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

2005－ＦＩ－ 80 (10)
2005－ＮＬ－169 (10)
2005／9／30

The Effect of Topic Sampling in 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, proposed in 2002, is useful for comparing the discrimination power of Information Retrieval (IR) and Question Answering (QA) metrics, given a test collection and a set of runs. However, the method has recently been criticised as having a flaw in that it draws two disjoint subsets of topics without replacement from the base topic set. This paper defends the swap method by using alternative topic sampling methods, one of which draws topics *with* replacement, and comparing the general outcome of metrics comparisons. Our IR and QA experiments show that the original swap method and its with-replacement variation do generally yield similar results. Thus, we do not believe there is a practical flaw in the original swap method, even though one advantage of the with-replacement method is that it can resample up to the size of the base topic set.

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 [12]. The TREC organisers [11, 13] and the present author [8, 9, 10] 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 takes two disjoint subsets of Q, which we denote by Q_i and Q_i'. 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 absolute performance difference between x and y. 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, however, the method has been criticised as having a flaw by an anonymous Reviewer at a conference. The Reviewer's comment, quoted verbatim, is: *The flaw arrises because the two sets are selected without replacement (i.e. the two sets are disjoint). This innevitably means that there is a dependecy between the two sets and this dependency I belive causes an overestimation of swap rates. When selecting in this way, there is a relatively high probability of having the first topic set rank one system over another while the other set ranks the systems in the opposite way; the probability of obtaining such a pair of topic sets, is much higher than one would expect to see if one was sampling the two sets from a much larger set of several hundred or several thousand topics.*

The present author discussed how to interpret the above comment with Stephen Robertson at Microsoft Research Cambridge, UK. Two kinds of dependency were mentioned during the discussion:

Dependency between Q_i and Q_i' This appears to be what the Reviewer saw as a problem. However, as mentioned earlier, Voorhees and Buckley deliberately used disjoint subsets in order to guarantee a given confidence level. Thus we are interested in obtaining a swap rate *upperbound*.

Dependency between Q_i and Q_j The 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 clearly does not hold when the size c of each subset is half that of Q. In this case, $Q_i - Q_j = Q_j' - Q_i'$ holds, since each trial represents how to divide Q in half.

The dependencies arise from the fact that the method draws two subsets from Q and not from the notional population P of all possible search requests, where

$|Q| << |P|$. If direct sampling from P is possible, then we would not have to worry about overlaps between Q_i and Q'_i and whether replacement takes place or not.

The present author does *not* claim that the Voorhees/Buckley method obtains the *true* swap rates, which, in theory, can be obtained by drawing Q_i and Q'_i ($1 \leq i \leq 1000$) from P instead of Q. Rather, the author is interested in testing the following hypotheses.

Hypothesis 1 The original Voorhees/Buckley method generally obtains *higher* swap rates than other topic sampling methods, and is therefore more useful for drawing careful conclusions for guaranteeing a given confidence level.

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.

This paper tests the above hypotheses by repeating the main experiments reported in [8, 9, 10], using two alternative topic sampling methods and comparing 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 [8, 9, 10], and Section 5 compares the three sets of results. Section 6 concludes this paper.

2 The Original Voorhees/Buckley Method

Let S denote a set of runs submitted to a particular 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 averaged across a topic set $Q_i (\subset Q)$. Let d denote a performance difference between two systems. The Voorhees/Buckley method [12] 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 [8, 10]. By plotting swap rates against the performance difference bins, 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, 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**.) However, as there is a claim that this topic sampling strategy is a flaw, we consider two alternative ways to sample topics from Q in the following section.

for each pair of runs $x, y \in S$
  for each trial from 1 to 1000
    select $Q_i \subset Q$ **and** $Q'_i \subset Q$ **s.t.**
      $Q_i \cap Q'_i == \phi$ **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(($d_M(Q_i) * d_M(Q'_i) < 0$) or
      ($d_M(Q_i) == 0$ and $d_M(Q'_i) \neq 0$) or
      ($d_M(Q_i) \neq 0$ and $d_M(Q'_i) == 0$))
      $swap\_counter(BIN(d_M(Q_i))) + +$;
for each bin b
  $swap\_rate(b) = swap\_counter(b)/counter(b)$;

Figure 1. The Original algorithm for computing the swap rates.

