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

2006－ＦＩ－82 (10)
2006－ＤＤ－54 (10)
2006／3／22

# A Further Note on Evaluation Metrics for the Task of Finding One Highly Relevant Document

Tetsuya Sakai

Knowledge Media Laboratory, Toshiba Corporate R&D Center
tetsuya.sakai@toshiba.co.jp

## Abstract

This paper proposes a new evaluation metric for Information Retrieval systems that aim at providing exactly one highly relevant document to the user. Such information retrieval tasks are especially important for modern large-scale retrieval environments (e.g. the Web) where recall is either unimportant or unknown. Existing metrics for the task of finding one relevant document assume that the user stops examining the ranked list of documents as soon as he finds one relevant document, even if it is a partially relevant one. In contrast, our proposed metric, called P-measure, assumes that the user looks for a highly relevant document even if it is ranked below partially relevant documents, and is probably suitable for retrieval situations such as known-item search or where it is easy for the user to spot a highly relevant document in the ranked output. Our main new findings, based on experiments using two sets of data comprising test collections and submitted runs from NTCIR, are: (a) P-measure is more stable and sensitive than Normalised Weighted Reciprocal Rank (NWRR) and Reciprocal Rank, and is at least as stable and sensitive as O-measure; and (b) Although O-measure and NWRR are highly correlated with each other, Q-measure may be more stable and sensitive than NWRR. In summary, P-measure and O-measure are probably the most reliable metrics for the task of finding one highly relevant document. Researchers can decide on which one to use by considering which better models user behaviour in the real retrieval environment.

## 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 [11] may require a *precision-oriented* metric. When we search the Web, we often stop going through the ranked list after finding *one* good Web page even though the list may contain some more relevant pages, knowing/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) [11] is commonly used for the task of finding one relevant document with high precision: $RR = 0$ 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. In light of this, Sakai [8] proposed a metric called *O-measure* for the task of finding one *highly* relevant document with high precision. O-measure is a variant of *Q-measure* which is very highly correlated with TREC Average Precision (AveP) but can handle graded relevance [5, 6]. O-measure can also be regarded as a generalisation of RR (See Section 2.2). Using well-known methods for evaluating IR metrics [1, 13], Sakai [8] showed that O-measure is more stable than (and at least as sensitive as) RR, and that system rankings based on graded relevance can be quite different from those based on binary relevance.

Eguchi *et al.* [3], the organisers of the NTCIR Web track, have also proposed an evaluation metric for the task of finding one highly relevant document with high precision, namely, *Weighted Reciprocal Rank* (WRR). WRR assumes that ranking a partially relevant document at Rank 1 is more important than ranking a highly relevant document at Rank 2. It has never actually been used for ranking the systems at NTCIR (See Section 2.2) and its reliability is unknown. Sakai [9] points out that, if WRR must be used, then it should be normalised before averaging across topics: he calls the normalised version *Normalised Weighted Reciprocal Rank* (NWRR).

As both O-measure and NWRR are generalisations of RR for handling graded relevance, they are also based 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 situations, but not always, as we shall discuss later. We therefore propose a variant of O-measure, called *P-measure*, which assumes that *the user looks for a highly relevant document even if it is ranked below partially relevant documents.* For some real-world retrieval situations such as known-item search or where it is easy for the user to spot a highly relevant document in the ranked output, P-measure may better model user behaviour than O-measure and NWRR do.

To evaluate and compare the reliability of P-measure and other retrieval metrics, we use two sets of data comprising test collections and submitted runs from NTCIR. Our main new findings are: (a) P-measure is more stable and sensitive than NWRR and RR, and is at least as stable and sensitive as O-measure; and (b) Although O-measure and NWRR are highly correlated with each other, O-measure may be more stable and sensitive than NWRR. In summary, P-measure and O-measure are probably the most reliable metrics for the task of finding one highly relevant document. Researchers can decide on which one to use by considering which better models user behaviour in the real retrieval environment.

The remainder of this paper is organised as follows. Section 2 defines the IR metrics considered in this study, and Section 3 describes methods for evaluating and comparing the metrics. Section 4 reports on the results of our experiments. Section 5 clarifies the contribution of this study through comparisons with related work. Finally, Section 6 concludes this paper. The Appendix explains why our definition of WRR is equivalent to the original definition by Eguchi *et al.* [3].

## 2 Metrics

This section defines the IR metrics considered in this study. Prior to discussing the metrics for the task of finding *one* relevant document, Section 2.1 defines TREC Average Precision (AveP) and Q-measure, both of which are metrics for the task of finding *all* relevant documents. These metrics have been studied extensively elsewhere [6], and are used only as references in this study. Section 2.2 defines the existing metrics for the task of finding one relevant document, namely, RR, O-measure and (N)WRR. Section 2.3 proposes P-measure.

