Overview of the TREC 2003 Robust Retrieval Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899

Abstract

The robust retrieval track is a new track in TREC 2003. The goal of the track is to improve the consistency of retrieval technology by focusing on poorly performing topics. In addition, the track brings back a classic, ad hoc retrieval task to TREC that provides a natural home for new participants.

An important component of effectiveness for commercial retrieval systems is the ability of the system to return reasonable results for every topic. Users remember abject failures. A relatively few such failures cause the user to mistrust the system and discontinue use. Yet the standard retrieval evaluation paradigm based on averages over sets of topics does not significantly penalize systems for failed topics. The robust retrieval track looks to improve the consistency of retrieval technology by focusing on poorly performing topics.

The task within the track was a traditional ad hoc task. An ad hoc task in TREC investigates the performance of systems that search a static set of documents using previously-unseen topics. For each topic, participants create a query and submit a ranking of the top 1000 documents for that topic. In addition to the standard evaluation by `trec_eval`, each run was also evaluated using two new effectiveness measures that focus on the effectiveness of the least-well-performing topics.

This paper presents an overview of the results of the track. The first section provides more details regarding the task and defines the new evaluation measures. The following section presents the systems' retrieval results, while Section 3 examines the new evaluation measures. Systems compare differently when evaluated on the new measures then when evaluated on standard measures such as MAP, suggesting that the new measures capture a different aspect of retrieval behavior. However, the measures are less stable than the traditional measures, and the margin of error associated with the new measures is large relative to the differences in scores observed in the track.

1 The Robust Retrieval Task

As noted above, the task within the robust retrieval track was a traditional ad hoc task. The topic set consisted of a total of 100 topics, 50 old topics taken from TREC topics 301–450 (TRECs 6–8) and 50 new topics. The document collection was the set of documents on TREC disks 4 and 5, minus the *Congressional Record*, since that is what was used for TRECs 6–8. This document set contains approximately 528,000 documents and 1,904 MB of text.

Since the focus of the track is on poorly performing topics, we wanted to ensure that there were topics that are generally difficult for systems to answer in the test set. We could not (purposely) construct a difficult topic set using only new topics since it is notoriously hard to predict whether or not a topic will be difficult a priori [5]. Instead, we used the effectiveness of the retrieval runs in TRECs 6–8 to construct a topic set of known-to-be-difficult topics. For each of topics 301–450, NIST created a box plot of the average precision scores for all runs (both automatic and manual) submitted to the ad hoc task in that topic's TREC. NIST then selected topics with low median average precision scores but with at least one (there was usually more than one) high outlier. The requirement for at least one system doing well on the topic was designed to eliminate flawed topics from the topic set. The set of old topics selected for the robust track is given in Figure 1.

While using old topics allowed NIST to construct a test set with certain properties, it also meant that full relevance data for these topics was available to the participants, and that systems were likely developed using those topics. NIST therefore created 50 new topics using the standard topic creation process as a type of control group. The 50 new topics are numbered 601–650. Since we could not control how the old topics had been used in the past, the assumption was that the old topics were fully exploited in any way desired in the construction of a participants' retrieval system. In

303	322	344	353	363	378	394	408	426	439
307	325	345	354	367	379	397	409	427	442
310	330	346	355	372	383	399	414	433	443
314	336	347	356	374	389	401	416	435	445
320	341	350	362	375	393	404	419	436	448

Figure 1: The set of old topics used in the robust track.

other words, participants were allowed to explicitly train on the 50 old topics in the test set if they desired to. The only restriction placed on the use of relevance data for the 50 old topics was that the relevance judgments could not be used during the processing of the submitted runs. This precluded such things as true (rather than pseudo) relevance feedback and computing weights based on the known relevant set.

