

On information retrieval metrics designed for evaluation with incomplete relevance assessments

Tetsuya Sakai · Noriko Kando

Received: 14 November 2007 / Accepted: 5 March 2008 / Published online: 20 March 2008
© The Author(s) 2008

Abstract Modern information retrieval (IR) test collections have grown in size, but the available manpower for relevance assessments has more or less remained constant. Hence, how to reliably evaluate and compare IR systems using *incomplete* relevance data, where many documents exist that were never examined by the relevance assessors, is receiving a lot of attention. This article compares the robustness of IR metrics to incomplete relevance assessments, using four different sets of graded-relevance test collections with submitted runs—the TREC 2003 and 2004 robust track data and the NTCIR-6 Japanese and Chinese IR data from the crosslingual task. Following previous work, we artificially reduce the original relevance data to simulate IR evaluation environments with extremely incomplete relevance data. We then investigate the effect of this reduction on *discriminative power*, which we define as the proportion of system pairs with a statistically significant difference for a given probability of Type I Error, and on *Kendall's rank correlation*, which reflects the overall resemblance of two system rankings according to two different metrics or two different relevance data sets. 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 point out some weaknesses of $bpref$ and Rank-Biased Precision by examining their formal definitions.

Keywords Evaluation metrics · Relevance assessments · Test collections · Incompleteness

1 Introduction

An information retrieval (IR) test collection comprises a document collection, a set of search requests and a set of manually judged relevant documents for each request.

T. Sakai (✉)
NewsWatch, Inc., Tokyo, Japan
e-mail: tetsuyasakai@acm.org

N. Kando
National Institute of Informatics, Tokyo, Japan
e-mail: kando@nii.ac.jp

Following TREC¹ parlance, hereafter the search requests will be referred to as *topics*, and the set of relevant documents will be referred to as *qrels*. The methodology of using a test collection for evaluating and comparing IR techniques was established in the 1960s, through the Cranfield 2 Test by Cleverdon (1967). Since then, laboratory experiments using test collection have always played a central role for the progress of IR techniques, as they are objective, efficient and repeatable (Voorhees 2002).

Now, in the twenty-first century, evaluation using test collections is still a necessity for most IR researchers. However, the 1990s actually saw a major departure from the original Cranfield 2 experiments, with the advent of TREC, NTCIR² and CLEF³ evaluation efforts for constructing very large-scale test collections. On the surface, the main difference between the Cranfield 2 test collection and the modern test collections is the document collection size: The Cranfield 2 test collection contained only 1,400 documents; The TREC and NTCIR test collections typically contain between half a million to one million documents. However, a more important difference is that while the small scale Cranfield 2 collection had *complete* relevance assessments, the modern test collections do not: It is simply not feasible to examine the documents exhaustively for these large scale collections.

For creating the qrels, TREC, CLEF and NTCIR all adopt a mechanism called *pooling*, which works as follows:

1. Participants submit their “runs” (collections of ranked lists for each topic, where each ranked list usually contains up to 1,000 documents) to the organisers;
2. For each topic, organisers take the top k (typically 100) documents from some of the submitted runs and obtain a list of unique documents, i.e, a document pool;
3. For each topic, assessors judge the relevance of all documents within the pool.

Hence, despite the fact that pooling is a very efficient way of collecting relevant documents, qrels formed through pooling are possibly *incomplete*. That is, there may exist relevant documents within the document collection, which none of the participating systems managed to retrieve and therefore are outside the qrels (Voorhees 2002).

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 remains more or less constant, and therefore test collections are destined to become more and more incomplete. Using an incomplete test collection for IR evaluation raises the following concerns at least:

- (a) Is it possible to reliably compare two participating systems and judge which is superior? Were the test collection less incomplete, would this judgement be the same?
- (b) Is it possible to reliably compare a participating system and a new system that never contributed to the pool? How about two new systems? In short, is the incomplete test collection *reusable*?

For these reasons, IR evaluation using incomplete relevance assessments is receiving more attention than ever.

One obvious approach to tackling these problems is to devise IR effectiveness metrics that are *robust* to relevance data incompleteness: We say that an IR metric is robust to incompleteness if system comparison results based on an incomplete set of relevance data

¹ <http://trec.nist.gov/>

² <http://ntcir.nii.ac.jp/>

³ <http://www.clef-campaign.org/>

are similar to those based on a less incomplete one. This article follows this approach, and more specifically, addresses the issues mentioned in (a) above. Below, we discuss three existing studies that are directly related to the present one. We shall discuss other related work in Sect. 2.

Buckley and Voorhees (2004) 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 used this metric along with AP. Bpref penalises a system if it ranks a *judged nonrelevant* document above a judged relevant one, and is independent of how the *unjudged* documents are retrieved.

More recently, Moffat et al. (2007) 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 always increases the RBP score.

Sakai (2007a) reported that applying Q-measure (or simply, 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 article 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—the TREC 2003 and 2004 robust track data and the NTCIR-6 Japanese and Chinese IR data from the crosslingual task. We believe that evaluating IR systems using graded relevance is important for the progress of IR because, if one adheres to IR evaluation based on binary relevance, it would be very difficult for him to devise an IR algorithm that can retrieve highly relevant documents on top of partially relevant ones. Following previous work, we artificially reduce the original relevance data to simulate IR evaluation environments with extremely incomplete relevance data. We then investigate the effect of this reduction on *discriminative power* (Sakai 2006b, 2007b), which we define as the proportion of system pairs with a statistically significant difference for a given probability of Type I Error, and on *Kendall's rank correlation* (Voorhees 2001), which reflects the overall resemblance of two system rankings according to two different metrics or two different qrels. According to these experiments, Q' , $nDCG'$ and AP' are superior to bpref and RBP.

This article generalises a recent study by Sakai (2007a), in that (1) While he used the NTCIR-3 and NTCIR-5 Japanese/Chinese data, we use TREC robust track data and NTCIR-6 Japanese/Chinese data to obtain more general and substantial conclusions; (2) We compare RBP with the other metrics, after discussing some properties of the metrics that immediately follow from their formal definitions.

