Vol. 48 No. SIG 14(TOD 35) IPSJ Transactions on Databases Sep. 2007

Regular Paper

On the Properties of Evaluation Metrics for Finding One Highly Relevant Document

Tetsuya Sakai[†]

Traditional information retrieval evaluation relies on both precision and recall. However, modern search environments such as the Web, in which recall is either unimportant or immeasurable, require precision-oriented evaluation. In particular, finding one highly relevant document is very important for practical tasks such as known-item search and suspected-item search. This paper compares the properties of five evaluation metrics that are applicable to the task of finding one highly relevant document in terms of the underlying assumptions, how the system rankings produced resemble each other, and discriminative power. We employ two existing methods for comparing the discriminative power of these metrics: The Swap Method proposed by Voorhees and Buckley at ACM SIGIR 2002, and the Bootstrap Sensitivity Method proposed by Sakai at SIGIR 2006. We use four data sets from NTCIR to show that, while P($^+$)-measure, O-measure and NWRR (Normalised Weighted Reciprocal Rank) are reasonably highly correlated to one another, P($^+$)-measure and O-measure are more discriminative than NWRR, which in turn is more discriminative than Reciprocal Rank. We therefore conclude that P($^+$)-measure and O-measure, each modelling a different user behaviour, are the most useful evaluation metrics for the task of finding one highly relevant document.

1. Introduction

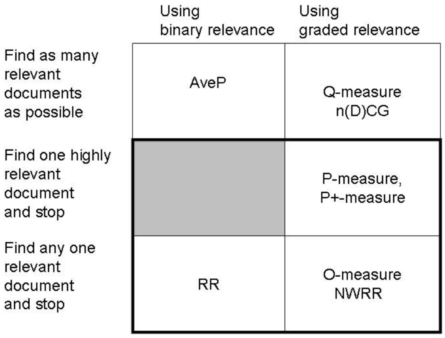
Different Information Retrieval (IR) tasks require different evaluation metrics. For example, a patent survey task may require a *recall-oriented* metric, while a *known-item search* task[20] may require a *precision-oriented* metric. When we search the Web, for example, we often stop going through the ranked list after finding *one* good Web page even though the list may contain some more relevant pages, either knowing or assuming that the rest of the retrieved pages lack *novelty*, or additional information that may be of use to him. Thus, finding exactly *one* relevant document with high precision is an important IR task.

Reciprocal Rank (RR)[3),20)] is commonly used for the task of finding one relevant document: $RR = 0$ if the ranked output does not contain a relevant document; otherwise, $RR = 1/r_1$, where r_1 is the rank of the retrieved relevant document that is nearest to the top of the list. However, RR is based on binary relevance and therefore cannot distinguish between a retrieved *highly* relevant document and a retrieved *partially* relevant document. Thus, as long as RR is used for evaluation, it is difficult for researchers to develop a system that can rank a highly relevant document above partially relevant ones. In light of this, Sakai[15)] proposed a metric called *O-measure* for the task of finding one *highly* relevant document. O-measure is a variant of *Q-measure*[10),18)] which is very highly correlated with Average Precision (AveP) but can handle graded relevance. O-measure can also be regarded as a generalisation of RR (See Section 3).

Eguchi, et al.[5)], who ran the currently-discontinued NTCIR Web track, also proposed a metric for the task of finding one highly relevant document, namely, *Weighted Reciprocal Rank* (WRR). WRR assumes that ranking a *partially* relevant document at (say) Rank 1 is more important than ranking a *highly* relevant document at Rank 2. It was never actually used for ranking the systems at NTCIR (See Section 3) and its discriminative power has not been reported. We point out in Section 3 that, if WRR must be used, then it should be normalised before averaging across topics: We call the normalised version *Normalised Weighted Reciprocal Rank* (NWRR).

Just like RR, both O-measure and NWRR rely on r_1, the rank of the first relevant document in the list. This means that all of these metrics assume that *the user stops examining the ranked list as soon as he finds one relevant document, even if it is only partially relevant.* This assumption may be valid in some retrieval

[†] NewsWatch, Inc. (This work was done when the author was at Toshiba.)



	Using binary relevance	Using graded relevance
Find as many relevant documents as possible	AveP	Q-measure n(D)CG
Find one highly relevant document and stop	▨	P-measure, P+-measure
Find any one relevant document and stop	RR	O-measure NWRR

Fig. 1 Categories of IR metrics, with some examples.

situations, but not always, as we shall discuss in Section 3. In contrast, *P-measure*, proposed at the ACM SIGIR 2006 poster session[14)], assumes that *the user looks for a highly relevant document even if it is ranked below partially relevant documents*. We shall also discuss its variant called P^{+}(pee-plus)-measure in Section 3.

Thus we have at least five evaluation metrics for the task of finding one relevant document: P^{+}-measure, P-measure, O-measure, NWRR and RR. All of them except for RR can handle graded relevance, as illustrated in **Fig. 1**. (Section 2 will touch upon all of the metrics shown in this figure.) This paper[☆] compares the properties of these five metrics that are applicable to the task of finding one highly relevant document, in terms of the underlying assumptions, how the system rankings produced resemble each other, and discriminative power. We employ two existing methods for comparing the discriminative power of these metrics: The Swap Method proposed at ACM SIGIR 2002[24)], and the Bootstrap Sensitivity Method proposed at SIGIR 2006[12),16)]. We use four data sets from NTCIR to show that, while $\mathrm{P}(^{+})$-measure, O-measure and NWRR are reasonably highly correlated to one another, $\mathrm{P}(^{+})$-measure and O-measure are more discriminative than NWRR, which in turn is more discriminative than RR. We therefore conclude that $\mathrm{P}(^{+})$-measure and O-measure, each modelling a different user behaviour, are the most useful evaluation metrics for the task of finding one highly relevant document.

The remainder of this paper is organised as follows. Section 2 discusses previous work on evaluating IR metrics or IR evaluation environments to clarify the contribution of this study. Section 3 formally defines and characterises the metrics we examine. Section 4 describes the methods we use for comparing the metrics, namely, Kendall's rank correlation for examining the resemblance between metrics, and the Swap Method and the Bootstrap Sensitivity Method for assessing the discriminative power of each metric. Section 5 describes our experiments for comparing $\mathrm{P}(^{+})$-measure, O-measure, NWRR and RR, and Section 6 concludes this paper. The Appendix provides some statistics on the actual ranks examined by $\mathrm{P}(^{+})$-measure and O-measure for the NTCIR data.

2. Related Work

This section discusses previous work on evaluating IR metrics or IR evaluation environments, with an emphasis on those based on graded relevance. IR metrics are often evaluated in terms of how they resemble each other in system ranking, and in terms of discriminative power and/or *stability*[1)] with respect to change in the test collection topic set.

At ACM SIGIR 2001, Voorhees[22)] used Kendall's rank correlation to compare system rankings produced by binary-relevance IR metrics such as Average Precision (AveP) and a graded-relevance IR metric known as *Discounted Cumulative Gain* (DCG)[7)]. Kekäläinen[9)] conducted a similar study, but compared *normalised* (Discounted) Cumulative Gain (n(D)CG) with Precision at a fixed document cutoff (PDoc). These studies examined the *resemblance* among different metrics, but did not discuss the discriminative power of each metric. Moreover, these studies considered IR metrics for the task of handling as many relevant documents as possible, in contrast to the present study which concerns metrics for finding one relevant document only.

At SIGIR 2002, Voorhees and Buckley[24)] proposed the Swap Method for assessing the discriminative power of IR metrics, and for estimating the overall performance difference between two systems for guaranteeing that one system is better than another with a given "confidence". The original Swap Method samples topics *without* replacement from the original topic set Q to generate B pairs of topic sets Q_i and Q'_i such that $Q_i \cap Q'_i = \phi$ $(1 \leq i \leq B)$, and establishes an empirical relationship between the overall performance difference between two

[☆] This paper extends one presented at the Asia Information Retrieval Symposium (AIRS) 2006, by including additional experiments using the Swap Method[24)] and Kendall's rank correlation.

systems and the *swap rate*, which represents the probability of the event that two experiments (each using a different topic set) are contradictory. (We let $B = 1000$ throughout this paper.) For example, if the original topic set contains 50 topics, new topic sets Q_i and Q'_i of size up to 25 are created, while ensuring that Q_i and Q'_i are disjoint. The Swap Method can estimate how much difference is required to guarantee a small swap rate, and also what percentage of system pairs actually satisfy the difference criterion. The latter quantity represents the discriminative power of a given metric with a given number of topics.

Several studies followed that used the Swap Method for assessing different IR metrics: Buckley and Voorhees[2] assessed their *bpref* (binary preference) metric to deal with incomplete relevance judgments; Voorhees[23] assessed the *area* measure and *Geometric Mean* AveP (G_AveP) to emphasise the effect of worst-performing topics. Soboroff[20] assessed RR for the TREC Web known-item search task. Sakai[18] assessed graded-relevance metrics *Q-measure*, *R-measure* and normalised (Discounted) Cumulative Gain (n(D)CG) and binary-relevance ones such as Average Precision (AveP) and Precision at a fixed document cut-off (PDoc). Moreover, Sanderson and Zobel[19] and Sakai[11] explored some variants of the Swap Method, and the latter showed that topic sampling *with* and *without* replacement for the Swap Method yield very similar results for the purpose of comparing different metrics. This implies that it is no longer necessary to take (say) 25 topics from 50 in order to ensure that Q_i and Q'_i are disjoint[12],[16]: One can resample 50 topics from the original set of 50 topics, by sampling *with* replacement, to directly estimate the overall performance difference required given 50 topics, instead of doing the original Swap Method experiments with 25 topics and then *extrapolating* to 50 topics[19],[24]. This paper therefore uses the sampling-*with*-replacement version of the Swap Method.

