Regular Paper

Evaluating Information Retrieval Metrics Based on Bootstrap Hypothesis Tests

Tetsuya Sakai†

This paper describes how the bootstrap approach to statistics can be applied to the evaluation of IR effectiveness metrics. More specifically, we describe straightforward methods for comparing the discriminative power of IR metrics based on Bootstrap Hypothesis Tests. Unlike the somewhat ad hoc Swap Method proposed by Voorhees and Buckley, our Bootstrap Sensitivity Methods estimate the overall performance difference required to achieve a given confidence level directly from Bootstrap Hypothesis Test results. We demonstrate the usefulness of our methods using four different data sets (i.e., test collections and submitted runs) from the NTCIR CLIR track series for comparing seven IR metrics, including those that can handle graded relevance and those based on the Geometric Mean. We also show that the Bootstrap Sensitivity results are generally consistent with those based on the more ad hoc methods.

1. Introduction

A typical IR paper claims that System X is better than System Y in terms of an effectiveness metric M computed based on a test collection C: How reliable is this paper? More specifically, (a) What happens if C is replaced with another set of data C'? (b) How good is M?

Question (a) posed above is usually dealt with as follows:

(1) Use two or more test collections of "respectable" size, and observe trends that are consistent across the different data;
(2) Make sure that the overall performance difference between X and Y is "relatively large";
(3) Conduct statistical significance tests to claim that the difference was not observed due to chance ☆.

All of the above are arguably *necessary* conditions for making a good IR paper that involves comparative experiments, although surprisingly many IR papers do not satisfy them [19].

Unfortunately, there has been a controversy as to which statistical significance tests should be used for IR evaluation, as well as whether such tests should be used at all [4),19),20]. It is known that a typical IR evaluation environment often violates the underlying assumptions of significance tests, but it is also known that some significance tests work well even when some of the assumptions are violated. Parametric tests rely on the *normality* assumption and generally have higher power than nonparametric ones. (That is, it is easier to detect significant differences with parametric tests.) But even nonparametric tests are not assumption-free: the Paired Wilcoxon Test depends on both the *symmetry* and the *continuity* assumptions [4]. An IR researcher who wants to be *conservative* (i.e., who wants to minimise the risk of jumping to wrong conclusions) might, for example, choose the two-tailed Sign Test, which generally has little power.

However, as Savoy [20] points out, there is a very attractive alternative called the *bootstrap* [3]. Invented in 1979, the bootstrap is the approach to statistics for the computer age, and has strong theoretical foundations. While classical statistics rely on mathematical derivations that often require several assumptions on the underlying distributions of data, the bootstrap tries to achieve the same goal by directly estimating the distributions through *resampling* from observed data. The Bootstrap Hypothesis Tests are free from the normality and symmetry assumptions, and it is known that they often show power comparable to that of traditional parametric significance tests. Moreover, the *Unpaired* Bootstrap Hypothesis Test is di-

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

☆ We are aware that there are negative arguments against statistical significance tests in general [6]. However, while we agree that significance tests are not The Perfect Methods that can validate an experiment, we believe that they can be useful if used correctly.

rectly applicable even to unconventional summary statistics that are not Arithmetic Means over a topic set (e.g., the "area" measure based on the worst N topics for each system [21]) and Geometric Means [11),18),21)]). We therefore believe that Bootstrap Hypothesis Tests deserve more attention from the IR community.

This paper ☆ concerns Question (b) posed above: How "good" is IR metric M? More specifically, we use the Bootstrap Hypothesis Tests to assess and compare the *discriminative power* (or *sensitivity*) of different IR metrics. This is related to the *Swap Method* proposed by Voorhees and Buckley [22], which derives the overall performance difference required for guaranteeing that a system is better than another, and that the chance of obtaining a contradictory result with another topic set (the *swap rate*) is below a given threshold. However, while the Swap Method is not directly related to statistical significance tests, our new methods estimate the overall performance difference required to achieve a given confidence level directly from Bootstrap Hypothesis Test results.

To demonstrate the usefulness of our Bootstrap Sensitivity Methods, we use four different data sets (i.e., test collections and submitted runs) from the NTCIR CLIR track series [7] and compare seven IR effectiveness metrics, including those based on graded relevance and those based on the Geometric Mean. Our methods agree with the more ad hoc ones such as the Swap Method that, for the majority of the data sets, the most sensitive IR metrics are Q-measure [17], normalised Discounted Cumulative Gain at cut-off 1000 [5),9)] and Average Precision (AveP), while the least sensitive one is Precision at cut-off 1000. In the middle lie normalised Cumulative Gain at cut-off 1000 and Geometric Mean AveP/Q-measure.

The remainder of this paper is organised as follows. Section 2 discusses previous work related to this study. Section 3 formally defines the seven IR effectiveness metrics we consider, and Section 4 describes our NTCIR data sets. Section 5 describes how Paired and Unpaired Bootstrap Hypothesis Tests can be conducted in IR evaluation, and Section 6 proposes and tests our new methods for comparing the discriminative power of IR metrics based on Bootstrap Hypothesis Tests. Section 7 compares our Bootstrap Sensitivity results with those based on the more ad hoc methods. Finally, Section 8 provides conclusions and prospects for our future research. In addition, the Appendix discusses the *similarities* among the metrics we consider using Kendall's rank correlation, in order to complement our results on the *discriminative power* of individual metrics, which is the focus of this study.

☆ An early, eight-page version of this paper was presented at ACM SIGIR 2006. While the SIGIR paper used the NTCIR-3 Chinese and Japanese data only, this paper uses the NTCIR-5 Chinese and Japanese data in addition, and provides more detailed analyses of the experimental results.

2. Related Work

Savoy [20] used the Paired Bootstrap Hypothesis Test and Confidence Intervals, along with traditional significance tests, for comparing two IR strategies. However, comparing the discriminative power of IR metrics based on many system pairs was beyond the scope of his work.

Our Bootstrap Sensitivity Methods have much in common with the Swap Method proposed by Voorhees and Buckley [22], in that we use a test collection and a set of submitted runs as input and generate sets of resampled topics from the original topic set to estimate the overall performance difference required between two systems in order to satisfy a given "confidence level". However, there are important differences. Our confidence level is defined as $1-\alpha$, where α is the probability of concluding that two systems are different even though they are in fact equivalent in a Bootstrap Hypothesis Test, also known as Type I Error. Thus, our methods are directly related to achieved significance levels. In contrast, the Swap Method is rather ad hoc (although Lin and Hauptmann [10] have discussed a theoretical justification of the method), and is not directly related to statistical significance tests: Sanderson and Zobel [19] used significance tests for filtering out some system pairs *before* applying the Swap Method; Sakai [17] reported that the system pair ranking according to significance tests and that according to the Swap Method are not very highly correlated.

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. The "confidence level" in this method is defined as one minus the swap rate, that is, the probabil-

ity that a pair of topic sets agree as to whether one system outperforms the other. The method estimates this probability by generating many pairs of new topic sets from Q, the original set of topics provided in the test collection. The *original* Swap Method samples topics *without* replacement from Q, each time generating two new topic sets that are disjoint from each other: Hence the new topic sets contain no duplicates. Sanderson and Zobel [19] also used sampling *without* replacement (although they called their method "selection *with* replacement" in their paper) but ensured that the two topic sets were independently drawn from Q: Hence the two topic sets were generally *not* disjoint from each other in their method. In both the original Swap Method and the Sanderson/Zobel variant, the sets of resampled topics were no greater than half the size of Q, due to the original idea that each pair of new sets should be disjoint. For this reason, these studies used *extrapolation* to discuss the reliability of IR evaluation using a topic set of size $|Q|$. Subsequently, however, Sakai [12] showed that sampling *with* and *without* replacement with the Swap Method yield similar results for comparing different IR metrics. In light of this, the present study uses *bootstrap samples* [3] with the Swap Method as well as with our proposed methods, where each bootstrap sample is obtained by sampling *with* replacement and is equal in size to Q. Therefore, our experiments do not require extrapolation: We can directly discuss the reliability of IR evaluation using exactly $|Q|$ topics. More details will follow in Section 7.

Recently, Cormack, Lynam and Cheriton [2] proposed a method for estimating the (statistical) precision of IR evaluation using bootstrap techniques. However, while our methods (and standard significance tests) focus on the effect of random error associated with the selection of *topics*, their concern is the effect of random error associated with the target *documents*.

The Bootstrap Sensitivity Method has also been used for comparing IR metrics for the task of finding *one* relevant document [13], and for investigating the effect of IR metric parameters for penalising late arrivals of relevant documents [15].