The existing relevance judgments were used for the old topics; no new judgments of any kind were made for these topics. The new topics were judged by creating pools from all runs submitted to the track and using the top 125 documents per run. There was an average of 959 documents judged for each new topic. The assessors made three-way judgments of not relevant, relevant, or highly relevant for the new topics. Seven of the 50 new topics had no highly relevant documents, and another 14 topics had fewer than 5 highly relevant documents. All the evaluation results reported for the track consider both relevant and highly relevant documents as the relevant set since there are no highly relevant judgments for the old set. The number of relevant documents per topic for the old topic set ranged from a low of 5 to a high of 361 and an average of 88. For the new topic set, the minimum number of relevant was 4, the maximum was 115, and the average was 33.

While no new judgments were made for the old topics, we did form pools for those topics (using the top 100 retrieved per run) to examine the coverage of the original judgment set. Across the set of 50 old topics, an average of 61.4 % (minimum 43.2 %, maximum 79.7 %) of the documents in the pools created using robust track runs were judged. A relatively low number of judged documents is to be expected since the old topics were chosen because they were difficult, and there is known to be less overlap among the retrieved sets for difficult topics than for easier topics. Across the 78 runs that were submitted to the track, there was an average of 0.4 unjudged documents in the top 10 documents retrieved and 11.6 unjudged documents in the top 100 retrieved. These averages are inflated by a set of five runs that had very poor effectiveness (a cursory examination confirmed that the poor effectiveness was caused by retrieving documents that were indeed not relevant). Without these five runs, there was an average of 0.2 unjudged documents in the top 10 documents retrieved and 8.7 unjudged documents in the top 100 retrieved. There is still a tendency toward poorer runs having larger numbers of unjudged documents in the retrieved set, but such a bias is expected and is caused by poorer runs retrieving different, really-not-relevant documents.

Runs were evaluated using `trec_eval`, with average scores computed over the set of 50 old topics, the set of 50 new topics, and the combined set of 100 topics. Two additional measures were computed over the same three topic sets. The first measure was the percentage of topics that retrieved no relevant documents in the top ten retrieved. If one accepts "no relevant documents in the top ten retrieved" as an adequate definition of poorly performing topic, then this is a direct measure of the behavior of interest and is therefore a very intuitive and easily understood measure. It has the drawback of being a very coarse measure. That is, there are relatively few discrete values the measure can assume in theory, and the actual range of values seen in practice is much smaller than the theoretical range.

The second measure was suggested by Chris Buckley. One of the initial proposals for a measure for the track was to compute the mean of the average precision scores (MAP) for the system's worst X topics (as measured by average precision) rather than the entire set of topics as `trec_eval` does. In an attempt to pick a suitable X—big enough to make the measure stable but small enough to emphasize the poorly performing topics—the mean average precision over the worst X topics, MAP(X), was plotted as a function of X for several runs. Chris suggested that instead of picking a single point on the curve to use as the measure, to use the area underneath the MAP(X) vs. X curve as the measure. Just as MAP (the area underneath the recall-precision curve) emphasizes high precision but has a recall component, the area under the MAP(X) vs. X curve measure emphasizes the worst-performing topics, but also gives a general measure of quality. The measure as implemented for the track computes the area under the MAP(X) vs. X curve, but limits X to the worst quarter topics. That is, X is set from $1 \ldots 12$ for the 50-topic sets and $1 \ldots 25$ for the combined set. This measure is not exactly intuitive (it doesn't even have a better name than "area underneath the MAP(X) vs. X curve" yet), but it incorporates much more information than the percentage of topics with no relevant

Table 1: Groups participating in the robust track.

Chinese Academy of Sciences (CAS-NLPR)	Tsinghua University (Ma)
Fondazione Ugo Bordoni	University of Amsterdam
Hummingbird	University of Glasgow
Johns Hopkins University/APL	University of Illinois at Chicago
OcE Technologies	University of Illinois at Urbana-Champaign
Queens College, CUNY	University of Melbourne
Rutgers University (Neu)	University of Waterloo (MultiText)
Sabir Research, Inc.	Virginia Tech

in the top 10 retrieved. Note that since the measure is computed over the individual system's worst X topics, different systems' scores are computed over a different set of topics in general.

