The Effect of Topic Sampling on Sensitivity Comparisons of Information Retrieval Metrics

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

Abstract

The Voorhees/Buckley swap method is useful for comparing the discrimination power of Information Retrieval (IR) and Question Answering (QA) metrics. Given a test collection, a set of runs and an evaluation metric, it derives the swap rate, the chance of observing inconsistencies when two completely different topic sets are used for comparing a pair of runs. Recently, however, Sanderson and Zobel claimed that the method overestimates swap rates as it samples topics without replacement. The main question we address in this paper is whether sampling with and without replacement produce any different results for the purpose of comparing the sensitivity of different metrics. Our IR and QA experiments show that the two methods do generally yield similar results, which suggests that the original Voorhees/Buckley method is valid.
Keywords: *evaluation metrics, sampling.*

1 Introduction

In 2002, Voorhees and Buckley proposed a method of estimating the *sensitivity* (i.e. discrimination power) of Information Retrieval (IR) metrics, given a test collection and a set of runs submitted to the task defined by that collection [13]. The TREC organisers [12, 14] and Sakai [7, 8, 9] have used this method (along with other methods) and have reported several findings for several tasks.

Given a topic set Q, the Voorhees/Buckley method creates two *disjoint* subsets $Q_i(\subset Q)$ and $Q'_i(\subset Q)$. That is, $Q_i \cap Q'_i = \phi$. Then, for a given metric M and a pair of runs x and y, it asks the following question: *Do Q_i and Q'_i agree with each other as to which run is better on average?* The pair of subsets are in fact drawn from Q, say, 1000 times (i.e. $1 \leq i \leq 1000$) and the comparison is performed for every trial and for every pair of runs. Every time a *swap* (i.e. an inconsistency between Q_i and Q'_i for runs x and y) occurs, this is recorded along with the performance difference between x and y based on Q_i. Thus, at the end of all computations, a decreasing curve that plots *swap rates* against *performance difference bins* can be obtained (See Section 2). Based on this graph, one can discuss how much performance differences are required in order to conclude that a run is better than another with a required confidence level. For example, if 95% confidence is required, one looks for the minumum performance difference that guarantees 5% swap rate or less. Moreover, by examining how many of the trials actually satisfied this condition, one can compare the sensitivity of different metrics.

The Voorhees/Buckley method uses two *disjoint* subsets because its purpose is to *guarantee* a given confidence level: a worst case, in which topics are *completely* replaced, is considered in order to estimate a swap rate *upperbound*. Recently, Sanderson and Zobel [10] claimed that the method *overestimates* swap rates because there is a dependency between the two sets as topics are sampled *without replacement*. (That is, once a topic is drawn from Q for trial i, it is not returned to Q until trial $i+1$.) They used sampling *with replacement* instead, and claimed that this gives swap rate *lowerbounds*. Ian Soboroff at NIST also conducted experiments using sampling with replacement (See Section 3.1).

We had a discussion on this issue with Stephen Robertson at Microsoft Research Cambridge, during which *two* kinds of dependency were mentioned:

Dependency between Q_i and Q'_i This is what Sanderson and Zobel saw as a problem.

Dependency between Q_i and Q_j This dependency across *trials* was first pointed out by Stephen Robertson as a *possible* problem. Even though the 1000 trials should ideally be independent of each other, this does not seem not hold when the size c of each subset is half that of Q. In this case, there is a constraint across trials i and j, namely $Q_i - Q_j = Q'_j - Q'_i$, since each trial represents how to divide Q in half.

(The above dependencies arise because two subsets are drawn from Q instead of P, the notional Population of all possible search requests, where $|Q| << |P|$. If direct sampling from P were possible, we would not have to worry about overlaps between Q_i and Q'_i and whether replacement takes place or not.)

This paper tests the following hypotheses.

Hypothesis 1 The original Voorhees/Buckley method yields *higher* swap rates than other topic sampling methods (as claimed by Sanderson and Zobel), and therefore yields more *conservative* (i.e. higher) difference thresholds for for determining whether a run is better than another.

Hypothesis 2 Even if Q_i and Q'_i are independently selected *with replacement* from Q, the general tendencies regarding the relative sensitivity of metrics would remain the same.

To this end, we repeat the Voorhees/Buckley-based experiments in [7, 8, 9], using two alternative topic sampling methods and compare the outcome with the original ones. Section 2 summarises the Voorhees/Buckley method, and Section 3 describes the two alternative methods. Section 4 describes the experimental settings duplicated from [7, 8, 9], and Section 5 compares the results. Section 6 concludes this paper.