Most of the abovementioned work that examined the discriminative power of metrics focussed on IR metrics for finding as many relevant documents as possible. An exception is the aforementioned study by Soboroff[20], which examined RR. However, as mentioned earlier, RR ignores the difference between a retrieved highly relevant document and a retrieved partially relevant document. In light of this, Sakai[15] proposed *O-measure*, which is a metric for the task of finding one relevant document and is based on *graded* relevance. He examined the rank correlation between O-measure and RR, and also showed that O-measure is more discriminative than RR according to the original Swap Method. Subsequently, Sakai[14] showed that a new metric called *P-measure* is at least as discriminative as O-measure according to the sampling-with-replacement version of the Swap Method[☆].

While the aforementioned studies demonstrated the usefulness of the Swap Method for comparing the discriminative power of different IR metrics, the method lacks a theoretical foundation, and is not highly correlated with statistical significance tests[18]. In light of this, Sakai[12],[16] proposed the *Bootstrap Sensitivity Method* for assessing the discriminative power of IR metrics, which relies on the time-honoured *Bootstrap Hypothesis Tests*[4]. This method obtains B *bootstrap samples* Q^{*b} $(1 \leq b \leq B = 1000)$ by sampling *with* replacement from the original topic set Q, such that $|Q^{*b}| = |Q|$, conducts a Bootstrap Hypothesis Test for every system pair, and estimates an absolute difference required to guarantee a given significance level α.

However, Sakai's work[12],[16] only dealt with IR metrics for finding as many relevant documents as possible, including AveP, Q-measure and *Geometric Mean Q-measure* (G_Q-measure). This paper focusses on the task of finding one relevant document, and compares P-measure, P^{+}-measure[14], O-measure[15], NWRR and RR, using both the Swap Method and the Bootstrap Sensitivity Method for comparing discriminative power, as well as Kendall's rank correlation for examining the resemblance among metrics. In short, previous work paid relatively little attention to the evaluation metrics for the task of finding one relevant document, but this is exactly the focus of this study.

3. IR Effectiveness Metrics

This section formally defines and characterises $\mathrm{P}(^{+})$-measure, O-measure and NWRR. (We have already defined RR in Section 1.)

[☆] Sakai's SIGIR poster[14] used the NTCIR-5 data only: The Swap Method experiments reported in this paper extends his work by (a) using the NTCIR-3 data as well; and (b) examining additional metrics, namely, NWRR and P^{+}-measure.

Prior to this, we also define AveP and Q-measure since we include them in our experiments just for comparison. Our experimental results of AveP and Q-measure have been copied from Refs. 12), 16), and are not part of this paper's contribution.

3.1 AveP and Q-measure

Let R denote the number of relevant documents for a topic, and let L (≤ 1000) denote the size of a ranked output. For each Rank r ($\leq L$), let $isrel(r)$ be 1 if the document at Rank r is relevant and 0 otherwise, and let $count(r) = \sum_{1 \leq i \leq r} isrel(i)$. Clearly, Precision at Rank r is given by $P(r) = count(r)/r$. Then, AveP is defined as:

$$AveP = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r)P(r) \ . \qquad (1)$$

Next, we define Q-measure[10),18)], which is very highly correlated with AveP but can handle graded relevance. Let $R(\mathcal{L})$ denote the number of $\mathcal{L}$-relevant documents so that $\sum_{\mathcal{L}} R(\mathcal{L}) = R$, and let $gain(\mathcal{L})$ denote the *gain value* for retrieving an $\mathcal{L}$-relevant document. In the case of NTCIR, $\mathcal{L} = S$ (highly relevant), $\mathcal{L} = A$ (relevant) or $\mathcal{L} = B$ (partially relevant), and we use $gain(S) = 3, gain(A) = 2, gain(B) = 1$ by default. Let $cg(r) = \sum_{1 \leq i \leq r} g(i)$ denote the *cumulative gain* at Rank r for a system output[9)], where $g(i) = gain(\mathcal{L})$ if the document at Rank i is $\mathcal{L}$-relevant and $g(i) = 0$ otherwise. Similarly, let $cg_I(r)$ denote the cumulative gain at Rank r for an *ideal* ranked output: For NTCIR, an ideal ranked output lists up all S-, A- and B-relevant documents in this order. Then, Q-measure is defined as:

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

where

$$BR(r) = \frac{cg(r) + count(r)}{cg_I(r) + r} \ . \qquad (2)$$

$BR(r)$ is called the *blended ratio*, which measures how a system output deviates from the ideal ranked output *and* penalises "late arrival" of relevant documents. (Unlike the blended ratio, it is known that *weighted precision* $WP(r) = cg(r)/cg_I(r)$ cannot properly penalise late arrival of relevant documents and is therefore not suitable for IR evaluation[15),18)].)

3.2 O-measure and NWRR

Traditional IR assumes that *recall* is important: Systems are expected to return as many relevant documents as possible. AveP and Q-measure, both of which are recall-oriented, are suitable for such tasks. (Note that the number of relevant documents R appear in their definitions.) However, as was discussed in Section 1, some IR situations do not necessarily require recall. More specifically, some IR situations require *one* relevant document only. Although RR is commonly used in such a case, it cannot reflect the fact that users prefer highly relevant documents to partially relevant ones. Below, we describe O-measure and Normalised Weighted Reciprocal Rank (NWRR), both of which can be regarded as graded-relevance versions of RR.

O-measure[15)] is defined to be zero if the ranked output does not contain a relevant document. Otherwise:

$$O\text{-}measure = BR(r_1) = \frac{g(r_1)+1}{cg_I(r_1)+r_1} \ . \quad (3)$$

That is, O-measure is the blended ratio at Rank r_1. (Since the document at r_1 is the *first* relevant one, note that $cg(r_1) = g(r_1)$ and $count(r_1) = 1$ hold.) In a binary relevance environment, $O\text{-}measure = RR$ holds iff $r_1 \leq R$, and $O\text{-}measure > RR$ holds otherwise. Moreover, if small gain values are used with O-measure, then it behaves like RR[15)].

Next, we define Weighted Reciprocal Rank (WRR) proposed by Eguchi, et al.[5)]. Our definition looks slightly different from their original one, but it is easy to show that the two are equivalent[13)]. In contrast to cumulative-gain-based metrics (including Q-measure and O-measure) which require the gain values ($gain(\mathcal{L})$) as parameters, WRR requires "penalty" values $\beta(\mathcal{L})$ (> 1) for each relevance level $\mathcal{L}$. We let $\beta(S) = 2, \beta(A) = 3, \beta(B) = 4$ throughout this paper: note that the smallest penalty value must be assigned to highly relevant documents. WRR is defined to be zero if the ranked output does not contain a relevant document. Otherwise:

$$WRR = \frac{1}{r_1 - 1/\beta(\mathcal{L}_1)} \qquad (4)$$

where $\mathcal{L}_1$ denotes the relevance level of the relevant document at Rank r_1.

WRR was designed for the NTCIR Web track, but the track organisers always used $\beta(\mathcal{L}) = \infty$ for all $\mathcal{L}$, so that WRR is reduced to binary RR. That is, the graded relevance capability of WRR has never actually been used.

WRR is not bounded by one: If the highest

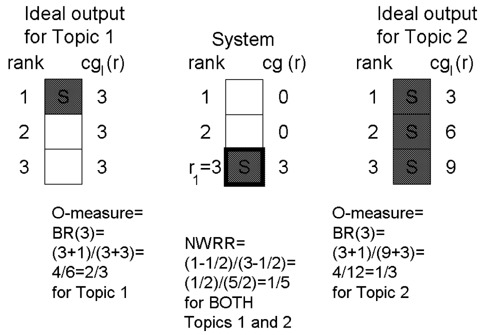


Fig. 2 O-measure vs NWRR: Topics 1 and 2.

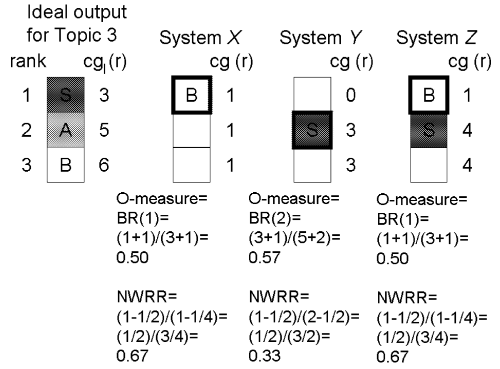


Fig. 3 O-measure vs NWRR: Topic 3.

relevance level for a given topic is denoted by $\mathcal{M}$, WRR is bounded above by $1/(1-1/\beta(\mathcal{M}))$. This is undesirable for two reasons: Firstly, a different set of penalty values yields a different range of WRR values, which is inconvenient for comparisons; Secondly, the highest relevance level $\mathcal{M}$ may not necessarily be the same across topics, so the upperbound of WRR may differ across topics. This means that WRR is not suitable for averaging across topics if $\mathcal{M}$ differs across the topic set of the test collection.

This paper therefore considers Normalised WRR (NWRR) instead. NWRR is defined to be zero if the ranked output does not contain a relevant document. Otherwise:

$$NWRR = \frac{1 - 1/\beta(\mathcal{M})}{r_1 - 1/\beta(\mathcal{L}_1)} \ . \tag{5}$$

The upperbound of NWRR is one for any topic and is therefore averageable.

There are two important differences between NWRR and O-measure.

(a) *Just like RR, NWRR disregards whether there are many relevant documents or not. In contrast, O-measure takes the number of relevant documents into account by comparing the system output with an ideal output.*

(b) *NWRR assumes that the rank of the first retrieved document is more important than the relevance levels.* Whereas, O-measure is free from this assumption.