2 Retrieval Results

The robust track received a total of 78 runs from the 16 groups listed in Table 1. All of the runs submitted to the track were automatic runs. Participants were allowed to submit up to 5 runs. One of the runs was required to use only the description portion of the topic statements; the other runs could use any portion of the topic statements. There was a noticeable difference in effectiveness depending on the portion of the topic statement used: runs using all of the topic statement were better than those using selected fields, and runs using only the title field were worse than those using other portions. The retrieval results reported here are restricted to the runs that used just the description portion of the topic since that was the required run. There were 44 description-only runs submitted to the track.

Table 2 gives the evaluation scores for one run for each of the groups that submitted a description-only run (one group did not submit such a run by mistake). The table gives the scores for the four main measures used in the track as computed over the old topics only, the new topics only, and the combined set of 100 topics. The four measures are mean average precision (MAP), the average of precision at 10 documents retrieved (P10), the percentage of topics with no relevant in the top 10 retrieved (%no), and the area underneath the MAP(X) vs. X curve (area). The run shown in the table is the run with the highest MAP score as computed over the combined topic set; the table is sorted by this same value.

As expected given the way the topic set was constructed, the results show that as a set the 50 old topics are clearly much more difficult than the 50 new topics. The scores for all measures and all runs are better, usually much better, for the new topics than for the old. While all systems score better on the new set than the old, the amount of improvement is not uniform, so the relative ordering of systems is different for the two topic sets. We can quantify how different the relative orderings are by computing the Kendall τ correlation between system rankings using each of the topics sets in turn. A system ranking is an ordering of the runs by decreasing score of an effectiveness measure. The Kendall τ correlation measures the similarity between two rankings as a function of the number of pairwise swaps needed to turn one ranking into the other. The τ ranges between -1.0 and 1.0 where the expected correlation between two randomly generated rankings is 0.0 [2]. Table 3 shows the system rankings for the 44 description-only runs for each of the four evaluation measures of Table 2 for both topics sets. The ranking for the old topic set is given on the top and the ranking for the new topic set on the bottom. Each run is represented by a single character in the rankings. When two runs have a tied score for one measure they are ranked according to their MAP scores for that topic set. The last column in Table 3 gives the Kendall correlation between the two rankings. The τ values confirm that the rankings are different. The precise cause for the differences cannot be determined from this data since there are (at least) two confounded factors: different systems doing different amounts of training on the old topics and different systems being relatively more effective for difficult topics.

Are current retrieval systems handling the difficult topics better now than when the topics first appeared? We can give an approximate answer to this question by comparing the median and maximum scores obtained for each topic when computed over the set of runs submitted to the TREC in which the topic first appeared and the set of runs submitted to the robust track. Figure 2 shows this comparison using average precision as the evaluation measure. Since there were few description-only runs submitted to the previous ad hoc tasks, the sets of runs used to compute

Table 2: Evaluation results for the best description-only run per group as measured by MAP over the combined topic set. Runs are ordered by MAP over the combined topic set. Values given are the mean average precision (MAP), precision at rank 10 averaged over topics (P10), the percentage of topics with no relevant in the top ten retrieved (%no), and the area underneath the MAP(X) vs. X curve (area) as computed for the set of 50 old topics, the set of 50 new topics, and the combined set of 100 topics.