The remainder of this article is organised as follows. Section 2 provides an overview of related studies. Section 3 defines and discusses the characteristics of AP' , Q' , $nDCG'$, bpref and RBP. Section 4 describes the TREC and NTCIR data we used for comparing the robustness of these metrics to relevance data incompleteness. Section 5 compares the metrics in terms of discriminative power based on statistical significance tests. Section 6 compares the metrics in terms of Kendall's rank correlation between the entire system rankings. Finally, Sect. 7 concludes this article.

2 Related work

This section provides an overview of previous work related to the present study.

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. Along the first line of research, methods for creating judgment pools efficiently were proposed by Cormack et al. (1998) and by Zobel (1998). Soboroff et al. (2001) proposed a method for ranking systems without any relevance assessments, but subsequently Aslam and Savell (2003) pointed out that the method tends to rank them by “popularity” rather than performance. More recently, Carterette et al. (2006) analyzed the distribution of AP over all possible assignments of relevance to all unjudged documents and proposed a method to construct a test collection with minimal relevance assessments; Büttcher et al. (2007) proposed to “expand” existing relevance assessments by treating them as training data for machine learning.

This article takes the latter approach, of choosing reliable IR metrics for handling relevance data incompleteness. The proposals of the aforementioned bpref (Buckley and Voorhees 2004), RBP (Moffat et al. 2007) and Q' , nDCG' and AP' (Sakai 2007a) fall into this category. Also along this line, Aslam et al. (2006) and Yilmaz and Aslam (2006) proposed *Induced AP*, *Subcollection AP* and *Inferred AP*. Induced AP is exactly what we call AP'. We do not consider Subcollection AP and Inferred AP in our present study, because (1) 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; (2) Both Subcollection AP and Inferred AP require knowledge of the *pooled but unjudged* documents, which limits their applicability;⁴ (3) According to Bompada et al. (2007), Inferred AP is not as robust as the original nDCG for evaluation with incomplete relevance data.

Another metric proposed for handling incomplete relevance data, called *RankEff* (Grönqvist 2005), has been examined by Büttcher et al. (2007). However, Sakai (2008b) points out that RankEff is in fact equivalent to an existing variant of bpref called *bpref_N*, also known as “bpref_allnonrel” implemented in *trec_eval*, the standard IR evaluation software for TREC. Sakai (2007a) showed both analytically and empirically that *bpref_N* is not a good evaluation metric. See Sect. 3.2 for more discussions.

Büttcher et al. (2007) also used Precision at l *judged* documents, which relies on condensed lists just like Q' , AP' and nDCG'. However, Precision is not a satisfactory metric for us because: (1) It ignores the ranks of retrieved relevant documents; (2) It does not average well, especially with a large document cut-off; (3) With a small document cut-off, it gives unreliable results as systems are evaluated based on a small number of observations, i.e., documents near the top of the ranked list (Sakai 2007f).

De Beer and Moens (2006) proposed *rpref*, a graded-relevance version of bpref. Sakai (2007a) pointed out that it has minor bugs, and proposed *rpref_{relative2}* by fixing them. However, he reported that it does not have any advantage over Q' , AP' and nDCG' despite its complexity.

Sakai (2007c) conducted a study similar to the present one, but focussed on the task of finding one highly relevant document. He showed that the application of Reciprocal Rank, O-measure (Sakai 2006c) and P+-measure (Sakai 2006a, 2007e) to condensed lists is an effective way of handling the relevance data incompleteness problem.

⁴ Subcollection AP requires even more knowledge, namely, how small the subcollection with relevance assessments is compared to the entire document collection.

3 Formal definitions of the IR metrics

This section formally defines the IR metrics we consider, namely, $Q(\prime)$, $AP(\prime)$, $nDCG(\prime)$, $bpref$ and RBP , and also discusses their properties that immediately follow from the definitions. Among these metrics, only $Q(\prime)$, $nDCG(\prime)$ and RBP can handle graded relevance.

3.1 Q, AP and nDCG

Let \mathcal{L} denote a *relevance level*, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an \mathcal{L} -relevant document for a particular topic. Following the NTCIR tradition (Kando 2007), this article assumes that we have S-relevant (i.e., highly relevant), A-relevant (i.e., relevant) and B-relevant (i.e., partially relevant) documents. Other documents, i.e., judged nonrelevant documents and unjudged documents, are considered nonrelevant and therefore do not carry a gain value. We let $gain(S) = 3$, $gain(A) = 2$ and $gain(B) = 1$ hereafter as metrics such as Q and nDCG are robust to the choice of gain values (Sakai 2007f). 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. The latter were treated as B-relevant rather than A-relevant because it has been reported that there are many marginally or partially relevant documents in the TREC qrels: Sormunen (2002) reported that about one half of their TREC qrels were only marginally relevant; Sakai and Sparck Jones (2001) reported that only about 56% of the TREC qrels were highly relevant for a subcollection of the early TREC document sets.

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 $cg_I(r)$ 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. Note that whether “nonrelevant” (i.e., either judged nonrelevant or unjudged) documents are retrieved below these relevant documents does not matter, as the nonrelevant documents do not carry gain values. Moreover, note that several “ideal” ranked outputs can exist in general, since documents can be interchanged within each relevance level.

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 $P(r) = count(r)/r$.

We first define Q-measure:

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

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

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 (2007d)

showed empirically that $\beta = 1, 10$ are good choices, so we take $\beta = 1$ throughout this article. Note also that $\beta = 0$ reduces Q-measure to AP:

$$AP = \frac{1}{R} \sum_r isrel(r) \frac{count(r)}{r} = \sum_r isrel(r) P(r). \tag{3}$$

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 $dg_l(r)$ denote the discounted gain for an ideal ranked list. nDCG at document cut-off l is defined as:

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

Throughout this article, we let $l = 1,000$ as it has been reported that nDCG with a small document cut-off is unreliable (Sakai 2007f). Moreover, we let $a = 2$ because it has been reported that nDCG with a large logarithm base is counterintuitive and lacks discriminative power (Sakai 2007d), despite the fact that this parameter was designed to reflect persistence just like RBP’s p and Q-measure’s β . We shall come back to this issue in Sect. 3.4.

3.2 Q', AP', nDCG' and bpref

Sakai (2007a) reported that Q', AP' and nDCG' are simpler and better solutions to the problem of evaluating IR systems with incomplete relevance data than *bpref* (Buckley and Voorhees 2004). Recall that these represent the application of Q, AP and nDCG to condensed lists, respectively.