3. IR Effectiveness Metrics

The basic IR metrics we consider in this paper are Average Precision (AveP), Precision at cut-off 1000 (PDoc_{1000}), Q-measure [17], and normalised (Discounted) Cumulative Gain at cut-off 1000 ($\mathrm{n(D)CG}_{1000}$) [5),9)].

AveP represents a very sensitive IR metric based on *binary* relevance, while PDoc_{1000} represents a very insensitive one. (PDoc_{1000} rewards a system with 10 relevant documents at Ranks 1-10 and one with 10 relevant documents at Ranks 991-1000 equally. Note also that it does not average well [17].) 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:

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

$$PDoc_l = P(l) \ . \quad (2)$$

We can also use IR metrics based on *graded* relevance, since the NTCIR data contain S-, A- and B-relevant (highly relevant, relevant and partially relevant) documents. 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* [5),9)] for retrieving an $\mathcal{L}$-relevant document. We use $gain(S) = 3$, $gain(A) = 2$, $gain(B) = 1$ thoughout this paper: The effect of using different gain values has been discussed elsewhere [15),17)]. Let $cg(r) = \sum_{1 \leq i \leq r} g(i)$ denote the *cumulative gain* [5)] at Rank r of the system's output, where $g(i) = gain(\mathcal{L})$ if the document at Rank i is $\mathcal{L}$-relevant and $g(i) = 0$ otherwise. In particular, consider an *ideal* ranked output, such that $isrel(r) = 1$ for $1 \leq r \leq R$ and $g(r) \leq g(r-1)$ for $r > 1$, and let $cg_I(r)$ denote the ideal cumulative gain at Rank r. Similarly, by using $dg(i) = g(i)/\log_a(i)$ instead of $g(i)$ for $i > a$, we can obtain the (ideal) *discounted* cumulative gain $dcg(r)$ and $dcg_I(r)$ [5),9)]. Then we have:

$$nCG_l = cg(l)/cg_I(l) \ . \quad (3)$$

$$nDCG_l = dcg(l)/dcg_I(l) \ . \quad (4)$$

$$Q\text{-}measure = \frac{1}{R} \sum_{1 \leq r \leq L} isrel(r) BR(r) \quad (5)$$

where $BR(r)$ is the *blended ratio* given by:

$$BR(r) = (cg(r) + count(r))/(cg_I(r) + r) \ . \quad (6)$$

It is known that nCG_l has a problem: it cannot penalise late arrival of relevant docu-

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

ments after Rank R as $cg_I(r) = cg_I(R)$ holds for $r \geq R$. nDCG$_l$ partially solves this problem by discounting the gains, but using a large logarithm base a with it would imply that it inherits the defect of nCG$_l$, since discounting cannot be applied until the rank is greater than a. For this reason, we let $a = 2$ throughout this paper. On the other hand, Q-measure solves the above problem by including r in the denominator of $BR(r)$. For more details on the "late arrival" problem, we refer the reader to Sakai's papers [15),17)].

nDCG$_l$ is a stable and sensitive metric provided that l is large and a is small [15),17)]. Q-measure is also stable and sensitive [15),17)], and it has been applied to XML retrieval [8)] as well as factoid question answering [16)]. It is more highly correlated with AveP than nDCG is; In a binary relevance environment, *Q-measure* = *AveP* holds for any ranked output if there is no relevant document below Rank R, and *Q-measure* > *AveP* holds otherwise ☆.

By default, we use the *Arithmetic* Mean over a given topic set with any IR metric. However, this paper also considers the *Geometric* Mean versions of AveP and Q-measure, which we denote by G_AveP and G_Q-measure. It has been argued that the Geometric Mean may be more useful than the Arithmetic Mean for building *robust* IR systems, i.e., those that can produce a decent output whatever the query is [11),18),21)]. This is because the Geometric Mean puts more emphasis on low-performing topics than the Arithmetic one: Note, for example, that while the Arithmetic Mean of 2 and 50 is 26, the corresponding Geometric mean is 10.

Let x_i denote the value of a metric for the i-th topic (down to four significant figures). Then, following 21), the actual method we use for obtaining the Geometric Mean (GM) is:

$$GM = \exp\left(\frac{\sum_{1\leq i\leq n}\log(x_i+0.00001)}{n}\right) - 0.00001\,. \quad (7)$$

☆ Q-measure can also be reduced completely to AveP by setting all gain values to zero [15)].

The 0.00001's are necessary because limiting the ranked output size to $L \leq 1000$ implies that x_i may be zero. Geometric Mean AveP was used at the TREC Robust Track in order to focus on the "hardest" topics. Sakai [18)] used Geometric Mean *Q-measure* for analysing their results at NTCIR-5.

4. Data

Our experiments use four different data sets (i.e., test collections and submitted runs) from the NTCIR CLIR track series [7)]. **Table 1** provides some statistics of the data such as the number of $\mathcal{L}$-relevant documents per topic. 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 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. 1**: 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. Bootstrap Hypothesis Tests

This section describes how existing Bootstrap Hypothesis Tests [3)] can be applied to IR evaluation. Our proposed methods, which will be described in Section 6, use these tests as the basis for comparing the discriminative power of IR metrics.

5.1 Paired Test: One Sample Problem

We first describe the *Paired* Bootstrap Hypothesis Test, which can be used for comparing two IR strategies run against a common test collection. This is similar to the one described earlier by Savoy [20)], except that we use a *Studentised* test statistic to enhance accuracy. It is based on fewer assumptions than standard significance tests such as Paired t- and Wilcoxon tests, is easy to apply, yet has high power.

Let Q be the set of topics provided in the test collection, and let $|Q| = n$. Let $\mathbf{x} =$

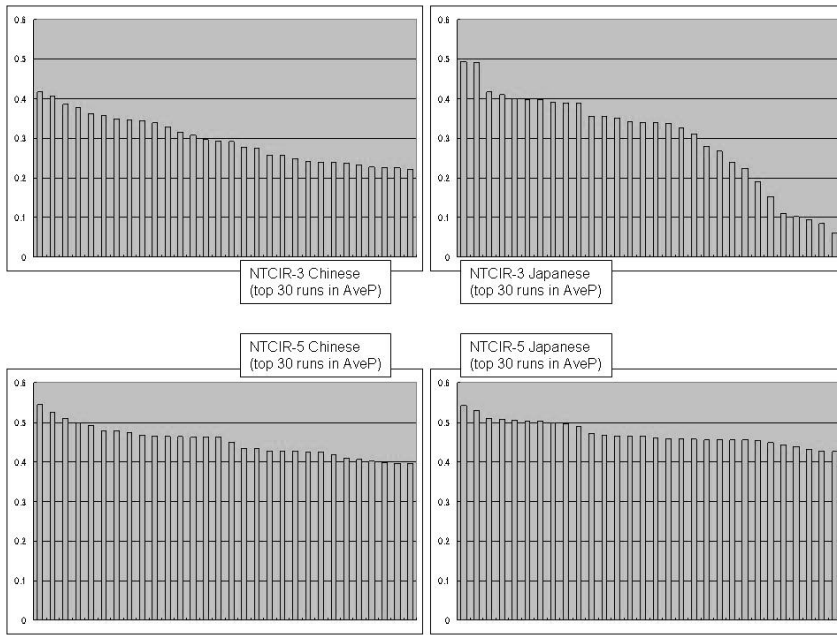

Fig. 1 Distribution of Mean AveP values for the runs used in this study.

