DRAFT
Overview of the TREC 2004 Robust Retrieval Track

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

Abstract

The robust retrieval track explores methods for improving the the consistency of retrieval technology by focusing on poorly performing topics. The task in the track is a traditional ad hoc retrieval task, but new evaluation measures emphasize a system's least effective topics. The 2004 edition of the track used 250 topics and required systems to rank the topics by predicted difficulty.

The ability to return at least passable results for any topic is an important feature of an operational retrieval system. While system effectiveness is generally reported as average effectiveness, an individual user does not see the average performance of the system, but only the effectiveness of the system on his or her requests. A user whose request is responded to by a set of documents that have no discernible connection with the information need is unlikely to be consoled by the fact that the system responds better to other people's requests.

The TREC robust retrieval track was started in TREC 2003 to investigate methods for improving the consistency of retrieval technology. The first year of the track had two main technical results:

1. The track provided ample evidence that optimizing average effectiveness using the standard Cranfield methodology and standard evaluation measures further improves the effectiveness of the already-effective topics, sometimes at the expense of the poor performers.

2. The track results demonstrated that measuring poor performance is intrinsically difficult because there is so little signal in the sea of noise for a poorly performing topic. New measures were devised that did emphasize poor performers, but because there is so little information the measures are unstable. Having confidence in the conclusion that one system is better than another using these measures requires larger differences in scores than are generally observed in practice when using 50 topics.

The retrieval task in the track is a traditional ad hoc task. In addition to calculating scores using `trec_eval`, each run is also evaluated using two measures introduced in the TREC 2003 track that focus more specifically on the least-well-performing topics. The TREC 2004 track differed from the initial track in two important ways. First, the test set of topics consisted of 249 topics, up from 100 topics. Second, systems were required to rank the *topics* by predicted difficulty, with the goal of eventually being able to use such predictions to do topic-specific processing.

This paper presents an overview of the results of the track. The first section describes the data used in the track, and the following section gives the retrieval results. Section 3 investigates how accurately systems can predict which topics are difficult. Since one of the main results of the TREC 2003 edition of the track was that the poor performance is hard to measure with 50 topics, section 4 examines the stability of the evaluation measures for larger topic set sizes.

1 The Robust Retrieval Task

As mentioned, the task within the robust retrieval track was a traditional ad hoc task. Since the TREC 2003 track had shown that 50 topics was not sufficient for a stable evaluation of poorly performing topics, the TREC 2004 track used a set of 250 topics (one of which was subsequently dropped due to having no relevant documents). The topic set consisted of 200 topics that had been used in some prior TREC plus 50 topics created for this year's track. The 200 old topics were the combined set of topics used in the ad hoc task in TRECs 6–8 (topics 301–450) plus the topics developed for the TREC 2003 robust track (topics 601–650). The 50 new topics created for this year's track are

Table 1: Relevant document statistics for topic sets.

Topic Set	Number of topics	Mean Relevant per Topic	Minimum # Relevant	Maximum # Relevant
Old	200	76.8	3	448
New	49	42.1	3	161
Hard	50	88.3	5	361
Combined	249	69.9	3	448

topics 651–700. The document collection was the set of documents on TREC disks 4 and 5, minus the *Congressional Record*, since that was the document set used with the old topics in the previous TREC tasks. This document set contains approximately 528,000 documents and 1,904 MB of text.

In the TREC 2003 robust track, 50 of the topics from the 301–450 set were distinguished as being particularly difficult for retrieval systems. These topics each had low median average precision scores but at least one high outlier score in the initial TREC in which they were used. Effectiveness scores over this topic set remained low in the 2003 robust track. This topic set is designated as the “hard” set in the discussion below.

While using old topics allows the test set to contain many topics with at least some of the topics known to be difficult, it also means that full relevance data for these topics is available to the participants. 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 other words, participants were allowed to explicitly train on the old topics if they desired to. The only restriction placed on the use of relevance data for the 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 three runs per group and using the top 100 documents per run. There was an average of 704 documents judged for each new topic. The assessors made three-way judgments of not relevant, relevant, or highly relevant for the new topics. As noted above, topic 672 had no documents judged relevant for it, so it was dropped from the evaluation. An additional 10 topics had no documents judged highly relevant. All the evaluation results reported for the track consider both relevant and highly relevant documents as the relevant set. Table 1 gives the total number of topics, the average number of relevant documents, and the minimum and maximum number of relevant documents for a topic for the four topic sets used in the track.