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 (Sakai 2007a):

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

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:

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

In fact, $R \leq N$ in holds for all topics used in our experiments (See also Table 2), so *bpref* is always *bpref_{-R}* in our study.

Sakai (2007a) 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

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

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

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 (Sakai 2007a). For example, suppose that there is a condensed list that has a judged nonrelevant document at Rank 1 and a relevant document at Rank 2. For this document at rank $r' = 2$, the misplacement penalty is $r' - count(r') = 2 - 1 = 1$, and $P(r') = 1/2$. Thus, the existence of the judged nonrelevant document at Rank 1 weighs heavily in the case of AP'. In contrast, this has very little impact on bpref, because the misplacement penalty is divided not by $r' = 2$ but by a large number, namely, R or N . 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 (Sakai 2007a), which always uses N for scaling the misplacement penalty, suffers severely from this problem, as N is generally a very large number: See, for example, Table 2 which we shall discuss later. Sakai (2007a) showed experimentally that bpref_N indeed performs very poorly.

It should be noted that AP' is actually implemented in trec_eval. However, it appears that it was never properly examined until Yilmaz and Aslam (2006) and Sakai (2007a) rediscovered it.

3.3 RBP

We now formally define RBP (Moffat et al. 2007; Moffat and Zobel 2008). Let \mathcal{H} denote the highest relevance level across all topics. RBP can be expressed as follows:

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

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 (2008) 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 . Whether this assumption is realistic or not is debatable, but this does make RBP easy to interpret and to compute. Moreover, Moffat et al. (2007) 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 $R = 1, 10, 100, 1,000$. When $p = 0.95$, for example, an ideal ranked output for a topic with $R = 10$ receives an RBP of .4013, while one for a topic with $R = 100$ receives .9941. Whether it is good to average such a measurement across topics is debatable, 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.

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
$R = 1,000$	1	1	1

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 $1 - p$. Thus the user's persistence, i.e., the probability of moving from a document from Rank r to that at Rank $(r + 1)$, influences the effectiveness value of the same ranked output quite drastically, even though only the document at Rank 1 is being examined. Whether this is a desirable feature for an IR metric is also debatable. In contrast, Q, nDCG and AP are by definition guaranteed to be one whenever the system output is an ideal ranked output regardless of the value of R , since they are based on comparing the system output with the ideal one.⁵

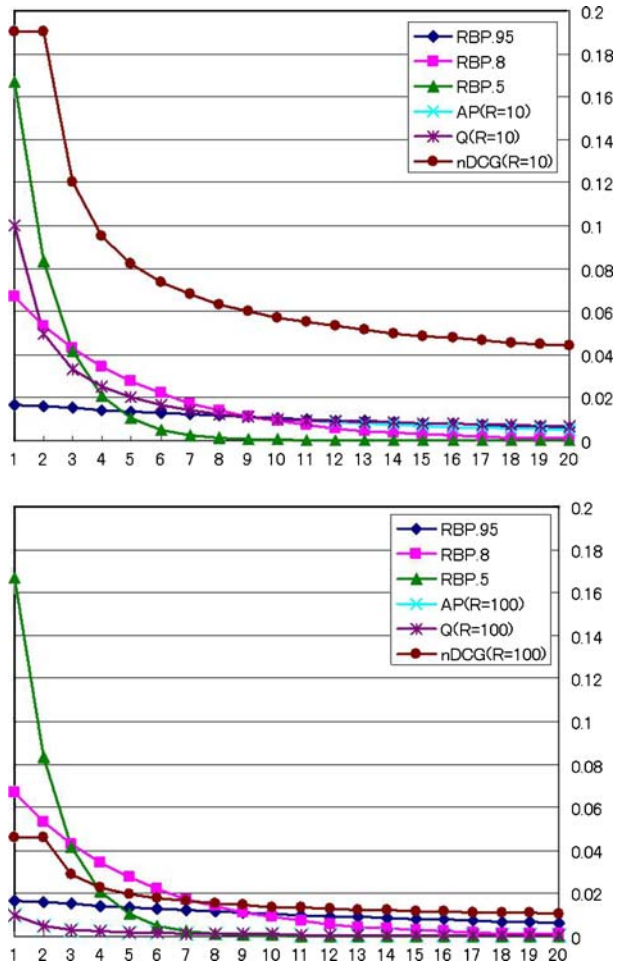
We further argue that depending on recall is not necessarily bad. The real user may have some idea of the number of relevant documents, due to his background knowledge, or if not, by looking at the total number of hits shown in the IR interface. Moreover, even if this is not the case, a good IR performance 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 which lists up all relevant documents.

3.4 Top-heaviness of RBP, AP, Q and nDCG

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 affected by the value of R . 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

⁵ AP, on the other hand, has a different weakness, in that it can be one for a suboptimal ranked output in a graded relevance environment. To be more specific, AP is one as long as all top R documents are at least somewhat relevant: It does not matter if partially relevant documents are retrieved above highly relevant ones.

Fig. 1 Comparison of “top-heaviness”



Sect. 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 Fig. 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 (Sakai 2006c). 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 (Sakai 2007d): it makes the top-heaviness curve even flatter.⁶ It should also be noted

⁶ One way to avoid this problem is to discount the gains by $\log_a(r + 1)$ for every rank r (Burges et al. 2005). However, this cancels out the logarithm base a , which was originally intended as a parameter for reflecting the user’s patience (Järvelin and Kekäläinen 2002).

Table 2 TREC and NTCIR data used in our experiments

	TREC03	TREC04	NTCIR-6J	NTCIR-6C
#Topics	50	49	50	50
#Documents	Approx. 529,000		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
#Teams	16	14	12	11
#All runs	78	110	73	45
#Runs used for rank correlation	30	30	30	30

that the top-heaviness curve for nCG (Järvelin and Kekäläinen 2002; Kekäläinen 2005), 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: and the same goes for Precision at l (Sakai 2007f).

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 assessments. 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 (Voorhees 2004; 2005), and the “NTCIR-6J” and “NTCIR-6C” data are from the NTCIR-6 crosslingual track (Kando 2007). The TREC runs are English monolingual runs, and the NTCIR-6J runs include both monolingual and crosslingual runs for the Japanese document retrieval subtask. Similarly, the NTCIR-6C runs include both monolingual and crosslingual runs for the Chinese document retrieval subtask.