We first discuss (a). From Eq. (5), it is clear that NWRR depends only on the rank and the relevance level of the first retrieved relevant document. For example, consider a system output shown in the middle of **Fig. 2**, which has an S-relevant document at Rank 3. The NWRR for this system is $(1 - 1/\beta(S))/(3 - 1/\beta(S)) = (1 - 1/2)/(3 - 1/2) = 1/5$ for *any* topic. Whereas, the value of O-measure for this system depends on how many $\mathcal{L}$-relevant documents there are. For example, if the system output was produced in response to Topic 1 which has only one S-relevant document (and no other relevant documents), then, as shown on the left hand side of Fig. 2, *O-measure* $= (g(3) + 1)/(cg_I(3) + 3) = (3 + 1)/(3 + 3) = 2/3$. On the other hand, if the system output was produced in response to Topic 3 which has at least three S-relevant documents, then, as shown in the right hand side of the figure, *O-measure* $= (3 + 1)/(9 + 3) = 1/3$. Thus, *O-measure assumes that it is relatively easy to retrieve an $\mathcal{L}$-relevant document if there are many $\mathcal{L}$-relevant documents in the database.* If the user has no idea as to whether a document relevant to his request exists or not, then one could argue that NWRR may be a better model. On the other hand, if the user has some idea about the number of relevant documents he might find, then O-measure may be more suitable. Put another way, O-measure is more *system-oriented* than NWRR.

Next, we discuss (b) using Topic 3 shown in **Fig. 3**, which has one S-relevant, one A-relevant and one B-relevant document. System X has a B-relevant document at Rank 1, while System Y has an S-relevant document at Rank 2. Regardless of the choice of penalty values ($\beta(\mathcal{L})$), X always outperforms Y according to NWRR. Thus, *NWRR is unsuitable for IR situations in which retrieving a highly relevant document is more important than retrieving any relevant document in the top ranks.* In contrast, O-measure is free from the assumption underlying NWRR: Fig. 3 shows that, with default gain values, Y outperforms X. But if X should be preferred, then a different gain value assignment (e.g. $gain(S) = 2, gain(A) = 1.5, gain(B) = 1$) can be used. In this respect, O-measure is more flexible than NWRR.

3.3 P-measure and P$^+$-measure

Despite the abovementioned differences, both NWRR and O-measure rely on r_1, the rank of the first retrieved relevant document. Thus, both *NWRR and O-measure assume that the user stops examining the ranked list as soon as he finds one relevant document, even if it is only a partially relevant one.* This assumption may be counterintuitive in some cases: Consider System Z in Fig. 3, which has a B-relevant document at Rank 1 *and* an S-relevant document at Rank 2. According to both NWRR and O-measure, System Z and System X are always equal in performance regardless of the parameter values, because only the B-relevant document at Rank $r_1 = 1$ is taken into account for Z. In short, both NWRR and O-measure ignore the fact that there is a better document at Rank 2.

This is not necessarily a flaw. NWRR and O-measure may be acceptable models for IR situations in which it is difficult for the user to spot a highly relevant document in the ranked list. For example, the user may be looking at a plain list of document IDs, or a list of vague titles and poor-quality text snippets of the retrieved documents. Or perhaps, he may be examining the content of each document one-by-one without ever looking at a ranked list, so that he has no idea what the next document will be like. However, if the system can show a high-quality ranked list that contain informative titles and abstracts, then perhaps it is fair to assess System Z by considering the fact that it has an S-relevant document at Rank 2, since a real-world user can probably spot this document. Similarly, in *known-item search*[20], the user probably knows that there exists a highly relevant document, so he may continue to examine the ranked list even after finding some partially relevant documents. (A "narrow" definition of known-item search would involve only one relevant document per topic. That is, the target document is defined to be the one that the user has seen before, and it is always highly relevant. However, we adopt a broader definition: In addition to the known highly relevant document, there may be unvisited documents which are in fact relevant to the topic. It is possible to treat these documents as partially relevant in evaluation. Moreover, there is a related task called *suspected-item* search[6], which does not require that the user has actually *seen* a relevant document. It is clear that more than one relevant document may exist in such cases too, possibly with different relevance levels.)

We now define P-measure[14] for the task of finding one highly relevant document, under the assumption that *the user continues to examine the ranked list until he finds a document with a satisfactory relevance level.* P-measure is defined to be zero if the system output does not contain a relevant document. Otherwise, let $\mathcal{L}_p$ be the highest relevance level observed within the system output, and let the *preferred rank* r_p be the rank of the first $\mathcal{L}_p$-relevant document found in it. Then:

$$\begin{aligned}\textit{P-measure} &= BR(r_p)\\ &= \frac{cg(r_p) + \textit{count}(r_p)}{cg_I(r_p) + r_p}\ . \qquad (6)\end{aligned}$$

That is, P-measure is simply the blended ratio at Rank r_p. For System Z in Fig. 3, $r_p = 2$. Therefore, $\textit{P-measure} = BR(2) = (cg(2) + 2)/(cg_I(2) + 2) = (4 + 2)/(5 + 2) = 0.86$. Whereas, since $r_p = r_1$ holds for systems X and Y, $\textit{P-measure} = \textit{O-measure} = 0.50$ for X and $\textit{P-measure} = \textit{O-measure} = 0.57$ for Y. Thus, only Z is handsomely rewarded, for retrieving both B- and S-relevant documents.

Because P-measure looks for a most highly relevant document in the ranked output and then evaluates by considering all (partially) relevant documents ranked above it, it is possible that P-measure may be more discriminative than O-measure, as we shall see later. Moreover, it is clear that P-measure inherits some properties of O-measure: It is a system-oriented metric, and is free from the assumption underlying NWRR, namely, that ranks are more important than relevance levels.

However, just like R-measure[18], P-measure is "forgiving", in that it can be one for a sub-optimal ranked output. For example, in Fig. 3, supppose that there is a fourth system output, which is a *perfect inverse* of the ideal output. For this system output, $r_p = 3$ and therefore $\textit{P-measure} = BR(3) = (6 + 3)/(6 + 3) = 1$. One could argue that this is counterintuitive. We therefore examine P$^+$(pee-plus)-measure in addition, which does not have this problem:

$$\begin{aligned}&\textit{P}^+\textit{-measure}\\ &\quad = \frac{1}{\textit{count}(r_p)} \sum_{1 \le r \le r_p} \textit{isrel}(r) BR(r)\ . \qquad (7)\end{aligned}$$

For example, for the above perfect inverse output, $BR(1) = (1 + 1)/(3 + 1)$, $BR(2) = (3 +$

$2)/(5+2)$ and $BR(3) = P\text{-}measure = 1$. Thus $P^{+}\text{-}measure = (2/4 + 5/7 + 1)/3 = 0.74$. Note also that $P^{+}\text{-}measure = P\text{-}measure = O\text{-}measure$ holds if there is no relevant document above Rank r_p, i.e., if $r_p = r_1$.

In practice, a document cut-off may be used with P($^{+}$)-measure, since these metrics assume that the user is willing to examine an "unlimited" number of documents. That is, in theory, r_p can be arbitrarily large. However, a small cut-off makes IR evaluation unstable, and requires a larger topic set 1),18). The Appendix examines the actual values of r_p for the NTCIR data.

4. Methods for Assessing Evaluation Metrics

This section describes three methods for comparing IR metrics. The first method is *Kendall's Rank Correlation*, which measures the resemblance of system rankings according to two different IR metrics. The other two methods, namely, the *Swap Method* and the *Bootstrap Sensitivity Method*, examine the discriminative power of individual metrics in a given IR evaluation environment.

4.1 Kendall's Rank Correlation

Following previous work 9),10),22) we examine the resemblance between a pair of metrics using *Kendall's rank correlation* between two system rankings, which computes the minimum number of adjacent swaps to turn one ranking into another. Kendall's rank correlation lies between 1 (identical rankings) and -1 (completely reversed rankings), and its expected value is zero for two rankings that are in fact not correlated with each other. Let n_s denote the number of systems that are to be ranked. Let a_i $(1 \leq i \leq n_s)$ denote the rank of the i-th system as measured by a metric, and let b_i denote the rank of the same system as measured by another. Then, clearly, there are $n_s(n_s-1)/2$ combinations of (a_i, b_i) and (a_j, b_j) $(i \neq j)$ in total. Among these combinations, let *pos* denote the number of combinations such that $a_i < a_j$ and $b_i < b_j$, or $a_i > a_j$ and $b_i > b_j$ (i.e., the number of agreements between two metrics regarding the i-th and the j-th systems). Likewise, let *neg* denote the number of combinations such that $a_i < a_j$ and $b_i > b_j$, or $a_i > a_j$ and $b_i < b_j$ (i.e., the number of disagreements). Then, Kendall's rank correlation (τ) can be expressed as:

$$\tau = \frac{2(pos - neg)}{n_s(n_s-1)} \; . \tag{8}$$

There is a standard significance test available for Kendall's τ: Given the number of systems n_s, it is known that

$$Z_0 = \frac{|\tau|}{((4n_s+10)/(9n_s(n_s-1)))^{\frac{1}{2}}} \tag{9}$$

obeys a normal distribution. Thus, a normal test can be applied. Note that the test statistic Z_0 is proportional to $|\tau|$ given n_s: In terms of a two-tailed test with $n_s = 30$ runs, the rank correlation is significant at $\alpha = 0.01$ if it is over 0.34.

4.2 Voorhees/Buckley Swap Method

The essence of the swap method is to estimate the *swap rate*, which represents the probability of the event that two experiments (each using a different topic set) are contradictory given an overall performance difference. Our version works as follows: First, we create pairs of *bootstrap samples* Q^{*b} and Q'^{*b} $(1 \leq b \leq B = 1000)$ by sampling with replacement from the original topic set Q. Thus, $|Q^{*b}| = |Q'^{*b}| = |Q|$. Let D denote the performance difference between two systems as measured by M based on a topic set; we prepare 21 *performance difference bins*, where the first bin represents performance differences such that $0 \leq D < 0.01$, the second bin represents those such that $0.01 \leq D < 0.02$, and so on, and the last bin represents those such that $0.20 \leq D$ 24). Let $BIN(D)$ denote the mapping from a difference D to one of the 21 bins where it belongs. The algorithm shown in **Fig. 4** calculates a *swap rate* for each bin: It compares systems X and Y using the first topic set Q^{*b} and records the overall performance difference according to Q^{*b}; It then compares the same system pair using the second topic set Q'^{*b}, and if this topic set disagrees with the

```
for each system pair $(X, Y) \in C$
    for $b = 1$ to $B$
        $D^{*b} = M(X, Q^{*b}) - M(Y, Q^{*b})$;
        $D'^{*b} = M(X, Q'^{*b}) - M(Y, Q'^{*b})$;
        $count(BIN(D^{*b}))++$;
        if( $D^{*b} * D'^{*b} > 0$ ) then
            continue
        else
            $swap\_count(BIN(D^{*b}))++$;
for each bin $i$
    $swap\_rate(i) = swap\_count(i)/count(i)$;
```

Fig. 4 Algorithm for computing the swap rates.