2 The Voorhees/Buckley Method

Let S denote a set of runs submitted to a task, and let x and y denote a pair of runs from S. Let $M(x, Q_i)$ denote the performance of run x in terms of metric M computed with a topic set $Q_i (\subset Q)$. Let d denote a performance difference between two systems. The Voorhees/Buckley method [13] begins by preparing 21 *performance difference bins*, where the first bin represents performance differences such that $0 \leq d < 0.01$, the second bin represents those such that $0.01 \leq d < 0.02$, and so on, and the last bin represents those such that $0.20 \leq d$. Let $BIN(d)$ denote a mapping from a difference d to one of the 21 bins where it belongs. Then, for a given constant $c (\leq |Q|/2)$, the algorithm shown in Figure 1 calculates a *swap rate* for each bin [7, 9]. By plotting swap rates against the performance difference bins, one can discuss how much performance differences are required to conclude that a run is better than another with a required confidence level, e.g. 95%.

As was discussed in Section 1, the Original Voorhees/Buckley method ensures that Q_i and Q'_i are disjoint to consider a worst case in which the properties of the two topic sets are completely different. Thus, the method is hereafter referred to as **Disjoint**.

3 Alternative Topic Sampling Methods

3.1 Drawing Topics with Replacement

Ian Soboroff at NIST, USA, has done experiments which borrow ideas from Efron's Bootstrap [1, 11].

```
for each pair of runs x, y ∈ S
    for each trial from 1 to 1000
        select Q_i ⊂ Q and Q'_i ⊂ Q s.t.
                Q_i ∩ Q'_i == φ and |Q_i| == |Q'_i| == c;
        d_M(Q_i) = M(x, Q_i) − M(y, Q_i);
        d_M(Q'_i) = M(x, Q'_i) − M(y, Q'_i);
        counter(BIN(d_M(Q_i))) + +;
        if( d_M(Q_i) ∗ d_M(Q'_i) > 0 )
            continue
        else
            swap_counter(BIN(d_M(Q_i))) + +;
for each bin b
    swap_rate(b) = swap_counter(b)/counter(b);
```

Figure 1. The Voorhees/Buckley algorithm for computing the swap rates.

This method creates Q_i and Q'_i *independently* from Q, and therefore the two sets may overlap. Moreover, it draws topics from Q *with replacement*, meaning that both Q_i and Q'_i can contain *duplicate* topics. Thus we refer to this method as **Replacement**. Note that, with **Replacement**, the number of *unique* topics in Q_i may be smaller than c.

Soboroff's motivation for using **Replacement** in place of **Disjoint** was to drop the constraint $c \leq |Q|/2$. That is, **Replacement** allows sampling up to the full topic set size $|Q|$. (In fact, Efron's *bootstrap sample* is of size exactly $|Q|$.) However, we stick to $c \leq |Q|/2$ for comparison with **Disjoint**. Recently, Sanderson and Zobel [10] also used sampling with replacement, and they also used $c \leq |Q|/2$.

The fact that Q_i and Q'_i may overlap with each other seems to suggest that **Replacement** may yield lower swap rates than **Disjoint**, as claimed by Sanderson and Zobel [10]. On the other hand, **Replacement** generally uses a smaller number of *unique* topics, and has duplicates within Q_i and within Q'_i. How would this affect the swap rate?

3.2 Creating Two Subsets Independently

The second alternative method, which we call **Independent**, simply replaces the subset selection process in Figure 1 (shown in bold) with the following:
select $Q_i \subset Q$ **and** $Q'_i \subset Q$ **independently, s.t.** $|Q_i| == |Q'_i| == c$**;**.

Thus both Q_i and Q'_i contain unique topics just like **Disjoint**, but the two subsets may overlap with each other just like **Replacement**. This should give higher swap rates than **Disjoint** due to the overlaps.

4 Experiments

Sakai used the **Disjoint** method for comparing IR metrics in [8, 9] and for comparing exact-answer Question Answering (QA) metrics in [7]. This paper repeats the main experiments from these papers using **Replacement** and **Independent** to test the two hypotheses mentioned in Section 1. In particular, if **Hypothesis 2** holds true, then **Disjoint** is valid, and so are the results of all previous publications that used this method.

Below we describe our three sets of experiments that correspond to Sakai's [7, 8, 9].

4.1 Binary vs Graded IR Metrics

In [9], Sakai used the **Disjoint** method for comparing *graded-relevance* IR metrics based on *cumulative gain* [2] and standard *binary-relevance* IR metrics.

The binary-relevance metrics considered were:

AveP TREC (noninterpolated) Average Precision;

R-Prec R-Precision;

PDoc$_l$ Precision at document cut-off l ($l = 10, 100, 1000$).

The graded-relevance metrics considered were:

Q-measure A metric similar to $AveP$, but can handle graded relevance [5, 6, 9];

R-measure A metric similar to $R\text{-}Prec$, but can handle graded relevance [5, 6, 9];

(A)n(D)CG$_l$ (Average) normalised (Discounted) Cumulative Gain at document cut-off l ($l = 10, 100, 1000$) [2, 9].