$(x_1, \ldots, x_n)$ 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* AveP 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* [3]. 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*. Consider:

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

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* $\mathbf{w}^{*b}$ of per-topic performance differences that obey H_0. **Figure 2** shows the algorithm for obtaining B bootstrap samples of topics (Q^{*b}) and the corresponding values for $\mathbf{w}^{*b}$. (We let $B = 1000$ throughout this paper.) For simplicity, 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 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*b_i = observed value in w for topic q;
```

Fig. 2 Algorithm for creating bootstrap samples Q^{*b} and $\mathbf{w}^{*b} = (w_1^{*b}, \ldots, w_n^{*b})$ for the Paired Test.

For each b, let $\bar{w}^{*b}$ and $\bar{\sigma}^{*b}$ denote the mean and the standard deviation of $\mathbf{w}^{*b}$. **Figure 3** shows how to compute the Achieved Significance Level (ASL) using $\mathbf{w}^{*b}$. In essence, we examine how *rare* the observed difference would

☆ One could also consider a one-tailed test, e.g., $H_0 : \mu > 0$. However, in order to do this, one needs prior knowledge that System X is better than Y. We prefer a two-tailed test, which is more *conservative* (See Section 1).

$count = 0$;
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. 3 Algorithm for estimating the Achieved Significance Level based on the Paired Test.

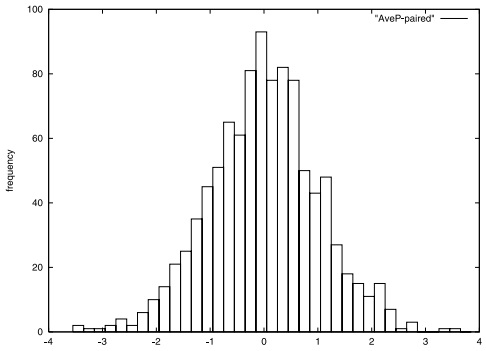

Fig. 4 Histogram of $t(\mathbf{w}^{*b})$ for the difference in Mean AveP between Run 1 and Run 16 from the NTCIR-5 Chinese data.

be under H_0. If $ASL < \alpha$, where typically $\alpha = 0.01$ (very strong evidence against H_0) or $\alpha = 0.05$ (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.

Figure 4 shows a histogram of 1000 *bootstrap replicates* $t(\mathbf{w}^{*b})$ for the difference in Mean AveP between the top run and a "median" (16th-best) run from the NTCIR-5 Chinese data. It can be observed that the graph looks quite *normal*. The observed value $t(\mathbf{z})$ for this run pair is 5.373, which is greater than any value of $t(\mathbf{w}^{*b})$ represented in this figure. Therefore, according to the algorithm in Fig. 3, the estimated ASL is 0.000. That is, it is extremely unlikely that the observed difference is simply due to chance. **Figure 5** shows a similar histogram for the same run pair, but for Mean PDoc_{1000}. The observed value $t(\mathbf{z})$ for this run pair is -0.633, and small values like this occur quite frequently according to Fig. 5. Hence, according to the algorithm in Fig. 3, the estimated ASL is 0.544. To sum up, the two runs are significantly different according to AveP, but not according to PDoc_{1000}. It can be observed that the AveP distribution is sharper than the PDoc_{1000} one, which gives the significance test with AveP higher *power* than that with PDoc_{1000}.

The Paired Test described above relies on the

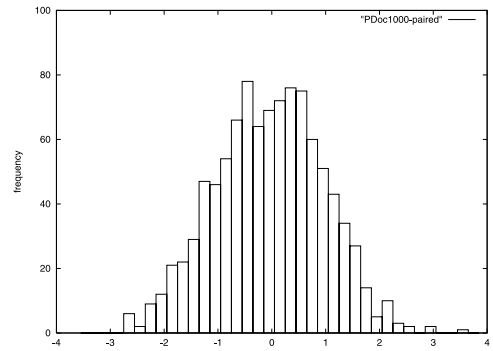

Fig. 5 Histogram of $t(\mathbf{w}^{*b})$ for the difference in Mean PDoc_{1000} between Run 1 and Run 16 from the NTCIR-5 Chinese data.

fact that the difference between two Arithmetic Means equals the Arithmetic Mean of individual differences. But then how should we discuss statistical significance in terms of *Geometric* Mean AveP/Q-measure?

There are at least two ways to handle the problem: One way is to use the *Unpaired* Bootstrap Hypothesis Test instead, as we shall describe in Section 5.2. Unlike the Paired Test, the Unpaired Test is directly applicable to virtually *any* metric, such as the "area" measure based on the worst N topics for each system [21]. Alternatively, we can stick to the Paired Test. Thus, instead of examining $z_i = x_i - y_i$ as mentioned earlier, we could examine $\log(x_i + 0.00001) - \log(y_i + 0.00001)$. This is because testing the significance in terms of the Arithmetic Mean inside Eq. (7) should be equivalent to testing that in terms of the entire Geometric Mean formula. For convenience, "Arithmetic Mean inside the Geometric Mean" will be denoted by the prefix "AG_": In Section 6.2, we test AG_AveP and AG_Q-measure to discuss the discriminative power of G_AveP and G_Q-measure.

5.2 Unpaired Test: Two Sample Problem

As mentioned above, the *Unpaired* Bootstrap Hypothesis Test is more widely applicable than the Paired one, and it can handle Geometric Means directly. The downside is that the Unpaired Test has much less *power* than the Paired one since it uses less information. To be more specific, the Unpaired Test does not utilise the fact that the performance values x_i and y_i (See Section 5.1) correspond to each other for each topic. For this reason, the Paired Test should be preferred wherever it is applicable.

The Unpaired Test treats $\mathbf{x}$ and $\mathbf{y}$ as unpaired

```
let v = (x_1, ..., x_n, y_1, ..., y_m);
for b = 1 to B
    from a set of integers (1, ..., n + m),
    obtain a random sample of size n + m
    by sampling with replacement;
    for i = 1 to n
        j = i-th element of the sample of integers;
        x*b_i = j-th element of v;
    for i = n + 1 to n + m
        j = i-th element of the sample of integers;
        y*b_{i-n} = j-th element of v;
```

Fig. 6 Algorithm for creating bootstrap samples $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ and $\mathbf{y}^{*b} = (y_1^{*b}, \ldots, y_m^{*b})$ for the Unpaired Test. (We let $m = n$ throughout this paper.)

data, naturally. (In general, the two sets of observations may come from different test collections, hence $|\mathbf{x}|$ and $|\mathbf{y}|$ may differ. For this reason, $|\mathbf{x}|$ and $|\mathbf{y}|$ are denoted by n and m, respectively. However, since we are dealing with data obtained by running two IR strategies against a *common* test collection, $|\mathbf{x}| = |\mathbf{y}| = n$ holds in our case.) As with standard significance tests, we assume that $\mathbf{x}$ and $\mathbf{y}$ are independently and identically distributed samples from unknown distributions F and G, respectively. The test statistic we consider in this case is

$$\hat{d} = M(\mathbf{x}) - M(\mathbf{y}) \tag{8}$$

where, for example, $M(\mathbf{x})$ is the value of metric M computed based on $\mathbf{x}$. (Note that M does not have to be an Arithmetic Mean metric.) But what we really want to know is d, which represents the "true" absolute performance difference between Systems X and Y when the whole population of topics is taken into account. Thus the hypotheses we can set up for a two-tailed test are:

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

We now need a null distribution for the data under H_0. A natural choice would be to assume that $F = G$, i.e., that the observed values x_i and y_i actually come from an identical distribution. (In fact, $F = G$ itself is commonly used as the null hypothesis.) First, let $\mathbf{v}$ denote a vector of size $n+m$ obtained by concatenating the two per-topic performance vectors $\mathbf{x}$ and $\mathbf{y}$. **Figure 6** shows how to generate bootstrap samples for the Unpaired Test. For simplicity, let us assume that $\mathbf{x} = (0.1, 0.3)$ and $\mathbf{y} = (0.2, 0.0)$, and therefore that $\mathbf{v} = (0.1, 0.3, 0.2, 0.0)$. Then we generate random integers that range between 1 and 4: Suppose that we have obtained (1,4,1,2) for $b = 1$. Then, by splitting this vector into (1,4) and (1,2), we obtain $\mathbf{x}^{*1} = (0.1, 0.0)$ and $\mathbf{y}^{*1} = (0.1, 0.3)$. In this way, Fig. 6 shuffles the observed values without looking at whether they come from $\mathbf{x}$ or $\mathbf{y}$.