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

Fig. 5 Algorithm for creating Bootstrap samples Q^{*b} and $\mathbf{w}^{*b} = (w_1^{*b}, \ldots, w_n^{*b})$ for the Paired Test.

first one, the *swap count* for the aforementioned performance difference is incremented.

We can thus plot swap rates against performance difference bins. By looking for bins whose swap rates do not exceed (say) 5%, we can estimate how much absolute difference is required in order to conclude that System X is better than Y with 95% "confidence": However, it should be noted that the swap method is not directly related to statistical significance tests: the "confidence" in this context is to do with the probability of observing a discrepancy between two experiments, whereas confidence in statistical significance tests is derived from the probability of *Type I error*[12),16)]. The next section describes a less ad hoc method for assessing the discriminative power of IR metrics, which relies on Bootstrap Hypothesis Tests.

4.3 Sakai's Bootstrap Sensitivity Method

This section briefly describes Sakai's Bootstrap Sensitivity Method for assessing the discriminative power of IR metrics[☆].

First, we describe the paired Bootstrap Hypothesis Test, which, unlike traditional significance tests, is free from the normality and symmetry assumptions and yet has high power[4)]. The strength of the Bootstrap lies in its reliance on the computer for directly estimating any data distribution through *resampling* from observed data. Let Q be the set of topics provided in the test collection, and let $|Q| = n$. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ denote the per-topic performance values of systems X and Y as measured by some performance metric M. A standard method for comparing X and Y is to measure the difference between *sample means* $\bar{x} = \sum_i x_i/n$ and $\bar{y} = \sum_i y_i/n$ such as *Mean* Average Precision values. But what we really want to know is whether the *population means* for X and Y (μ_X and μ_Y), computed based on the population P of topics, are any different. Since we can regard $\mathbf{x}$ and $\mathbf{y}$ as *paired* data, we let $\mathbf{z} = (z_1, \ldots, z_n)$ where $z_i = x_i - y_i$, let $\mu = \mu_X - \mu_Y$ and set up the following hypotheses for a two-tailed test:

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

Thus the problem has been reduced to a *one-sample problem*[4)]. As with standard significance tests, we assume that $\mathbf{z}$ is an independent and identically distributed sample drawn from an unknown distribution.

In order to conduct a Hypothesis Test, we need a *test statistic* t and a *null hypothesis distribution.* Here, let us consider a Studentised statistic:

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

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

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

Moreover, let $\mathbf{w} = (w_1, \ldots, w_n)$ where $w_i = z_i - \bar{z}$, in order to create *bootstrap samples* of per-topic performance differences $\mathbf{w}^{*b}$ that obey H_0. **Figure 5** shows the algorithm for obtaining B bootstrap samples of topics (Q^{*b}) and the corresponding values of $\mathbf{w}^{*b}$. (We let $B = 1000$ throughout this paper.) For example, let us assume that we only have five topics Q = (001, 002, 003, 004, 005) and that $\mathbf{w} = (0.2, 0.0, 0.1, 0.4, 0.0)$. Suppose that, for trial b, sampling with replacement from Q yields Q^{*b} = (001, 003, 001, 002, 005). Then, $\mathbf{w}^{*b} = (0.2, 0.1, 0.2, 0.0, 0.0)$.

For each b, let $\bar{w}^{*b}$ and $\bar{\sigma}^{*b}$ denote the mean and the standard deviation of $\mathbf{w}^{*b}$. **Figure 6** shows how to compute the Achieved Significance Level (ASL) using $\mathbf{w}^{*b}$. In essence, we examine how *rare* the observed difference would be under H_0. If $ASL < \alpha$, where typically $\alpha = 0.01$ (very strong evidence against H_0) or

[☆] This paper uses the *Paired-Test* version of the Bootstrap Sensitivity Method. The *Unpaired-Test* version is also available for non-arithmetic-mean summary metrics such as Geometric Mean Average Precision and the "area" measure[12),16)].

Table 1 Statistics of the NTCIR CLIR data.

	$\|Q\|$	R/topic	$R(S)$/topic	$R(A)$/topic	$R(B)$/topic	runs used
NTCIR-3 Chinese	42	78.2	21.0	24.9	32.3	30
NTCIR-3 Japanese	42	60.4	7.9	31.5	21.0	30
NTCIR-5 Chinese	50	61.0	7.0	30.7	23.3	30
NTCIR-5 Japanese	47	89.1	3.2	41.8	44.2	30

```
$count = 0$;
for $b = 1$ to $B$
    $t(\mathbf{w}^{*b}) = \bar{w}^{*b}/(\bar{\sigma}^{*b}/\sqrt{n})$;
    if( $|t(\mathbf{w}^{*b})| \geq |t(\mathbf{z})|$ ) then $count$++;
$ASL = count/B$;
```

Fig. 6 Algorithm for estimating the Achieved Significance Level based on the Paired Test.

```
$DIFF = \phi$;
for each system pair $(X, Y) \in C$
    sort $|t(\mathbf{w}^{*1}_{X,Y})|, \ldots, |t(\mathbf{w}^{*B}_{X,Y})|$;
    if $|t(\mathbf{w}^{*b'}_{X,Y})|$ is the $B\alpha$-th largest value,
    then add $|\bar{w}^{*b'}_{X,Y}|$ to $DIFF$;
$estimated\_diff = \max\{diff \in DIFF\}$
(rounded to two significant figures);
```

Fig. 7 Algorithm for estimating the performance difference required for achieving a given significance level with the Paired Test.

$\alpha = 0.05$ (reasonably strong evidence against H_0), then we reject H_0. That is, we have enough evidence to state that μ_X and μ_Y are probably different.

We now describe Sakai's Bootstrap Sensitivity Method for assessing the discriminative power of IR metrics. Let C denote the set of all possible combinations of two systems. First, perform a Bootstrap Hypothesis Test for every system pair in C and count how many of the pairs satisfy $ASL < \alpha$: The result represents the discriminative power of a given IR metric. We can thus compare different IR metrics while holding the probability of Type I error (α) constant. We thereby obtain the values of $\bar{w}^{*b}$ and $t(\mathbf{w}^{*b})$ for each system pair (X, Y), which we shall denote explicitly by $\bar{w}^{*b}_{X,Y}$ and $t(\mathbf{w}^{*b}_{X,Y})$. Since each $\bar{w}^{*b}_{X,Y}$ is a performance difference computed based on $|\mathbf{w}^{*b}_{X,Y}| = |Q| = n$ topics, we can use the algorithm shown in **Fig. 7** to obtain a natural estimate of the minimum performance difference required for guaranteeing $ASL < \alpha$, given the topic set size n. For example, if $\alpha = 0.05$ is chosen, the algorithm looks for the $B\alpha = 1000 * 0.05 = 50$-th largest value among $|t(\mathbf{w}^{*b}_{X,Y})|$ and takes the corresponding value of $|\bar{w}^{*b}_{X,Y}|$ for each (X, Y). Among the $|C|$ values thus obtained, the algorithm takes the maximum value just to be conservative.

Note that the estimated differences themselves are not necessarily suitable for comparing metrics, since some metrics tend to take small values while others tend to take large values. The discriminative power of metrics should primarily be compared in terms of how many system pairs satisfy $ASL < \alpha$, that is, how many pairs show a statistically significant difference.

5. Experiments

This section describes our experiments for comparing P($^+$)-measure, O-measure, NWRR and RR using Kendall's rank correlation, the Swap Method and the Bootstrap Sensitivity Method. Section 5.1 describes the data sets we used. Section 5.2 reports on our Kendall's rank correlation results to discuss how the metrics resemble one another. Sections 5.3 and 5.4 report on our Swap Method and Bootstrap Sensitivity Method results to discuss which are the most useful metrics from the viewpoint of discriminative power. Finally, Section 5.5 discusses the effect of changing the gain values for P$^+$-measure and O-measure on Kendall's rank correlation and Bootstrap Sensitivity, since we find these metrics to be the most discriminative.

As mentioned earlier, our experiments include AveP and Q-measure, which are metrics for the task of finding as many relevant documents as possible, not for the task of finding one relevant document, just for comparison. The AveP and Q-measure results have been copied from Refs. 12), 16), and are not part of this paper's contribution.

5.1 Data

Our experiments use four different data sets (i.e., test collections and submitted runs) from the NTCIR CLIR track series[8]. **Table 1** provides some statistics of the data. From each data set, only the top 30 runs as measured by Mean *relaxed* AveP (i.e., AveP that treats S-, A- and B-relevant documents just

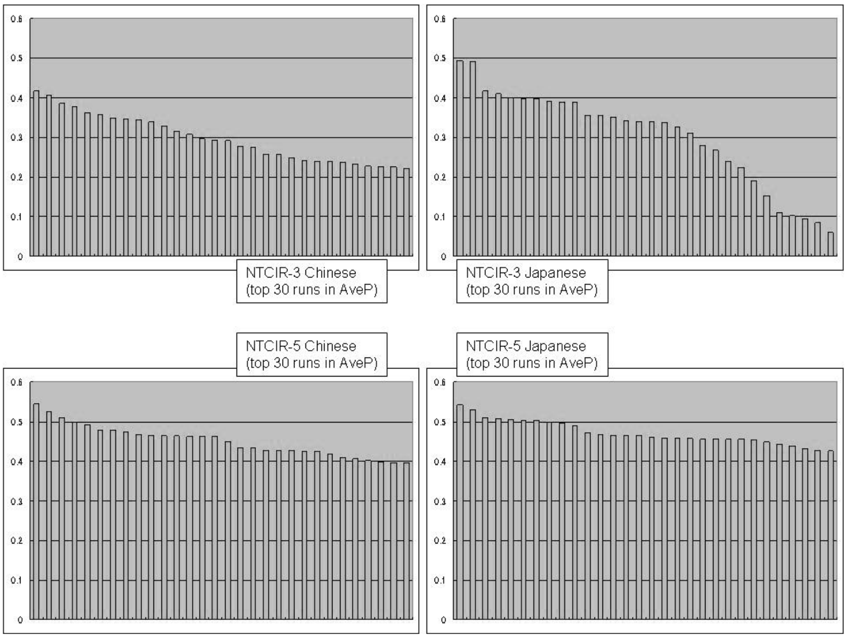


Fig. 8 Distribution of Mean AveP values for the runs used in this study.