3 Alternative Topic Sampling Methods

3.1 Drawing Topics with Replacement

Ian Soboroff, a TREC organiser at NIST, USA, has done experiments which borrow ideas from Efron's Bootstrap [2]. 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**. (We can still treat Q_i and Q'_i as *sets*: for example, if Topic 001 is included twice in Q_i, we can formally treat them as two different topics 001-1 and 001-2. Note that, with **Replacement**, c is the number of topic samples including duplicates, and that Q_i generally contains a smaller number of *unique* topics.)

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**.

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**. On the other hand, allowing duplicate topics implies that a smaller number of *unique* topics may be used within each trial and throughout the experiment. How this would affect the swap rate was not clear to the author before the experiment.

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

The present author used the **Disjoint** method for comparing IR metrics in [9, 10] and for comparing exact-answer Question Answering (QA) metrics in [8]. 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 (e.g. [8, 9, 10, 11, 12, 13]).

Below we describe the three sets of experiments that correspond to three previous publications [8, 9, 10].

4.1 Binary vs Graded IR Metrics

In [10], Sakai used the **Disjoint** method for comparing *graded-relevance* IR metrics based on *cumulative gain* [3] 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 [6, 7, 10];

R-measure A metric similar to R-$Prec$, but can handle graded relevance [6, 7, 10];

(A)n(D)CG$_l$ (Average) normalised (Discounted) Cumulative Gain at document cut-off l (l = 10, 100, 1000) [3, 10].

Sakai [10] used two test collections (Chinese and English) and the runs from the NTCIR-3 CLIR track [4]. 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* of 3,2,1 are given for each retrieved S-,A-,B-relevant document, respectively [3].

Since $|Q| = 42$ for this data set, we let $c = 20$ 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 [9], Sakai conducted experiments similar to those in [10], 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 [9].

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 [8], Sakai conducted experiments using the **Disjoint** method for comparing *QA* metrics for NTCIR-4 QAC2 Subtask 1 [4], 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 [6] 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 $c = 97$. 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 [5] but representing a variety of performances [8]. Note that our QA experiment uses more topics (i.e. questions) than the IR ones (97 vs 20), but fewer runs (10 vs 30).

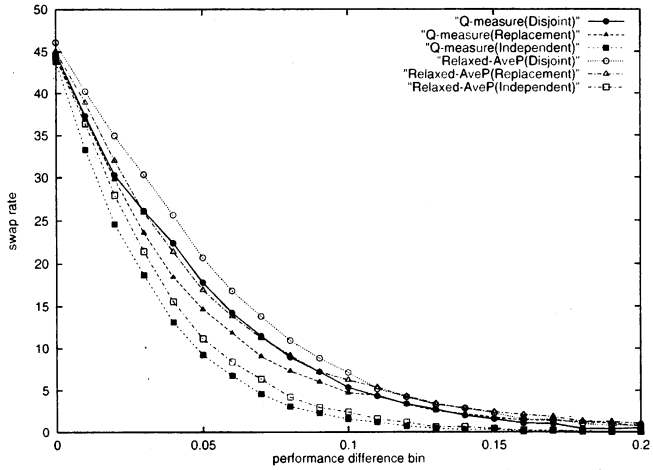

Figure 2. IR Experiment 1: Swap Rates for Q-measure and Relaxed-AveP.

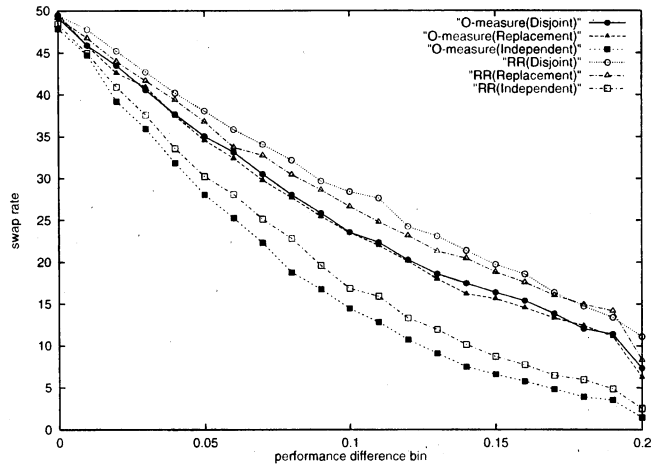

Figure 3. IR Experiment 2: Swap Rates for O-measure and RR.