Sakai used two test collections (Chinese and English) and the runs from the NTCIR-3 CLIR track [3]. This paper repeats the experiments with the Chinese-document runs, since the Chinese data set is the largest data available. (Currently, only the NTCIR-3 CLIR runs are available to non-organisers of NTCIR.) Following the NTCIR tradition, we use both "Relaxed" and "Rigid" versions of the binary-relevance metrics, where the former treats S-, A-, and B-relevant (i.e. highly-relevant, relevant and partially relevant) documents as relevant and the latter ignores the B-relevant ones. By default, *gain values* [2] of 3,2,1 are given for each retrieved S-,A-,B-relevant document, respectively.

Since $|Q| = 42$ for this data set, we let $c = 20 (< |Q|/2)$ throughout our experiments. Among the 45 Chinese-document runs that are available from NTCIR, the top 30 runs in terms of Relaxed-AveP were used for the experiments. This set of experiments will be referred to as "IR Experiment 1".

4.2 O-measure and RR as IR Metrics

In [8], Sakai conducted experiments similar to those in [9], but focused on the metrics for the task of finding *one* relevant document. In addition to AveP and Q-measure, which are metrics for the task of finding *all* relevant documents in the sense that they are computed by averaging over all relevant documents, Sakai examined the following:

RR Reciprocal Rank of the first relevant document retrieved;

O-measure A variant of Q-measure, that handles graded relevance but examines only the first relevant document retrieved [8].

The experimental setting for these metrics is identical to that of IR Experiment 1. This set of experiments will be referred to as "IR Experiment 2".

4.3 QA Metrics

In [7], Sakai conducted experiments using the **Disjoint** method for comparing *QA* metrics for NTCIR-4 QAC2 Subtask 1 [3], which required the systems to output a ranked list of exact answer strings (along with IDs of supporting documents, which are ignored throughout this study), containing up to five candidate answers. The official evaluation metric used was RR, but the QAC organisers also considered the use of "NQcorrect5" and "NQcorrect1" (number of questions for which the system managed to return a correct answer within top 5/1). But because neither of these metrics can handle *multiple correct answers* and *answer correctness levels*, Sakai [5] proposed the application of the aforementioned Q-measure to QA evaluation at NTCIR. He showed that, by (a) assigning a *correctness level* (S,A,B) to each answer string; and (b) forming *answer equivalence classes* for ignoring duplicate answers in the list, Q-measure can be applied to QA evaluation successfully. The official QAC2 data already had equivalence classes, but lacked the correctness level data. We therefore use our own correctness level assessment data.

As in the IR case, gain values of 3,2,1 are given for each S-,A-,B-correct answer by default to calculate Q-measure. When gain values of a, b, c are given instead, this is denoted by "Q$a : b : c$".

Our "QA experiment" uses the official 195 QAC2 Subtask 1 questions, and therefore lets $c = 97 (< |Q|/2)$. Whereas, because the official run files are currently *not* available to non-organisers of NTCIR-4 QAC2 (unlike the case with NTCIR-3 CLIR), we use 10 runs generated by a single system [4] but representing a variety of performances [7]. Note that our QA experiment uses more topics (i.e. questions) than the IR ones (97 vs 20), but fewer runs (10 vs 30).

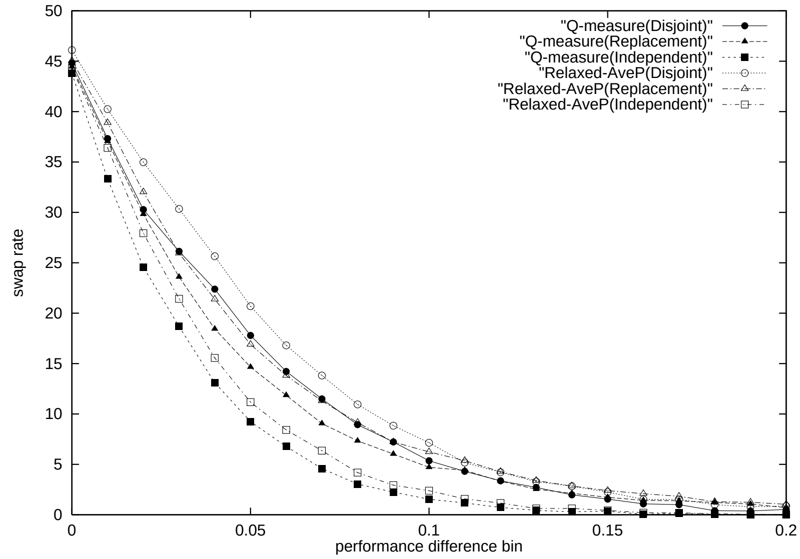

Figure 2. IR Experiment 1: Swap Rates for Q-measure and Relaxed-AveP.