### 2.1 Existing Metrics for the Task of Finding All Relevant Documents

We first define AveP, which is probably the most widely-used IR metric today despite its inability to handle graded relevance. Let $R$ denote the number of relevant documents for a topic, and $count(r)$ denote the number of relevant documents within top $r$ of a system output of size $L$ ($\leq$ 1000). Clearly, Precision at Rank $r$ can be expressed as $P(r) = count(r)/r$. Let $isrel(r)$ be 1 if the document at Rank $r$ is relevant and 0 otherwise. Then, AveP can be defined as:

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

Next, we define Q-measure [5, 6], which is very highly correlated with AveP but can handle graded relevance. Let $gain(X)$ denote the *gain value* for retrieving an $X$-relevant document, where, in the case of NTCIR, $X = S$ (highly relevant), $X = A$ (relevant) or $X = B$ (partially relevant). 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 [4], where $g(i) = gain(X)$ if the document at Rank $i$ is $X$-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

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

$BR(r)$ is called the *blended ratio*, which measures how a system output deviates from the ideal ranked output *and* penalises "late arrieval" 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. For more details, we refer the reader to [8, 9].)

We have the following theorems for the blended ratio and Q-measure:

- In a binary relevance environment, $BR(r) = P(r)$ holds iff $r \leq R$, and $BR(r) > P(r)$ holds otherwise.
- In a binary relevance environment, $Q\text{-}measure = AveP$ holds iff there is no relevant document below Rank $R$, and $Q\text{-}measure > AveP$ holds otherwise.

Moreover, if small gain values ($gain(X)$) are used with Q-measure, then it behaves like AveP [5]. Q-measure is at least as stable and sensitive as AveP as measured by Buckley / Voorhees and Voorhees / Buckley methods (See Section 3) [6].

### 2.2 Existing Metrics for the Task of Finding One Relevant Document

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 with high precision. As was mentioned earlier, Reciprocal Rank (RR) is often used for the task of finding one relevant document [11]. However, RR cannot handle graded relevance, even though it is clear that users prefer highly relevant documents to partially relevant ones. Below, we describe O-measure and Normalised Weighted Reciprocal Rank (NWRR), both of which are generalised versions of RR for handling graded relevance.

Sakai's O-measure [8] is defined to be zero if the ranked output does not contain a relevant document. Otherwise:

$$O\text{-}measure = BR(r_1) = \frac{g(r_1) + 1}{cg_I(r_1) + r_1} . \quad (2)$$

That is, O-measure is the blended ratio at Rank $r_1$. (Since the document at $r_1$ is the *first* relevant one, $cg(r_1) = g(r_1)$ and $count(r_1) = 1$ hold.) Thus, while Q-measure examines the blended ratio for all relevant documents, O-measure examines that of the first retrieved relevant document only. For this reason, O-measure is less stable and less sensitive than Q-measure and AveP. However, it is at least as stable and sensitive as RR [8].

We have the following theorem for O-measure:

- 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 [5].

Next, we define Weighted Reciprocal Rank (WRR) without normalisation, proposed by Eguchi *et al.* [3]. Our definition looks different from their original one, but the Appendix proves that the two are equivalent. In contrast to cumulative-gain-based metrics (including Q-measure and O-measure) which require the gain values ($gain(X)$) as parameters, WRR requires "penalty" values $\beta(X)$ ($> 1$) for each relevance level $X$. 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:

$$WRR = \frac{1}{r_1 - 1/\beta(X_1)} \quad (3)$$

where $X_1$ denotes the relevance level of the relevant document at Rank $r_1$.

Eguchi *et al.* [3] proposed WRR for the NTCIR Web track, but they always set $\beta(X) = \infty$ for all $X$, so that WRR is reduced to binary RR. That is, the graded relevance capability of WRR has never actually been used.

It is easy to see that WRR is not bounded by one: if the highest relevance level for a given topic is denoted by $Y$, WRR is bounded above by $1/(1-1/\beta(Y))$ (See the Appendix). 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 $Y$ may not necessarily be the same across topics, so the upperbound of WRR may differ across topics. For example, the upperbound for a topic that has at least one highly relevant document is $1/(1-1/\beta(S)) = 1/(1-1/2) = 2$, but that for a topic with only relevant and partially relevant documents is $1/(1-1/\beta(A)) = 1/(1-1/3) = 1.5$. This means that *WRR is not suitable for averaging across topics* if $Y$ differs across the topic set of the test collection.