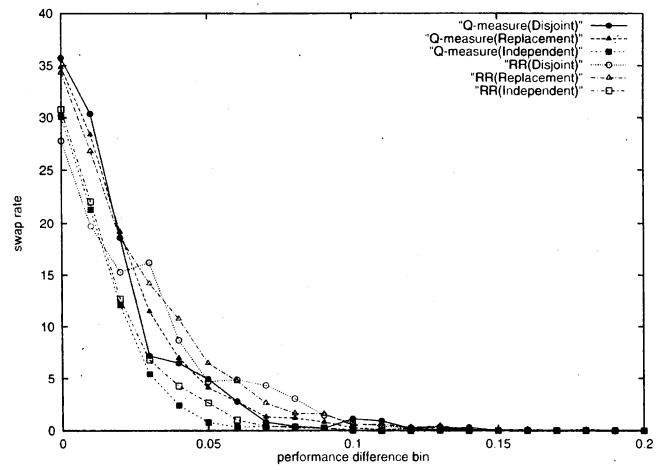

Figure 4. QA Experiment: Swap Rates for Q-measure and RR.

5 Results and Discussions

5.1 Swap Rate Curves

We first test **Hypothesis 1**. Figures 2-4 show the swap-rate/performance-difference-bin curves for a few metrics selected from IR Experiments 1&2 and 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.

From Figure 2, it can be observed that:

- For both Q-measure and Relaxed-AveP, **Independent** yields considerably lower swap rates than **Disjoint** and **Replacement**.
- For both Q-measure and Relaxed-AveP, **Disjoint** yields slightly higer swap rates than **Replacement** when the performance differences between Q_i and Q'_i are small (say, less than 0.1), but the curves overlap for larger performance differences, when swap rates are lower than, say 5%.
- 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 both O-measure and RR, **Independent** yields considerably lower swap rates than **Disjoint** and **Replacement**.
- For O-measure, the **Disjoint** and **Replacement** curves are almost identical. For RR, the results are similar, though **Disjoint** appears to yield slightly higer swap rates than **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** yield similar swap rates at least for the larger performance difference bins. Thus, **Hypothesis 1** is only partially supported: **Replacement** may actually be as good as **Disjoint** for drawing careful conclusions for guaranteeing a given condifence level.

5.2 Sensitivity Comparisons

Next, we test **Hypothesis 2**. 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 [10], which used the **Disjoint** method. 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)).

Figures 5 and 6 visualise Column (iv): It can be observed 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 topic sampling method is used but with different sets of randomly selected topic samples.) That is, the following observations we made in [10] do seem to hold true 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. Relaxed-AveP).

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

As for **Independent**, it appears that the impact of topic overlaps overshadows the differences across metrics, and that 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.

Table 3 provides a summary of our sensitivity comparisons in IR Experiment 2 in a way similar to those for Experiment 1. The table 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 [9]. Figure 7 visualises Column (iv) of this table, for 80% confidence only. Again, it is clear that **Disjoint** and **Replacement** yield similar results. Thus, the following observations we made in [9] 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.

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

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 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 [10]]				
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%

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 [10]]				
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%

Table 4(a) is a duplication from [8]. Figure 8 visualises Column (iv) of Table 4. Again, it is clear that **Disjoint** and **Replacement** yield similar results. Thus, the following observations we made in [8] 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

While all of our results support **Hypothesis 2**, **Hypothesis 1** is only partially supported. Thus, while **Independent** clearly underestimates swap rates and is not a good topic sampling strategy, **Replacement** may actually be as useful as **Disjoint** for carefully setting a minumum performance difference required to guarantee a given confidence level. This is a little surprising, given that topic overlaps do occur in **Replacement**.