	Old Topic Set				New Topic Set				Combined Topic Set			
Tag	MAP	P10	%no	area	MAP	P10	%no	area	MAP	P10	%no	area
pircRBd2	0.177	0.382	12	0.009	0.403	0.532	4	0.082	0.290	0.457	8	0.022
uwmtCR0	0.150	0.370	14	0.011	0.403	0.536	8	0.052	0.276	0.453	11	0.018
aplrob03d	0.162	0.290	28	0.005	0.383	0.496	16	0.027	0.273	0.393	22	0.008
humR03de	0.148	0.352	24	0.004	0.377	0.484	14	0.023	0.263	0.418	19	0.009
VTDokrcgp5	0.130	0.314	18	0.005	0.382	0.502	12	0.023	0.256	0.408	15	0.008
fub03InOLe3	0.134	0.338	24	0.007	0.370	0.488	10	0.015	0.252	0.413	17	0.008
UIUC03Rd3	0.125	0.282	26	0.003	0.375	0.498	16	0.019	0.250	0.390	21	0.006
Sel78QE	0.124	0.292	32	0.003	0.363	0.452	18	0.007	0.243	0.372	25	0.003
THUIRr0305	0.117	0.312	16	0.009	0.370	0.508	8	0.048	0.243	0.410	12	0.015
SABIR03BF	0.106	0.220	34	0.003	0.346	0.464	12	0.030	0.226	0.342	23	0.006
UAmsT03RDesc	0.107	0.264	14	0.006	0.306	0.442	16	0.014	0.206	0.353	15	0.008
oce03noXbmD	0.092	0.240	24	0.003	0.305	0.446	16	0.013	0.199	0.343	20	0.005
MU03rob01	0.089	0.268	18	0.004	0.297	0.448	10	0.021	0.193	0.358	14	0.009
NLPR03vb50	0.095	0.334	6	0.005	0.259	0.460	8	0.013	0.177	0.397	7	0.007
rutcor0375	0.028	0.104	48	0.000	0.129	0.208	36	0.001	0.078	0.156	42	0.000

Table 3: System rankings and corresponding Kendall τ scores for the old and new topic sets.

Measure	Rankings $\frac{\text{Old Set}}{\text{New Set}}$	τ
MAP	WXCVoDLAqBHIFErhJnimNjpGlkegfMdRUOTQKSPcbZaY qWoVXCrLnljIEBmiHNADFpGhMJfegdkUORTQKSPcZbaY	0.772
P10	WXoLqIFERQPHVrjGpTSJhiCNgnDBmAMOlkUdefKcZbaY oWXqVjrFnClBImLGENpJMHeRQPifAhOUgkDTSdKZcbaY	0.562
% no	DRQPTSWXoGkgjArpOKqMJLBHIEmUFhnldCViNefZbcaY WVXpqojGMJgRQPIFfdOTSrlEBAeKLNDCnmHhkUiZcbaY	0.427
area	qojWpBIADXkEHMCgdRrGJFVhLOTfQmnileUSKPNcZabY WVqXojBApfDeCdJMGrLlOmFNngIkURKEHTQihPSZcbaY	0.560

the median and maximum average precision scores consisted of all automatic runs (i.e., runs using any combination of the fields in the topic statement). In the figure the median scores are plotted using filled symbols while the maximum scores are plotted using hollow symbols. The values computed using the set of runs submitted to the TREC in which the topic first appeared, called the Original TREC, are plotted as ovals; the values computed using robust track runs are plotted as triangles. The topics are sorted by decreasing median average precision score as computed using the robust track submissions. Median effectiveness for the robust track runs is generally better than for the original TREC runs, though for about 10 topics the original runs have a better median. The difference between medians is generally small (with a few notable exceptions). The maximum scores have larger differences and there are more topics for which the original runs had the better maximum score than for the robust runs. There were more different systems contributing to the Original TREC runs set than for the robust track runs set which may account for the better maximum scores. Nonetheless, it is clear that this old topic set remains a difficult set of topics.

Many of the participants used the robust track as a place to try new techniques for general ad hoc retrieval, without particularly focusing on the question of poorly performing topics. Both of the two groups with top-scoring runs, Queens College, CUNY and the Multitext group at the University of Waterloo, expanded the query using terms extracted from the Web (and possibly other document sets). Other groups experimented with new retrieval mod-