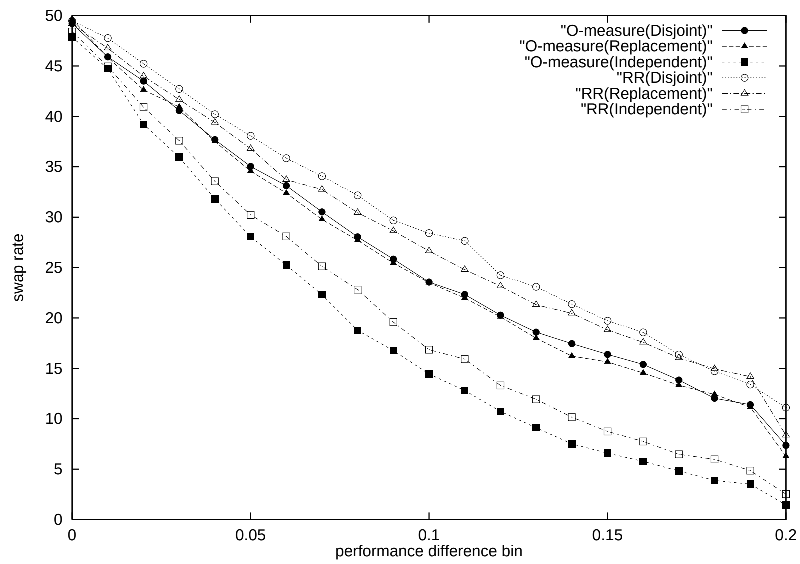

Figure 3. IR Experiment 2: Swap Rates for O-measure and RR.

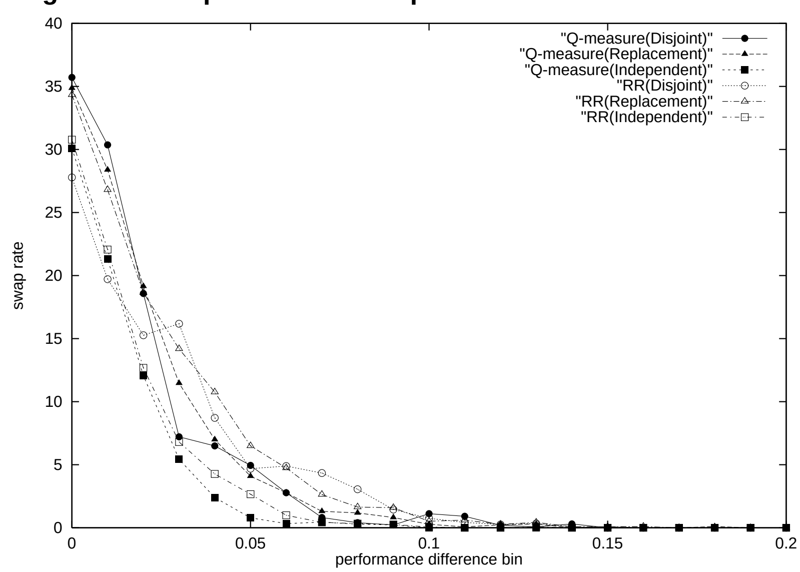

Figure 4. QA Experiment: Swap Rates for Q-measure and RR.

5 Results and Discussions

Figures 2-4 plot swap rates against performance difference bins for a few metrics selected from IR Experiments 1&2 and the QA Experiment, respectively. For example, "Q-measure(Disjoint)" in Figure 2 represents the swap rate curve of Q-measure obtained using the **Disjoint** method in IR Experiement 1.

Based on swap rate curves including those shown in Figure 2, Tables 1 and 2 provide a summary of our sensitivity comparisons in IR Experiment 1. Table 1(a) and Table 2(a) are exact duplications from [9], which used **Disjoint**. The rest of the tables show the new results with **Replacement** and **Independent**. For example, Table 1(a) shows that, when 20 topics are used for ranking the C-runs with Relaxed-AveP, an absolute difference of at least 0.11 (or 20% in terms of relative difference) is required in order to conclude that a run is better than another with 95% confidence. Of the 435,000 comparisons (30*29/2=435 system pairs, each with 1000 trials), 23.7% actually had this difference. The metrics have been sorted by this measure of discrimination power (Column (iv)).

Table 3 provides a similar table for IR Experiment 2. It compares O-measure and RR (i.e. metrics for finding one relevant document) with Q-measure and Relaxed-AveP (i.e. metrics for finding as many relevant documents as possible), for 95%, 90% and 80% confidence levels. Tables 3(a) is a duplication from [8].

Table 4 provides a summary of our sensitivity comparisons in the QA Experiment, which includes Q-measure with "flat" and "mild" gain value assignments ("Q1:1:1" and "Q2:1.5:1") as well as default Q-measure. Table 4(a) is a duplication from [7].

5.1 Testing Hypothesis 1

We first discuss **Hypothesis 1** by examining the swap rate curves, as well as the difference thresholds shown in the aforementioned tables.

From Figure 2, it can be observed that:

- For both Q-measure and Relaxed-AveP, the **Disjoint** and **Replacement** curves overlap with each other when the swap rates are less than 5% (which correspond to practically useful confidence levels), although the **Disjoint** curves are slightly above the **Replacement** ones when the performance differences are less than 0.1.
- For both Q-measure and Relaxed-AveP, **Independent** yields considerably lower swap rates than **Disjoint** and **Replacement**.
- Regardless of topic sampling methods, Q-measure yields slightly but consistently lower swap rates than Relaxed-AveP.

