Evaluating Evaluation Metrics based on the Bootstrap

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

ABSTRACT

This paper describes how the Bootstrap approach to statistics can be applied to the evaluation of IR effectiveness metrics. First, we argue that Bootstrap Hypothesis Tests deserve more attention from the IR community, as they are based on fewer assumptions than traditional statistical significance tests. We then describe straightforward methods for comparing the sensitivity of IR metrics based on Bootstrap Hypothesis Tests. Unlike the heuristics-based "swap" method proposed by Voorhees and Buckley, our method estimates the performance difference required to achieve a given significance level directly from Bootstrap Hypothesis Test results. In addition, we describe a simple way of examining the accuracy of rank correlation between two metrics based on the Bootstrap Estimate of Standard Error. We demonstrate the usefulness of our methods using test collections and runs from the NTCIR CLIR track for comparing seven IR metrics, including those that can handle graded relevance and those based on the Geometric Mean.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

Test Collection, Evaluation, Bootstrap, Graded Relevance

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 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 [13].

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, 6, 13, 14]. 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 [14] 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 directly 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 [16] and Geometric Means [11, 16]). 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 *sensitivity* of

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.

different IR metrics. This is related to the "swap" method proposed by Voorhees and Buckley [12, 13, 15] (and the "stability" method proposed by Buckley and Voorhees [2, 12]): The swap method derives the performance difference required for declaring 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 (See Section 3.5), our method estimates the performance difference required to achieve a given significance level directly from Bootstrap Hypothesis Test results. In addition, we describe a simple way of examining the accuracy of rank correlation between two metrics based on the Bootstrap Estimate of Standard Error.

To demonstrate the usefulness of our Bootstrap-based methods, we use two test collections with submitted runs from NTCIR [9] and compare seven IR metrics, including those based on graded relevance and those based on the Geometric Mean [11, 16]. The swap, stability and our Bootstrap-based methods agree that, for these data sets, the most sensitive IR metrics are Q-measure [12], normalised Discounted Cumulative Gain at cut-off 1000 [5, 7] 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.

Section 2 describes the IR metrics and the NTCIR data we use. Section 3 discusses our new methods for comparing the sensitivity of IR metrics based on the Bootstrap Hypothesis Tests. Section 4 describes a method for examining the accuracy of rank correlation between two IR metrics based on the Bootstrap Estimate of Standard Error. Section 5 discusses previous work and Section 6 concludes this paper.

2. METRICS AND DATA

2.1 IR Effectiveness Metrics

The IR metrics we consider in this paper are Average Precision (AveP), Precision at cut-off 1000 (PDoc_{1000}), Q-measure [12, 17], and normalised (Discounted) Cumulative Gain at cut-off 1000 ($\mathrm{n(D)CG}_{1000}$) [5, 7].

AveP represents a very sensitive IR metric based on *binary* relevance, while PDoc_{1000} represents a very insensitive one [12]. (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.) Let R denote the number of relevant documents for a topic, and let L (≤ 1000) denote the size of a ranked output. Let $count(r)$ denote the number of relevant documents within top r ($\leq L$), and let $isrel(r)$ be 1 if the document at Rank r is relevant and 0 otherwise. Clearly, Precision at Rank r is given by $P(r) = count(r)/r$. Hence:

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

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

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* for retrieving an $\mathcal{L}$-relevant document [5]. (Throughout this paper, we use $gain(S) = 3, gain(A) = 2, gain(B) = 1$.) 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]. (We use $a = 2$ throughout this paper.) 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 documents after Rank R (as $cg_I(r) = cg_I(R)$ holds for $r \geq R$) [12]. nDCG_l solves this problem by discounting the gains, while Q-measure solves it by including r in the denominator of $BR(r)$. nDCG_l is a stable and sensitive metric provided that l is large [12]. Microsoft reportedly uses a variant of this metric for improving their Web search engine [1]. Q-measure is also stable and sensitive [12], and it has been applied to XML retrieval [17] as well as Question Answering. It is more highly correlated with AveP than nDCG is; In a binary relevance environment, $Q\text{-}measure = AveP$ holds for any ranked output if there is no relevant document below Rank R, and $Q\text{-}measure > AveP$ 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. Let x_i denote the value of a metric for the i-th topic (down to four significant figures). Then the actual method we use for obtaining the Geometric Mean (GM) is as follows [16]:

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

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 *et al.* [11] used Geometric Mean *Q-measure* for analysing their results at NTCIR-5. Geometric Means are arguably more practical than Arithmetic Means for building *robust* IR systems, which can produce a decent output whatever the query is.