```
count = 0;
for b = 1 to B
    d*b = M(x*b) − M(y*b);
    if( |d*b| ≥ |d̂| ) then count++;
ASL = count/B;
```

Fig. 7 Algorithm for estimating the Achieved Significance Level based on the Unpaired Test.

Figure 7 shows how to compute the two-tailed ASL based on the Unpaired Test, in a way similar to Fig. 3.

6. Discriminative Power Comparison Using Bootstrap Sensitivity Methods

6.1 Proposed Methods

We now propose straightforward methods for assessing and comparing the discriminative power of IR effectiveness metrics, which we call the Bootstrap Sensitivity Methods. The idea is simple: Perform a Bootstrap Hypothesis Test for *every* system pair, and count how many of the pairs satisfy $ASL < \alpha$. That is, we compare the sensitivity of metrics while holding the probability of Type I error (α) constant. Moreover, since each *bootstrap replicate* of the difference between two summary statistics (i.e., $\bar{w}^{*b}$ for the Paired Test and d^{*b} for the Unpaired Test) is derived from exactly $n = |Q|$ topics, we can obtain a natural estimate of the overall performance difference required for guaranteeing $ASL < \alpha$, given n. This may be useful for informally guessing whether two systems are significantly different by just looking at the difference between two summary statistics.

Let $\bar{w}^{*b}_{X,Y}$ and $d^{*b}_{X,Y}$ explicitly denote the above bootstrap replicates for a particular system pair (X, Y). **Figures 8** and **9** show our algorithms for estimating the overall performance difference required for achieving $ASL < \alpha$ given n, based on the Paired Test and the Unpaired Test, respectively. For example, if $\alpha = 0.05$ is chosen for Fig. 9, the algorithm obtains the $B\alpha = 1000 * 0.05 = 50$-th largest value among $|d^{*b}_{X,Y}|$ for each (X, Y). Among the $|C| = 435$ values thus obtained, the algorithm takes the maximum value just to be conservative. Fig. 8 is almost identical to Fig. 9, although it looks slightly more complicated: Since we used Studentisation with the Paired Test, the bootstrap

$DIFF = \phi$;
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. 8 Algorithm for estimating the overall performance difference required for achieving a given significance level with the Paired Test.

$DIFF = \phi$;
for each system pair $(X, Y) \in C$
  sort $|d^{*1}_{X,Y}|, \ldots, |d^{*B}_{X,Y}|$ and
  add the $B\alpha$-th largest value to $DIFF$;
$estimated\_diff = \max\{diff \in DIFF\}$
(rounded to two significant figures);

Fig. 9 Algorithm for estimating the overall performance difference required for achieving a given significance level with the Unpaired Test.

replicate $\bar{w}^{*b}_{X,Y}$ is not equal to the test statistic $t(\mathbf{w}^{*b}_{X,Y})$ (See Fig. 3) which we are using as the sort key .

6.2 Experimental Results

Figures 10, **11**, **12** and **13** plot, for each IR metric, the Paired/Unpaired Bootstrap ASLs of all system pairs for the NTCIR-3 Chinese/Japanese data. (Similar graphs for the NTCIR-5 data are omitted due to space limitation.) The horizontal axis represents 435 system pairs sorted by the ASL. As the curves may be difficult for the reader to distinguish from one another, below we report the results in words, using the symbol "$\geq$" to represent the relationship "is at least as sensitive as".

- Figures 10, 11 and 12 all agree that "Q-measure, nDCG_{1000}, AveP $\geq$ nCG_{1000}, AG_AveP, AG_Q-measure $\geq$ PDoc_{1000}". (In Fig. 11, PDoc_{1000} fails to achieve $ASL < \alpha = 0.05$ for all system pairs.)
- Figure 13 suggests that PDoc_{1000} is far less sensitive than the other six metrics, which does not contradict with our first observation.

The ASL curves for the NTCIR-5 data, not shown here, generally agree with the above trends, except that G_Q-measure and G_AveP appear to be more sensitive than Q-measure and AveP according to the Unpaired Tests with the NTCIR-5 Japanese data. We will discuss this anomaly later.

Table 2 was obtained by cutting the Paired Test ASL curves (Fig. 10, Fig. 12, and similar graphs for the NTCIR-5 data not shown in this paper) in half at $ASL < \alpha = 0.05$. For each data set, the metrics have been sorted by Column (ii), which represents our proposed sensitivity criterion based on the Paired Bootstrap Hypothesis Tests. Column (iii) shows, for each metric, the estimated difference required for satisfying $ASL < \alpha = 0.05$, given the topic set size. The algorithm shown in Fig. 8 was used to obtain these estimates. For example, the first entry in Table 2 (a) shows that Q-measure managed to detect a significant difference for 242 system pairs out of 435 NTCIR-3 Chinese runs (i.e., 56%) and therefore is the most sensitive metric for this data set, and that an overall difference of around 0.10 is required to achieve $ASL < \alpha = 0.05$ using 42 topics. For AG_AveP and AG_Q-measure, the estimates represent differences between two Arithmetic Means of *logs* rather than differences between two Geometric Means (See Section 5.1).

Again, Table 2 (a)-(c) all suggest that Q-measure, nDCG_{1000} and AveP are the most sensitive metrics, that PDoc_{1000} is the least sensitive metric, and that nCG_{1000}, G_AveP and G_Q-measure lie somewhere in the middle. Table 2 (d), which represents the NTCIR-5 Japanese results, agrees with this trend except that nCG_{1000} appears to be exceptionally insensitive.

It is clear that the Bootstrap Sensitivity values are heavily dependent on the distribution of the run performances. For example, the sensitivity values for the NTCIR-5 Japanese data (Table 2 (d)) are much lower than those for the NTCIR-3 Japanese data (Table 2 (b)), reflecting how similar the run performances are for the NTCIR-5 Japanese data (See Fig. 1). On the other hand, the estimated overall difference required for guaranteeing $ASL < \alpha = 0.05$ appears to be relatively stable for a given topic set size: For example, the overall difference required in terms of AveP for the NTCIR-3 Chinese data and that for the NTCIR-3 Japanese data are both 0.11 (Compare Table 2 (a) and (b)). Note, however, that Column (iii) is not for comparing the discriminative power of different IR metrics: some metrics tend to take small values while others tend to take large values, so such comparisons are not necessarily valid.

Similarly, **Table 3** (a)-(d) were obtained by cutting the Unpaired Test ASL curves (Fig. 11,

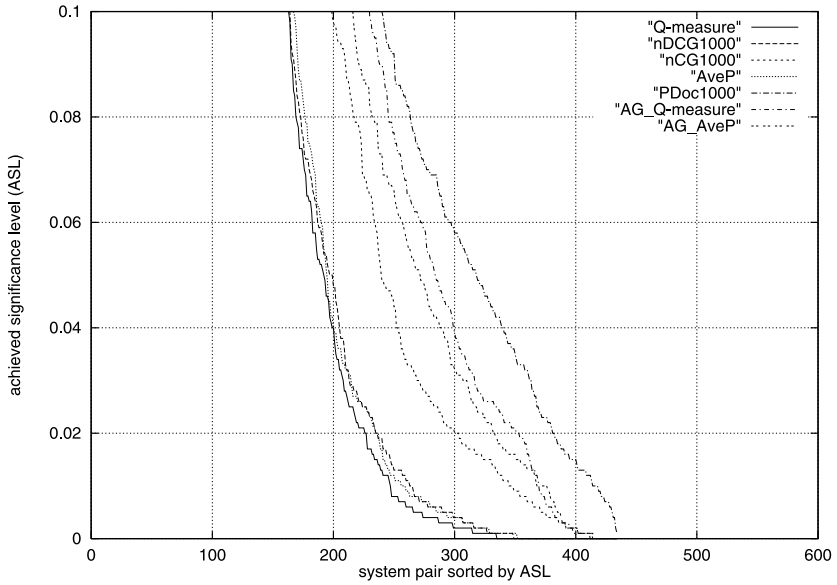

Fig. 10 Paired Test ASL curves for the NTCIR-3 Chinese data.

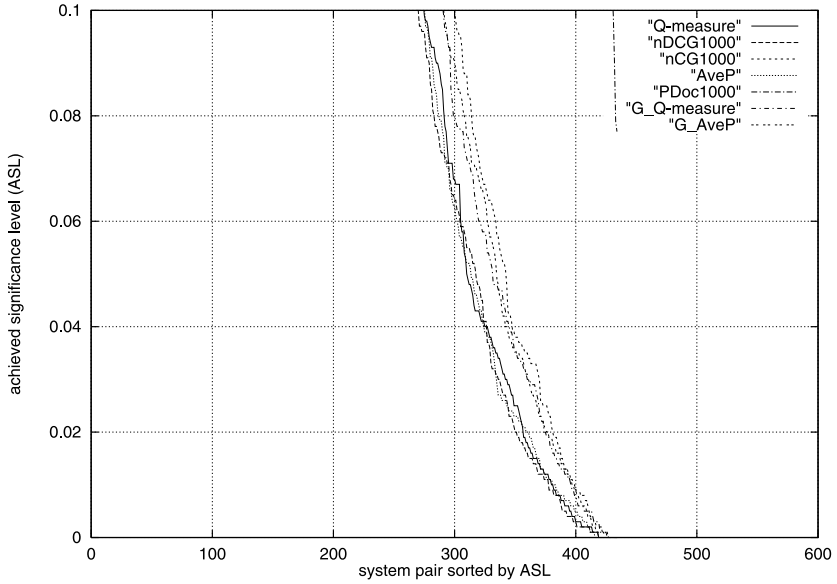

Fig. 11 Unpaired Test ASL curves for the NTCIR-3 Chinese data.