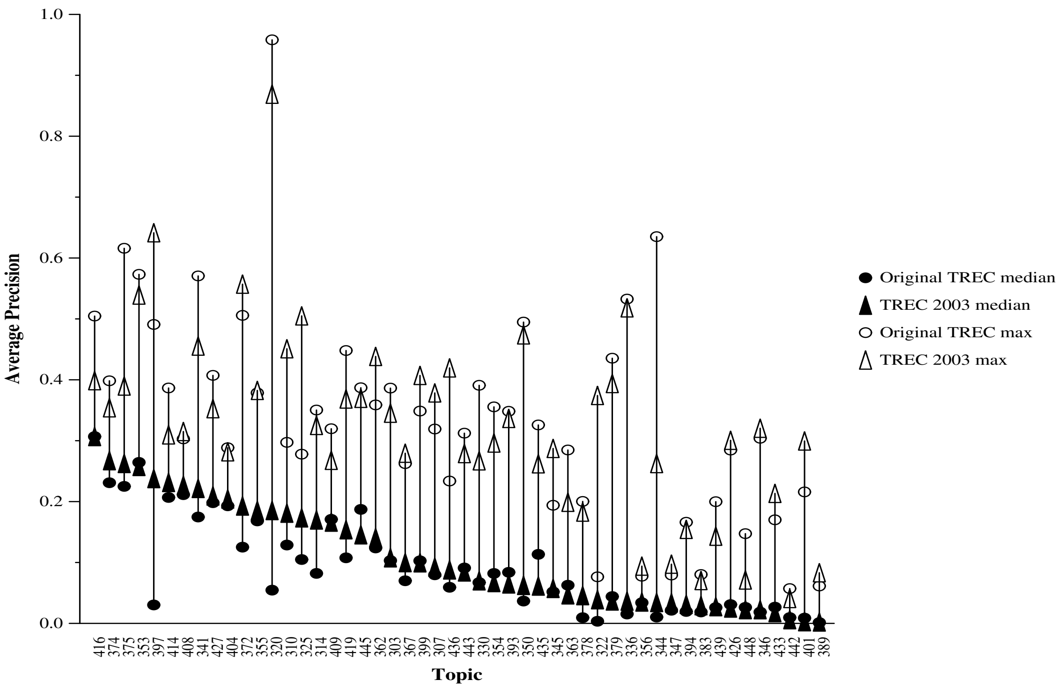


Figure 2: Median and maximum per-topic average precision scores for the old set of topics as computed using the runs submitted to the first TREC the topic was used in (Original TREC) and the TREC 2003 robust track runs (TREC 2003).

els or ranking functions (CAS-NLPR, Tsinghua University, University of Glasgow, University of Illinois Urbana-Champaign, Virginia Tech); with weighting schemes (OcE, Rutgers University, University of Illinois at Chicago, University of Melbourne); and with tokenization techniques (John Hopkins/APL, University of Amsterdam).

Almost all groups tried some version of query expansion based on pseudo-feedback. The query expansion improved average effectiveness, but did not help (and frequently hurt) the worst performing topics except when the expansion was done using a different corpus. This is not particularly surprising since the poorly performing topics are unlikely to have relevant documents in the top retrieved documents, and thus the feedback is as likely to harm as to help the results. After the qrels were published, the group from Fondazione Ugo Bordoni ran a series of experiments to see if they could predict when expansion would be beneficial based on an estimate of the MAP score of the initial retrieval result. When expanding only if the prediction determined it would be beneficial, they were able to both increase the MAP score and decrease the number of topics with no relevant retrieved as compared to their baseline.

Other approaches to increasing the effectiveness of the poorly performing topics included per-topic merging of results from different component runs and reordering the similarity-ranked list to maximize the number of retrieved-set document clusters with representatives in the top 10 ranks. Johns Hopkins/APL found modest success in decreasing the number of poorly performing topics by merging multiple runs, but also found that their results were far below the optimum theoretically obtainable from merging. Hummingbird tried to increase the diversity of the documents in the top 10 ranks by clustering the retrieved set and reranking the top 100 documents such that the top 10 documents were from different clusters. Unfortunately, the reranking did not lead to a significant increase in the number of topics with a relevant document in the top 10 retrieved.