2.2 NTCIR CLIR Chinese and Japanese data

Our experiments use two sets of data (test collections and submitted runs) from the NTCIR-3 CLIR track [9], provided by National Institute of Informatics, Japan. Table 1 provides some statistics of the data. From each data set, only the top 30 runs as measured by *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 sensitivity of metrics.

Table 1: Statistics of the NTCIR-3 CLIR Chinese and Japanese data.

	$\|Q\|$	R	$R(S)$	$R(A)$	$R(B)$	runs
		per topic				(used)
Chinese	42	78.2	21.0	24.9	32.3	45 (30)
Japanese	42	60.4	7.9	31.5	21.0	33 (30)

Thus, for each data set, we have a set of $30 * 29/2 = 435$ system combinations, which we shall denote by C.

3. BOOTSTRAP HYPOTHESIS TESTS AND IR METRICS SENSITIVITY

Sections 3.1 and 3.2 describe how *Paired* and *Unpaired* Bootstrap Hypothesis Tests can be applied to IR evaluation. Section 3.3 proposes methods for comparing the sensitivity of IR metrics and for estimating the performance difference required for achieving a given significance level. Section 3.4 presents our experimental results, and Section 3.5 compares them with those based on the stability and swap methods.

3.1 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 [14], except that we use a *Studentised* test statistic to enhance accuracy. We encourage other IR researchers to try it as 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 $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ denote the per-topic performance values of System 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 1 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 example, let us assume that we only have five topics Q = (001, 002, 003, 004, 005) and that $\mathbf{w} = (0.2, 0.0, 0.1, 0.4, 0.0)$. Suppose that, for trial b, sampling with replacement from Q yields Q^{*b} = (001, 003, 001, 002, 005). Then, $\mathbf{w}^{*b} = (0.2, 0.1, 0.2, 0.0, 0.0)$.

```
for 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;
```

Figure 1: Algorithm for creating Bootstrap samples Q^{*b} and $\mathbf{w}^{*b} = (w_1^{*b}, \ldots, w_n^{*b})$ for the Paired Test.

```
count = 0;
for b = 1 to B
    t(w*b) = w̄*b / (σ̄*b / √n);
    if( |t(w*b)| ≥ |t(z)| ) then count++;
ASL = count/B;
```

Figure 2: Algorithm for estimating the Achieved Significance Level based on 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 2 shows how to compute the Achieved Significance Level (ASL) using $\mathbf{w}^{*b}$. In essence, we examine how *rare* the observed difference would be under H_0. If $ASL < \alpha$, where typically $\alpha = 0.01$ (very strong evidence against H_0) or $\alpha = 0.05$ (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.

The above test relies on the 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 is to use the *Unpaired* Bootstrap Hypothesis Test described in Section 3.2 instead, which is directly applicable to virtually *any* metric, such as the "area" measure based on the worst N topics for each system [16]. The other is to stick to the Paired Test: 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_": We shall test AG_AveP and AG_Q-measure to discuss the sensitivity of G_AveP and G_Q-measure in Section 3.4.

3.2 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. For this reason, the Paired Test should be preferred wherever it is applicable.

The Unpaired Test treats $\mathbf{x}$ and $\mathbf{y}$ as unpaired data, naturally. (In this paper, $\mathbf{x}$ and $\mathbf{y}$ are data obtained by running two IR strategies against a *common* test collection and therefore $|\mathbf{x}| = |\mathbf{y}| = n$. But more generally, the two sets of obser-

let $\mathbf{v} = (x_1, \ldots, x_n, y_1, \ldots, y_m)$;
for $b = 1$ to B
    from a set of integers $(1, \ldots, n+m)$, 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;
        $x_i^{*b} = j$-th element of $\mathbf{v}$;
    for $i = n+1$ to $n+m$
        $j = i$-th element of the sample of integers;
        $y_{i-n}^{*b} = j$-th element of $\mathbf{v}$;

Figure 3: 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.)