Fig. 13, and similar graphs for the NTCIR-5 data not shown in this paper) in half at $ASL < \alpha = 0.05$. It can be observed that, because the Unpaired Test has considerably less power than the Paired one no matter what metric is used, our Unpaired Bootstrap Sensitivity Method is less useful than the Paired one for discussing which metrics are more sensitive than others. Note that the sensitivity values in Column (ii) are naturally much lower compared to the corresponding Paired Test values. On the other hand, the advantage of using the Unpaired version of our method is that it can directly estimate the overall performance difference required even for non-Arithmetic Mean metrics such as G_AveP and G_Q-measure. For example, Table 3 (c) shows that, for the NTCIR-5 Chinese runs, a difference of 0.23 is required in terms of G_AveP, which is much larger than 0.12, the difference required in terms of Arithmetic Mean AveP.

Figures 14 and **15** summarise the general trends of our Paired and Unpaired Test results, by visualising all the sensitivity values

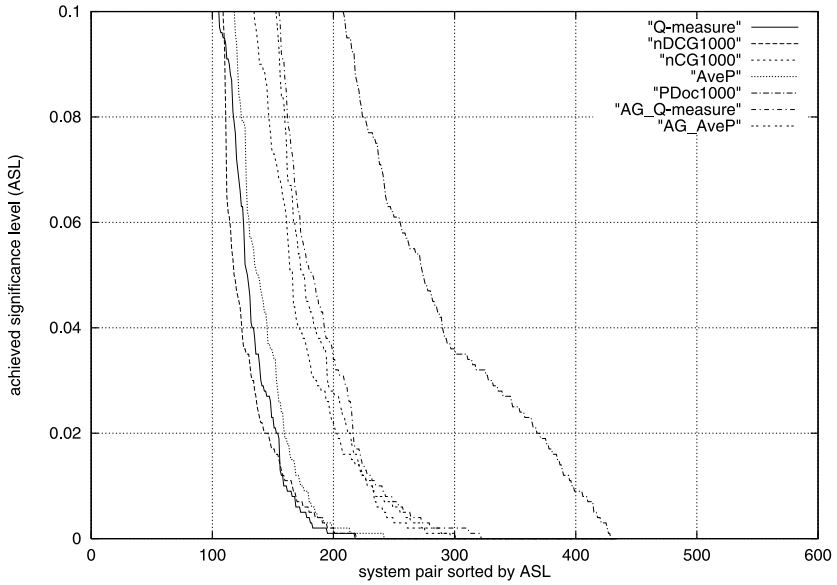

Fig. 12 Paired Test ASL curves for the NTCIR-3 Japanese data.

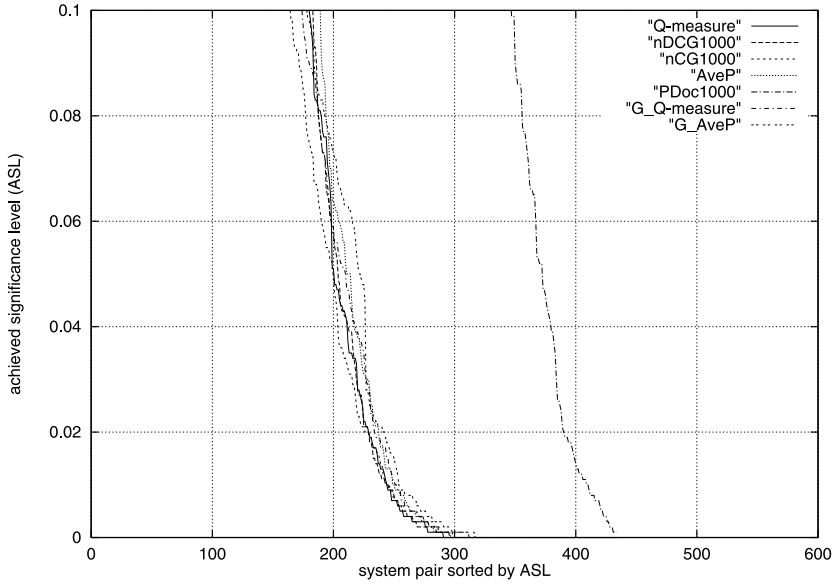

Fig. 13 Unpaired Test ASL curves for the NTCIR-3 CLIR Japanese data.

shown in Column (ii) of Table 2 and Table 3. Again, the Paired Test results in Fig. 14 suggest that Q-measure, $nDCG_{1000}$ and AveP are the most sensitive metrics while $PDoc_{1000}$ is the least sensitive one (although nCG_{1000} is exceptionally insensitive for the NTCIR-5 Japanese data). Whereas, the Unpaired Test results in Fig. 15 just suggest that $PDoc_{1000}$ is less sensitive than others. Thus, as mentioned earlier, the Unpaired Bootstrap Sensitivity Method is less useful than the Paired one for comparing metrics.

We now discuss the anomalous results with the NTCIR-5 Japanese data represented at the very front of Fig. 15, which show that G_Q-measure and G_AveP are exceptionally sensitive for this data set. While we do not have a good explanation for the discrepancy between the Paired Test results and the Upaired Test ones, we conjecture that the NTCIR-5 Japanese runs themselves have some characteristics that are different from the other run sets. To begin with, the run performances are very similar, as Fig. 1 shows. Moreover, the fact that the

Table 2 Results based on the Paired Bootstrap Sensitivity Method ($\alpha = 0.05$; NTCIR-3 and NTCIR-5 CLIR data).

(i) metric	(ii) sensitivity ($ASL < \alpha$)	(iii) estimated difference
(a) NTCIR-3 Chinese (42 topics)		
Q-measure	242/435 = **56**%	0.10
AveP	240/435 = **55**%	0.11
$nDCG_{1000}$	235/435 = **54**%	0.13
nCG_{1000}	195/435 = **45**%	0.15
AG_AveP	163/435 = **37**%	1.42
AG_Q-measure	150/435 = **34**%	1.76
$PDoc_{1000}$	116/435 = **27**%	0.02
(b) NTCIR-3 Japanese (42 topics)		
$nDCG_{1000}$	316/435 = **73**%	0.14
Q-measure	305/435 = **70**%	0.13
AveP	296/435 = **68**%	0.11
nCG_{1000}	268/435 = **62**%	0.18
AG_AveP	258/435 = **59**%	1.62
AG_Q-measure	251/435 = **58**%	1.74
$PDoc_{1000}$	160/435 = **37**%	0.04
(c) NTCIR-5 Chinese (50 topics)		
Q-measure	174/435 = **40**%	0.11
$nDCG_{1000}$	163/435 = **37**%	0.10
AveP	159/435 = **37**%	0.11
nCG_{1000}	92/435 = **21**%	0.17
AG_AveP	65/435 = **15**%	1.14
AG_Q-measure	64/435 = **15**%	1.02
$PDoc_{1000}$	51/435 = **12**%	0.01
(d) NTCIR-5 Japanese (47 topics)		
Q-measure	136/435 = **31**%	0.09
$nDCG_{1000}$	120/435 = **28**%	0.13
AveP	113/435 = **26**%	0.10
AG_AveP	93/435 = **21**%	0.59
AG_Q-measure	90/435 = **21**%	0.57
$PDoc_{1000}$	53/435 = **12**%	0.01
nCG_{1000}	35/435 = **8**%	0.12

Table 3 Results based on the Unpaired Bootstrap Sensitivity Method ($\alpha = 0.05$; NTCIR-3 and NTCIR-5 CLIR data).

(i) metric	(ii) sensitivity ($ASL < \alpha$)	(iii) estimated difference
(a) NTCIR-3 Chinese (42 topics)		
Q-measure	124/435 = **29**%	0.12
AveP	121/435 = **28**%	0.12
$nDCG_{1000}$	117/435 = **27**%	0.13
G_Q-measure	103/435 = **24**%	0.17
G_AveP	100/435 = **23**%	0.16
nCG_{1000}	92/435 = **21**%	0.15
$PDoc_{1000}$	0/435 = **0**%	0.03
(b) NTCIR-3 Japanese (42 topics)		
nCG_{1000}	235/435 = **54**%	0.19
Q-measure	234/435 = **54**%	0.13
$nDCG_{1000}$	231/435 = **53**%	0.15
G_Q-measure	224/435 = **51**%	0.20
AveP	220/435 = **51**%	0.14
G_AveP	212/435 = **49**%	0.19
$PDoc_{1000}$	62/435 = **14**%	0.03
(c) NTCIR-5 Chinese (50 topics)		
$nDCG_{1000}$	44/435 = **10**%	0.11
nCG_{1000}	44/435 = **10**%	0.11
Q-measure	43/435 = **10**%	0.11
G_Q-measure	43/435 = **10**%	0.25
G_AveP	39/435 = **9**%	0.23
AveP	35/435 = **8**%	0.12
$PDoc_{1000}$	0/435 = **0**%	0.03
(d) NTCIR-5 Japanese (47 topics)		
G_Q-measure	55/435 = **13**%	0.19
G_AveP	49/435 = **11**%	0.20
$nDCG_{1000}$	34/435 = **8**%	0.10
nCG_{1000}	17/435 = **4**%	0.09
Q-measure	14/435 = **3**%	0.11
AveP	7/435 = **2**%	0.12
$PDoc_{1000}$	0/435 = **0**%	0.04

Geometric Means do well suggests that some "hard" topics (i.e., those with very low performances) are playing an important role in system discrimination [11),18),21)]. Note also that, in Fig. 15, $nDCG_{1000}$ does relatively well for the NTCIR-5 Japanese data, though not as well as G_Q-measure and G_AveP. This may be because $nDCG_{1000}$ is relatively highly correlated with G_Q-measure and G_AveP, as is discussed in the Appendix.

7. Comparison with Stability and Swap Methods

Readers familiar with the *Stability Method* proposed at ACM SIGIR 2000 [1)] and the Swap Method proposed at SIGIR 2002 [22)] will note that our Bootstrap Sensitivity Methods are related to them. The crucial difference is that our methods are based on the bootstrap which has time-honoured theoretical foundations. At the implementation level, the main difference is that the original Stability and Swap Methods use sampling *without* replacement whereas the bootstrap samples are obtained by sampling *with* replacement. However, as was mentioned in Section 2, Sakai [12)] showed that sampling with and without replacement yield similar results for the purpose of ranking metrics according to discriminative power. In light of this, we conduct Stability and Swap experiments by reusing our Paired Test Bootstrap samples Q^{*b}, which have been sampled with re-

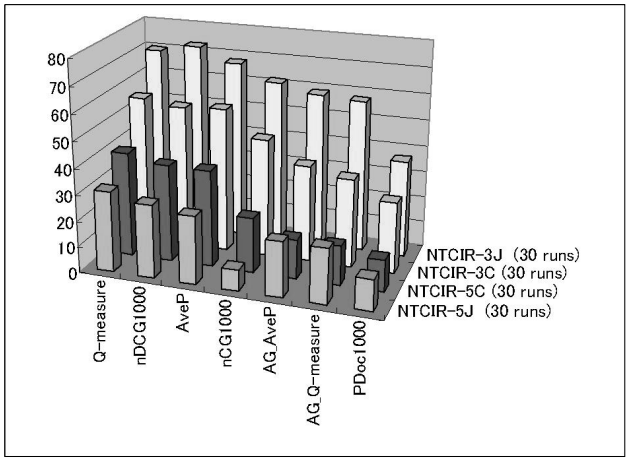

Fig. 14 Summary of Paired Bootstrap Sensitivity results.

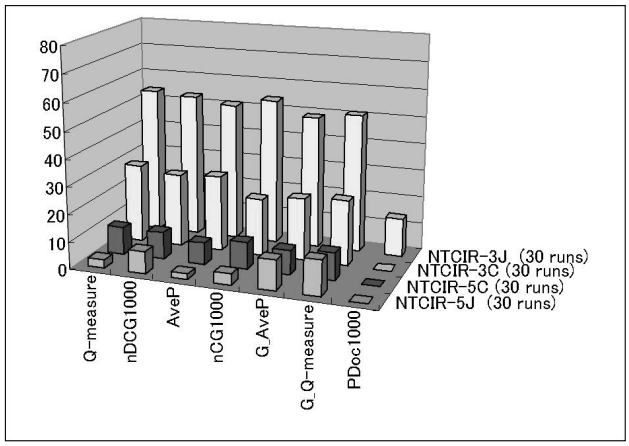

Fig. 15 Summary of Unpaired Bootstrap Sensitivity results.

placement from the original topic set Q. Recall that these samples are equal in size to the original topic set Q.

7.1 Comparison with the Stability Method

The essence of the Stability Method is to compare systems X and Y in terms of metric M using B different topic sets and count how often X outperforms Y, how often Y outperforms X and how often the two are regarded as equivalent. Our version works as follows: Let $M(X, Q^{*b})$ denote the value of metric M for System X computed based on Q^{*b}. Given a *fuzziness value* f [1], we count $GT(X,Y)$, $GT(Y,X)$ and $EQ(X,Y)$ as shown in **Fig. 16**. From the algorithm, it is clear that $GT(X,Y)+GT(Y,X)+EQ(X,Y) = B$, where $GT(X,Y)$ is the number of times System X outperforms System Y, and $EQ(X,Y)$ is the number of times System X and System Y are