In light of this, Sakai [9] defined Normalised WRR (NWRR). NWRR is defined to be zero if the ranked output does not contain a relevant document. Otherwise:

$$NWRR = \frac{1 - 1/\beta(Y)}{r_1 - 1/\beta(X_1)} \quad (4)$$

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

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.* (The Appendix shows why this is true.) Whereas, O-measure is free from this assumption.

We first discuss (a). From Eq. (4), 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 Figure 1, 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 $X$-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 Figure 1, $O\text{-}measure = (g(3) + 1)/(cg_I(3) + 3) = (3+1)/(3+3) = 2/3$. 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, $O\text{-}measure = (3 + 1)/(9 + 3) = 1/3$. Thus, *O-measure assumes that it is relatively easy to retrieve an $X$-relevant document if there are many $X$-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.

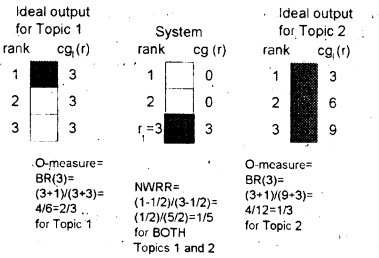


**Figure 1. O-measure vs NWRR: Topics 1 and 2.**

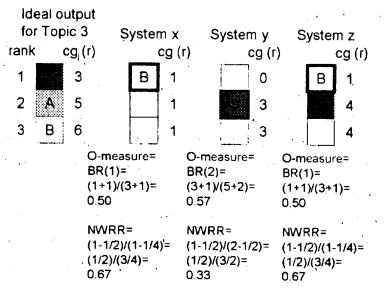


**Figure 2. O-measure vs NWRR: Topic 3.**

Next, we discuss (b) using Topic 3 shown in Figure 2, which has one S-relevant, one A-relevant and one B-relevant document. In this figure, 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 ($\beta(X)$), System A always outperforms System B 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: Figure 2 shows that, with default gain values, System $y$ outperforms System $x$. But if System $x$ should be preferred, then a different gain value assignment (e.g. $gain(S) = 2, gain(A) = 1.5, gain(B) = 1$) can be used [9]. In this respect, O-measure is more flexible than NWRR.

## 2.3 A New Metric for the Task of Finding One Relevant Document

Despite the abovementioned differences, both NWRR and O-measure rely on $r_1$, the rank of the first retrieved relevant document. This means that 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 Figure 2, 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 Sys-

tem $x$ 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 System $z$. 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 output that contain informative titles and abstracts, then perhaps it is fair to assess System C 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* [11], 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.

We now define *P-measure* for the task of finding one highly relevant document with high precision, 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 the *Preferred rank* $r_p$ be the rank of the first record obtained by sorting the system output, using the relevance level as the primary sort key (preferring higher relevance levels) and the rank as the secondary sort key (preferring the top ranks). Then:

$$P\text{-}measure = BR(r_p) = \frac{cg(r_p) + count(r_p)}{cg_I(r_p) + r_p} . \quad (5)$$

That is, P-measure is simply the blended ratio at Rank $r_p$. For System $z$ in Figure 2, $r_p = 2$. Therefore, $P\text{-}measure = BR(2) = (cg(2) + 2)/(cg_I(2) + 2) = (4 + 2)/(5 + 2) = 0.86$. Whereas, since $r_p = r_1$ holds for Systems A and B, $P\text{-}measure = O\text{-}measure = 0.50$ for System $x$ and $P\text{-}measure = O\text{-}measure = 0.57$ for System $y$. Thus, System $z$ 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 stable and sensitive as 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. Furthermore, the following clearly holds:

- In a binary relevance environment, $P\text{-}measure = O\text{-}measure$ holds.

# 3 Methods for Comparing the Reliability of Metrics

This section describes three methods for assessing and comparing the reliability of IR metrics.

## 3.1 Kendall's Rank Correlation

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 [12]. 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 ($\tau$) can be expressed as:

$$\tau = \frac{2(pos - neg)}{n_s(n_s - 1)} . \quad (6)$$

It is known that

$$Z_0 = \frac{|\tau|}{((4n_s + 10)/(9n_s(n_s - 1)))^{\frac{1}{2}}} \quad (7)$$

obeys a normal distribution, and therefore a normal test can easily be applied. Note that the test statistic $Z_0$ is proportional to $|\tau|$, given $n_s$: When $n_s = 30$ (See Section 4.1), Kendall's rank correlation is statistically significant at $\alpha = 0.01$ if it is over 0.34 (two-sided test).