$count = 0$;
for $b = 1$ to B
    $d^{*b} = M(\mathbf{x}^{*b}) - M(\mathbf{y}^{*b})$;
    if($|d^{*b}| \geq |\hat{d}|$) then $count$++;
$ASL = count/B$;

Figure 4: Algorithm for estimating the Achieved Significance Level based on the Unpaired Test.

vations may come from different test collections, so $|\mathbf{y}|$ will be denoted by m rather than n hereafter.) 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

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

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 3 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, Figure 3 shuffles the observed values without looking at whether they come from $\mathbf{x}$ or $\mathbf{y}$.

Figure 4 shows how to compute the two-tailed ASL based on the Unpaired Test, in a way similar to Figure 2.

3.3 Sensitivity Comparison Methods

We now propose straightforward methods for assessing and comparing the sensitivity of IR effectiveness metrics.

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

Figure 5: Algorithm for estimating the performance difference required for achieving a given significance level with the Paired Test.

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

Figure 6: Algorithm for estimating the performance difference required for achieving a given significance level with the Unpaired Test.

The idea is simple: Perform a Bootstrap Hypothesis Test for *every* system pair, and count how many of the pairs satisfy $ASL < \alpha$. 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 $n = |Q|$ topics, we can obtain a natural estimate of the 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 $\bar{w}_{X,Y}^{*b}$ and $d_{X,Y}^{*b}$ explicitly denote the above bootstrap replicates for a particular system pair (X, Y). Figures 5 and 6 show our algorithms for estimating the 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 Figure 6, the algorithm obtains the $B\alpha = 1000 * 0.05 = 50$-th largest value among $|d_{X,Y}^{*b}|$ for each (X, Y). Among the $|C| = 435$ values thus obtained, the algorithm takes the maximum value just to be conservative. Figure 5 is almost identical to Figure 6, although it looks slightly more complicated: Since we used Studentisation with the Paired Test, the bootstrap replicate $\bar{w}_{X,Y}^{*b}$ is not equal to the test statistic $t(\mathbf{w}_{X,Y}^{*b})$ (See Figure 2), which we are using as the sort key .

3.4 Experimental Results

Figures 7 and 8 plot, for each IR metric, the Paired and Unpaired Bootstrap ASLs of system pairs from the NTCIR-3 Chinese data. The horizontal axis represents 435 system pairs sorted by the ASL. Since the figures may be too small for the reader, here we summarise the results in words: According to Figure 7, Q-measure, AveP and nDCG_{1000} are the most sensitive metrics, while PDoc_{1000} is clearly the least sensitive. Whereas, nCG_{1000}, AG_AveP and AG_Q-measure (which represent G_AveP and G_Q-measure, respectively) lie in the middle.

The trends are similar in Figure 8, but it can be observed that the Unpaired Test has considerably less power for any metric. (The "outlier" curve represents PDoc_{1000}: it fails to

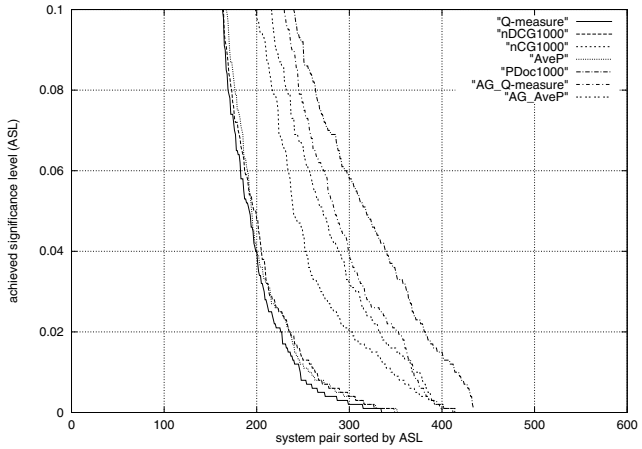

Figure 7: ASL curves based on Paired Bootstrap Hypothesis Tests (NTCIR-3 CLIR Chinese data).