For conducting our discriminative power experiments described in Sect. 5, we randomly selected one run from each participating team. Thus, with the TREC03 data, for example, we used 16 runs, which yields $16 \cdot 15/2 = 120$ combinations of teams for significance testing. Figure 2 shows the distribution of AP values over the runs thus selected. For computing Kendall’s rank correlation, we wanted more runs, and we also wanted the same number of runs across all four data sets.⁷ We therefore randomly sampled 30 runs from each data set, disregarding which team each run comes from.

⁷ The statistical significance of Kendall’s rank correlation depends directly on the number of runs (Sakai 2006b).

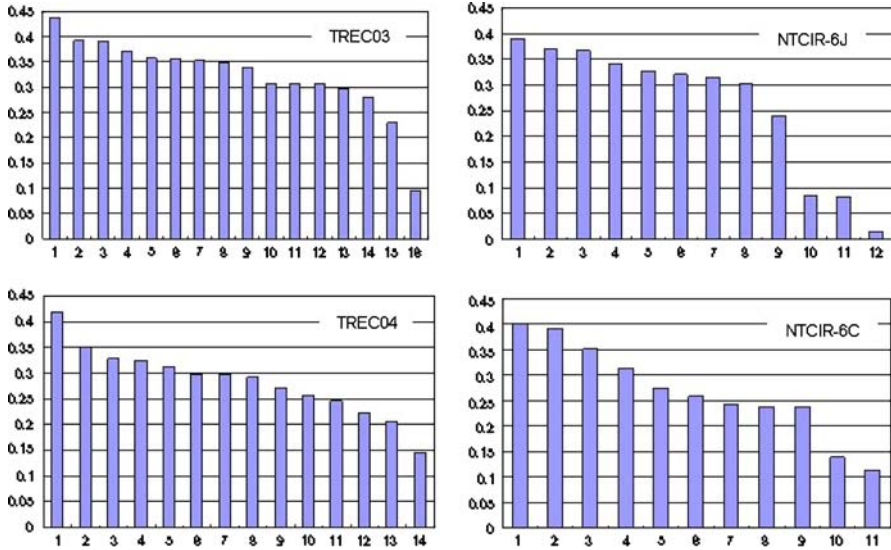


Fig. 2 AP values of the participating teams

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 methodology by Buckley and Voorhees (2004): 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 $R_j = \max(1, \text{truncate}(R*j/100))$ and $N_j = \max(10, \text{truncate}(N*j/100))$. The contents 1 and 10 have been copied from Buckley and Voorhees (2004), representing the minimum number of judged relevant and nonrelevant documents required for a topic, respectively. 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 (Yilmaz and Aslam 2006).

The above method of random sampling from the original qrels may be criticised: Possibly, a better method of studying the effect of incompleteness would be to use the actual pools for each topic, and vary the pool depth. However, we prefer to be faithful to the methodology by Buckley and Voorhees as one of the main goals of this study is to contrast their claims regarding bpref (Buckley and Voorhees 2004) with our new findings. We will report on our “shallow pool” experiments elsewhere (Sakai 2008a).

Figures 3 and 4 show 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 horizontal axis represents the reduction rate j . It is clear that the values of the metrics based on the original ranked lists, i.e., 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 (Buckley and Voorhees 2004). However, it is also clear that the Q' , AP' and $nDCG'$ curves are just as flat as the bpref one.

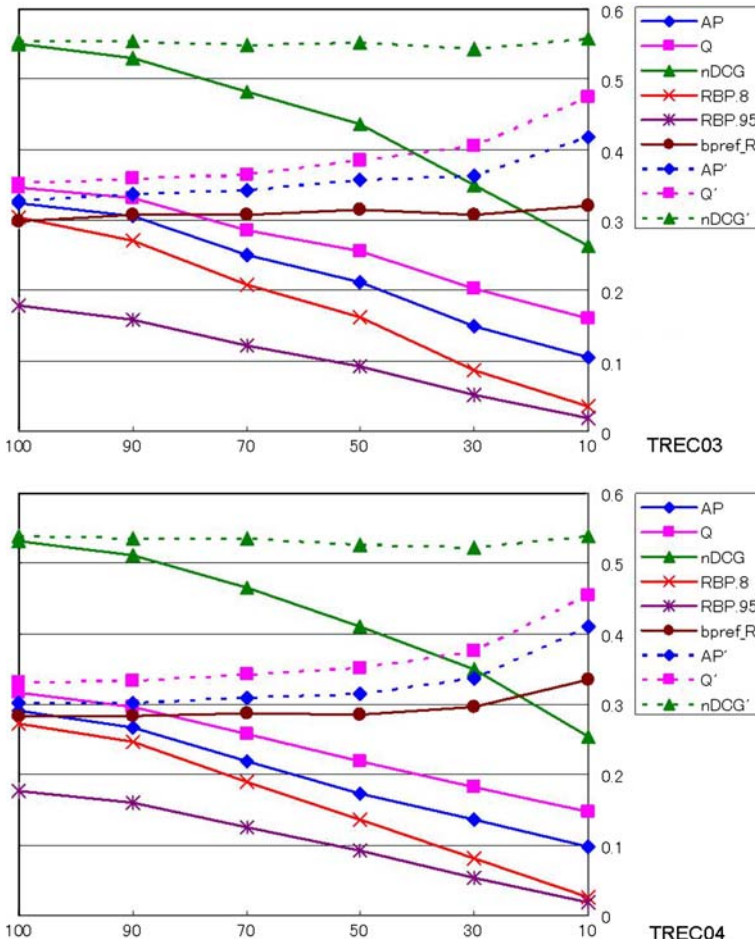


Fig. 3 Reduction rate (x-axis) versus absolute performance values averaged over 30 runs (y-axis) —TREC

5 Discriminative power

This section compares the robustness of IR metrics to incomplete relevance assessments in terms of discriminative power using Sakai’s Bootstrap Sensitivity Method (Sakai 2006b, 2007b). 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* (Efron and Tibshirani 1993). Using resampled topic sets, the method conducts a paired 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. For this purpose, one thousand bootstrap samples of topics were created for each data set, by sampling with replacement from the original topic set. Details of Sakai’s Bootstrap Sensitivity Method can be found elsewhere (Sakai 2006b; 2007b).