While no new judgments were made for the old topics, NIST did form pools for those topics to examine the coverage of the original judgment set. Across the set of 200 old topics, an average of 70.8% (minimum 36.6%, maximum 93.7%) of the documents in the pools created using robust track runs were judged. Across the 110 runs that were submitted to the track, there was an average of 0.3 (min 0.0, max 2.9) unjudged documents in the top 10 documents retrieved, and 11.2 (min 2.9, max 37.5) unjudged documents in the top 100 retrieved. The runs with the largest number of unjudged documents were also the runs that performed the least well. This make sense in that the irrelevant documents retrieved by these runs are unlikely to be in the pools. While it is possible that the runs were scored as being ineffective *because* they had large numbers of unjudged documents, this is unlikely to be the case since the same runs were ineffective when evaluated over just the new set of topics.

Runs were evaluated using `trec_eval`, with average scores computed over the set of 200 old topics, the set of 49 new topics, the set of 50 hard topics, and the combined set of 249 topics. Two additional measures that were introduced in the TREC 2003 track were computed over the same four topic sets [6]. The *%no* measure is the percentage of topics that retrieved no relevant documents in the top ten retrieved. The *area* measure is the area under the curve produced by plotting MAP(X) vs. X when X ranges over the worst quarter topics. Note that since the area measure is computed over the individual system’s worst X topics, different systems’ scores are computed over a different set of topics in general.

Table 2: Groups participating in the robust track.

Chinese Academy of Sciences (CAS-NLPR)	Fondazione Ugo Bordoni
Hong Kong Polytechnic University	Hummingbird
IBM Research, Haifa	Indiana University
Johns Hopkins University/APL	Max-Planck Institute for Computer Science
Peking University	Queens College, CUNY
Sabir Research, Inc.	University of Glasgow
University of Illinois at Chicago	Virginia Tech

2 Retrieval Results

The robust track received a total of 110 runs from the 14 groups listed in Table 2. All of the runs submitted to the track were automatic runs, (most likely because there were 250 topics in the test set). Participants were allowed to submit up to 10 runs. To have comparable runs across participating sites, one run was required to use just the description field of the topic statements, one run was required to use just the title field of the topic statements, and the remaining runs could use any combination of fields. There were 31 title-only runs and 32 description-only runs submitted to the track. There was a noticeable difference in effectiveness depending on the portion of the topic statement used: runs using both the title and description fields were better than using either field in isolation.

Table 3 gives the evaluation scores for the best run for the top 10 groups who submitted either a title-only run or a description-only run. 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, the difficult topics, and all 249 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.

One obvious aspect of the results is that the hard topics remain hard. Evaluation scores when computed over just the hard topics are approximately half as good as they are when computed over all topics for all measures except P(10) which doesn't degrade quite as badly. While the robust track results don't say anything about why these topics are hard, the 2003 NRRC RIA workshop [3] performed failure analysis on 45 topics from the 301–450 topic set. As one of the results of the failure analysis, Buckley assigned each of the 45 topics into 10 failure categories [1]. He ordered the categories by the amount of natural language understanding (NLU) he thought would be required to get good effectiveness for the topics in that category, and suggested that topics in categories 1–5 should be amenable to today's technology if systems could detect what category the topic was in. More than half of the 45 topics studied during RIA were placed in the first 5 categories.

Twenty-six topics are in the intersection of the robust track's hard set and the RIA failure analysis set. Table 4 shows how the topics in the intersection were categorized by Buckley. Seventeen of the 26 topics in the intersection are in the earlier categories, suggesting that the hard topic set should not be a hopelessly difficult topic set.