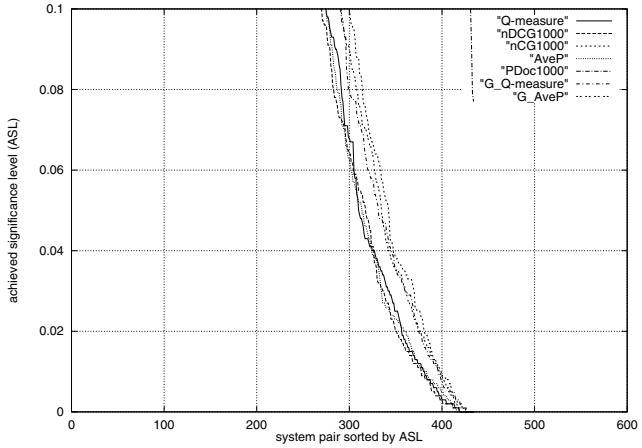

Figure 8: ASL curves based on Unpaired Bootstrap Hypothesis Tests (NTCIR-3 CLIR Chinese data).

satisfy $ASL < \alpha = 0.05$ for all system pairs.) By comparing the two figures, it can be observed that our AG_AveP / AG_Q-measure results based on the Paired Test (i.e., our "indirect" assessments of G_AveP and G_Q-measure) are consistent with our G_AveP / G_Q-measure results based on the Unpaired Test. We have similar graphs for the NTCIR-3 Japanese data, which we omit due to lack of space.

Table 2(a) cuts Figure 7 in half to show, for each IR metric, how many pairs of Chinese systems satisfied $ASL < \alpha = 0.05$ with the Paired Test. The metrics have been sorted by this measure of *sensitivity* (Column 2). Table 2(b) shows similar results for the *Japanese* data, based on a Japanese version of Figure 7 not shown in this paper. It can be observed that Q-measure, nDCG_{1000} and AveP are the most sensitive while PDoc_{1000} is the least sensitive for both data sets. Moreover, the sensitivity values are substantially higher for the Japanese data: This is because the performance variance for the Japanese run set is greater than that for the Chinese run set. For example, the maximum/mininum observed AveP values for the Japanese run set are .4932 and .0617, while those for the Chinese run set are .4166 and .2224.

Column 3 of Table 2 shows, for each metric, the estimated

Table 2: Sensitivity and the estimated differences based on Paired Tests Bootstrap Tests ($\alpha = 0.05$; NTCIR-3 CLIR data).

(a) Chinese	$ASL < \alpha$	estimated diff.
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) Japanese	$ASL < \alpha$	estimated diff.
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

difference required for satisfying $ASL < \alpha = 0.05$, given the topic set size $n = 42$. The algorithm we described in Figure 5 was used to obtain these estimates. For AG_AveP and AG_Q-measure, the estimates represent differences between two Arithmetic Means of *logs* rather than differences between two Geometric Means. Hence we tried to translate these estimates back into the latter by looking up the system pair (X, Y) and the bootstrap replicate number b' that actually yielded the log-space differences, but our results did not always look accurate: For the purpose of estimating the required difference between Geometric Means, the Unpaired Test may be more reliable (See below).

Our table based on the Unpaired Test, which we omit due to lack of space, is generally similar to Table 2: The sensitivity values (Column 2) are naturally much lower, but the estimated differences (Column 3) are comparable to the Paired case. According to the Unpaired Test, the estimated differences for G_AveP and G_Q-measure based on the Chinese data are 0.16 and 0.17, respectively. However, as Figure 8 suggests, the Unpaired Test table appears to be less useful for discussing which metrics are more sensitive than others.

Note that Column 3 is not for comparing different IR metrics: some metrics tend to take small values while others tend to take large values, so such comparisons are not necessarily valid. *Relative* performance differences [13] could be used for comparison across metrics, but this is beyond the scope of our work. We prefer to regard, for example, $M(\mathbf{x}) = 0.0001, M(\mathbf{y}) = 0.0002$ and $M(\mathbf{x}) = 0.1000, M(\mathbf{y}) = 0.2000$ as quite different situations.

3.5 Comparison with Stability/Swap Methods

Readers familiar with the stability method [2] and the swap method [15] will note that our Bootstrap-based methods for comparing the sensitivity of metrics is 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 stability and swap methods use sampling *without* replacement whereas the Bootstrap samples are obtained by sampling *with* replacement. Sanderson and Zobel [13] and Sakai [10] explored variations of the swap method and the latter showed that with- and without-replacement yield similar results for the purpose of ranking metrics according

```
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) + +;
```

Figure 9: Algorithm for computing $EQ(X,Y)$, $GT(X,Y)$ and $GT(Y,X)$.