Table 5 shows the average number of topics actually shared 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, the number of unique topics in a subset was 16.1 on average, and 6.2 unique topics were shared across two subsets on average. It is remarkable that **Replacement** yields results that are very 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. For example, sampling with replacement can be applied to *Buckley/Voorhees stability computation* as well [1, 10]. At the same time, since **Hypothesis 2** is supported, we are happy to conclude that our previous findings using **Disjoint** [8, 9, 10] are still valid. There is no evidence that the possible dependency *across trials* has any ill effect.

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 [9]]				
95% confidence				
Q-measure	0.10	.5490	18%	25.4%
Relaxed-AveP	0.11	.5392	20%	23.7%
O-measure	-	.8792	-	-
RR	-	.9750	-	-
90% confidence				
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% confidence				
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% confidence				
Q-measure	0.10	.6005	17%	27.1%
Relaxed-AveP	0.12	.5998	20%	21.3%
O-measure	-	.9313	-	-
RR	-	1.000	-	-
90% confidence				
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% confidence				
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% confidence				
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% confidence				
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% confidence				
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%

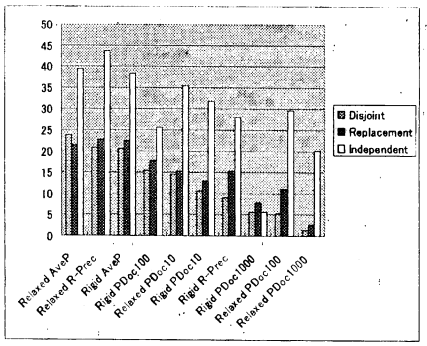

Figure 5. IR Experiment 1: Discrimination power at 95% confidence (binary relevance metrics).

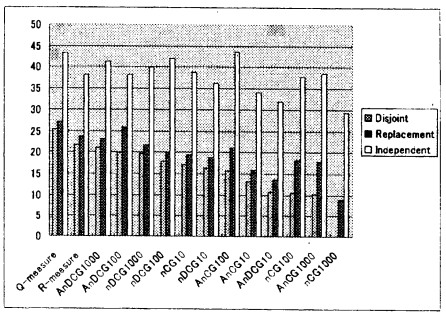

Figure 6. IR Experiment 1: Discrimination power at 95% confidence (graded relevance metrics).

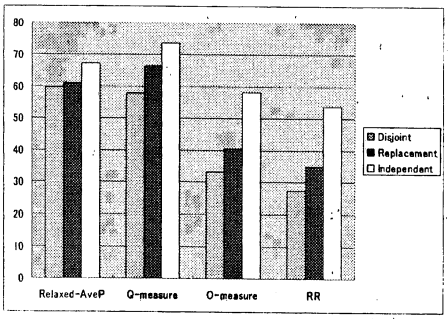

Figure 7. IR Experiment 2: Discrimination power at 80% confidence.

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 [8]]				
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. Number of unique overlapping topics between Q_i and Q'_i.

	IR	QA
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 do not believe there is a practical flaw in the original swap method, even though one advantage of the with-replacement method is that it can resample up to the size of the base topic set. We plan to carry out more experiments with other data and with new IR metrics, such as the *geometric mean* [13] version of Q-measure.

Acknowledgments

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

References

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

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

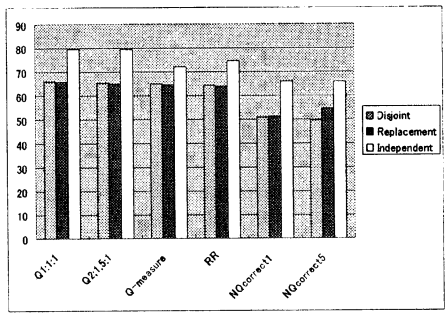

Figure 8. QA Experiment: Discrimination power at 95% confidence.

[3] 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.

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

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

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

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

[8] Sakai, T.: A Note on the Reliability of Japanese Question Answering Evaluation, IPSJ SIG Technical Reports FI-77-7, pp.57-64 / Digital Libraries No.25&26, pp.59-66, 2004.

[9] 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, 2005.

[10] Sakai, T.: The Reliability of Metrics based on Graded Relevance, *AIRS 2005 Proceedings*, to appear, 2005.

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

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

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