from. Note that the statistical significance of Kendall's rank correlation depends on the number of runs [11].

To examine the effect of relevance data incompleteness on the IR metrics, we created *reduced relevance data* from the full relevance data, following the original Buckley/Voorhees methodology [3]: First, for each topic, we created a randomised list of judged relevant documents of size R, and a separate randomised list of judged nonrelevant documents of size N. Then, for each *reduction rate* $j \in \{90, 70, 50, 30, 10\}$, we created a reduced set of relevance data by taking the first R_j and N_j documents from the two lists, respectively, where $R_j = \max(1, truncate(R * j/100))$ and $N_j = \max(10, truncate(N * j/100))$. The contants 1 and 10 have been copied from [3], representing the minimum number of judged (non)relevant documents required for a topic. (In practice, the constant 10 was seldom used since N was generally very large.) This stratified sampling is essentially equivalent to random sampling from the entire set of judged documents [17].

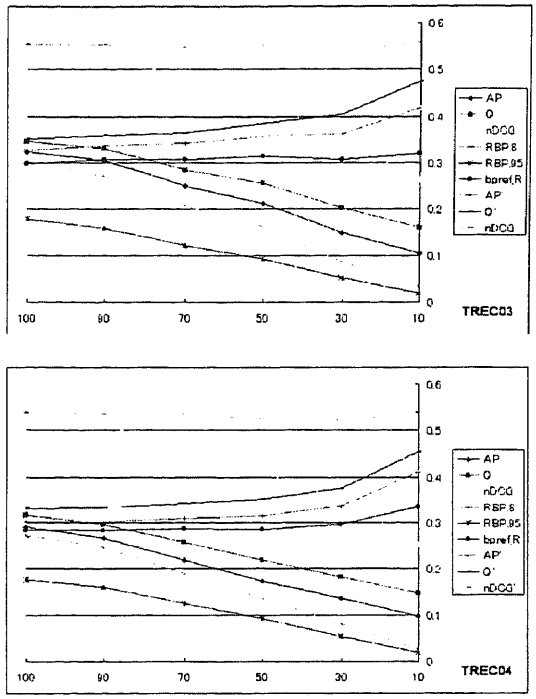

Figure 2. Reduction rate (x axis) vs. absolute performance values averaged over 30 runs (y axis).

Figure 2 shows the effect of relevance data reduction on the absolute overall performances (e.g., Mean AP) averaged across all 30 runs for each data set. (The NTCIR graphs are omitted due to lack of space.) The horizontal axis represents the reduction rate j. It is clear that the values of the metrics based on the original ranked lists (AP, Q, nDCG and RBP) quickly diminish as the relevance data becomes more and more incomplete. (This is not necessarily a flaw: RBP has been designed to behave this way.) In contrast, the bpref_R (i.e., bpref) curve is relatively flat, and this much supports what Buckley and Voorhees reported [3]. However, it is also clear that the Q′, AP′ and nDCG′ curves are just as flat as the bpref one.

5 Discriminative Power

This section compares the robustness of IR metrics to incomplete relevance assessments in terms of discriminative power (or sensitivity) using Sakai's Bootstrap Sensitivity Method [11]. The input to this method are a test collection, a set of runs, an IR metric, and the significance level α for bootstrap hypothesis tests. Using resampled topic sets, the method conducts a bootstrap hypothesis test for every system pair, and computes the discriminative power, i.e., for how many system pairs the IR metric was able to detect a significant difference, and the estimated overall performance difference required to achieve that significance.

Table 3 compares the discriminative power of Q(′), AP(′), nDCG(′), bpref_R and RBP *with the original 100% relevance data*. For example, Table 3(a) shows that Q-measure manages to detect a statistical significance for 80 pairs out of 120

Table 3. Discriminative power at $\alpha = 0.05$ with 100% qrels.