```
for each system pair (X, Y) ∈ C
    for b = 1 to B
        margin = f * max(M(X, Q*b), M(Y, Q*b));
        if( |M(X, Q*b) − M(Y, Q*b)| < margin )
            EQ(X, Y) + +
        else if( M(X, Q*b) > M(Y, Q*b) )
            GT(X, Y) + +
        else
            GT(Y, X) + +;
```

Fig. 16 Algorithm for computing $EQ(X,Y)$, $GT(X,Y)$ and $GT(Y,X)$.

"almost" equal, where "almost" is defined by the fuzziness value. Then, the *minority rate* (MR) and the *proportion of ties* (PT) for M are computed as:

$$MR = \frac{\sum_C \min(GT(X,Y), GT(Y,X))}{B \sum_C} . \quad (9)$$

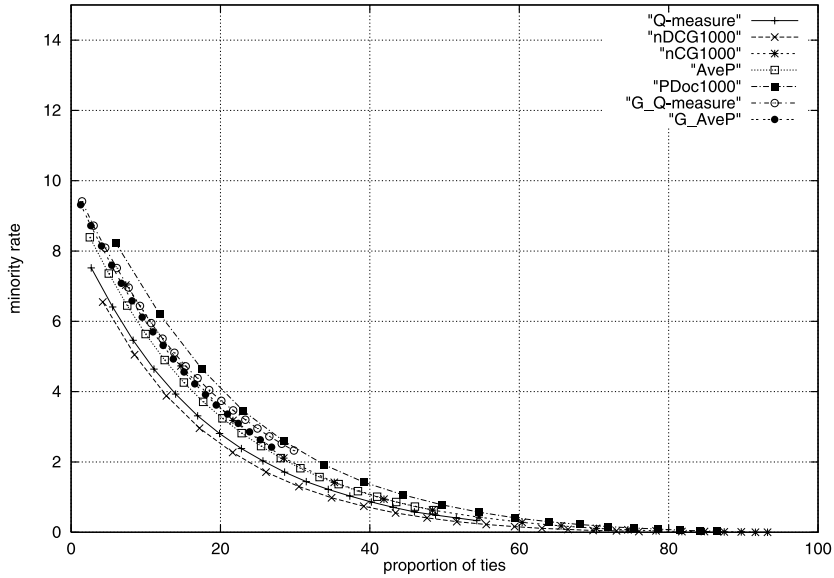

Fig. 17 MR-PT curves based on 30 runs for the NTCIR-3 Chinese data.

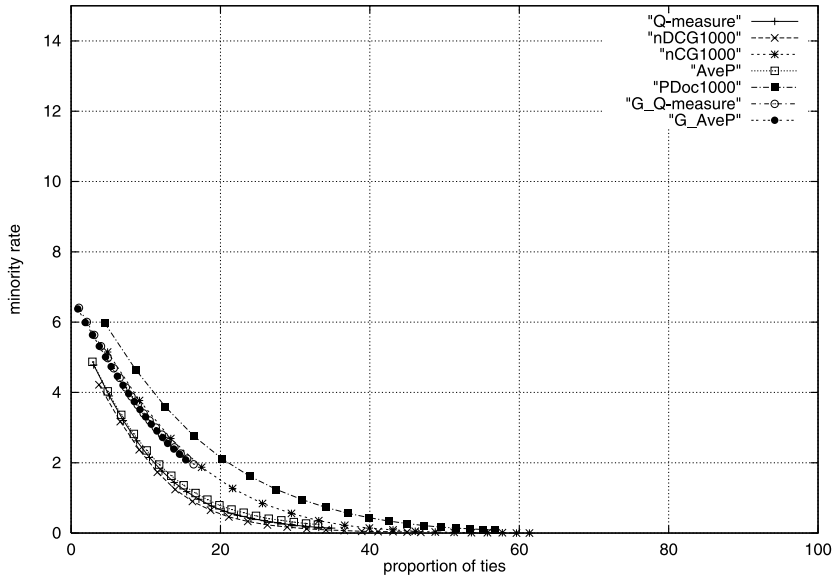

Fig. 18 MR-PT curves based on 30 runs for the NTCIR-3 Japanese data.

$$PT = \frac{\sum_C EQ(X,Y)}{B\sum_C} \ . \qquad (10)$$

MR estimates the chance of reaching a wrong conclusion about a system pair, while *PT* reflects lack of discriminative power. Thus, for a good performance metric, both of these values should be small. As a fixed fuzziness value implies different trade-offs for different metrics, we vary f $(= 0.01, 0.02, \ldots, 0.20)$ for comparing the stability. We refer to the trade-off curves as *MR-PT curves* [12),17)].

Figures 17 and **18** show the *MR-PT* curves for the NTCIR3 Chinese and Japanese data, respectively. (Similar graphs for the NTCIR-5 data are omitted due to space limitation.) It can be observed that the Stability Method results are consistent with the Bootstrap Sensitivity results which suggested that "Q-measure, nDCG_{1000}, AveP $\geq$ nCG_{1000}, G_AveP, G_Q-measure $\geq$ PDoc_{1000}". The two methods agree with each other for the NTCIR-5 data as well: they even agree that G_AveP, G_Q-measure do exceptionally well for the NTCIR-5 Japanese data.