3 Effectiveness Measures

One of the common themes of the participants' results was that query expansion improved MAP scores while not improving or even degrading the effectiveness of the worst topics. This demonstrates that MAP scores are essentially unaffected by the poorly performing topics. Mathematically, a poorly performing topic would have to improve dramatically to affect the MAP score since the magnitude of the MAP score is so much larger than an individual poorly

MAP	`WXVoqCLrDABIEnHFjilmhNJpGefMgkdUORTQKSPcZbaY`
P10	`WoXqVLFIjrEHRQPGCpnJBmNlihMTSgADOkUefdKZcbaY`
% no	`RQPWXDTSoGgjpqMOrAJkKVIBEFdLlHmUnhCefNiZcbaY`
area	`WqoXjpVBADGMOLJFdCrIkRgmfEnleHUTQNhiKPSZcbaY`

a) Rankings computed using combined topic set

	Old Topics			New Topics			All Topics		
	P10	% no	area	P10	% no	area	P10	% no	area
MAP	0.560	0.171	0.558	0.753	0.334	0.588	0.592	0.180	0.584
P10		0.433	0.444		0.463	0.535		0.397	0.493
% no			0.393			0.518			0.457

b) Kendall τ scores computed between rankings for all pairs of measures

Figure 3: Agreement among system rankings produced by different measures.

performing topic's average precision score.

This section examines the behavior of the two new measures that were introduced in the track. It shows that the new measures do emphasize poorly performing topics as designed, but because their scores are based on relatively few topics, they are more unstable than traditional measures and the margin of error associated with the new measures is large relative to the differences in scores observed in the track. More reliable measures are needed to support research on developing techniques for consistent retrieval.

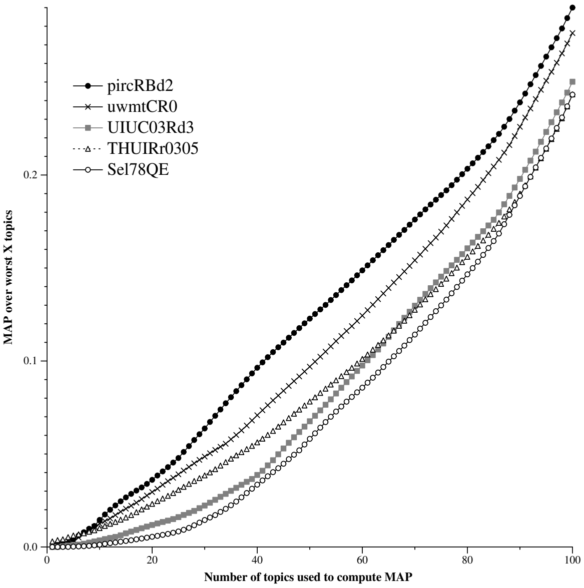
3.1 Agreement among measures

One way to show that different measures emphasize different factors is to examine whether they rank systems differently. We can produce systems rankings as above (using description-only runs), except that now instead of comparing rankings produced using different topic sets, we compare rankings produced using different evaluation measures. Figure 3 shows the agreement among system rankings for MAP, the average of precision at 10 documents retrieved, the percentage of topics with no relevant in the top 10 retrieved, and the area under the MAP(X) vs. X curve as computed over the set of old topics, the set of new topics, and the combined set of 100 topics. The system rankings themselves as computed over the combined set of topics are given at the top of the figure. The bottom of the figure shows the Kendall τ score computed between the rankings for each pair of measures.

The correlations are quite low, providing support for the contention that the measures are affected by different aspects of retrieval behavior. The correlation between MAP and the percentage of topics with no relevant documents in the top 10 documents is only slightly better than chance. While in theory such a low correlation with MAP means only that the two measures are emphasizing different aspects of retrieval, MAP has been shown to be an effective, stable measure [1] so in practice a low correlation with MAP can be a sign of an unstable measure. The stability of the new measures is investigated below.

The area under the MAP(X) vs. X curve measure depends on the greatest value that X assumes. This value reflects the trade-off in emphasis given to the worst-performing topics and the overall effectiveness of the system. The graphs in Figure 4 illustrate how the relative effectiveness among systems changes as X changes. The graphs plot MAP(X) vs. X using the combined topic set for a subset of the runs shown in Table 2. The left side of the figure shows the plot for all 100 values of X, and the right side of the figure shows the same plot restricted to $X = 1 \ldots 25$ so more detail can be seen. The value of the official area measure is the area underneath the curve plotted in the right side of the figure.