From Figure 3, it can be observed that:

- For O-measure, the **Disjoint** and **Replacement** curves are almost identical. For RR, **Disjoint** does seem to yield higher swap rates than **Replacement**, but the differences are very small.
- For both O-measure and RR, **Independent** yields considerably lower swap rates than **Disjoint** and **Replacement**.
- Regardless of topic sampling methods, O-measure yields lower swap rates than RR.

Unfortunately, Figure 4 is not as stable as Figures 2 and 3 as only 10 runs were used in the experiment. However, we can still observe that **Independent** tends to *underestimate* swap rates for the QA task as well.

Similar results were obtained for metrics not included in the graphs. Thus, **Independent** yields lower swap rates than **Disjoint** and **Replacement**, but **Disjoint** and **Replacement** often yield similar swap rates. Moreover, Tables 1-4 show that the actual difference thresholds obtained by these two methods are almost identical (although the *sensitivity* values in Column (iv) are often *slightly* higher with **Replacement**). Thus, our results do not really support **Hypothesis 1**, contrary to Sanderson and Zobel's view that **Disjoint** yields swap rate upperbounds while **Replacement** yields lowerbounds.

The above inconsistency may be attributable to the differences in the data used (NTCIR vs TREC; the former is admittedly much smaller, but has graded relevance data). Another possible cause is that Sanderson and Zobel examined AveP and PDoc_{10} only: Looking into other metrics may (or may not) produce results that are more in line with ours. Moreover, while they used *extrapolation* for larger topic sets, we stuck to the swap rates actually measured because extrapolation can easily magnify errors. Another difference is that we were faithful to the original method: bins of *absolute* differences were used, and these were translated into *relative* differences based on the maximum values observed as shown in Tables 1-4. Whereas, Sanderson and Zobel created bins of *relative* differences, so that, for example, $M(x, Q_i) = 0.01, M(y, Q_i) = 0.02$ and $M(z, Q_j) = 0.10, M(w, Q_j) = 0.20$ concern the same bin.

5.2 Testing Hypothesis 2

Next, we discuss **Hypothesis 2** by focussing on Column (iv) of Tables 1-4, visualised in Figures 5-8.

Figures 5 and 6 show that **Disjoint** and **Replacement** generally yield similar results as to relative sensitivity of metrics, even though the ranking of the metrics are not identical. (We get minor inconsistencies of this kind even when a single sampling method is

Table 1. IR Experiment 1: The sensitivity of binary IR metrics at 95% confidence.

(i): Absolute difference required; (ii): Maximum performance observed; (iii): Relative difference required ((i)/(ii)); (iv): %comparisons with the required difference. The rows have been sorted by (iv).

	(i)	(ii)	(iii)	(iv)
(a) **Disjoint** [duplicated from [9]]				
Relaxed-AveP	0.11	0.5392	20%	23.7%
Relaxed-R-Prec	0.11	0.5554	20%	20.8%
Rigid-AveP	0.10	0.4698	21%	20.6%
Rigid-PDoc$_{100}$	0.05	0.2860	17%	15.4%
Relaxed-PDoc$_{10}$	0.17	0.7400	23%	14.6%
Rigid-PDoc$_{10}$	0.16	0.5900	27%	10.5%
Rigid-R-Prec	0.12	0.4660	26%	9.2%
Rigid-PDoc$_{1000}$	0.01	0.0628	16%	5.7%
Relaxed-PDoc$_{100}$	0.09	0.3940	23%	5.3%
Relaxed-PDoc$_{1000}$	0.02	0.1009	20%	1.4%
(b) **Replacement**				
Relaxed-R-Prec	0.11	.5966	18%	22.7%
Rigid-AveP	0.10	.5203	19%	22.5%
Relaxed-AveP	0.12	.5998	20%	21.3%
Rigid-PDoc100	0.05	.3550	14%	17.7%
Relaxed-PDoc10	0.18	.7850	23%	15.3%
Rigid-R-Prec	0.11	.5156	21%	15.2%
Rigid-PDoc10	0.16	.6800	24%	12.9%
Relaxed-PDoc100	0.08	.4685	17%	11.1%
Rigid-PDoc1000	0.01	.0777	13%	7.9%
Relaxed-PDoc1000	0.02	.1182	17%	2.7%
(c) **Independent**				
Relaxed-R-Prec	0.07	.5554	13%	43.6%
Relaxed-AveP	0.08	.5527	14%	39.5%
Rigid-AveP	0.07	.4931	14%	38.4%
Relaxed-PDoc10	0.11	.7500	15%	35.4%
Rigid-PDoc10	0.10	.5850	17%	31.7%
Relaxed-PDoc100	0.05	.3925	13%	29.6%
Rigid-R-Prec	0.08	.4624	17%	27.9%
Rigid-PDoc100	0.04	.2885	14%	25.7%
Relaxed-PDoc1000	0.01	.0962	10%	20.1%
Rigid-PDoc1000	0.01	.0632	16%	5.7%

used but with different sets of randomly selected topics.) Thus, the following observations we made in [9] do seem to hold true even when **Replacement** is used instead of **Disjoint**:

- Q-measure, R-measure and (A)nDCG$_l$ (with large l) are generally more sensitive than (A)nCG$_l$.
- The best graded-relevance metrics (e.g. Q-measure) may be slightly more sensitive than the best binary-relevance metrics (e.g. AveP).