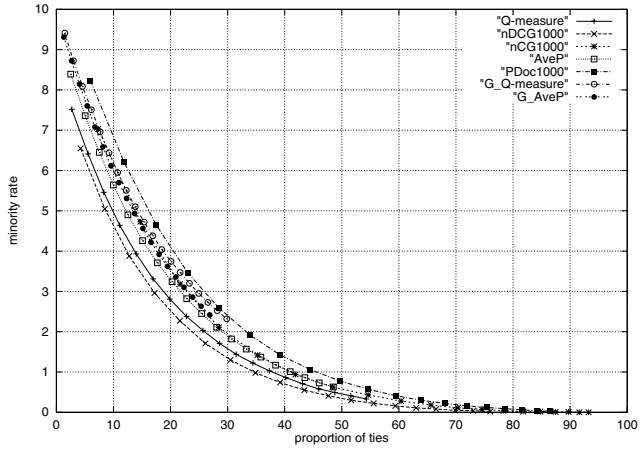

Figure 10: MR-PT curves based on the stability method (NTCIR-3 CLIR Chinese data).

to sensitivity (See Section 5). In light of this, this section uses the *with*-replacement versions of the stability and swap methods, and show that they yield results that are in agreement with our Bootstrap-based results. More specifically, we reuse our paired-test Bootstrap samples Q^{*b} with these two methods.

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 [2, 12], we count $GT(X,Y)$, $GT(Y,X)$ and $EQ(X,Y)$ as shown in Figure 9, where $GT(X,Y)+GT(Y,X)+EQ(X,Y)=B$. Then, the minority rate (MR) and the proportion of ties (PT) for M are computed as:

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

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

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.

Figure 10 shows our *MR-PT* curves based on the NTCIR-3 CLIR Chinese data. (A similar graph for the Japanese data is omitted due to lack of space.) The results are generally in agreement with those based on Bootstrap Hypothesis Tests: nDCG_{1000}, Q-measure and AveP are the most stable, while PDoc_{1000} is the least stable. In the middle lie nCG_{1000}, G_AveP and G_Q-measure.

```
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);
```

Figure 11: Algorithm for computing the swap rates.

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 performance difference between two systems as measured by M based on a topic set; we prepare 21 *performance difference bins* [12, 15], 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 Figure 11 calculates a *swap rate* for each bin. Note that D^{*b} is not the same as our d^{*b} from Figure 4: D^{*b} is the 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 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.

Table 3 summarises the results of our "swap" experiments using the NTCIR-3 CLIR Chinese and Japanese data. The "abs" column shows the absolute difference required for guaranteeing that the swap rate does not exceed 5%. The "rel" column translates these values into relative values using the maximum performance recorded among all trials (the "max" column). The "sensitivity" column, by which the IR metrics have been sorted, shows the percentage of comparisons (among the total of 435*1000 comparisons) that actually satisfied the difference threshold shown in the "abs" column. Again, the results are generally consistent with the Bootstrap ones: For both data sets, nDCG_{1000}, Q-measure and AveP are the most sensitive metrics, while PDoc_{1000} is the least sensitive. In the middle lie nCG_{1000}, G_AveP and G_Q-measure. The results generalise those by Voorhees [16] who

Table 3: Differences and sensitivity based on the swap method (swap rate $\leq$ 5%; NTCIR-3 CLIR Chinese and Japanese data).

Chinese	abs	max	rel	sensitivity
$nDCG_{1000}$	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%**
Japanese	abs	max	rel	sensitivity
$nDCG_{1000}$	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%**

compared AveP and G_AveP. Note also that the estimated 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.

4. QUANTIFYING THE ACCURACY OF RANK CORRELATION BETWEEN TWO IR METRICS

The previous section discussed how the Bootstrap Hypothesis Tests can be used to assess the sensitivity of individual IR metrics. We now show a different application of the Bootstrap to the evaluation of IR metrics.

Kendall's rank correlation is often used for comparing the system rankings according to two different IR metrics [7, 12, 17], but the statistical significance of rank correlation is rarely discussed in the IR community. In fact, there is a traditional significance test available for Kendall's rank correlation: Let the number of systems be n_s, and let τ be the value of Kendall's rank correlation based on two system rankings. Then, it is known that

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

obeys a normal distribution. Thus, a normal test can easily be applied. (In the case of *Spearman's* rank correlation s, the test statistic would be $t_0 = (|s| * (n_s-2)^{1/2})/(1-s^2)^{1/2}$ which obeys a t-distribution with n_s-2 degrees of freedom.) Note that the test statistic Z_0 is proportional to $|\tau|$ given n_s: In terms of a two-tailed test with $n_s = 30$ runs, the rank correlation is significant at $\alpha = 0.01$ if it is over 0.34.