## 3.2 Buckley / Voorhees Stability

We measure the *stability* of each IR metric with respect to change in the topic set using our adaptation of the *Buckley / Voorhees method* [1]. The input to this method are:

- An IR test collection;
- A set of *systems* (or *runs*) submitted to the task defined by the above test collection;
- An IR evaluation metric;
- A *fuzziness value* $f$, which determines how much relative performance difference between a system pair should be regarded as negligible.

The output of the method are:

- *Minority Rate*, which represents lack of stability with respect to the change in topic sets;
- *Proportion of Ties*, which represents lack of discrimination power.

More specifically, our adaptation of the Buckley / Voorhees method works as follows. Let $S$ denote a set of systems submitted to a particular task, and let $x$ and $y$ denote a pair of systems from $S$. Let $Q$ denote the topic set of the test collection, and let $M(x, Q)$ denote the value of metric $M$ for System $x$ calculated based on $Q$. Then, using the algorithm shown in Figure 3, the minority rate and the proportion of ties

```
for i = 1 to 1000
    create Q_i s.t. |Q_i| = |Q| by
    random sampling with replacement from Q;
for each pair of systems x, y ∈ S
    for i = 1 to 1000
        margin = f * max(M(x, Q_i), M(y, Q_i));
        if( |M(x, Q_i) − M(y, Q_i)| < margin )
            EQ_M(x, y) + +
        else if( M(x, Q_i) > M(y, Q_i) )
            GT_M(x, y) + +
        else
            GT_M(y, x) + +;
```

**Figure 3. The algorithm for computing $EQ_M(x,y)$,$GT_M(x,y)$ and $GT_M(y,x)$.**

of $M$ ($MR_M$ and $PT_M$), given a fuzziness value $f$, can be computed as:

$$MR_M = \frac{\sum_{x,y \in S} \min(GT_M(x,y), GT_M(y,x))}{\sum_{x,y \in S} 1000}. \quad (8)$$

$$PT_M = \frac{\sum_{x,y \in S} EQ_M(x,y)}{\sum_{x,y \in S} 1000}. \quad (9)$$

From the algorithm, it is clear that $GT_M(x,y) + GT_M(y,x) + EQ_M(x,y) = 1000$ holds for each system pair, and that a larger $f$ yields larger $EQ_M(x,y)$ values, and therefore a larger PT and a smaller MR. As a fixed value of $f$ implies different trade-offs for different metrics, we vary $f(= 0.01, 0.02, \ldots, 0.20)$ to draw *MR-PT curves* [6, 8] for the purpose of comparing different metrics.

Our method differs from the *original* Buckley / Voorhees method in that we use sampling *with* replacement from $Q$ to create *bootstrap samples* $Q_i$ [2]. We shall discuss this issue in Section 3.3.

### 3.3 Voorhees / Buckley Sensitivity

We measure the *discrimination power* (or *sensitivity*) of each IR metric using our adaptation of the *Voorhees / Buckley method* [13]. The input to this method are:

- An IR test collection;
- A set of systems submitted to the task defined by the above test collection;
- An IR evaluation metric;
- The required *confidence level* of a conclusion as to which of the given two systems $x$ and $y$ is better.

The output of the method are:

- The mininum absolute/relative performance difference required in order to guarantee the given confidence level;
- How often system pairs actually satisfy the above requirement, which represents the discrimination power of the metric.

More specifically, our adaptation of the Voorhees / Buckley method works as follows. Let $d$ denote a performance difference between two systems computed based on a topic set. We

```
for i = 1 to 1000
    create Q_i and Q'_i s.t. |Q_i| = |Q'_i| = |Q| by
    random sampling with replacement from Q;
for each pair of systems x, y ∈ S
    for i = 1 to 1000
        d_M(Q_i) = M(x, Q_i) − M(y, Q_i);
        d_M(Q'_i) = M(x, Q'_i) − M(y, Q'_i);
        count(BIN(d_M(Q_i))) + +;
        if( d_M(Q_i) * d_M(Q'_i) > 0 )
            continue
        else
            swap_count(BIN(d_M(Q_i))) + +;
for each bin b
    swap_rate(b) = swap_count(b)/count(b);
```

**Figure 4. The algorithm for computing the swap rates.**

first 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 $0.20 \leq d$. Let $BIN(d)$ denote a mapping from a difference $d$ to one of the 21 bins where it belongs. The algorithm shown in Figure 4 calculates a *swap rate* for each bin.

By plotting swap rates against the performance difference bins, one can discuss how much performance difference is required in order to conclude that a run is better than another with a required confidence level. For example, if 95% confidence is required, one looks for the minumum performance difference that guarantees 5% swap rate or less. Moreover, by examining how often this condition is satisfied among all pairwise comparisons from all the trials, one can compare the discrimination power of different metrics.