In summary, IR Experiment 1 supports **Hypothesis 2**.

As for **Independent**, the impact of topic overlaps overshadows the differences across metrics, and it is not very useful for comparing metrics. The large intersection between Q_i and Q'_i reduces the chance of swaps, no matter what metric is used.

Figure 7 also shows that **Disjoint** and **Replacement** yield similar results. Thus, the following observations we made in [8] do hold true:

- O-measure and RR are less sensitive than Q-measure and Relaxed-AveP.
- But O-measure may be slightly more sensitive than RR.

Table 2. IR Experiment 1: The sensitivity of graded IR metrics at 95% confidence.

(i): Absolute difference required; (ii): Maximum performance observed; (iii): Relative difference required ((i)/(ii)); (iv): %comparisons with the required difference. The rows have been sorted by (iv).

	(i)	(ii)	(iii)	(iv)
(a) **Disjoint** [duplicated from [9]]				
Q-measure	0.10	0.5490	18%	25.4%
R-measure	0.11	0.5777	19%	21.8%
AnDCG$_{1000}$	0.12	0.7067	17%	21.0%
AnDCG$_{100}$	0.13	0.6237	21%	19.8%
nDCG$_{1000}$	0.12	0.7461	16%	19.6%
nDCG$_{100}$	0.13	0.6440	20%	17.9%
nCG$_{10}$	0.14	0.5967	23%	17.1%
nDCG$_{10}$	0.15	0.6262	24%	16.3%
AnCG$_{100}$	0.14	0.6662	21%	15.8%
AnCG$_{10}$	0.17	0.6613	26%	13.2%
AnDCG$_{10}$	0.19	0.6869	28%	10.7%
nCG$_{100}$	0.16	0.7377	22%	10.5%
AnCG$_{1000}$	0.15	0.8770	17%	10.1%
nCG$_{1000}$	-	0.9632	-	-
(b) **Replacement**				
Q-measure	0.10	.6005	17%	27.1%
AnDCG100	0.12	.6787	18%	25.8%
R-measure	0.11	.6061	18%	23.8%
AnDCG1000	0.12	.7395	16%	23.1%
nDCG1000	0.12	.7791	15%	21.8%
AnCG100	0.13	.7526	17%	21.2%
nDCG100	0.13	.7071	18%	20.0%
nCG10	0.14	.6661	21%	19.4%
nDCG10	0.15	.6869	22%	18.8%
nCG100	0.14	.8661	16%	18.3%
AnCG1000	0.13	.9338	14%	17.9%
AnCG10	0.17	.7346	23%	16.0%
AnDCG10	0.19	.7634	25%	13.7%
nCG1000	0.16	.9845	16%	8.9%
(c) **Independent**				
AnCG100	0.08	.6660	12%	43.6%
Q-measure	0.07	.5666	12%	43.2%
nDCG100	0.08	.6469	12%	42.0%
AnDCG1000	0.08	.7215	11%	41.2%
nDCG1000	0.08	.7556	11%	39.8%
nCG10	0.09	.5967	15%	38.7%
AnCG1000	0.08	.8893	9%	38.6%
R-measure	0.08	.5777	14%	38.1%
AnDCG100	0.09	.6267	14%	38.1%
nCG100	0.09	.7538	12%	37.7%
nDCG10	0.10	.6262	16%	36.2%
AnCG10	0.11	.6613	17%	34.0%
AnDCG10	0.12	.6869	17%	31.9%
nCG1000	0.09	.9674	9%	29.3%

In summary, IR Experiment 2 also supports **Hypothesis 2**. Note that even **Independent** agrees with the above observations.

Figure 8 also shows that **Disjoint** and **Replacement** yield similar results. Thus, the following observations we made in [7] do hold true:

- Q-measure (preferrably with "mild" gain values) is at least as sensitive as RR;
- NQcorrect1 and NQcorrect5 are not as sensitive as RR and Q-measure.

Thus our QA Experiment also supports **Hypothesis 2**.

5.3 Discussions

Surprisingly, our experimental results do not support **Hypothesis 1**, suggesting that **Replacement** may

Table 3. IR Experiment 2: The sensitivity of metrics at 80-95% confidence.

(i): Absolute difference required; (ii): Maximum performance observed; (iii): Relative difference required ((i)/(ii)); (iv): %comparisons with the required difference. The rows have been sorted by (iv).