	disc. power	diff. required
(a)TREC03		
Q	80/120=66.7%	0.07
Q′	77/120=64.2%	0.07
AP	77/120=64.2%	0.07
AP′	77/120=64.2%	0.09
nDCG	71/120=59.2%	0.08
nDCG′	71/120=59.2%	0.08
bpref_R	69/120=57.5%	0.08
RBP.8	57/120=47.5%	0.08
RBP.95	55/120=45.8%	0.04
RBP.5	45/120=37.5%	0.12
(b)TREC04		
Q	63/91=69.2%	0.08
Q′	62/91=68.1%	0.08
AP	61/91=67.0%	0.07
AP′	61/91=67.0%	0.07
nDCG	58/91=63.7%	0.08
nDCG′	58/91=63.7%	0.09
bpref_R	57/91=62.6%	0.09
RBP.95	45/91=49.5%	0.05
RBP.8	36/91=39.6%	0.09
RBP.5	30/91=33.0%	0.12
(c)NTCIR-6J		
nDCG	48/66=72.7%	0.09
nDCG′	47/66=71.2%	0.10
Q	47/66=71.2%	0.08
Q′	47/66=71.2%	0.09
AP	46/66=69.7%	0.10
AP′	46/66=69.7%	0.09
bpref_R	42/66=63.6%	0.12
RBP.95	42/66=63.6%	0.07
RBP.8	40/66=60.6%	0.08
RBP.5	36/66=54.5%	0.10
(d)NTCIR-6C		
nDCG′	43/55=78.2%	0.10
Q	42/55=76.4%	0.07
nDCG	42/55=76.4%	0.09
RBP.95	42/55=76.4%	0.06
AP′	42/55=76.4%	0.07
bpref_R	42/55=76.4%	0.08
AP	41/55=74.5%	0.08
Q′	40/55=72.7%	0.08
RBP.8	35/55=63.6%	0.09
RBP.5	27/55=49.1%	0.13

(16*15/2) combinations of teams at $\alpha = 0.05$, and that a difference of around 0.07 is required in order to achieve significance given 50 topics.

We first summarise Table 3 in words:

- For TREC03 and TREC04, Q(′), AP(′), nDCG(′) and bpref_R are more discriminative than RBP.
- For NTCIR-6J, Q(′), AP(′) and nDCG(′) are more discriminative than bpref_R and RBP.
- For NTCIR-6C, Q(′), AP(′), nDCG(′) bpref_R and RBP.95 are more discriminative than RBP.8 and RBP.5.
- To sum up, the overall winners given 100% relevance data are Q(′), AP(′) and nDCG(′).

It is clear from Table 3 that small values of p for RBP hurt discriminative power. This is probably because a small p makes RBP too "top heavy": As we have seen in Figure 1, using $p = 0.5$ implies that IR systems are more or less evaluated based on the top 10 documents only, which makes evaluation very unstable [12]. For this reason, we drop RBP.05 from our experiments henceforth.

Figure 3 shows the effect of relevance data reduction on discriminative power for AP(′), Q(′), nDCG(′), bpref_R and RBP.95 and RBP.8. The results are similar to those reported by Sakai [14], who used four data sets from NTCIR-3 and

Table 4. Discriminative power at $\alpha = 0.05$ with 10% qrels.

	disc. power	diff. required
(a)TREC03		
AP′	63/120=52.5%	0.14
Q′	61/120=50.8%	0.13
nDCG′	60/120=50.0%	0.14
bpref_R	47/120=39.2%	0.16
Q	32/120=26.7%	0.09
nDCG	29/120=24.2%	0.09
RBP.95	26/120=21.7%	0.01
AP	13/120=10.8%	0.08
RBP.8	6/120=5.0%	0.03
(b)TREC04		
Q′	50/91=54.9%	0.11
AP′	46/91=50.5%	0.12
nDCG′	43/91=47.3%	0.12
nDCG	42/91=46.2%	0.09
bpref_R	37/91=40.7%	0.15
Q	29/91=31.9%	0.11
RBP.95	24/91=26.4%	0.01
AP	15/91=16.5%	0.09
RBP.8	10/91=11.0%	0.04
(c)NTCIR-6J		
Q′	46/66=69.7%	0.10
nDCG	45/66=68.2%	0.06
AveP′	44/66=66.7%	0.11
nDCG′	44/66=66.7%	0.10
Q	43/66=65.2%	0.05
bpref_R	39/66=59.1%	0.11
RBP.95	36/66=54.5%	0.01
AP	34/66=51.5%	0.04
RBP.8	21/66=31.8%	0.03
(d)NTCIR-6C		
Q′	39/55=70.9%	0.11
AP′	39/55=70.9%	0.11
bpref_R	39/55=70.9%	0.12
nDCG′	38/55=69.1%	0.12
nDCG	37/55=67.3%	0.06
Q	33/55=60.0%	0.04
RBP.95	31/55=56.4%	0.02
AP	28/55=50.9%	0.04
RBP.8	12/55=21.8%	0.03