```
for each system pair (X, Y) ∈ 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);
        counter(BIN(D*b)) + +;
        if( D*b ∗ D'*b > 0 ) then
            continue
        else
            swap_counter(BIN(D*b)) + +;
for each bin i
    swap_rate(i) = swap_counter(i)/counter(i);
```

Fig. 19 Algorithm for computing the swap rates.

7.2 Comparison with the Swap Method

As was mentioned in Section 2, the essence of the Swap Method is to estimate the *swap rate*, which represents the probability of the event that two experiments are contradictory given an overall performance difference. Our version works as follows: First, in addition to the set of B bootstrap samples $\{Q^{*b}\}$, we create *another* set of B bootstrap samples $\{Q'^{*b}\}$ by sampling with replacement from Q. Let D denote the overall performance difference between two systems as measured by M based on a topic set; we prepare 21 *performance difference bins* [22], 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 the mapping from a difference D to one of the 21 bins where it belongs. The algorithm shown in **Fig. 19** calculates a *swap rate* for each bin. Note that D^{*b} is not the same as our d^{*b} from Fig. 7: D^{*b} is the overall performance difference between X and Y as measured using the bootstrap topic sample Q^{*b}; whereas, d^{*b} is the bootstrap replicate of the observed overall performance difference *under the assumption that the per-topic values of X and Y come from an identical distribution*.

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": But, as mentioned earlier, 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 directly from the probability of Type I error.

Table 4 Swap Method results (swap rate $\leq$ 5%; NTCIR-3 and NTCIR-5 CLIR Chinese and Japanese data).

(i) metric	(ii) diff.	(iii) max.	(ii)/(iii)	%pairs satisfying (ii)
(a) NTCIR-3 Chinese (42 topics)				
$nDCG_{1000}$	0.07	.7414	9%	**47%**
Q-measure	0.07	.5374	13%	**43%**
AveP	0.08	.5295	15%	**40%**
nCG_{1000}	0.08	.9514	8%	**35%**
G_AveP	0.09	.4739	18%	**33%**
G_Q -measure	0.10	.4967	20%	**33%**
$PDoc_{1000}$	0.01	.0983	10%	**20%**
(b) NTCIR-3 Japanese (42 topics)				
$nDCG_{1000}$	0.07	.7994	9%	**69%**
Q-measure	0.07	.6433	11%	**67%**
AveP	0.07	.6449	11%	**66%**
nCG_{1000}	0.10	.9913	10%	**56%**
G_AveP	0.10	.5699	18%	**54%**
G_Q -measure	0.12	.5981	20%	**53%**
$PDoc_{1000}$	0.01	.0982	10%	**41%**
(c) NTCIR-5 Chinese (50 topics)				
Q-measure	0.07	.6757	10%	**26%**
AveP	0.07	.6480	11%	**26%**
$nDCG_{1000}$	0.07	.8334	8%	**23%**
G_AveP	0.14	.6065	23%	**23%**
G_Q -measure	0.15	.6501	23%	**22%**
nCG_{1000}	0.07	.9927	7%	**16%**
$PDoc_{1000}$	0.01	.0830	12%	**0.4%**
(d) NTCIR-5 Japanese (47 topics)				
G_AveP	0.13	.6203	21%	**17%**
Q-measure	0.07	.6652	11%	**16%**
G_Q -measure	0.13	.6480	20%	**15%**
$nDCG_{1000}$	0.07	.8389	8%	**15%**
AveP	0.08	.6438	12%	**11%**
nCG_{1000}	0.07	.9973	7%	**7%**
$PDoc_{1000}$	0.01	.1324	8%	**0.1%**

Table 4 summarises the results of our swap experiments for the four data sets. Column (ii) shows the absolute difference required for guaranteeing that the swap rate does not exceed 5%. The column labelled with "(ii)/(iii)" translates these values into relative values using the maximum performance recorded among all trials (Column (iii)). The last column, by which the IR metrics have been sorted, shows the

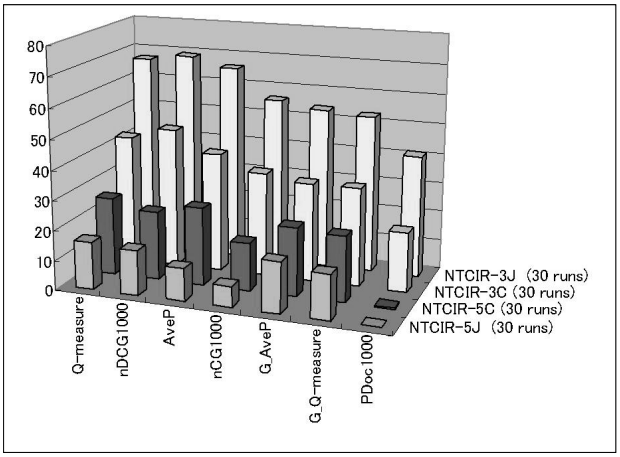

Fig. 20 Summary of Swap Method results.

percentage of comparisons (among the total of 435*1000 comparisons) that actually satisfied the difference threshold shown in Column (ii). **Figure 20** visualises the last column of Table 4.

Again, the results are generally consistent with the Bootstrap Sensitivity ones: According to the Swap Method, nDCG_{1000}, Q-measure and AveP are the best metrics and PDoc_{1000} is the worst metric for the NTCIR-3 Chinese/Japanese data and the NTCIR-5 Chinese data. In the middle lie nCG_{1000}, G_AveP and G_Q-measure. (The results generalise those by Voorhees [21] who compared AveP and G_AveP.) Yet again, the NTCIR-5 Japanese results (Table 4 (d)) look somewhat anomalous, in that G_AveP and G_Q-measure do as well as their Arithmetic Mean counterparts. As mentioned earlier, this suggests that "hard" topics (i.e., those with low performance values) are playing an important role for discriminating some of the NTCIR-5 Japanese runs from others. To sum up, the Swap Method results are generally consistent with both the Stability Method and the Bootstrap Sensitivity results.

Finally, note that the estimated overall performance differences for guaranteeing 5% swap rate or less are lower than those required for achieving $ASL < \alpha = 0.05$ with the Bootstrap Hypothesis Tests. For example, Table 4 (c) shows that, given 50 topics, the overall performance difference in Q-measure (or AveP) required for guaranteeing 5% swap rate or less is 0.07. Whereas, the Paired Test result in Table 2 (c) and the Unpaired one in Table 3 (c) agree that, under the same circumstance, the estimated difference in Q-measure (or AveP) required for achieving $ASL < \alpha = 0.05$ is 0.11. Thus the requirement of $ASL < \alpha = 0.05$ based on our Bootstrap Sensitivity Methods is more demanding.

8. Conclusions and Future Work

This paper showed that Bootstrap Hypothesis Tests are useful not only for comparing IR strategies, but also for comparing the discriminative power of IR metrics. The Paired Bootstrap Test is directly applicable to any Arithmetic Mean metric. The Unpaired Bootstrap Test has less power, but is directly applicable even to unconventional metrics. Our experiments with the NTCIR-3 Chinese/Japanese and the NTCIR-5 Chinese data showed that Q-measure, nDCG_{1000} and AveP are all very sensitive metrics; PDoc_{1000} is naturally extremely insensitive; and that nCG_{1000} and Geometric Mean AveP and Geometric Mean Q-measure lie in the middle. Whereas, for the NTCIR-5 Japanese data, the Geometric Mean metrics appear to do at least as well as the Arithmetic Mean ones, possibly because some difficult topics are playing an important role for system discrimination for this particular data set. More importantly, however, these Bootstrap Sensitivity results are generally consistent with those based on the somewhat ad hoc Stability and Swap Methods.

Finally, it should be noted that the bootstrap is *not* assumption-free: the most basic assumption that it relies on is that the original topics of the test collection are indepent and identically distributed samples from the population P. We are aware that not all IR researchers are happy even with this assumption [2]. Moreover, the bootstrap is known to fail when the empir-

ical distribution based on the observed data is a poor approximation of the true distribution. Clarifying the limitations of our approach will be one of the subjects of our future work.

References

1) Buckley, C. and Voorhees, E.M.: Evaluating Evaluation Measure Stability, *Proc. ACM SIGIR 2000*, pp.33–40 (2000).