Our method differs from the *original* Voorhees / Buckley method in that we use sampling *with* replacement from $Q$ to create *boostrap samples* $Q_i$ and $Q'_i$. The original method used sampling *without* replacement from $Q$ and ensured that $Q_i$ and $Q'_i$ are disjoint (i.e. $Q_i \cap Q'_i = \phi$), but this implies that (i) $Q_i$ and $Q'_i$ are not independent of each other [10]; (ii) $Q_i$ and $Q'_i$ can only be half the size of $Q$. Regarding (i), Sakai [7] showed that sampling with and without replacement yield similar results for the purpose of comparing the sensitivity of different IR metrics. Regarding (ii), *extrapolation* can be used to estimate the sensitivity of metrics when $Q_i$ is as large as the original topic set $Q$, but the accuracy of such an approach is not clear. Hence, we chose sampling *with* replacement, so that we can directly measure the sensitivity of an IR metric given the topic set size $|Q_i| = |Q|$.

## 4 Experiments

This section describes our experiments using data from NTCIR for comparing the reliability of IR metrics.

### 4.1 Data

Table 1 shows some statistics of the NTCIR data we used for comparing the reliability of IR metrics [5, 8]. For example, The NTCIR-3 Chinese data set contains 42 topics, 45

**Table 1. Statistics of the NTCIR-3 CLIR data**

| | $\|Q\|$ | #runs | $R(S)$ | $R(A)$ | $R(B)$ | $\bar{R}$ |
|---|---|---|---|---|---|---|
| Chinese | 42 | 30 (45) | 21.0 | 24.9 | 32.3 | 78.2 |
| Japanese | 42 | 30 (33) | 7.9 | 31.5 | 21.0 | 60.4 |

submitted runs (of which we used the top 30 as measured by Relaxed-AveP for all experiments), 21.0 S-relevant documents per topic, and so on.

## 4.2 Results: Rank Correlation

Table 2 shows the Kendall's rank correlation values for each pair of metrics, based on the Chinese and Japanese data. Table 3 shows similar data between P-measure with default gain values and P-measure with alternative gain value assignments: For example, "P30:20:10" represents P-measure with $gain(S) = 30, gain(A) = 20, gain(B) = 10$. All the correlation values exceed 0.34 and therefore are statistically highly significant (See Section 3.1), but values higher than 0.9 are shown in bold. We can observe that:

- O-measure and NWRR are consistently highly correlated with each other. This is because they are both based on $r_1$, while taking graded relevance into account.
- P-measure is highly correlated with O-measure and NWRR for the Japanese data, but the correlation values are somewhat lower for the Chinese data. This reflects the fact that P-measure relies on $r_p$ rather than $r_1$.
- P-measure, O-measure and NWRR are not so highly correlated with RR. Hence, finding one *highly* relevant document is not the same as finding *any* one relevant document. (This generalises a finding in [8], which considered neither P-measure nor NWRR.)
- P-measure, O-measure, NWRR and RR are not highly correlated with AveP and Q-measure. Hence, finding *one* relevant document is not the same as finding as many relevant documents as possible. (This also generalises a finding in [8].)
- P-measure is fairly robust to the choice of gain values. "P30:20:10" and "P10:5:1" produce rankings similar to the default P-measure (i.e. "P3:2:1").

## 4.3 Results: Buckley / Voorhees Stability

Figures 5 and 6 show the Buckley / Voorhees MR-PT curves for the Chinese and Japanese data, respectively. Recall that good IR metrics should show low minority rates *and* low proportion of ties. We can observe that:

- P-measure is possibly more stable than O-measure. This difference probably arises from the fact that P-measure considers all relevant documents ranked above $r_p$.
- P-measure and O-measure are more stable than NWRR. This difference probably arises from the fact that the two compare the system output with an ideal one, while NWRR looks at the system output only.
- P-measure, O-measure and NWRR are all more stable than RR. Hence, the use of graded relevance improves stability. However, these metrics, designed for the task of finding *one* relevant document, are not as stable as Q-measure and AveP, which are for the task of finding *all* relevant documents. (This generalises a finding in [8].)