NTCIR-5. Table 4 is similar to Table 3 *but uses the 10% relevance data*, thus representing the "tails" of the curves. We summarise Figure 3 and Table 4 in words:

- For TREC03 and TREC04, Q′, AP′ and nDCG′ are more robust than other metrics to incomplete relevance assessments. The original nDCG does well for TREC04 but not for TREC03.
- Similarly, for NTCIR-6J and NTCIR-6C, Q′, AP′, nDCG′ and nDCG are the most robust. (Bpref appears to do well for NTCIR-6C, but it has a problem, as we shall discuss later using Table 5.)
- RBP.95, AP and RBP.8 are at the bottom of the list, exactly in this order for all four data sets.
- To sum up, the overall winners in terms of robustness to incomplete relevance assessments are Q′, AP′ and nDCG′. AP and RBP clearly lack the robustness. nDCG, Q and bpref_R lie in the middle.

The above analysis was based on the number of statistically significant differences detected given incompleteness relevance data. The basic assumption here is that the set of significantly different pairs at X% reduction rate is basically a subset of one with the full relevance data. However, it might be the case that most of these conclusions at X% reduction rate are in fact inconsistent with the original conclusions with the 100% relevance data. We thus provide an additional analysis in Table 5, which is similar in spirit to the "accuracy" of Bompada *et al.* [2]. The table compares, for each metric, the set

Table 5. Number of significant differences detected with 10% qrels but not with 100% qrels.

	#significant	#inconsistent	percentage
(a)TREC03			
AP	13	2	15%
Q	32	2	6%
nDCG	29	0	0%
RBP.8	6	0	0%
RBP.95	26	1	4%
bpref_R	47	7	15%
AP′	63	3	5%
Q′	61	5	8%
nDCG′	60	5	8%
(b)TREC04			
AP	15	1	7%
Q	29	0	0%
nDCG	42	0	0%
RBP.8	10	0	0%
RBP.95	24	0	0%
bpref_R	37	4	11%
AP′	46	5	11%
Q′	50	3	6%
nDCG′	43	1	2%
(c)NTCIR-6J			
AP	34	2	6%
Q	43	4	9%
nDCG	45	1	2%
RBP.8	21	0	0%
RBP.95	36	2	6%
bpref_R	39	1	3%
AP′	44	2	5%
Q′	46	1	2%
nDCG′	44	0	0%
(d)NTCIR-6C			
AP	28	1	4%
Q	33	0	0%
nDCG	37	0	0%
RBP.8	12	0	0%
RBP.95	31	1	3%
bpref_R	39	10	26%
AP′	39	1	3%
Q′	39	1	3%
nDCG′	38	0	0%

of significantly different pairs at 10% reduction rate with that with the full relevance data. For example, Table 5(a) shows that, for TREC03, AP detected a statistical significance for 13 cases with the 10% relevance data, but two of them (15%) are *not* among the set of cases detected by AP with the 100% relevance data. Assuming that the conclusions with the 100% relevance data are the ground truth, the numbers presented in the table represent "errors". As can be seen, the number of errors are generally small, supporting the aforementioned assumption. Bpref_R, however, appears to be quite unreliable from this viewpoint as well: For example, Table 5(d) shows that as many as 10 cases out of the 39 significant differences detected by bpref_R at 10% reduction rate (See also Table 4(d)) are inconsistent with the original bpref results. This, again, is not good news for bpref.