However, there is another way to critically assess the rank correlation values: We could use the *Bootstrap Estimate of Standard Error*, which reflects the accuracy of an observed statistic. Let τ^{*b} be the Bootstrap replicate of τ computed using the Bootstrap sample Q^{*b} instead of Q, and let $m_\tau = \sum_{b=1}^{B} \tau^{*b}/B$. Then the Bootstrap estimate of standard error for Kendall's rank correlation is given by:

$$\hat{se}_{boot} = (\sum_{b=1}^{B} \frac{(\tau^{*b} - m_\tau)^2}{B-1})^{\frac{1}{2}} . \quad (12)$$

We can thus quantify the accuracy of an observed rank correlation using $\hat{se}_{boot}$.

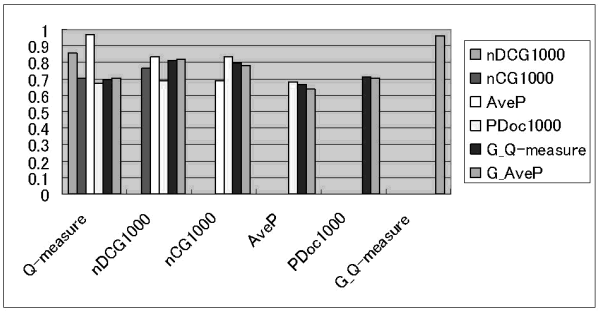

Figure 12: Kendall's rank correlations with NTCIR-3 CLIR Chinese data.

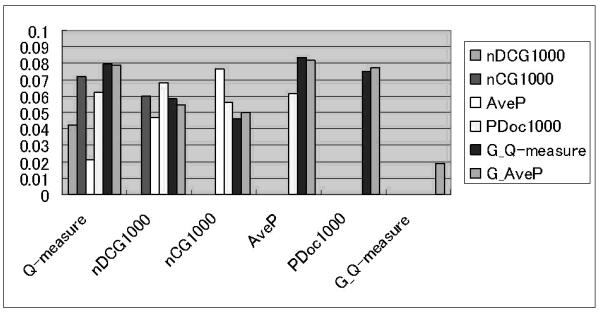

Figure 13: Bootstrap estimates of standard error for the rank correlations shown in Figure 12.

Figure 12 shows Kendall's rank correlations with the NTCIR-3 CLIR Chinese data for all pairs of IR metrics considered in this study. All of the correlation values exceed 0.6, and therefore are statistically highly significant. In particular, the most highly correlated pairs are (Q-measure, AveP) and (G_Q-measure, G_AveP); Also, (Q-measure, $nDCG_{1000}$), ($nDCG_{1000}$, AveP) and (nCG_{1000}, $PDoc_{1000}$) are highly correlated. Figure 13 shows the corresponding Bootstrap Estimates of Standard Error. These values are approximately ten times smaller than the observed rank correlation values, suggesting that the observed values are quite accurate. The results are generally consistent with the traditional significance tests based on Z_0: For example, the graph shows that the correlations between (Q-measure, AveP) and (G_Q-measure, G_AveP) are highly accurate, while those between (AveP, G_Q-measure), (AveP, G_AveP), (Q-measure, G_Q-measure) and (Q-measure, G_AveP) are less accurate. These results probably reflect the fact that, unlike Arithmetic Means, Geometric Means produce quite different system rankings when there is change in the topic set as this change yields a "new" set of worst-performing topics that will be emphasised, and that the new worst-performing topics are different across systems.

The Bootstrap Estimate of Standard Error is widely applicable and useful for assessing the accuracy of any computable statistic. For example, it is applicable to the observed performance difference $\hat{d}$ between two IR systems (See Eq. 8). In this case, one can compute

$$\hat{se}_{boot} = (\sum_{b=1}^{B} \frac{(D^{*b} - m_D)^2}{B-1})^{\frac{1}{2}} . \quad (13)$$

using D^{*b} from Figure 11 and $m_D = \sum_{b=1}^{B} D^{*b}/B$. If, for

example, $\hat{d}$ is not much larger than $\hat{se}_{boot}$, then the difference is probably not substantial. However, Bootstrap Hypothesis Tests address such questions more formally.