Table 2 Kendall's rank correlations based on the top 30 runs from each data set.

NTCIR-3	metric	(b)	(c)	(d)	(e)	(f)	(g)
Chinese	(a)RR	**.8575**	.7977	.7425	.7747	.5264	.5494
	(b)NWRR	-	**.9126**	**.8575**	**.8989**	.5494	.5632
	(c)O-measure	-	-	**.8621**	**.9310**	.5264	.5402
	(d)P-measure	-	-	-	**.9126**	.5540	.5678
	(e)P$^+$-measure	-	-	-	-	.5126	.5356
	(f)AveP	-	-	-	-	-	**.9678**
	(g)Q-measure	-	-	-	-	-	-
NTCIR-3	metric	(b)	(c)	(d)	(e)	(f)	(g)
Japanese	(a)RR	**.8759**	**.8207**	**.8253**	.7977	.7701	.7701
	(b)NWRR	-	**.9356**	**.9126**	**.9126**	.7011	.7287
	(c)O-measure	-	-	**.9218**	**.9310**	.6920	.7011
	(d)P-measure	-	-	-	**.9540**	.7333	.7517
	(e)P$^+$-measure	-	-	-	-	.7149	.7425
	(f)AveP	-	-	-	-	-	**.9540**
	(g)Q-measure	-	-	-	-	-	-
NTCIR-5	metric	(b)	(c)	(d)	(e)	(f)	(g)
Chinese	(a)RR	**.8805**	**.8391**	.7793	.7977	.5172	.5540
	(b)NWRR	-	**.9034**	**.8253**	**.8529**	.4713	.5080
	(c)O-measure	-	-	**.8391**	**.8851**	.4851	.5126
	(d)P-measure	-	-	-	**.9080**	.5632	.6000
	(e)P$^+$-measure	-	-	-	-	.5356	.5724
	(f)AveP	-	-	-	-	-	**.9172**
	(g)Q-measure	-	-	-	-	-	-
NTCIR-5	metric	(b)	(c)	(d)	(e)	(f)	(g)
Japanese	(a)RR	.7839	.6506	.6138	.7057	.4667	.4529
	(b)NWRR	-	**.8391**	.7103	**.8207**	.4069	.4115
	(c)O-measure	-	-	.6874	.7885	.3563	.3793
	(d)P-measure	-	-	-	**.8437**	.5310	.5632
	(e)P$^+$-measure	-	-	-	-	.4667	.4989
	(f)AveP	-	-	-	-	-	**.8851**
	(g)Q-measure	-	-	-	-	-	-

as “relevant”) were used in our experiments, since “near-zero” runs are unlikely to be useful for discussing the discriminative power of metrics. Thus, for each data set, we have a set of $30 * 29/2 = 435$ system combinations, which we shall denote by C. The distribution of Mean AveP values for each data set is shown in **Fig. 8**: It can be observed, for example, that the top 30 NTCIR-5 Japanese runs are quite similar to one another in terms of Mean AveP.

5.2 Rank Correlation Results

Table 2 shows the Kendall’s rank correlation values between two IR metrics when the aforementioned 30 runs from each data set are ranked. For example, the table shows that the rank correlation between RR and NWRR is .8575 for the NTCIR-3 Chinese data. All the correlation values exceed 0.34, and therefore are statistically significant (See Section 4.1). For convenience, values higher than 0.8 are shown in bold, although the choice of this threshold is arbitrary.

In terms of rank correlations, our main observations are:

- P($^+$)-measure, O-measure and NWRR are relatively highly correlated with one another, reflecting the fact that they were all designed for the task of finding one highly relevant document. In particular, O-measure and NWRR are consistently highly correlated with each other, reflecting the fact that they both rely on r_1, the rank of the first relevant document found in the ranked output. On the other hand, the rank correlations between O-measure and P($^+$)-measure are below 0.8 for the NTCIR-5 Japanese data, reflecting the fact that P($^+$)-measure rely on r_p, the preferred rank (See Section 3.3). This example suggests that different user models sometimes lead to different system rankings.
- P($^+$)-measure, O-measure and NWRR are *not* very highly correlated with RR. For example, the correlation between RR and the other four metrics lie between .6138 and .7839 for the NTCIR-5 Japanese data. These results demonstrate that the task of finding one *highly* relevant document is *not* the same as that of finding *any* one relevant document. This generalises a finding by Sakai[15] who compared the rank correlation between O-measure and RR only.
- P($^+$)-measure, O-measure, NWRR and RR are *not* very highly correlated with AveP and Q-measure. For example, note that the rank correlation between O-measure and AveP is as low as .3563 for the NTCIR-5 Japanese data, which is statistically barely significant. These results demonstrate that the task of finding *one* relevant document is *not* the same as that of finding as many relevant documents as possible. This also generalises one of Sakai’s findings[15], as well as earlier findings which only considered binary-relevance metrics such as AveP and RR[3].

Table 3 Swap Method results (swap rate ≤ 5%; NTCIR-3 and NTCIR-5 CLIR Chinese and Japanese data).

(i) metric	(ii) abs. diff.	(iii) max.	(ii)/(iii)	%pairs satisfying (ii)
(a) NTCIR-3 Chinese (42 topics)				
Q-measure	0.07	.5374	13%	**43%**
AveP	0.08	.5295	15%	**40%**
P-measure	0.15	.8636	17%	**31%**
P$^+$-measure	0.15	.8632	17%	**30%**
O-measure	0.17	.8674	20%	**24%**
NWRR	0.18	.8633	21%	**22%**
RR	0.19	.9524	20%	**20%**
(b) NTCIR-3 Japanese (42 topics)				
Q-measure	0.07	.6433	11%	**67%**
AveP	0.07	.6449	11%	**66%**
P$^+$-measure	0.13	.8703	15%	**59%**
P-measure	0.13	.8759	15%	**59%**
O-measure	0.14	.8690	16%	**56%**
NWRR	0.16	.8757	18%	**51%**
RR	0.17	.9524	18%	**47%**
(c) NTCIR-5 Chinese (50 topics)				
Q-measure	0.07	.6757	10%	**26%**
AveP	0.07	.6480	11%	**26%**
P$^+$-measure	0.13	.9012	14%	**17%**
O-measure	0.14	.8901	16%	**17%**
P-measure	0.13	.9220	14%	**16%**
NWRR	0.15	.9075	17%	**15%**
RR	0.15	.9900	15%	**13%**
(d) NTCIR-5 Japanese (47 topics)				
Q-measure	0.07	.6652	11%	**16%**
AveP	0.08	.6438	12%	**11%**
P-measure	0.15	.8925	17%	**5.6%**
P$^+$-measure	0.15	.8874	17%	**5.5%**
O-measure	0.16	.8819	18%	**5.3%**
NWRR	0.17	.8899	19%	**5.1%**
RR	0.18	1.0000	18%	**4.5%**

5.3 Swap Results

Table 3 summarises the results of our Swap Method experiments using the aforementioned four data sets, each with 30 runs. For example, Table 3 (a) shows that, with the 42 topics and the 30 runs of the NTCIR-3 Chinese data, P-measure can guarantee that the swap rate is no greater than 5% if the overall absolute difference between two systems is at least

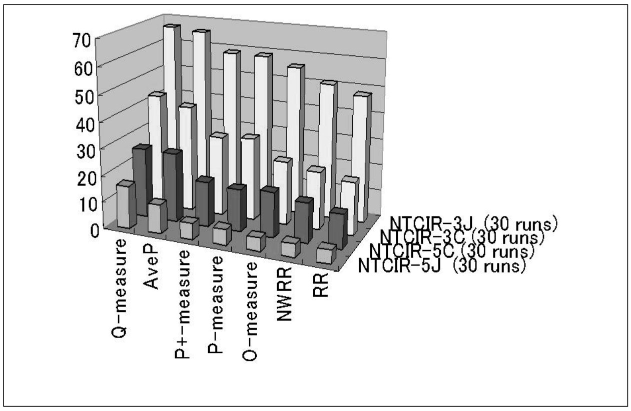


Fig. 9 Summary of the Swap Method results.

0.15. Since the maximum overall absolute difference observed among the 2,000 values (1,000 trials, each with two topic sets Q^{*b} and Q'^{*b}) is .8636, this translates to a relative difference of 17%. Moreover, the percentage of comparisons that actually satisfied this absolute difference requirement among all comparisons (435 run pairs times 1,000 trials) is 31%. **Fig. 9** visualises this last column, which represents the discriminative power of each metric, showing clearly that the results are consistent across our four data sets.