6 Rank Correlation

Table 6 shows the Kendall's rank correlation values between each pair of metrics given the original relevance data. (The NTCIR results are omitted due to lack of space.) As mentioned earlier, we randomly sampled 30 runs from each data set for computing the values: With 30 runs, the correlation is statistically significant if it is over 0.34 [11]; values over 0.9 are shown in bold to indicate high correlations. It can be observed that while the system rankings by AP(′), Q(′), nDCG(′) and bpref_R can be quite similar given the full relevance data, the RBP rankings can be quite different. This alone is not necessarily a flaw: it just means that RBP is measuring something substantially different from the other metrics. Recall that RBP disregards recall.

Table 7. Kendall's rank correlation: 100% vs 10% qrels for each metric.

(a)TREC03		(c)NTCIR-6J	
AP′	.807	AP′	.899
RBP.95	.802	Q′	.894
Q′	.793	nDCG′	.867
Q	.738	RBP.95	.839
nDCG′	.724	nDCG	.821
bpref_R	.724	bpref_R	.802
nDCG	.715	Q	.743
AP	.664	RBP.8	.655
RBP.8	.503	AP	.563
(b)TREC04		(d)NTCIR-6C	
nDCG′	.890	Q′	.949
Q′	.880	nDCG′	.936
bpref_R	.871	nDCG	.917
AP′	.839	Q	.885
nDCG	.798	AP′	.880
RBP.95	.752	bpref_R	.853
Q	.706	AP	.789
AP	.559	RBP.8	.775
RBP.8	.559	RBP.95	.756

Figure 4 shows the effect of relevance data reduction on the system ranking for each metric: Thus, the AP ranking at X% reduction rate is compared with the original AP ranking, and so on. Table 7 summarises the figures by sorting the metrics by Kendall's rank correlation at 10% reduction rate. Figure 4 and Table 7 show that:

- Q′, AP′ and nDCG′ are consistently among the most robust metrics in terms of system ranking stability. Bpref_R does well for TREC04.
- As Figure 4 shows, the system rankings by AP and RBP.8 collapse as relevance data is reduced. RBP.95 is also not very good: at 30% reduction rate, its Kendall's rank correlation with the original ranking is as low as that of AP for TREC04 and for NTCIR-6J; it performs as poorly as RBP.8 for NTCIR-6C.
- To sum up, Q′, AP′ and nDCG′ are again the overall winners, and the advantage of introducing a new metric like bpref is not clear in terms of system ranking stability either. RBP is not as good as Q′, AP′ and nDCG′ in terms of system ranking stability, even with $p = 0.95$. Again, nDCG, Q and bpref_R lie in the middle.

7 Conclusions

This paper compared the robustness of IR metrics to incomplete relevance data, using four different sets of graded-relevance test collections with submitted runs – two from TREC and two from NTCIR. Our discriminative power experiments and rank correlation experiments agreed that Q′, AP′ and nDCG′, the application of Q-measure, AP and nDCG to condensed lists, are more robust than other metrics to relevance data incompleteness; that AP and RBP lack the robustness; and that nDCG, Q and bpref_R lie in the middle. As these results hold across two different evaluation efforts, namely TREC and NTCIR, we believe that these findings are very general. It is also interesting that Q′, nDCG′ and AP′ are comparable in terms of robustness to incomplete relevance data even though Q and nDCG are clearly superior to AP. In

Table 6. Kendall's rank correlation between different metrics, given 100% qrels.

TREC03	Q	nDCG	RBP.8	RBP.95	bpref_R	AP′	Q′	nDCG′
AP	.931	.857	.706	.848	.922	.982	.931	.867
Q	-	.844	.655	.807	.871	.949	.991	.853
nDCG	-	-	.775	.853	.844	.857	.844	.991
RBP.8	-	-	-	.821	.747	.697	.655	.775
RBP.95	-	-	-	-	.899	.839	.798	.853
bpref_R	-	-	-	-	-	.913	.862	.844
AP′	-	-	-	-	-	-	.949	.867
Q′	-	-	-	-	-	-	-	.853

TREC04	Q	nDCG	RBP.8	RBP.95	bpref_R	AP′	Q′	nDCG′
AP	.968	.940	.747	.890	.968	.977	.945	.945
Q	-	.936	.733	.876	.954	.972	.977	.940
nDCG	-	-	.770	.903	.936	.936	.922	.977
RBP.8	-	-	-	.821	.770	.733	.710	.756
RBP.95	-	-	-	-	.913	.876	.853	.890
bpref_R	-	-	-	-	-	.945	.931	.931
AP′	-	-	-	-	-	-	.959	.940
Q′	-	-	-	-	-	-	-	.945

other words, the advantage of using graded relevance seems to disappear when condensed lists are used with very incomplete relevance data.