		(i)	(ii)	(iii)	(iv)
(a) **Disjoint** [duplicated from [8]]					
95%	Q-measure	0.10	.5490	18%	25.4%
	Relaxed-AveP	0.11	.5392	20%	23.7%
	O-measure	-	.8792	-	-
	RR	-	.9750	-	-
90%	Q-measure	0.08	.5490	15%	36.7%
	Relaxed-AveP	0.09	.5392	17%	33.8%
	O-measure	0.20	.8792	23%	16.5%
	RR	-	.9750	-	-
80%	Relaxed-AveP	0.05	.5392	9%	59.7%
	Q-measure	0.05	.5490	9%	57.7%
	O-measure	0.14	.8792	16%	33.2%
	RR	0.16	.9750	16%	27.5%
(b) **Replacement**					
95%	Q-measure	0.10	.6005	17%	27.1%
	Relaxed-AveP	0.12	.5998	20%	21.3%
	O-measure	-	.9313	-	-
	RR	-	1.000	-	-
90%	Q-measure	0.07	.6005	12%	44.6%
	Relaxed-AveP	0.08	.5998	13%	41.0%
	RR	0.20	1.000	20%	21.7%
	O-measure	0.20	.9313	21%	20.4%
80%	Q-measure	0.04	.6005	7%	66.4%
	Relaxed-AveP	0.05	.5998	8%	60.8%
	O-measure	0.13	.9313	14%	40.5%
	RR	0.15	1.000	15%	35.0%
(c) **Independent**					
95%	Q-measure	0.07	.5666	12%	43.2%
	Relaxed-AveP	0.08	.5527	14%	39.5%
	O-measure	0.17	.8792	19%	23.9%
	RR	0.19	.9583	20%	19.5%
90%	Q-measure	0.05	.5666	9%	57.7%
	Relaxed-AveP	0.06	.5527	11%	52.6%
	O-measure	0.13	.8792	15%	36.8%
	RR	0.15	.9583	16%	30.6%
80%	Q-measure	0.03	.5666	5%	73.7%
	Relaxed-AveP	0.04	.5527	7%	67.2%
	O-measure	0.08	.8792	9%	58.1%
	RR	0.09	.9583	9%	53.6%

be used instead of **Disjoint** for setting a *conservative* difference threshold for determining whether a run is better than another.

Table 5 shows the average degree of overlap between Q_i and Q'_i for each topic sampling method in our IR and QA experiments. For **Replacement**, the values are based on unique topics: For example, for the IR experiments, Q_i and Q'_i contained 16.1 unique topics on average, of which 6.2 topics were shared across the two sets. It is remarkable that **Replacement** yields results similar to those of **Disjoint** despite the substantial overlap. Since **Replacement** can resample topics up to $|Q_i| = |Q|$, it is probably a good alternative to the original **Disjoint** method, and the bootstrap approach is probably worth exploring further.

On the other hand, since the results in Section 5.2 generally support **Hypothesis 2**, we believe that the previous findings using **Disjoint** [7, 8, 9] are valid. There is no evidence that the dependencies inherent in the original Voorhees/Buckley method have any ill effect on sensitivity comparison of metrics.

Table 4. QA Experiment: The sensitivity of metrics at 95% confidence..

(i): Absolute difference required; (ii): Maximum performance observed; (iii): Relative difference required ((i)/(ii)); (iv): %comparisons with the required difference. The rows have been sorted by (iv).

	(i)	(ii)	(iii)	(iv)
(a) **Disjoint** [duplicated from [7]]				
Q1:1:1	0.05	.6967	7%	66.2%
Q2:1.5:1	0.05	.6890	7%	65.2%
Q-measure	0.05	.6860	7%	65.1%
RR	0.06	.7940	8%	64.3%
NQcorrect1	0.09	.7423	12%	51.0%
NQcorrect5	0.09	.8866	10%	49.5%
(b) **Replacement**				
Q1:1:1	0.05	.7315	7%	65.8%
Q2:1.5:1	0.05	.7211	7%	65.1%
Q-measure	0.05	.7166	7%	64.8%
RR	0.06	.8247	7%	64.0%
NQcorrect5	0.08	.8969	9%	54.5%
NQcorrect1	0.09	.7835	11%	51.3%
(c) **Independent**				
Q1:1:1	0.03	.7121	4%	79.8%
Q2:1.5:1	0.03	.6928	4%	79.4%
RR	0.04	.7940	5%	74.7%
Q-measure	0.04	.6860	6%	72.2%
NQcorrect1	0.06	.7423	8%	65.9%
NQcorrect5	0.06	.8866	7%	65.7%

Table 5. The degree of overlap between Q_i and Q'_i.