It should be stressed that we are comparing the discriminative power of different IR metrics while holding α constant in a common experimental environment, where α is the

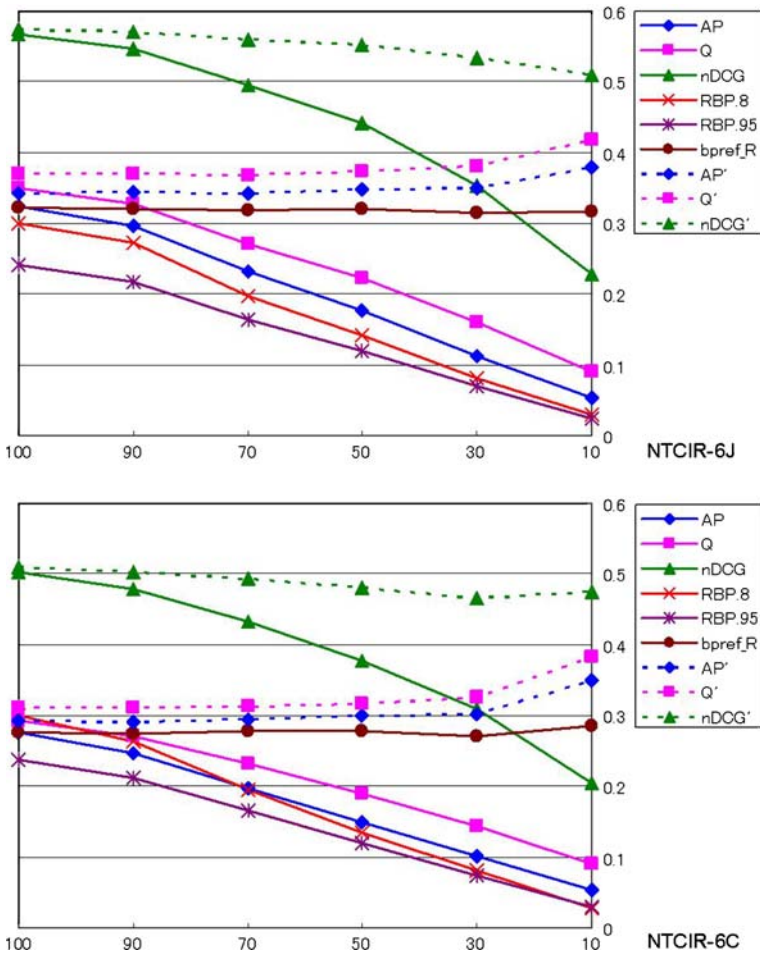


Fig. 4 Reduction rate (x-axis) versus absolute performance values averaged over 30 runs (y-axis)—NTCIR

probability of Type I Error, representing the chance of concluding that two systems are different even though the truth is they are equivalent. In other words, IR metrics with high discriminative power are those that can reliably detect meaningful differences between two systems. The Bootstrap Sensitivity Method is closely related to the *swap method* proposed by Voorhees and Buckley (2002) which compares two systems using two different topics sets and examines whether the two outcomes are the same. Unlike the Bootstrap Sensitivity Method, the swap method lacks a theoretical foundation and is not directly related to statistical significance tests. However, Sakai (2006b, 2007b) reported that the two methods yield very similar results when used for comparing different IR metrics.

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 (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 Fig. 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 unreliable (Sakai 2007f). For this reason, we drop RBP.05 from our experiments henceforth.

Figures 5 and 6 show 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 (2007a), who used four data sets from NTCIR-3 and NTCIR-5. Table 4 is similar to Table 3 but uses the 10% relevance data, thus representing the “tails” of the curves. We summarise Figs. 5, 6 and Table 4 in words:

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

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

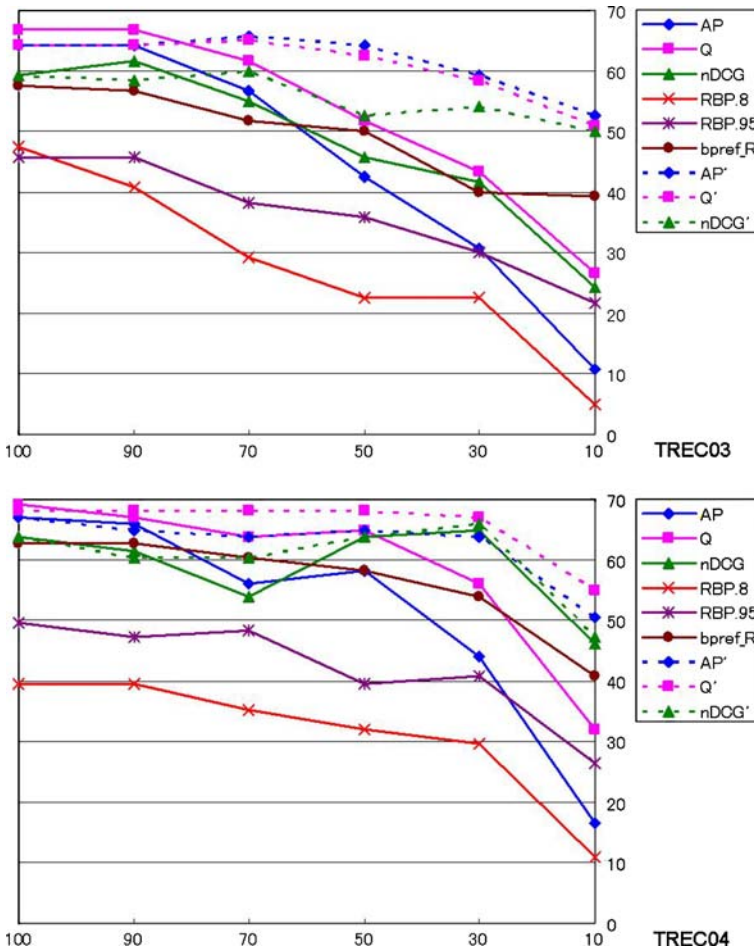


Fig. 5 Reduction rate (*x*-axis) versus discriminative power at $\alpha = 0.05$ (*y*-axis)—TREC