3 Predicting difficulty

A necessary first step in determining the problem with a topic is the ability to recognize whether or not it will be effective. Obviously, to be useful the system needs to be able to make this determination at run time and without any explicit relevance information. Cronen-Townsend, Zhou, and Croft suggested the *clarity measure*, the relative entropy between a query language model and the corresponding collection language model, as one way of predicting the effectiveness of a query [2]. The robust track required systems to rank the topics in the test set by predicted difficulty to explore how capable systems are at recognizing difficult topics. A similar investigation in the TREC 2002 question answering track demonstrated that accurately predicting whether a correct answer was retrieved is a challenging problem [5].

In addition to including the retrieval results for each topic, a robust track run ranked the topics in strict order from 1 to 250 such that the topic at rank 1 was the topic the system predicted it had done best on, the topic at rank 2 was the topic the system predicted it had done next best on, etc. This ranking was the *predicted* ranking. Once the

Table 3: Evaluation results for the best title-only run (a), and best description-only run (b) for the top 10 groups 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 200 old topics, the set of 49 new topics, the set of 50 hard topics, and the combined set of 249 topics.

	Old Topic Set				New Topic Set				Hard Topic Set				Combined Topic Set			
Tag	MAP	P10	%no	area	MAP	P10	%no	area	MAP	P10	%no	area	MAP	P10	%no	area
pircRB04t3	0.317	0.505	5	0.033	0.401	0.545	6	0.089	0.183	0.374	12	0.016	0.333	0.513	5	0.038
fub04Tge	0.298	0.484	13	0.019	0.351	0.480	12	0.046	0.145	0.338	22	0.008	0.309	0.483	12	0.021
uic0401	0.305	0.490	5	0.026	0.325	0.441	6	0.047	0.194	0.376	4	0.026	0.309	0.480	5	0.028
uogRobSWR10	0.296	0.461	16	0.010	0.322	0.453	12	0.021	0.136	0.316	26	0.003	0.301	0.459	15	0.011
vtumtitle	0.278	0.440	20	0.007	0.299	0.429	14	0.015	0.136	0.272	36	0.001	0.282	0.437	19	0.008
humR04t5e1	0.272	0.462	13	0.016	0.298	0.457	12	0.029	0.136	0.332	20	0.009	0.277	0.461	13	0.017
JuruTitSwQE	0.255	0.443	10	0.017	0.271	0.412	10	0.019	0.116	0.282	12	0.009	0.258	0.437	10	0.017
SABIR04BT	0.244	0.416	18	0.008	0.290	0.392	20	0.010	0.115	0.238	32	0.002	0.253	0.411	18	0.008
apl04rsTs	0.239	0.408	13	0.013	0.270	0.386	10	0.021	0.113	0.264	14	0.009	0.245	0.404	12	0.014
polyutp3	0.225	0.420	14	0.006	0.255	0.388	10	0.019	0.083	0.244	24	0.002	0.231	0.414	13	0.007