The graphs in Figure 4 make it clear that the relative order of systems ranked by their area scores does change depending on the maximal value of X. For example, the THUIRr0305 run has the best area score when $X \leq 6$, and is ranked third until approximately $X = 65$. However, the area measure is not highly sensitive to the maximal value of X, provided X is greater than about 10. We created system rankings based on the value of the area measure using the combined topic set and all description-only runs as the maximal value of X varied from 1 (i.e., the worst topic



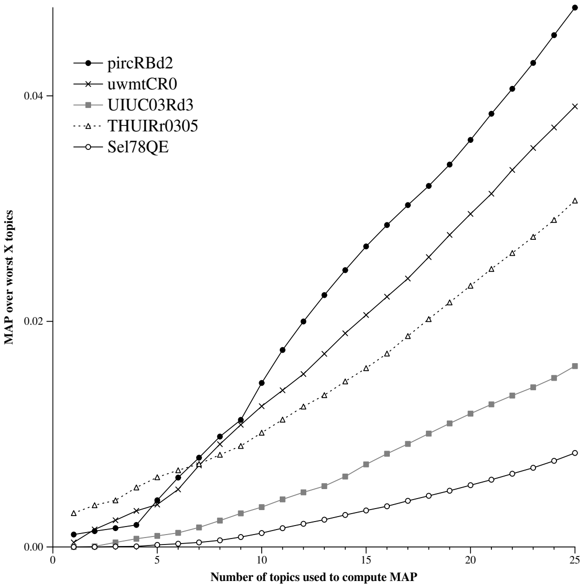


Figure 4: Plot of MAP(X) vs. X. The graph on the left side of the figure shows the entire range of $X = 1 \ldots 100$; the graph on the right is restricted to $X = 1 \ldots 25$ so more detail can be seen.

Table 4: Error rate and proportion of ties for different measures.

	Error Rate (%)	Proportion of Ties
MAP	1.4	0.171
P10	2.6	0.224
% no	9.1	0.090
area	8.4	0.040

determines the score) to 25. The Kendall τ correlations for $X < 10$ are small—in the 0.4 range when $X < 5$—but this is to be expected since measures based on the effectiveness of very few topics are known to be unstable. For $X > 10$, the τ values were greater then 0.85, and were generally greater than 0.95 when the X values being compared were within 5 of one another.

3.2 Stability of measures

The stability of the evaluation measures for topic sets containing 50 topics can be examined using a procedure similar to the one introduced by Buckley and Voorhees [1]. This procedure computes an error rate for an evaluation measure by counting how often the measure disagrees with respect to which of two systems being compared is preferred. Larger error rates imply a less stable measure.

We generated 1000 different test sets of size 50 topics each by randomly selecting 50 topics from the set of 100 topics used in the track. We evaluated all 78 runs submitted to the track on each of the 1000 test sets. For all pairs of runs A and B, we counted the number of test sets for which A evaluated as better than B ($A > B$), B evaluated as better than A ($B > A$), and A and B evaluated as equivalent ($A = B$). Two runs were considered equivalent if the difference in their scores was less than 5 % of the larger score. The error rate is defined as the sum over all run pairs of min($A > B, B > A$), divided by the total number of comparisons. The proportion of ties, $A = B$ divided by the total number of comparisons is also of interest since it indicates how much discrimination power a measure has. A measure with a low error rate but a high proportion of ties has little power.

Table 4 shows the error rate and proportion of ties computed for the four different measures. The numbers for MAP and P10 are close to the numbers reported by Buckley and Voorhees despite the different collection and slightly different methodology. As suspected, the error rates for the two new measures are substantially greater than for MAP and P10, though the proportion of ties for the new measures is substantially smaller than for the traditional measures.

The relative instability of the area and topics-with-no-relevant-retrieved measures is not difficult to understand. Numerically, a very low proportion of ties is likely to increase the error rate—the more decisions you make the more likely some of them are wrong, especially since fewer ties implies finer distinctions. In addition, the new measures are defined over a subset of the topics in the test set. For a test set of a given size, the score for the new measures will always be based on fewer topics than for the traditional measures.