**Table 2. Kendall's rank correlation values based on 30 runs (NTCIR-3 CLIR).**

| Chinese | (b) | (c) | (d) | (e) | (f) |
|---|---|---|---|---|---|
| (a) RR | .8575 | .7977 | .7425 | .5264 | .5494 |
| (b) NWRR | - | **.9126** | .8575 | .5494 | .5632 |
| (c) O-measure | - | - | .8621 | .5264 | .5402 |
| (d) P-measure | - | - | - | .5540 | .5678 |
| (e) AveP | - | - | - | - | **.9678** |
| (f) Q-measure | - | - | - | - | - |

| Japanese | (b) | (c) | (d) | (e) | (f) |
|---|---|---|---|---|---|
| (a) RR | .8759 | .8207 | .8253 | .7701 | .7701 |
| (b) NWRR | - | **.9356** | **.9126** | .7011 | .7287 |
| (c) O-measure | - | - | **.9218** | .6920 | .7011 |
| (d) P-measure | - | - | - | .7333 | .7517 |
| (e) AveP | - | - | - | - | **.9540** |
| (f) Q-measure | - | - | - | - | - |

**Table 3. Kendall' rank correlation based on 30 runs: default P-measure vs other gain value assignments (NTCIR-3 CLIR)**

| Chinese | P30:20:10 | P0.3:0.2:0.1 | P1:1:1 | P10:5:1 |
|---|---|---|---|---|
| P-measure | **.9540** | .8713 | .8667 | **.9126** |
| Japanese | P30:20:10 | P0.3:0.2:0.1 | P1:1:1 | P10:5:1 |
| P-measure | **.9862** | **.9632** | **.9402** | **.9632** |

**Table 4. Voorhees/Buckley sensitivity (swap rate $\leq$ 5%; NTCIR-3 CLIR).**

| Chinese | abs | max | rel | sensitivity |
|---|---|---|---|---|
| Q-measure | 0.07 | .5374 | 13% | **43.2%** |
| AveP | 0.08 | .5295 | 15% | **39.5%** |
| P-measure | 0.15 | .8636 | 17% | **30.5%** |
| O-measure | 0.17 | .8674 | 20% | **24.4%** |
| NWRR | 0.18 | .8633 | 21% | **22.2%** |
| RR | 0.19 | .9524 | 20% | **20.0%** |

| Japanese | abs | max | rel | sensitivity |
|---|---|---|---|---|
| Q-measure | 0.07 | .6433 | 11% | **67.1%** |
| AveP | 0.07 | .6449 | 11% | **66.1%** |
| P-measure | 0.13 | .8759 | 15% | **59.3%** |
| O-measure | 0.14 | .8690 | 16% | **56.2%** |
| NWRR | 0.16 | .8757 | 18% | **51.0%** |
| RR | 0.17 | .9524 | 18% | **47.6%** |

## 4.4 Results: Voorhees / Buckley Sensitivity

Table 4 summarises the results of our Voorhees / Buckley sensitivity experiments. For example, with 42 topics and 30 Chinese runs, P-measure is 95% confident that a system is better than another provided that the absolute performance difference is at least 0.15, which translates to a relative difference of 17% [6, 8]. Among 435,000 comparisons (30*29/2 run pairs times 1,000 trials), 30.5% actually satisfied the absolute difference requirement. From the table, we can observe that:

- P-measure is more sensitive than NWRR and RR; P-measure may be more sensitive than O-measure, which in turn may be more sensitive than NWRR.
- Even P-measure is not as sensitive as Q-measure and AveP, as it does not examine *all* relevant documents as Q-measure and AveP do.