Our TREC03, TREC04 and NTCIR-6 results, together with the NTCIR-3 and NTCIR-5 results reported by Sakai [14], provide ample evidence that Q′, AP′ and nDCG′ are not only simpler than but also superior to bpref. Although we have no intention of claiming that Q′, AP′ and nDCG′ are the perfect solution to the problem of relevance data incompleteness, we believe that they are more elegant than introducing metrics like bpref and RankEff that lack the "top-heaviness" property of AP by definition.

Even though Moffat, Webber and Zobel [8] claimed that RBP is suitable for evaluation with incomplete relevance data, we demonstrated that it has weaknesses. While RBP is interesting in that it is independent of recall, because of this very feature, it often cannot give 1 even to an ideal ranked output. As we have discussed using Table 1, an ideal output for a topic with 10 (regular) relevant documents may receive an RBP of .4013, while an ideal output for a topic with 100 (regular) relevant documents may receive an RBP of .9941. This is exactly because RBP denies the existence of an ideal ranked output, and whether it is good to average such a measurement across topics can be debated. Our experimental results showed that small values of p make RBP unreliable, and that RBP is not as robust to incomplete relevance data as Q′, AP′ and nDCG′ in terms of discriminative power and system ranking stability, even with $p = 0.95$.

Acknowledgments

We thank Ellen Voorhees for letting us use the TREC robust track data, and Alistair Moffat and Justin Zobel for providing their unpublished manuscript [9].

References

[1] Aslam, J. A., Pavlu, V. and Yilmaz, E.: A Statistical Method for System Evaluation Using Incomplete Judgments, *ACM SIGIR 2006 Proceedings*, pp. 541-548, 2006.

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

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

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

[5] Grönqvist, L.: Evaluating Latent Semantic Vector Models with Synonym Tests and Document Retrieval, ELECTRA Workshop - Methodologies and Evaluating of Lexical Cohesion Techniques in Real-World Applications, pp. 86-88, 2005.

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

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

[8] Moffat, A., Webber, W. and Zobel, J.: Strategic System Comparisons via Targeted Relevance Judgments, *ACM SIGIR 2007 Proceedings*, pp. 375-382, 2007.

[9] Moffat, A. and Zobel, J.: Rank-biased precision for Measurement of Retrieval Effectiveness, *under review*.

[10] Sakai, T.: On the Task of Finding One Highly Relevant Document with High Precision, Information Processing Society of Japan Digital Courier, Vol. 2, pp. 174-188, 2006.

[11] Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *ACM SIGIR 2006 Proceedings*, pp. 525-532, 2006.

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

[13] 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 Acess (EVIA 2007)*, pp. 32-43, 2007.

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

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

[16] Voorhees, E. M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *ACM SIGIR 2002 Proceedings*, pp. 316-323, 2002.