- 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_R 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 $j\%$ reduction rate is basically a subset of one with the full

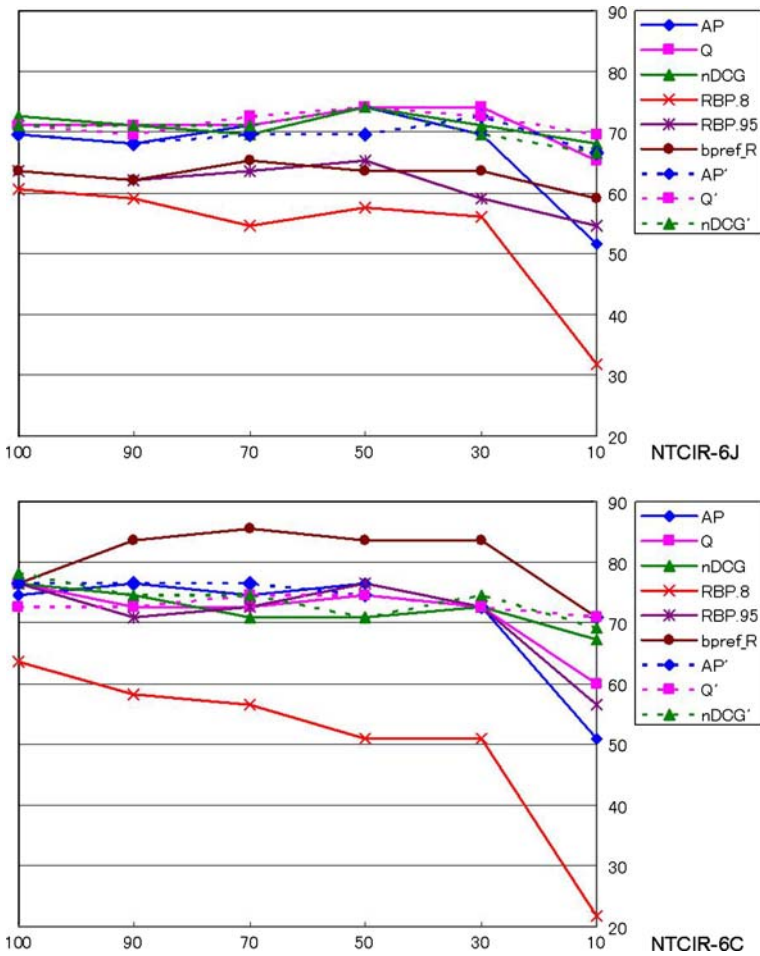


Fig. 6 Reduction rate (x -axis) versus discriminative power at $\alpha = 0.05$ (y -axis)—NTCIR

relevance data. However, it might be the case that most of these conclusions at $j\%$ 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. (2007). The table compares, for each metric, the set 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_R results. This, again, is not good news for bpref.

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

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

6 Rank correlation

The previous section compared the discriminative power of IR metrics, by focussing on the differences between individual *pairs* of runs. We now compare the *entire* system rankings according to two different IR metrics given the original relevance data, as well as two different qrels, namely, the original one and a reduced one.

Table 6 shows the Kendall’s rank correlation values (Voorhees 2001) between each pair of metrics given the original relevance data. 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 (Sakai 2006b); values over 0.9 are shown in bold to indicate high correlations. Note that Kendall’ rank correlation is 1 for a pair of identical rankings and -1 if one ranking is a perfect inverse of the other. It can be observed that while the system rankings by AP(\prime), Q(\prime), nDCG(\prime) 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.

Figures 7 and 8 show 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. Figures 7, 8 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.

Table 5 Number of significant differences detected with 10% qrels but not with 100% qrels: (i) #significant; (ii) #inconsistent; (iii) percentage

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

- As Figs. 7 and 8 show, 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 article compared the robustness of IR metrics to incomplete relevance data, using four different sets of graded-relevance test collections with submitted runs—the TREC 2003 and 2004 robust track data and the NTCIR-6 Japanese and Chinese IR data from the crosslingual task. Our discriminative power experiments and rank correlation experiments agreed that Q', AP' and nDCG', the application of Q, 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 general. It is also interesting that Q', nDCG' and AP' are comparable to one another in

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

	Q	nDCG	RBP.8	RBP.95	bpref_R	AP'	Q'	nDCG'
(a) TREC03								
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
(b) TREC04								
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
(c) NTCIR–6J								
AP	.968	.834	.770	.857	.908	.936	.903	.811
Q	–	.857	.793	.880	.913	.940	.936	.834
nDCG	–	–	.862	.885	.834	.853	.894	.968
RBP.8	–	–	–	.830	.770	.789	.802	.839
RBP.95	–	–	–	–	.867	.885	.890	.853
bpref_R	–	–	–	–	–	.972	.922	.830
AP'	–	–	–	–	–	–	.940	.848
Q'	–	–	–	–	–	–	–	.880
(d) NTCIR–6C								
AP	.963	.903	.885	.949	.972	.986	.936	.908
Q	–	.922	.857	.922	.936	.949	.972	.926
nDCG	–	–	.880	.945	.913	.899	.949	.995
RBP.8	–	–	–	.926	.903	.880	.839	.876
RBP.95	–	–	–	–	.968	.945	.913	.940
bpref_R	–	–	–	–	–	.977	.917	.908
AP'	–	–	–	–	–	–	.922	.903
Q'	–	–	–	–	–	–	–	.954

terms of robustness to incomplete relevance data, even though Q and nDCG are clearly superior to AP. In 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 (2007a), provide ample evidence that Q', AP' and nDCG' are not only simpler than but also superior to bpref, at least in terms of discriminative power and