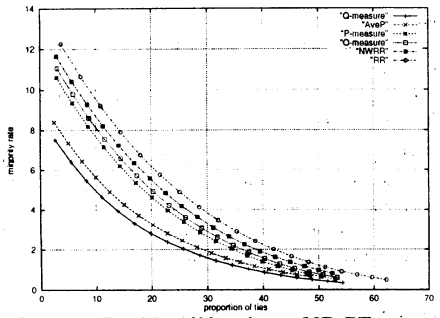


**Figure 5. Buckley / Voorhees MR-PT curves (NTCIR-3 CLIR Chinese).**

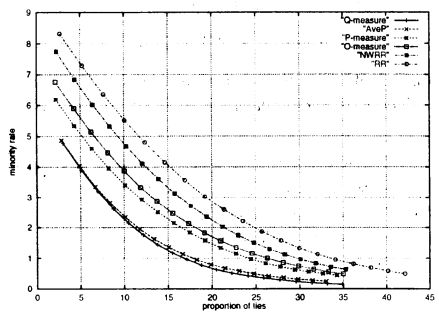


**Figure 6. Buckley / Voorhees MR-PT curves (NTCIR-3 CLIR Japanese).**

### 4.5 Per-topic Analysis

As was discussed earlier, the stability and sensitivity of P-measure probably arises from the fact that it relies on $r_p$ rather than $r_1$. We now examine some actual values of $r_p$ and $r_1$ per topic to see what values they take and how often they equal each other. Figure 7 shows the values of $r_p$ and $r_1$ for a "median" Chinese run: This run had the 15th highest P-measure value among the 30 Chinese runs we used. Values of $r_p$ and $r_1$ are represented by crosses and circles, respectively. It can be observed that $r_p$ often differs from $r_1$. For example, for Topic 2, $r_1 = 3$ but $r_p = 20$. The ranked output for this topic had A-relevant documents at Ranks 3, 11 and 12, B-relevant documents at Ranks 7 and 17, and its first S-relevant document at Rank 20. As a result, while the RR, NWRR and O-measure values are 0.33, 0.19 and 0.25, respectively, the P-measure value is 0.21. As mentioned earlier, P-measure models the user behaviour in IR tasks such as known-item search: Because the user knows (or believes) that a highly relevant document exists, he examines the ranked list until he finds one at Rank 20. Whereas, RR, NWRR and O-measure assume that the user is satisfied when he sees the A-relevant document at Rank 3.

Since our experiments showed that P-measure and O-measure are probably the most reliable metrics for the task of finding one highly relevant document, researchers can decide on which one to use by considering which better models user behaviour in the real retrieval environment.

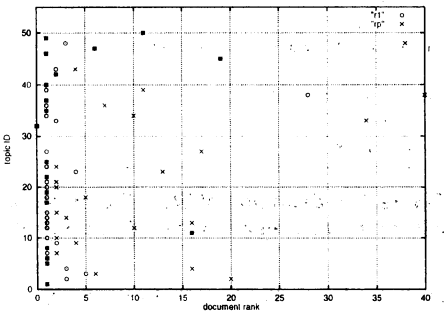


**Figure 7. A Comparison of $r_p$ and $r_1$ for the Chinese run with the 15th highest P-measure.**

## 5 Related Work

Voorhees [12] considered the problem of finding a few highly relevant Web pages, but used the unnormalised Discounted Cumulative Gain (DCG) [4] as an evaluation metric, which is not specifically designed for the problem of finding *one* highly relevant document. Moreover, using DCG without normalisation is not suitable for averaging across topics [6, 9]. Soboroff [11] reports on the Voorhees / Buckley sensitivity of RR for the TREC Web track, but his experiments are limited to binary-relevance metrics.

Sakai [8] proposed O-measure for the task of finding one highly relevant document and compared it with RR, AveP and Q-measure in terms of rank correlation, stability and sensitivity. This paper extends his work in that P-measure and NWRR are also included in the experiments. Moreover, while Sakai [8] used sampling *without* replacement in his stability and sensitivity experiments, the present study used sampling *with* replacement, which enabled us to use resampled topic sets $Q_i$ that are equal in size to the original topic set $Q$ [7].

## 6 Conclusions

This paper proposed a new evaluation metric called P-measure for the task of finding one highly relevant document. While Reciprocal Rank, (Normalised) Weighted Reciprocal Rank and O-measure assume that the user stops examining the ranked list of documents as soon as he finds *any* relevant document, P-measure assumes that the user looks for a highly relevant document even if it is ranked below partially relevant documents. This assumption is probably valid for retrieval situations such as known-item search or where it is easy for the user to spot a highly relevant document in the ranked output. Our experiments using two sets of data from NTCIR showed that:

- P-measure is more stable and sensitive than Normalised Weighted Reciprocal Rank (NWRR) and Reciprocal Rank, and is at least as stable and sensitive as O-measure;
- Although O-measure and NWRR are highly correlated with each other, O-measure may be more stable and sensitive than NWRR.

In summary, P-measure and O-measure are probably the most reliable metrics for the task of finding one highly relevant doc-

```
for i = 1 to l
    if document at Rank i is X-relevant (X ∈ {S, A, B})
        f_l(i) = δ(X)/(i − 1/β(X))
    else /* nonrelevant */
        f_l(i) = 0;
WRR_l = max_i f_l(i);
```

**Figure 8. A definition of WRR that is more faithful to the original one by Eguchi et al.**

ument. Researchers can decide on which one to use by considering which better models user behaviour in the real retrieval environment.

## Appendix: Proof that Our Definition of WRR is Equivalent to that by Eguchi et al.

Let $l$ denote a document cut-off value, and let $\delta(X) = 1$ if each $X$-relevant document should be counted as relevant and $\delta(X) = 0$ otherwise. Figure 8 provides a definition of WRR (at cut-off $l$) that is more faithful to the original one by Eguchi *et al.* [3] than Eq. (3). Since $l$ is just a cut-off value, we can let $l = L$ for the purpose of assessing the reliability of WRR (i.e. the whole ranked ouput is examined). This should provide the upperbounds of its stability and sensitivity, since it is known that using a small value of $l$ reduces stability and sensitivity [6]. Moreover, note that the parameters $\delta(X)$ merely define which relevance levels should be counted as relevant. Thus, in essence, the definition of WRR by Eguchi *et al.* (when the system output contains at least one relevant document) computes $f(i) = 1/(i - 1/\beta_i)$ for each $X$-relevant document at Rank $i$, where $\beta_i = \beta(X)$, and finally takes the highest value.