Our main observations based on the Swap Method are:

- P(+)-measure and O-measure are consistently more discriminative than NWRR, which in turn is consistently more discriminative than RR. Moreover, with the exception of the NTCIR-5 Chinese results, P$^+$-measure and P-measure are more discriminative than O-measure. This difference arises from the fact that P$^+$-measure and P-measure consider all relevant documents ranked above r_p, in contrast to O-measure which only considers the first retrieved relevant document. The difference in discriminative power between O-measure and NWRR arises from the fact that O-measure compares the system output with an ideal ranked output. Finally, the difference in discriminative power between NWRR and RR arises from the use of graded relevance. In short, P(+)-measure and O-measure are the best metrics in terms of discriminative power.
- Even P$^+$-measure and P-measure are not as discriminative as Q-measure and AveP. This is because the metrics that rely on one or a small number of relevant documents are inherently less stable than those that rely on all relevant documents, since they are based on a smaller number of observations. Therefore, one should prepare a larger topic set if metrics such as P(+)-measure and O-measure are to be used instead of more discriminative metrics such as Q-measure and AveP.

5.4 Bootstrap Sensitivity Results

Table 4 summarises the results of our Bootstrap Sensitivity experiments. It shows, for example, that if P-measure is used for assessing 30 systems that were submitted to the NTCIR-3 Chinese document retrieval subtask, it can detect a statistically significant difference at $\alpha = 0.05$ for 39% of the system pairs; The estimated overall performance difference required for detecting a statistical significance is 0.18. **Figure 10**, which visualises the sensitivity column of this table, clearly shows that the results are consistent across the four data sets, and that they agree very well with the Swap Method results (Compare Fig. 9 and Fig. 10). Thus, our Bootstrap Sensitivity results show that:

- P(+)-measure and O-measure are consistently more discriminative than NWRR, which in turn is consistently more discriminative than RR.
- But even P(+)-measure and O-measure are not as discriminative as the best metrics for the task of finding as many relevant documents as possible, namely, Q-measure and AveP.

It can be observed that the *absolute* discriminative power values depend heavily on the set of runs: For example, P-measure can detect a significant difference for 62% of the system pairs

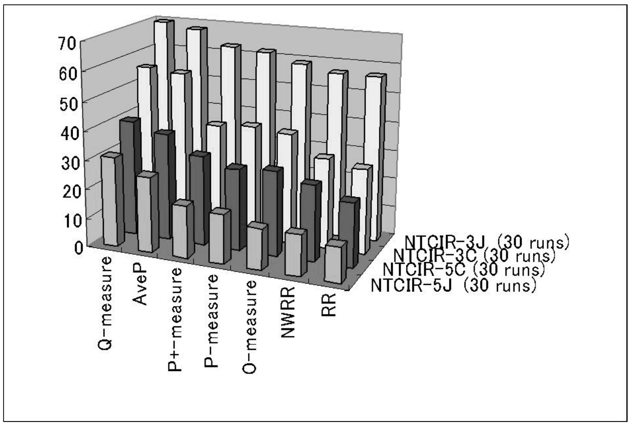


Fig. 10 Summary of the Bootstrap Sensitivity results.

Table 4 Bootstrap Sensitivity results ($\alpha = 0.05$).

metric	sensitivity (ASL $< \alpha$)	estimated diff.
(a) NTCIR-3 Chinese (42 topics)		
Q-measure	242/435=**56%**	0.10
AveP	240/435=**55%**	0.11
P-measure	170/435=**39%**	0.18
P^{+}-measure	167/435=**38%**	0.18
O-measure	165/435=**38%**	0.19
NWRR	136/435=**31%**	0.20
RR	126/435=**29%**	0.22
(b) NTCIR-3 Japanese (42 topics)		
Q-measure	305/435=**70%**	0.13
AveP	296/435=**68%**	0.11
P^{+}-measure	272/435=**63%**	0.18
P-measure	271/435=**62%**	0.20
O-measure	255/435=**59%**	0.22
NWRR	247/435=**57%**	0.19
RR	246/435=**57%**	0.23
(c) NTCIR-5 Chinese (50 topics)		
Q-measure	174/435=**40%**	0.11
AveP	159/435=**37%**	0.11
P^{+}-measure	134/435=**31%**	0.15
O-measure	125/435=**29%**	0.15
P-measure	123/435=**28%**	0.16
NWRR	114/435=**26%**	0.16
RR	94/435=**22%**	0.16
(d) NTCIR-5 Japanese (47 topics)		
Q-measure	136/435=**31%**	0.09
AveP	113/435=**26%**	0.10
P^{+}-measure	77/435=**18%**	0.14
P-measure	73/435=**17%**	0.15
O-measure	63/435=**14%**	0.16
NWRR	63/435=**14%**	0.16
RR	54/435=**12%**	0.17

for the NTCIR-3 Japanese data, but for only 17% for the NTCIR-5 Japanese data. That is, the NTCIR-5 Japanese runs are much harder to distinguish from one another because a larger number of teams performed equally well at NTCIR-5 than at NTCIR-3. (For both NTCIR-3 and NTCIR-5, the top 30 Japanese runs we used came from 10 different teams; but the two sets of teams are quite different.) However, it can be observed that the ranking of metrics according to discriminative power is quite consistent across data sets.

We also note that the overall absolute differences required according to the Bootstrap Sensitivity Method are generally higher than those according to the Swap Method. That is, the Bootstrap Sensitivity Method is more demanding [12),16)]. For example, while Table 3 (c) suggests that, if we have 50 topics, the overall difference required in terms of P($^{+}$)-measure or O-measure would be around 0.13-0.14 in order to ensure that the swap rate does not exceed 5%, Table 4 (c) suggests that, under the same circumstance, the overall difference required would be around 0.15-0.16 in order to detect a significant difference at $\alpha = 0.05$. This arises from the different definitions of “confidence”: the Swap Method concerns the probability of observing consistent results across two experiments given an overall performance difference; the Bootstrap Sensitivity Method concerns the probability of *correctly* concluding that two systems are equivalent, i.e., $1 - \alpha$, in a statistical significance test.

5.5 Changing Gain Values

We finally focus on P($^{+}$)-measure and O-measure, which we have shown to be the three most discriminative metrics for the task of finding one relevant document, and study the effect of changing *gain values* (See Section 3) on both rank correlation and the Bootstrap Sensitivity.

Table 5 shows the Kendall’s rank correlation values between the “default” metric and those with different gain values. Recall that the default gain values we use are $gain(S) =$

Table 5 Kendall's rank correlations: default gain values (3:2:1) versus others.

	30:20:10	0.3:0.2:0.1	1:1:1	10:5:1
(a) NTCIR-3 Chinese				
P^+-measure	**.9586**	**.8805**	**.8621**	**.9126**
P-measure	**.8989**	**.8713**	**.8667**	**.9126**
O-measure	**.8115**	**.8943**	**.8023**	**.8621**
(b) NTCIR-3 Japanese				
P^+-measure	**.9724**	**.9402**	**.8943**	**.9448**
P-measure	**.9862**	**.9632**	**.9402**	**.9632**
O-measure	**.9678**	**.8943**	**.8207**	**.9172**
(c) NTCIR-5 Chinese				
P^+-measure	**.9632**	**.9034**	**.8713**	**.9172**
P-measure	**.9448**	**.9264**	**.8759**	**.9218**
O-measure	**.9862**	**.8897**	**.8345**	**.9448**
(d) NTCIR-5 Japanese				
P^+-measure	**.8943**	**.8023**	.7655	.7517
P-measure	**.8713**	**.9126**	**.8575**	**.8621**
O-measure	**.8529**	.7471	.6506	.7195

$3, gain(A) = 2, gain(B) = 1$. The column labelled with "30:20:10" represents the gain value assignment $gain(S) = 30, gain(A) = 20, gain(B) = 10$, and so on[☆]. Using small gain values such as "0.3:0.2:0.1" implies high penalty on late arrival of relevant documents [17]; "1:1:1" represents binary relevance; and "10:5:1" represents a strong emphasis of relevance levels. It can be observed that the system rankings according to P($^+$)-measure and O-measure are relatively robust to the choice of gain values, although O-measure may be less robust than P($^+$)-measure.

In practice, we encourage researchers to try out several choices of gain values, since there is no theoretical justification for pre-setting the gain values. It is always useful to examine IR results from several different angles: For example, one could discuss the trends that are consistent across (several variations of) P($^+$)-measure and O-measure; and phenomena that are observed with only one of the metrics.

Table 6 The effect of changing gain values on the Bootstrap Sensitivity ($\alpha = 0.05$). The default results have been copied from Table 4.

metric	sensitivity	estimated
	(ASL $< \alpha$)	diff.
(a) NTCIR-5 Chinese (50 topics)		
P^+3:2:1 (default)	134/435=**31%**	0.15
P^+0.3:0.2:0.1	132/435=**30%**	0.15
P^+30:20:10	128/435=**29%**	0.16
P^+10:5:1	125/435=**29%**	0.14
P^+1:1:1	124/435=**29%**	0.15
P10:5:1	125/435=**29%**	0.14
P3:2:1 (default)	123/435=**28%**	0.16
P30:20:10	121/435=**28%**	0.15
P0.3:0.2:0.1	120/435=**28%**	0.14
P1:1:1	113/435=**26%**	0.15
O3:2:1 (default)	125/435=**29%**	0.15
O30:20:10	123/435=**28%**	0.15
O10:5:1	118/435=**27%**	0.16
O0.3:0.2:0.1	107/435=**25%**	0.17
O1:1:1	94/435=**22%**	0.16
(b) NTCIR-5 Japanese (47 topics)		
P^+3:2:1 (default)	77/435=**18%**	0.14
P^+30:20:10	73/435=**17%**	0.15
P^+0.3:0.2:0.1	69/435=**16%**	0.17
P^+1:1:1	67/435=**15%**	0.14
P^+10:5:1	67/435=**15%**	0.15
P10:5:1	85/435=**20%**	0.14
P30:20:10	81/435=**19%**	0.15
P3:2:1 (default)	73/435=**17%**	0.15
P0.3:0.2:0.1	64/435=**15%**	0.16
P1:1:1	59/435=**14%**	0.15
O30:20:10	65/435=**15%**	0.16
O3:2:1 (default)	63/435=**14%**	0.16
O10:5:1	62/435=**14%**	0.17
O0.3:0.2:0.1	59/435=**14%**	0.16
O1:1:1	54/435=**12%**	0.18

Table 6 summarises the effect of changing gain values on the discriminative power, based on the Bootstrap Sensitivity Method (with the NTCIR-5 data only). For example, "P10:5:1" represents P-measure with $gain(S) = 10, gain(A) = 5, gain(B) = 1$. (Hence the default gain value results, labelled with "3:2:1", have been copied from Table 4.) Recall that: (a) Using small gain values makes O-measure resemble RR (See Eq. (3)); (b) Given that the gain values are all one, *O-measure* $=$ *RR* holds iff $r_1 \leq R$; (c) *P*($^+$)*-measure* $=$ *O-measure* holds if there is no relevant document above Rank r_p. Thus, we can expect "flat" and small gain values to reduce the discriminative power of these metrics. Indeed, Table 6 shows that the gain value assignments $gain(S) = 1, gain(A) = 1, gain(B) = 1$ and $gain(S) = 0.3, gain(A) = 0.2, gain(B) = 0.1$ tend to hurt discriminative power, especially the former. On the other hand, using large gain values (which implies less penalty on late arrival of relevant documents) and using "steeper" gain values (which emphasises the relevance levels) generally do not seem to have a substantial impact on discriminative power. In summary, P($^+$)-measure and O-measure are fairly robust to the choice of gain values as long as graded relevance is properly utilised.