3.3 Sensitivity of measures

While the higher error rates for the new measures are understandable, they do mean that there is much more uncertainty associated with a comparison of two systems when using one of these measures. Voorhees and Buckley introduced a procedure to empirically determine the relationship between the number of topics in a test set, the observed difference in scores of a particular measure (called Δ), and the likelihood that a single comparison of two runs leads to the correct conclusion [4]. Once established, the relationship can be used to derive the minimum difference in scores required for a certain level of confidence in the conclusion.

With 100 topics in the robust track test set, we can directly compute the relationship for topic set sizes up to 50 topics. Robust track runs should require somewhat smaller Δ's for the same level of confidence since they contain 100 topics. Voorhees and Buckley's original procedure used extrapolation to derive minimum differences for topic set sizes larger than those that could be directly computed, but extrapolation is not appropriate for the new measures since their values depend directly on the number of topics in the test set.

For topic sets of size 50, a run needs at least 11 fewer topics with no relevant in the top 10 retrieved to have 95% confidence that it is better than a second run. Over the 1000 topic sets of size 50 generated to estimate the error rate and comparing all pairs of runs submitted to the track, only 11.0 % of the comparisons had a difference at least this large. This is a small percentage that confirms that the measure is only able to distinguish grossly different systems. The area measure could distinguish even fewer systems. For the area measure, the minimum Δ computed for 95 % confidence was 0.025; only 4.6% of the comparisons across all run pairs and the 1000 test sets had a difference in area score greater than 0.025.

Note that the best area *score* (not difference) obtained by a run in the robust track over the old set of 50 topics was 0.0203, so all robust track runs would be considered to be in a single equivalence class if only the old set of topics were used. This set of topics is known to be difficult, and all systems did sufficiently poorly on it that the area measure is not sensitive enough to distinguish one run from another. The best score obtained by a robust track run over the 50 new topics was 0.1062 with 38.6 % of the comparisons between pairs of systems having a difference greater than 0.025, so the measure can distinguish among systems for this topic set. But the new topic set appears to be unusually good: over the 1000 randomly selected 50-topic test sets, the best area score obtained by any run was only 0.043, and as stated above only 4.6 % of the comparisons across all run pairs had a difference greater than 0.025. The topics-with-no-relevant-retrieved measure was much less affected by the particular topic set. For the old topic set, 13.9 % of run pairs had a difference of at least 11 topics; for the new topic set, 11.4 % of run pairs had a difference of at least 11 topics; and over the 1000 randomly selected sets, 11.0 % of run pairs had a difference of at least 11 topics.

4 Conclusion

The TREC 2003 robust retrieval tracks was an initial effort to improve the consistency of retrieval performance by focusing on poorly performing topics. The results of the track provide strong confirmation that average values of traditional effectiveness measures do not reflect poorly performing topics. New measures introduced in the track do emphasize systems' worst topics as designed. The new measures are defined over a subset of the topics in the test set, however, causing them to be much less stable than traditional measures for a given test set size. In turn, the instability causes the margin of error associated with the measures to be large relative to the differences in scores commonly observed.

The robust track will continue in TREC 2004. The current plan for the track is to repeat this year's task using the same fifty old topics (they remain difficult topics) and another set of 50 new topics. A new aspect of the evaluation in the track will be to test whether a system can predict which topics it will perform most poorly on. A similar evaluation strategy in the TREC 2002 question answering track demonstrated that accurately predicting whether a correct answer was retrieved is a challenging problem [3].

References

[1] Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In N. Belkin, P. Ingwersen, and M.K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, 2000.

[2] Alan Stuart. Kendall's tau. In Samuel Kotz and Norman L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 4, pages 367–369. John Wiley & Sons, 1983.

[3] Ellen M. Voorhees. Overview of the TREC 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, number NIST Special Publication 500-251, pages 57–68, 2002.

[4] Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, 2002.

[5] Ellen M. Voorhees and Donna Harman. Overview of the sixth Text REtrieval Conference (TREC-6). In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 1–24, August 1998. NIST Special Publication 500-240. Electronic version available at http://trec.nist.gov/pubs.html.