	IR Exps. 1 and 2	QA Exp
Disjoint	0 / 20	0 / 97
Replacement	6.2 / 16.1	30.0 / 76.5
Independent	9.5 /20	48.0 / 97

6 Conclusions and Future Work

This paper showed, through experimentation, that the Voorhees/Buckley swap method and its variation, which uses topic sampling with replacement, yield similar results in relative sensitivity comparison of metrics. Thus, we believe that the results reported in [7, 8, 9] are all valid. However, sampling with replacement is certainly attractive in that it can resample up to the size of the base topic set. We plan to explore more direct applications of the bootstrap [1, 11] to the evaluation of stability and sensitivity of IR metrics. We also plan to carry out more experiments with other data and with new IR metrics.

Acknowledgments

The author thanks Stephen Robertson (Microsoft Research Cambridge, UK), Ellen Voorhees and Ian Soboroff (NIST, USA) for stimulating discussions.

References

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

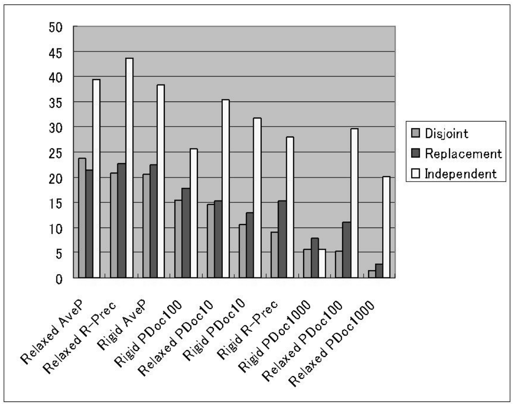

Figure 5. IR Experiment 1: Discrimination power at 95% confidence (binary relevance metrics).

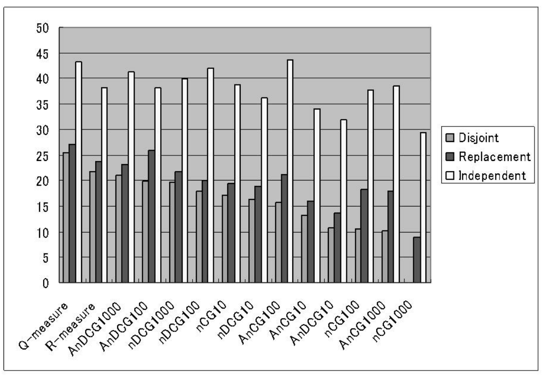

Figure 6. IR Experiment 1: Discrimination power at 95% confidence (graded relevance metrics).

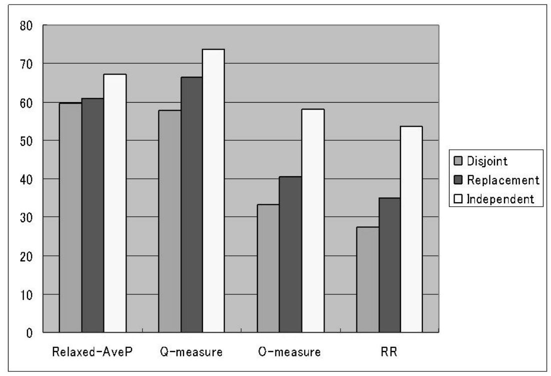

Figure 7. IR Experiment 2: Discrimination power at 80% confidence.

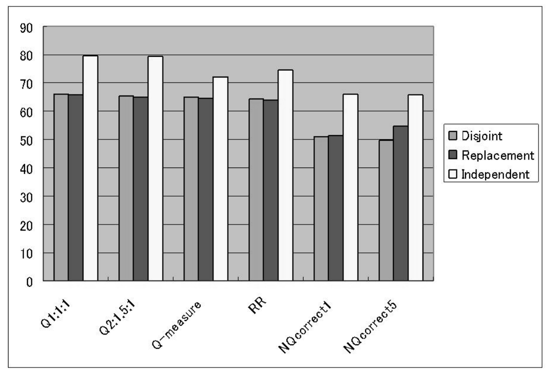

Figure 8. QA Experiment: Discrimination power at 95% confidence.

[2] 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.

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

[4] Sakai, T. *et al.*: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis, *RIAO 2004 Proceedings*, pp. 215-231, 2004.

[5] Sakai, T.: New Performance Metrics based on Multigrade Relevance: Their Application to Question Answering, *NTCIR-4 Proceedings*, 2004.

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

[7] Sakai, T.: A Note on the Reliability of Japanese Question Answering Evaluation, *IPSJ SIG Technical Reports 2004–FI-7-7*, pp.57-64 / Digital Libraries No.25&26, pp. 59-66, 2004.

[8] Sakai, T.: An Evaluation Metric for the Task of Retrieving One Highly Relevant Document with High Precision (*in Japanese*), *Forum on Information Technology 2005 Information Technology Letters*, LD-002, pp. 69-72, 2005.

[9] Sakai, T.: The Reliability of Metrics based on Graded Relevance, *AIRS 2005 Proceedings*, LNCS 3689, pp. 1-16, Springer-Verlag, 2005.

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

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

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

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

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