[☆] It should be noted that the Blended Ratio (Eq. (2)) is affected not only by the gain value *ratio* but also by the *absolute* gain values. To express this feature more explicitly, the Blended Ratio can be rewritten as $BR(r) = (\beta cg(r) + count(r))/(\beta cg_I(r) + r)$ [17].

6. Conclusions and Future Work

This paper compared the properties of five evaluation metrics that are applicable to the task of finding one highly relevant document, in terms of the underlying assumptions, how the system rankings produced resemble each other, and discriminative power. Our extensitve experiments using four different data sets from NTCIR showed that:

- P($^+$)-measure, O-measure and NWRR are relatively highly correlated to one another, since they were all designed for the task of finding one highly relevant document. They are not necessarily highly correlated with RR, which is a metric for the task of finding *any* one relevant document. Moreover, these metrics for the task of finding one (highly) relevant document are not necessarily highly correlated with those for the task of finding as many relevant documents as possible, namely, Q-measure and AveP.
- P($^+$)-measure and O-measure are more discriminative than NWRR, which in turn is more discriminative than RR. Moreover, P$^+$-measure and P-measure tend to be more discriminative than O-measure, as they examine all relevant documents ranked above the preferred rank r_p. However, even P$^+$-measure and P-measure are not as discriminative as Q-measure and AveP, as they generally examine only the very top of a ranked output.
- P($^+$)-measure and O-measure are fairly robust to the choice of gain values in terms of both rank correlation and discriminative power.

Based on our discussions on the underlying assumptions of each metric as well as experimental evidence, we conclude that P($^+$)-measure and O-measure are the most flexible and useful metrics for the task of finding one highly relevant document. Just like RR and NWRR, O-measure assumes that the user stops scanning the ranked list as soon as he finds *any* one relevant document, even if it is only partially relevant. Whereas, P$^+$-measure and P-measure mimic the user who continues to scan the ranked list until he finds a "satisfactory" document. Such user behaviours probably depend on the IR interface (whether a ranked list is shown at all to the user; and if it is, whether the list is informative enough for the user to spot highly relevant documents, and so on) and the kind of information need. Whether recent criticisms of Average Precision from the viewpoint of user satisfaction[21] applies to different IR environments and different metrics such as P($^+$)-measure and O-measure is an open question, and an important one too.

References

1) Buckley, C. and Voorhees, E.M.: Evaluating Evaluation Measure Stability, *Proc. ACM SIGIR 2000*, pp.33–40 (2000).

2) Buckley, C. and Voorhees, E.M.: Retrieval Evaluation with Incomplete Information, *Proc. ACM SIGIR 2004*, pp.25–32 (2004).

3) Buckley, C. and Voorhees, E.M.: Retrieval System Evaluation, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, pp.53–75 (2005).

4) Efron, B. and Tibshirani, R.: *An Introduction to the Bootstrap*, Chapman & Hall/CRC (1993).

5) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuruyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop, Technical Report NII-2003-002E, National Institute of Informatics (2003).

6) Hawking, D. and Craswell, N.: The Very Large Collection and Web Tracks, *TREC: Experiment and Evaluation in Information Retrieval*, MIT Press, pp.199–231 (2005).

7) Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (2002).

8) Kando, N.: Overview of the Fifth NTCIR Workshop, *Proc. NTCIR-5* (2005).

9) Kekäläinen, J.: Binary and Graded Relevance in IR evaluations — Comparison of the Effects on Ranking of IR Systems, *Information Processing and Management*, Vol.41, pp.1019–1033 (2005).

10) Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance, *Proc. AIRS 2004, Lecture Notes in Computer Science 3411*, pp.251–262 (2004).

11) Sakai, T.: The Effect of Topic Sampling on Sensitivity Comparisons of Information Retrieval Metrics, *Proc. NTCIR-5* (2005).

12) Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *Proc. ACM SIGIR 2006*, pp.525–532 (2006).

13) Sakai, T.: A Further Note on Evaluation Metrics for the Task of Finding One Highly Relevant Document, *Information Processing Society of Japan SIG Technical Reports 2006-FI-82/2006-DD-54*, pp.69–76 (2006).

14) Sakai, T.: Give Me Just One Highly Rele-

Table 7 The maximum preferred rank among the 30 runs and the number of relevant documents above the preferred rank (NTCIR-3 data).

NTCIR-3 Chinese					NTCIR-3 Japanese				
Topic ID	$maxr_p$	S	A	B	Topic ID	$maxr_p$	S	A	B
001	63	1(14)	1(12)	0(16)	002	506	0(0)	1(5)	1(4)
002	285	1(64)	2(73)	1(103)	004	660	0(0)	1(6)	4(48)
003	84	1(6)	0(2)	2(18)	005	870	0(0)	1(17)	9(64)
004	67	1(10)	4(44)	12(81)	007	617	0(0)	1(4)	0(2)
005	749	1(51)	0(63)	4(92)	008	183	0(5)	1(5)	2(12)
006	25	1(26)	0(3)	0(3)	010	86	1(5)	1(3)	3(5)
007	40	1(53)	0(26)	1(18)	012	349	1(9)	16(68)	44(189)
008	275	1(3)	1(1)	0(9)	014	868	1(14)	0(15)	0(2)
009	80	1(54)	0(14)	0(14)	015	9	1(11)	3(14)	1(11)
010	46	1(8)	0(2)	0(19)	016	69	1(16)	0(5)	0(6)
011	64	1(8)	0(0)	0(3)	017	499	0(7)	1(7)	0(14)
012	101	1(95)	0(21)	0(49)	018	681	1(27)	22(111)	1(12)
013	198	0(75)	1(75)	0(96)	019	218	1(23)	47(155)	12(30)
014	26	1(33)	9(52)	0(16)	020	226	1(7)	22(118)	23(80)
015	6	1(9)	1(4)	4(17)	021	11	0(0)	1(91)	0(25)
017	361	1(10)	0(1)	0(2)	022	137	0(0)	1(18)	1(10)
018	777	0(9)	1(27)	2(29)	023	393	0(0)	1(296)	0(49)
019	5	1(61)	1(77)	2(51)	024	244	1(2)	3(12)	4(13)
020	913	1(10)	0(6)	0(36)	025	64	0(0)	1(35)	4(20)
021	168	0(10)	0(8)	1(13)	026	365	0(0)	1(32)	1(14)
022	64	1(3)	7(10)	15(18)	027	602	1(7)	1(17)	0(36)
023	28	1(8)	4(20)	18(67)	028	856	1(6)	0(24)	0(34)
024	64	1(7)	1(2)	0(7)	029	75	1(3)	3(12)	4(31)
025	16	1(3)	1(2)	2(2)	030	226	1(2)	0(5)	1(20)
027	24	1(9)	1(17)	4(43)	031	791	1(1)	3(12)	0(3)
032	568	1(10)	1(6)	0(9)	032	516	1(3)	1(12)	0(9)
033	234	1(10)	3(35)	2(11)	033	666	0(5)	0(10)	1(6)
034	655	0(4)	0(8)	1(7)	034	977	0(7)	1(19)	3(18)
035	217	1(6)	2(22)	0(8)	035	38	1(46)	1(9)	1(12)
036	91	0(33)	1(74)	0(29)	036	162	0(90)	0(48)	1(13)
037	75	1(15)	0(2)	0(3)	037	6	1(28)	1(7)	1(6)
038	250	0(1)	1(2)	3(3)	038	187	0(0)	1(11)	0(5)
039	632	1(17)	6(30)	8(43)	039	898	1(1)	2(18)	3(18)
040	253	1(9)	1(3)	1(5)	040	602	1(1)	2(2)	6(6)
042	219	1(6)	6(19)	49(111)	041	867	1(2)	4(12)	2(10)
043	44	1(75)	2(66)	2(108)	042	595	0(1)	1(19)	0(7)
045	747	1(2)	3(6)	8(18)	043	221	0(0)	1(13)	0(12)
046	3	1(22)	0(147)	1(73)	044	960	0(0)	1(6)	0(0)
047	555	0(0)	1(4)	5(9)	045	731	0(1)	1(17)	0(8)
048	580	1(10)	7(35)	16(72)	046	303	0(0)	1(25)	2(8)
049	802	1(11)	2(11)	3(10)	047	980	0(0)	1(3)	6(9)
050	239	1(12)	1(14)	2(15)	050	366	0(0)	1(6)	0(3)

vant Document: P-measure, *Proc. ACM SIGIR 2006*, pp.695–696 (2006).

15) Sakai, T.: On the Task of Finding One Highly Relevant Document with High Precision, *IPSJ Trans. Databases*, Vol.47, No.SIG 4 (TOD29), pp.13–27 (2006).