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

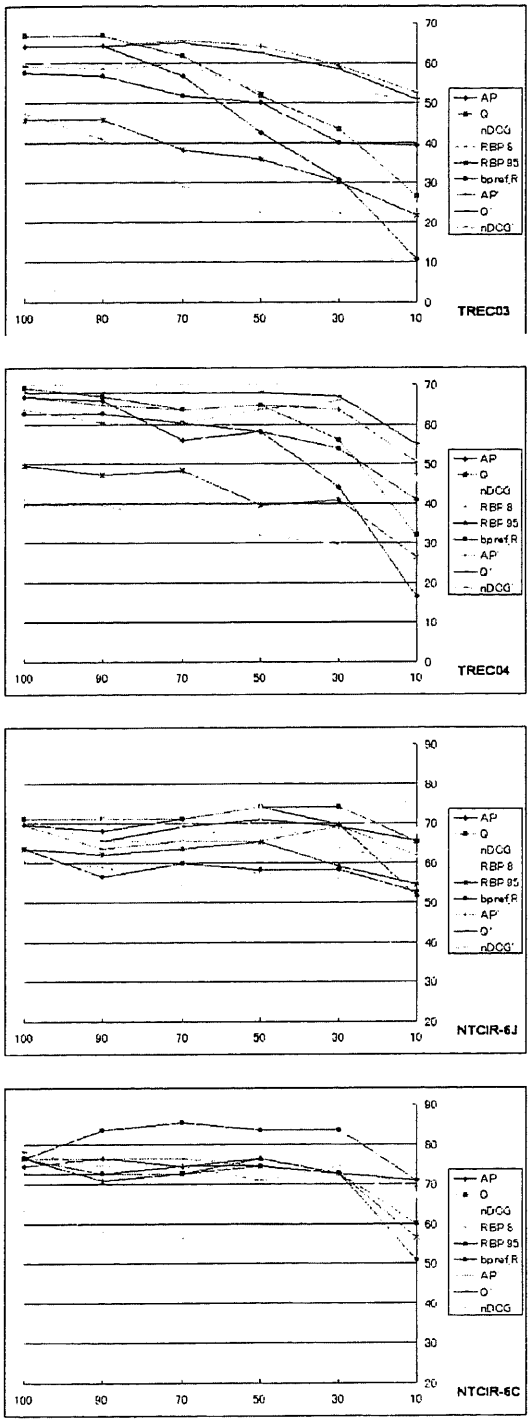

Figure 3. Reduction rate (x axis) vs. discriminative power at $\alpha = 0.05$ (y axis).

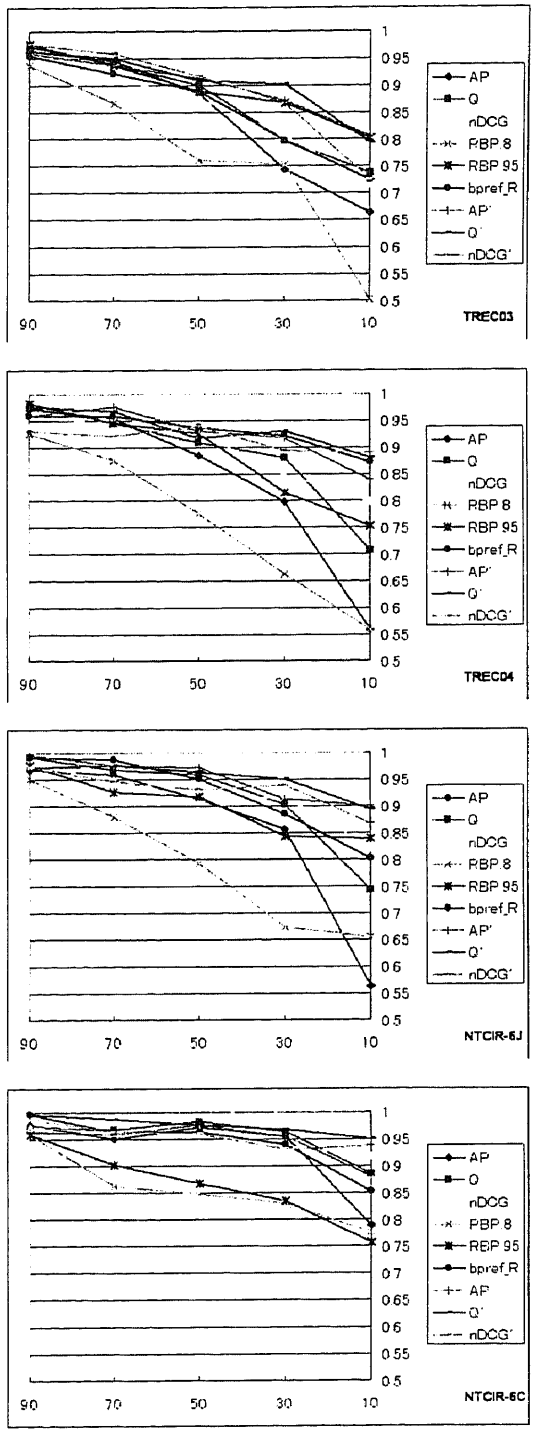

Figure 4. Reduction rate (x axis) vs. Kendall's rank correlation with the 100%-qrels ranking (y axis).