2) Cormack, G., Lynam, T. and Cheriton, D.: Statistical Precision of Information Retrieval Evaluation, *Proc. ACM SIGIR 2006*, pp.533–540 (2006).
3) Efron, B. and Tibshirani, R.: *An Introduction to the Bootstrap*, Chapman & Hall/CRC (1993).
4) Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments, *Proc. ACM SIGIR '93*, pp.329–338 (1993).
5) Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (2002).
6) Johnson, D.H.: The Insignificance of Statistical Significance Testing, *Journal of Wildlife Management*, Vol.63, No.3, pp.763–772 (1999).
7) Kando, N.: Overview of the Fifth NTCIR Workshop, *Proc. NTCIR-5* (2005).
8) Kazai, G. and Lalmas, M.: eXtended Cumulated Gain Measures for the Evaluation of Content-oriented XML Retrieval, *ACM Trans. Inf. Syst.*, Vol.24, No.4, pp. 503–542 (2006).
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) Lin, W.-H. and Hauptmann, A.: Revisiting the Effect of Topic Set Size on Retrieval Error, *Proc. ACM SIGIR 2005*, pp.637–638 (2005).
11) Robertson, S.: On GMAP — And Other Transformations, *Proc. ACM CIKM 2006*, pp.78–83 (2006).
12) Sakai, T.: The Effect of Topic Sampling on Sensitivity Comparisons of Information Retrieval Metrics, *Proc. NTCIR-5* (2005).
13) Sakai, T.: Bootstrap-Based Comparisons of IR Metrics for Finding One Relevant Document, *Proc. Asia Information Retrieval Symposium 2006 LNCS 4182*, pp.374–389 (2006).
14) Sakai, T.: Evaluating Evaluation Metrics based on the Bootstrap, *Proc. ACM SIGIR 2006*, pp.525–532 (2006).
15) 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).
16) Sakai, T.: On the Reliability of Factoid Question Answering Evaluation, *ACM Trans. Asian Language Information Processing*, Vol.6, No.1, p.Article 3 (2007).
17) 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).
18) Sakai, T., Manabe, T., Kumano, A., Koyama, M. and Kokubu, T.: Toshiba BRIDJE at NTCIR-5 CLIR: Evaluation using Geometric Means, *Proc. NTCIR-5* (2005).
19) Sanderson, M. and Zobel, J.: Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability, *Proc. ACM SIGIR 2005*, pp.162–169 (2005).
20) Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation, *Information Processing and Management*, Vol.33, No.4, pp.495–512 (1997).
21) Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track, *Proc. TREC 2004* (2005).
22) 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

This paper examined seven IR effectiveness metrics from the viewpoint of discriminative power in order to demonstrate the usefulness of our proposed Bootstrap Sensitivity Methods. However, there are other aspects of IR metrics that need to be examined, including *how the system rankings according to each metric resemble each other*. Following previous work [9),17),21)], we use Kendall's rank correlation for this purpose. For each of the four data sets we used, we use the same 30 runs for computing Kendall's rank correlation τ between two system rankings according to two different IR metrics. Note that Kendall's τ lie between -1 and 1, where the former represents a pair of rankings that are the perfect inverse of each other, and the latter represents two identical rankings.

For Kendall's τ, there is a standard significance test available: Given the number of systems n_s,

$$Z_0 = \frac{|\tau|}{((4n_s + 10)/(9n_s(n_s - 1)))^{\frac{1}{2}}} \quad (11)$$

obeys a normal distribution. Thus, a normal test can easily be applied. Note that the test statistic Z_0 is proportional to $|\tau|$ given n_s: In

Table 5 Kendall's rank correlations based on the top 30 runs from each data set.

NTCIR-3 Chinese	metric	(b)	(c)	(d)	(e)	(f)	(g)
	(a) Q-measure	**.8575**	.7057	**.9678**	.6736	.6966	.7057
	(b) $nDCG_{1000}$	-	.7655	**.8345**	.6874	**.8115**	**.8207**
	(c) nCG_{1000}	-	-	.6920	.8299	.7977	.7793
	(d) AveP	-	-	-	.6782	.6644	.6736
	(e) $PDoc_{1000}$	-	-	-	-	.7103	.7011
	(f) G_Q-measure	-	-	-	-	-	**.9632**
	(g) G_AveP	-	-	-	-	-	-
NTCIR-3 Japanese	metric	(b)	(c)	(d)	(e)	(f)	(g)
	(a) Q-measure	**.9126**	**.8345**	**.9540**	**.8345**	**.8345**	**.8391**
	(b) $nDCG_{1000}$	-	**.8667**	**.8851**	**.8115**	**.9126**	**.9172**
	(c) nCG_{1000}	-	-	**.8345**	.7977	**.8621**	**.8759**
	(d) AveP	-	-	-	**.8345**	**.8069**	**.8299**
	(e) $PDoc_{1000}$	-	-	-	-	.7609	.7563
	(f) G_Q-measure	-	-	-	-	-	**.9586**
	(g) G_AveP	-	-	-	-	-	-
NTCIR-5 Chinese	metric	(b)	(c)	(d)	(e)	(f)	(g)
	(a) Q-measure	**.8621**	.5126	**.9172**	.6368	.5494	.5448
	(b) $nDCG_{1000}$	-	.5402	.7977	.6460	.6414	.6552
	(c) nCG_{1000}	-	-	.4759	.7287	.5494	.5172
	(d) AveP	-	-	-	.5632	.5126	.5264
	(e) $PDoc_{1000}$	-	-	-	-	.4897	.4667
	(f) G_Q-measure	-	-	-	-	-	**.9218**
	(g) G_AveP	-	-	-	-	-	-
NTCIR-5 Japanese	metric	(b)	(c)	(d)	(e)	(f)	(g)
	(a) Q-measure	**.8161**	.5034	**.8851**	.5540	.6276	.6138
	(b) $nDCG_{1000}$	-	.5862	.7563	.5816	.7287	.7241
	(c) nCG_{1000}	-	-	.4529	.6368	.7379	.7149
	(d) AveP	-	-	-	.5494	.5862	.5816
	(e) $PDoc_{1000}$	-	-	-	-	.5402	.5448
	(f) G_Q-measure	-	-	-	-	-	**.9494**
	(g) G_AveP	-	-	-	-	-	-

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. (For an alternative approach to quantifying the accuracy of rank correlation, we refer the reader to [14].)

Table 5 shows Kendall's rank correlations for the four data sets. For example, Row (a) Column (b) represents the correlation between Q-measure and $nDCG_{1000}$. It can be observed that all the values easily exceed 0.34 and therefore are statistically highly significant. For convenience, values over 0.8 are shown in bold, although the choice of this threshold is arbitrary. It can be observed that:

- Q-measure is consistently highly correlated with AveP, and is also highly correlated with $nDCG_{1000}$ [17]. On the other hand, the rank correlation between AveP and $nDCG_{1000}$ is below 0.8 for the NTCIR-5 Chinese and the Japanese data.
- Not surprisingly, G_Q-measure and G_AveP are very highly correlated.
- Somewhat surprisingly, $nDCG_{1000}$ is consistently more highly correlated with G_Q-measure/G_AveP than Q-measure and AveP are.

Our explanation for the above third observation is as follows. Recall that, while $nDCG_{1000}$ is a rank-based metric, Q-measure and AveP are recall-based metrics computed by dividing either the Blended Ratio or Precision by R, the total number of relevant documents (See Eqs. (1) and (5)). This suggests that $nDCG_{1000}$ is more "forgiving" for low-recall topics than Q-measure and AveP. That is, the $nDCG_{1000}$ values for these topics tend to be relatively large, and therefore these topics may have a significant impact on the overall $nDCG_{1000}$ perfor-

mance. This tendency coincides with those of G_Q-measure and G_AveP, because they have been designed to boost the contribution of low-performing (i.e., low-recall) topics to the overall performance. However, it is known that $nDCG_{1000}$ is both counterintuitive and insensitive if a large logarithm base is used [15),17)].

(Received March 8, 2007)
(Accepted May 11, 2007)

(Editor in Charge: *Norihiko Uda*)

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.