16) Sakai, T.: Evaluating Information Retrieval Metrics based on Bootstrap Hypothesis Tests, *IPSJ Trans. Databases*, Vol.48, No.SIG 9 (TOD35) (2007).

17) Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proc. First International Workshop on Evaluating Information Access* (*EVIA 2007*), pp.32–43 (2007).

18) Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, Vol.43, No.2, pp.531–548 (2007).

19) Sanderson, M. and Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *Proc. ACM SIGIR 2005*, pp.162–169 (2005).

20) Soboroff, I.: On Evaluating Web Search with Very Few Relevant Documents, *Proc. ACM SIGIR 2004*, pp.530–531 (2004).

21) Turpin, A. and Scholer, F.: User Performance versus Precision Measures for Simple Search Tasks, *Proc. ACM SIGIR 2006*, pp.11–18 (2006).

Table 8 The maximum preferred rank among the 30 runs and the number of relevant documents above the preferred rank (NTCIR-5 data: Part I).

NTCIR-5 Chinese					NTCIR-5 Japanese				
Topic ID	$maxr_p$	S	A	B	Topic ID	$maxr_p$	S	A	B
001	19	1(3)	3(36)	9(17)	001	78	1(2)	5(6)	3(3)
002	20	1(7)	4(16)	7(15)	002	27	1(9)	2(36)	2(194)
003	9	1(8)	4(68)	1(30)	003	14	1(15)	1(46)	5(196)
004	4	1(6)	0(15)	1(5)	004	19	1(2)	10(11)	0(1)
005	12	1(8)	3(21)	1(7)	005	269	1(12)	3(112)	2(32)
006	5	1(6)	3(25)	1(16)	006	29	1(8)	1(18)	0(19)
007	6	1(2)	2(23)	1(14)	007	62	1(2)	0(4)	0(1)
008	13	1(9)	4(18)	0(24)	008	20	1(5)	1(20)	0(19)
009	496	0(9)	1(108)	0(41)	009	501	1(8)	1(90)	0(18)
010	14	1(12)	4(116)	1(15)	010	33	1(8)	12(130)	5(60)
011	143	1(3)	10(47)	14(23)	011	210	1(3)	1(7)	7(25)
012	105	1(3)	31(61)	15(51)	012	8	0(0)	1(19)	0(10)
013	7	0(0)	1(15)	2(45)	013	11	1(3)	7(44)	3(57)
014	22	1(3)	3(35)	15(81)	014	27	1(1)	21(88)	2(18)
015	3	1(20)	1(19)	1(225)	015	3	0(0)	1(78)	2(28)
016	2	1(15)	1(40)	0(51)	016	51	1(2)	22(33)	6(13)
017	3	1(15)	1(63)	1(70)	017	1	0(0)	1(90)	0(18)
018	2	1(24)	0(19)	0(41)	018	852	1(1)	0(23)	0(16)
019	14	1(2)	1(10)	4(23)	019	6	0(0)	1(21)	3(22)
020	758	0(0)	1(6)	0(0)	020	156	1(1)	16(22)	0(2)
021	1	1(3)	0(7)	0(8)	022	8	0(0)	1(51)	5(246)
022	21	1(2)	7(26)	4(45)	024	3	0(0)	1(16)	2(14)
023	125	1(4)	6(14)	0(6)	025	254	1(4)	1(6)	0(10)
024	4	1(16)	1(12)	0(16)	026	772	1(1)	7(8)	38(71)
025	4	0(0)	1(140)	0(11)	027	731	1(2)	19(57)	7(15)

22) Voorhees, E.M.: Evaluation by Highly Relevant Documents, *Proc. ACM SIGIR 2001*, pp.74–82 (2001).

23) Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track, *Proc. TREC 2004* (2005).

24) Voorhees, E.M. and Buckley, C.: The Effect of Topic Set Size on Retrieval Experiment Error, *Proc. ACM SIGIR 2002*, pp.316–323 (2002).

Appendix

As we have mentioned in Section 3.3, P-measure and P$^+$-measure assume that the user is willing to examine an "unlimited" number of documents. This appendix therefore examines the actual range of the preferred rank r_p using the NTCIR-3 and 5 Chinese and Japanese data. Recall that r_p is the rank of the first $\mathcal{L}_p$-relevant document in the system output, where $\mathcal{L}_p$ is the highest relevance level observed within the output.

Table 7 show the results for the NTCIR-3 Chinese and Japanese data. For each topic, we first selected a run that has the *largest* value of r_p among all runs that contain at least one relevant document, which we denote as $maxr_p$. Then, for this particular run, we counted the number of S-, A- and B-relevant documents at or above this rank. The total number of S-, A- and B-relevant documents are shown in addition. For example, for NTCIR-3 Chinese Topic 020, $maxr_p$ is as large as 913, and the document at Rank 913 in the ranked output examined is S-relevant. There are no A- and B-relevant documents above this rank, so $r_p = r_1$ holds in this case. Since there are 10 S-relevant, 6 A-relevant and 36 B-relevant documents for this topic, the ideal cumulative gain is $cg_I(r) = 3 * 10 + 2 * 6 + 1 * 36 = 48$ for $r \geq 52 (= 10 + 6 + 36)$. Thus $P^+\text{-}measure = P\text{-}measure = O\text{-}measure = BR(913) = (3 + 1)/(48 + 913) = 0.004$.

Table 8 and **Table 9** provides similar statistics for the NTCIR-5 Chinese and Japanese data. For example, for NTCIR-5 Japanese Topic 034, $maxr_p$ is as large as 424, and the document at Rank 424 in the ranked output examined is S-relevant. Unlike the example mentioned above, this ranked output has as many as 137 A-relevant documents and 29 B-relevant documents above this rank. Hence $cg(424) = 3 * 1 + 2 * 137 + 1 * 29 = 306$, and $count(424) = 1 + 137 + 29 = 167$. Whereas, since there are 5 S-relevant, 288 A-relevant and 61 B-relevant documents for this topic, $cg_I(r) = 3 * 5 + 2 * 288 + 1 * 61 = 652$ for $r \geq 354 (= 5 +$

Table 9 The maximum preferred rank among the 30 runs and the number of relevant documents above the preferred rank (NTCIR-5 data: Part II).

NTCIR-5 Chinese					NTCIR-5 Japanese				
Topic ID	$maxr_p$	S	A	B	Topic ID	$maxr_p$	S	A	B
026	880	1(7)	4(27)	5(32)	028	7	1(2)	0(5)	0(5)
027	779	0(0)	1(80)	0(14)	029	217	1(7)	0(22)	1(44)
028	16	0(0)	1(12)	3(34)	030	4	1(4)	0(25)	3(30)
029	20	1(22)	4(165)	0(9)	031	375	1(1)	1(4)	4(15)
030	30	1(4)	23(50)	3(6)	032	633	1(1)	10(13)	6(6)
031	701	0(0)	1(7)	0(47)	033	94	1(1)	2(31)	11(86)
032	202	1(2)	0(9)	0(16)	034	424	1(5)	137(288)	29(61)
033	328	1(9)	5(9)	1(8)	035	9	1(5)	2(3)	2(5)
034	72	1(18)	1(33)	2(3)	036	2	0(0)	1(23)	0(0)
035	167	1(5)	2(8)	0(3)	037	16	1(7)	1(10)	1(24)
036	853	1(4)	14(21)	0(9)	038	200	1(8)	25(43)	20(43)
037	13	1(3)	0(3)	0(2)	040	93	1(2)	4(6)	0(0)
038	50	1(1)	0(2)	0(4)	041	203	1(2)	2(11)	41(168)
039	93	1(3)	2(10)	1(1)	042	70	1(1)	1(8)	56(157)
040	59	1(5)	3(10)	4(5)	043	45	1(1)	3(4)	17(34)
041	3	1(3)	0(3)	0(1)	044	47	0(0)	1(23)	2(21)
042	81	1(7)	6(24)	0(2)	045	3	0(0)	1(7)	2(4)
043	346	1(2)	0(1)	0(2)	046	10	1(3)	7(26)	1(34)
044	54	1(5)	0(2)	3(5)	047	8	1(4)	2(11)	1(20)
045	930	1(3)	0(0)	1(1)	048	78	1(3)	37(116)	11(77)
046	4	1(4)	0(4)	0(0)	049	15	0(0)	1(136)	6(46)
047	497	1(4)	0(1)	0(2)	050	45	1(3)	19(112)	13(90)
048	8	1(43)	0(19)	3(22)	-	-	-	-	-
049	107	1(2)	5(23)	6(15)	-	-	-	-	-
050	10	1(14)	1(52)	0(44)	-	-	-	-	-

288+61). Therefore, $P\text{-}measure = BR(424) = (306+167)/(652+424) = 0.440$. Similarly, it turns out that $P^{+}\text{-}measure = 0.454$. In contrast, although not shown in the table, $r_1 = 2$ for the same ranked output, and the document at this rank is only B-relevant. Therefore, $O\text{-}measure = BR(2) = (1+1)/(3*2+2) = 0.250$. If the real user is unwilling to examine documents down to Rank 424, P-measure and P^{+}-measure may be counterintuitive for this particular case. As we have mentioned in Section 3.3, however, it is possible to compute these metrics based on a smaller number of retrieved documents only, provided that a large number of topics is used to ensure evaluation stability.

(Received March 8, 2007)
(Accepted May 28, 2007)

(Editor in Charge: *Toshikazu Fukushima*)

Tetsuya Sakai was born in 1968. He received a Master's degree from Waseda University in 1993 and joined Toshiba Corporate R&D Center in the same year. He received a Ph.D. from Waseda University in 2000 for his work on information retrieval and filtering systems. From 2000 to 2001, he was a visiting researcher at University of Cambridge Computer Laboratory. In 2007, he left Toshiba and became the Director of the Natural Language Processing Laboratory at NewsWatch, Inc. He received a FIT 2005 Paper Award, an IPSJ 2006 Yamashita SIG Research Award and an IPSJ 2006 Best Paper Award. He is a member of ACM, BCS-IRSG, IPSJ and IEICE.