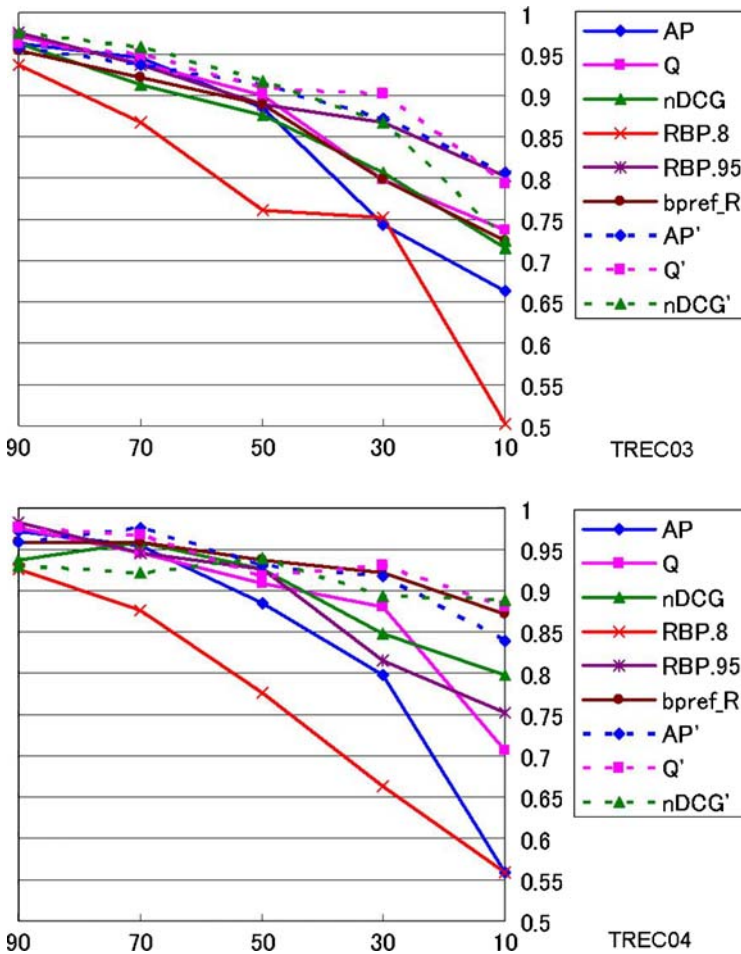


Fig. 7 Reduction rate (*x*-axis) versus Kendall's rank correlation with the 100%-qrels ranking (*y*-axis) — TREC

system ranking stability. 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 $bpref_N$ (i.e., RankEff) that lack the “top-heaviness” property of AP by definition.

Even though Moffat et al. (2007) claimed that RBP is suitable for evaluation with incomplete relevance data as its error due to unjudged documents can be quantified, we demonstrated that it has weaknesses. While RBP is interesting in that it is independent of recall, because of this very feature, it often does not equal one even for an ideal ranked output. For example, 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. Whether it is good to average such a measurement across topics is debatable. Moreover, our experimental results showed that small values of p make RBP unreliable, and that RBP is not as

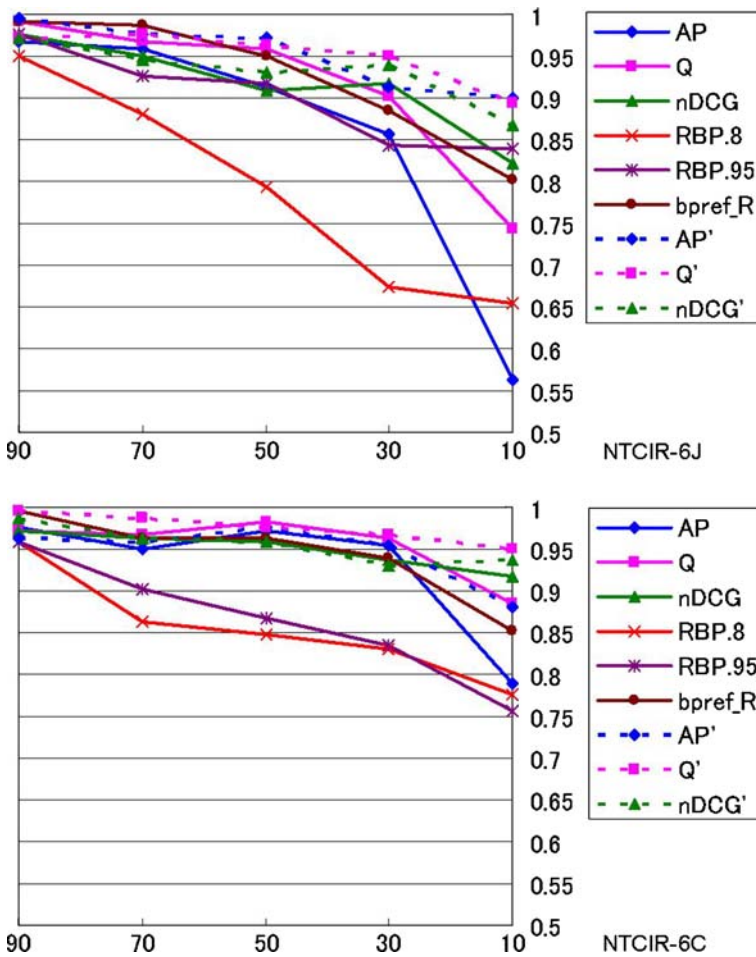


Fig. 8 Reduction rate (x-axis) versus Kendall's rank correlation with the 100%-qrels ranking (y-axis)—NTCIR

robust to incomplete relevance data as Q' , AP' and $nDCG'$ in terms of discriminative power and system ranking stability, even with $p = 0.95$.

The fact that Q' , AP' and $nDCG'$ perform clearly and consistently better than the original Q , AP and $nDCG$ in an incomplete relevance environment implies the following: The assumption that all unjudged documents are nonrelevant is not good; It is much better to treat all unjudged documents as if they never existed, in order to let judged relevant and judged nonrelevant documents move up the ranks and hence serve as stronger pieces of evidence for computing system effectiveness.

It should be recalled, however, that we used random sampling from the original qrels in order to artificially create very incomplete test collections. We shall discuss the effect of using shallow pools and that of using fewer participating teams for forming relevance assessments elsewhere Sakai (2008a, 2008b). Moreover, although we examined the IR metrics in terms of discriminative power and Kendall's rank correlation, there may be other criteria for choosing "good" metrics. "Simplicity" and "intuitiveness" are but a few

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

examples, although they are difficult to quantify. Establishing a standard set of criteria for metric selection is an important goal of our future research.