We now prove a simple theorem:

**Theorem 1** *If $i < j$ and $\beta_i, \beta_j > 1$, then $f(i) > f(j)$ holds (regardless of whether $\beta_i > \beta_j$, $\beta_i = \beta_j$ or $\beta_i < \beta_j$).*

*Proof:* Let us *assume* the contrary, i.e.:

$$1/(i - 1/\beta_i) \leq 1/(j - 1/\beta_j) .$$

Then,

$$j - 1/\beta_j \leq i - 1/\beta_i$$

and therefore

$$j - i \leq (\beta_i - \beta_j)/\beta_i\beta_j .$$

Thus,

$$\beta_i\beta_j(j - i) \leq \beta_i - \beta_j . \quad (10)$$

But since $i < j$ and $\beta_i, \beta_j > 1$,

$$\beta_i\beta_j \leq \beta_i\beta_j(j - i) \quad (11)$$

should hold. Therefore, From Eqs. (10) and (11),

$$\beta_i\beta_j \leq \beta_i - \beta_j .$$

Hence,

$$\beta_j(\beta_i + 1) \leq \beta_i < \beta_i + 1 .$$

Therefore $\beta_j < 1$, but this is a contradiction.

**Corollary 1** *The maximum value of $f(i)$ for a given ranked output is given by $1/(r_1 - \beta(X_1))$, where $r_1$ is the rank of a relevant document that is nearest to the top of the list and $X_1$ is the relevance level of this document. That is, the original definition of WRR is essentially equivalent to ours (Eq. (3)).*

*Proof:* Theorem 1 implies that, if documents at Ranks $i$ and $j$ are (at least partially) relevant and $i < j$, then $f(i) > f(j)$ holds *regardless of the relevance levels of these two documents*. Thus, the maximum value of $f(i)$ is obtained when $f(i)$ is computed for the *first* relevant document in the ranked output.

**Corollary 2** *If System $x$ has its first relevant document at Rank $i$ while System $y$ has its first relevant document at Rank $j$ ($> i$), then System $x$ outperforms System $y$ regardless of the relevance levels of these two documents.*

*Proof:* This is also a corollary of Theorem 1.

**Corollary 3** *WRR is bounded above by $1/(1 - 1/\beta(Y))$, where $Y$ is the highest relevance level for the topic in question.*

*Proof:* From Theorem 1, it is clear that the highest possible value of WRR for a topic is given by $1/(1 - 1/\beta(X))$ for some $X$. Moreover, it is clear that this value gets large as $\beta(X)$ gets small. Thus $X$ should be $Y$, which is given the smallest penalty.

## References


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

[2] Efron, B. and Tibshirani, R. J.: *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1993.

[3] Eguchi, K. *et al.*: Overview of the Web Retrieval Task at the Third NTCIR Workshop, *Technical Report NII-2003-002E*, National Institute of Informatics, 2003.

[4] 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.

[5] Sakai, T.: Ranking the NTCIR Systems based on Multigrade Relevance, *AIRS 2004 Proceedings*, pp.170-177, 2004. Also available in Myaeng, S. H. et al. (Eds.): AIRS 2004, Lecture Notes in Computer Science 3411, pp. 251–262, Springer-Verlag, 2005.

[6] Sakai, T.: The Reliability of Metrics based on Graded Relevance, *AIRS 2005 Proceedings*, Lecture Notes in Computer Science 3689, pp. 1-16, 2005.

[7] Sakai, T.: The Effect of Topic Sampling on Sensitivity Comparisons of Information Retrieval Metrics, *NTCIR-5 Proceedings*, pp. 505-512, 2005.

[8] Sakai, T.: On the Task of Finding One Highly Relevant Document with High Precision, *IPSJ Transactions on Databases*, TOD-29, 2006.

[9] Sakai, T.: For Building Better Retrieval Systems: Trends in Information Retrieval Evaluation based on Graded Relevance (*in Japanese*), *IPSJ Magazine*, Vol. 47, No. 2, 2006.

[10] Sanderson, M. and Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *ACM SIGIR 2005 Proceedings*, pp. 162-169, 2005.

[11] Soboroff, I.: On Evaluating Web Search with Very Few Relevant Documents, *ACM SIGIR 2004 Proceedings*, pp. 530-531, 2004.

[12] Voorhees, E. M.: Evaluation by Highly Relevant Documents, *ACM SIGIR 2001 Proceedings*, pp. 74-82, 2001.

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