5. RELATED WORK

Savoy [14] used the paired Bootstrap Hypothesis Test and Confidence Intervals, along with traditional significance tests, for comparing two IR strategies. Vu and Gallinari [17] compared, for an XML retrieval task, the Bootstrap Confidence Interval of the *top run* with those of the rest in order to compare the sensitivity of metrics such as Q-measure. In contrast, our methods perform a Bootstrap Hypothesis Test for every system pair and is less dependent on a single run. Furthermore, our methods can estimate the performance difference required for achieving a given significance level, and we examined a wider variety of metrics.

Unlike our new methods, the swap method [8, 15] is not directly related to significance tests: Sanderson and Zobel [13] used significance tests for filtering out some system pairs *before* applying the swap method; Sakai [12] reported that the system pair ranking according to significance tests and that according to the swap method are not very highly correlated.

The original swap method used sampling *without* replacement; Sanderson and Zobel [13] also used sampling *without* replacement (although they called their method "selection *with* replacement" in their paper) but ensured that the two topic sets Q_i and Q_i' were independent; Sakai [10] showed that sampling with and without replacement yield similar results for comparing different IR metrics. While all of these studies used resampled topic sets that are half the size of Q, we used Bootstrap samples with the swap method, so that $|Q^{*b}| = |Q|$. Hence extrapolation was not required.

As for rank correlation: Previous work that used rank correlation (e.g., [7, 16, 17]) often did not question statisitical significance, possibly because the correlations values are generally very high. This paper showed that the Bootstrap Estimate of Standard Error provides a simple way of examining the accuracy of rank correlation.

6. CONCLUSIONS

This paper showed that Bootstrap Hypothesis Tests are very useful not only for comparing IR strategies, but also for comparing the sensitivity 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. This paper also showed that the Bootstrap Estimate of Standard Error provides a simple way of quantifying the accuracy of rank correlation between two IR metrics. Through experiments with the NTCIR-3 data, we showed that Q-measure, nDCG_{1000} and AveP are all very sensitive metrics, while PDoc_{1000} is very insensitive. nCG_{1000} and Geometric Mean AveP / Q-measure lie in the middle. (But, as mentioned in Section 2.1, nCG_l has a defect.) Moreover, these Bootstrap-based results are in agreement with those based on the heuristics-based 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. Moreover, the Bootstrap is known to fail when the empirical 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.

Acknowledgements

I thank Steve Robertson for his insightful comments and advice on an early version of this paper, Ian Soboroff for guiding me to the Bootstrap Book [3], and the anonymous reviewers for their suggestions.

7. REFERENCES

[1] Asakawa, S. and Selberg, E.: The New MSN Search Engine Developed by Microsoft (*in Japanese*), *Information Processing Society of Japan Magazine*, Vol 46, No. 9, pp. 1008-1015, 2005.

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

[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, *ACM SIGIR '93 Proceedings*, pp. 329-338, 1993.

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

[6] Johnson, D. H.: The Insignificance of Statistical Significance Testing, *Journal of Wildlife Management* Vol. 63, Issue. 3, pp. 763-772, 1999.

[7] 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.

[8] Lin, W.-H. and Hauptmann, A.: Revisiting the Effect of Topic Set Size on Retrieval Error, *ACM SIGIR 2005 Proceedings*, pp. 637-638, 2005.

[9] NTCIR: `http://research.nii.ac.jp/ntcir/`

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

[11] Sakai, T. *et al.*: Toshiba BRIDJE at NTCIR-5 CLIR: Evaluation using Geometric Means, *NTCIR-5 Proceedings*, pp. 56-63, 2005.

[12] Sakai, T.: On the Reliability of Information Retrieval Metrics based on Graded Relevance, *Information Processing and Management*, to appear, 2006.

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

[14] Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation, *Information Processing and Management*, Vol. 33, No. 4, pp. 495-512, 1997.

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

[16] Voorhees, E. M.: Overview of the TREC 2004 Robust Retrieval Track, *TREC 2004 Proceedings*, 2005.

[17] Vu, H.-T. and Gallinari, P.: On Effectiveness Measures and Relevance Functions in Ranking INEX Systems, *AIRS 2005 Proceedings*, LNCS 3689, pp. 312-327, 2005.