As we mentioned in Sect. 2, our present study takes the approach of choosing IR metrics *given* a test collection with incomplete relevance data. However, the approach of constructing reliable test collections efficiently (e.g., work by Carterette et al. 2006) is equally important, and combining these two approaches is probably even more so. That is, IR metrics should perhaps be designed by taking the process of test collection construction into account. This is another research topic that needs to be explored.

Acknowledgments We thank William Webber, Justin Zobel and Alistair Moffat for their criticisms and constructive comments. We also thank Ellen Voorhees for letting us use the TREC robust track data, and Alistair Moffat and Justin Zobel for providing their unpublished manuscript (Moffat and Zobel 2008).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Aslam, J. A., & Savell, R. (2003). On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *ACM SIGIR 2003 Proceedings* (pp. 361–362).
- Aslam, J. A., Pavlu, V., & Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In *ACM SIGIR 2006 Proceedings* (pp. 541–548).
- Bompada, T., Chang, C.-C., Chen, J., Kumar, R., & Shenoy, R. (2007). On the robustness of relevance measures with incomplete judgments. In *ACM SIGIR 2007 Proceedings* (pp. 359–366).

- Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *ACM SIGIR 2004 Proceedings* (pp. 25–32).
- Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. (2005). Learning to rank using gradient descent. In *ACM ICML 2005 Proceedings* (pp. 89–96).
- Büttcher, S., Clarke, C. L. A., Yeung, P. C. K., & Soboroff, I. (2007). Reliable information retrieval evaluation with incomplete and biased judgements. In *ACM SIGIR 2007 Proceedings* (pp. 63–70).
- Carterette, B., Allan, J., & Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *ACM SIGIR 2006 Proceedings* (pp. 268–275).
- Cleverdon, C. W. (1967). The cranfield tests on index language devices. In *Aslib Proceedings* (Vol. 19, pp. 173–192).
- Cormack, G. V., Palmer, C. R., & Clarke, C. L. A. (1998). Efficient construction of large test collections. In *ACM SIGIR '98 Proceedings* (pp. 282–289).
- De Beer, J., & Moens, M.-F. (2006). Rpref—a generalization of bpref towards graded relevance judgments. In *ACM SIGIR 2006 Proceedings* (pp. 637–638).
- Efron, B., & Tibshirani, R. (1993). *Introduction to the bootstrap*. Chapman & Hall/CRC.
- Grönqvist, L. (2005). Evaluating latent semantic vector models with synonym tests and document retrieval. In *ELECTRA Workshop—Methodologies and Evaluating of Lexical Cohesion Techniques in Real-World Applications* (pp. 86–88).
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4), 422–446.
- Kando, N. (2007). *Overview of the Sixth NTCIR Workshop*. In *NTCIR-6 Proceedings* (pp. i–ix).
- Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information Processing and Management*, 41, 1019–1033.
- Moffat, A., Webber, W., & Zobel, J. (2007). Strategic system comparisons via targeted relevance judgments. In *ACM SIGIR 2007 Proceedings* (pp. 375–382).
- Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness (under review).
- Sakai, T. (2006a). Bootstrap-based comparisons of IR metrics for finding one relevant document. In *AIRS 2006: Lecture Notes in Computer Science 4182* (pp. 374–389). Springer-Verlag.
- Sakai, T. (2006b). Evaluating evaluation metrics based on the bootstrap. In *ACM SIGIR 2006 Proceedings* (pp. 525–532).
- Sakai, T. (2006c). On the task of finding one highly relevant document with high precision. *Information Processing Society of Japan Transactions on Databases*, 47(SIG4(TOD29)), 13–27.
- Sakai, T. (2007a). Alternatives to Bpref. In *ACM SIGIR 2007 Proceedings* (pp. 71–78).
- Sakai, T. (2007b). Evaluating information retrieval metrics based on bootstrap hypothesis tests. *Information Processing Society of Japan Transactions on Databases* 48(SIG9(TOD35)), 11–28.
- Sakai, T. (2007c). evaluating the task of finding one relevant document using incomplete relevance data. In *Forum on Information Technology 2007 Information Technology Letters* (pp. 91–94).
- Sakai, T. (2007d). On penalising late arrival of relevant documents in information retrieval evaluation with graded relevance. In *Proceedings of the First International Workshop on Evaluating Information Access (EVIA 2007)* (pp. 32–43).
- Sakai, T. (2007e). On the properties of evaluation metrics for finding one highly relevant document. *Information Processing Society of Japan Transactions on Databases*, 48(SIG9(TOD35)), 29–46.
- Sakai, T. (2007f). On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management*, 43(2), 531–548.
- Sakai, T. (2008a). Comparing metrics across TREC and NTCIR: The robustness to pool depth bias (under review).
- Sakai, T. (2008b). Comparing metrics across TREC and NTCIR: The robustness to system bias (under review).
- Sakai, T., & Sparck Jones, K. (2001). Generic summaries for indexing in information retrieval. In *ACM SIGIR 2001 Proceedings* (pp. 190–198).
- Soboroff, I., Nicholas, C., & Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In *ACM SIGIR 2001 Proceedings* (pp. 66–73).
- Sormunen, E. (2002). Liberal relevance criteria of TREC—counting on negligible documents? In *ACM SIGIR 2002 Proceedings* (pp. 324–330).
- Voorhees, E. M. (2001). Evaluation by highly relevant documents. In *ACM SIGIR 2001 Proceedings* (pp. 74–82).
- Voorhees, E. M. (2002). The effect of topic set size on retrieval experiment error. In *ACM SIGIR 2002 Proceedings* (pp. 316–323).
- Voorhees, E. M. (2004). Overview of the TREC 2003 robust retrieval Track. In *TREC 2003 Proceedings*.
- Voorhees, E. M. (2005). Overview of the TREC 2004 robust retrieval track. In *TREC 2004 proceedings*.

- Voorhees, E. M., & Buckley, C. (2002). The philosophy of information retrieval evaluation. In *Proceedings of CLEF 2001, Lecture Notes in Computer Science 2406* (pp. 355–370).
- Yilmaz, E., & Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments. In *ACM CIKM 2006 Proceedings* (pp. 102–111).
- Zobel, J. (1998). How Reliable are the results of large-scale information retrieval experiments? In *ACM SIGIR '98 Proceedings* (pp. 307–314).