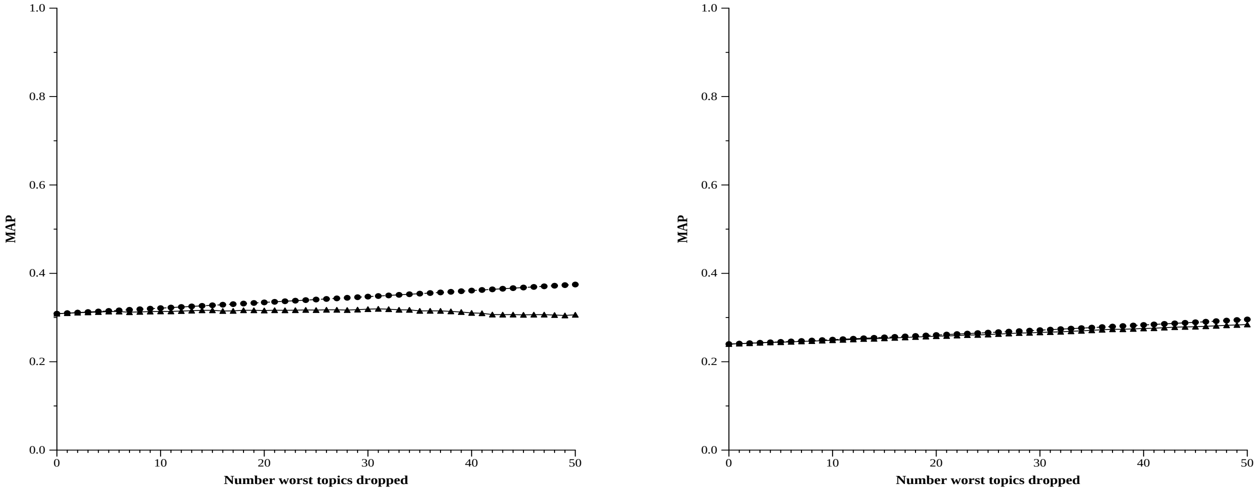
(a) title-only runs

pircRB04d4	0.316	0.507	8	0.023	0.407	0.547	2	0.074	0.162	0.382	12	0.013	0.334	0.515	7	0.028
fub04Dge	0.309	0.508	9	0.025	0.382	0.535	8	0.044	0.147	0.336	18	0.017	0.324	0.513	9	0.027
uogRobDWR10	0.286	0.454	16	0.007	0.374	0.529	12	0.023	0.131	0.296	28	0.002	0.303	0.468	15	0.008
vtumdesc	0.283	0.449	15	0.007	0.340	0.478	12	0.021	0.132	0.304	20	0.005	0.294	0.455	14	0.008
JuruTitDes	0.276	0.484	8	0.015	0.299	0.445	8	0.033	0.156	0.368	8	0.010	0.280	0.476	8	0.017
humR04d4e5	0.265	0.436	18	0.008	0.320	0.480	16	0.023	0.140	0.340	20	0.007	0.276	0.445	17	0.009
SABIR04BD	0.243	0.429	18	0.007	0.342	0.488	10	0.033	0.114	0.276	32	0.003	0.263	0.441	16	0.009
NLPR04OKapi	0.257	0.451	8	0.020	0.281	0.418	6	0.025	0.115	0.296	8	0.013	0.262	0.445	7	0.020
wdoqdn1	0.248	0.461	10	0.016	0.262	0.412	10	0.028	0.126	0.322	18	0.010	0.251	0.451	10	0.017
apl04rsDw	0.192	0.351	15	0.007	0.237	0.363	8	0.022	0.107	0.264	16	0.005	0.201	0.353	13	0.008

(b) description-only runs

Table 4: Failure categories of hard topics.

Category number	Category gloss	Topics
2	general technical failures such as stemming	353, 378
3	systems all emphasize one aspect, miss another required term	322, 419, 445
4	systems all emphasize one aspect, miss another aspect	350, 355, 372, 408, 409, 435, 443
5	some systems emphasize one aspect, some another, need both	307, 310, 330, 363, 436
6	systems all emphasize some irrelevant aspect, missing point of topic	347
7	need outside expansion of “general” term (e.g., expand Europe to individual countries)	401, 443, 448
8	need query analysis to determine relationship between query terms	414
9	systems missed difficult aspect	362, 367, 389, 393, 401, 404



(a) run `uic0404`

(b) run `humR04d5`

Figure 1: Effect of differences in actual and predicted rankings on MAP scores.

evaluation was complete, the topics were ranked from best to worst by average precision score; this ranking was the *actual* ranking.

One measure for how well two rankings agree is Kendall's τ [4]. Kendall's τ 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, and a τ of 1.0 indicates perfect agreement. The run with the largest τ between the predicted and actual ranking was the `uic0401` run with a τ of 0.623. Fourteen of the 110 runs submitted to the track had a negative correlation between the predicted and actual rankings. (The topic that was dropped from the evaluation was also removed from the rankings before the τ was computed.)

Unfortunately, Kendall's τ between the entire predicted and actual rankings is not a very good measure of whether a system can recognize poorly performing topics. The main problem is that Kendall's τ is sensitive to any difference in the rankings (by design). But for the purposes of predicting when a topic will be a poor performer, small differences in average precision don't matter, nor does the actual ranking of the very effective topics.

A more accurate representation of how well systems predict poorly performing topics is to look at how MAP scores change when successively greater numbers of topics are eliminated from the evaluation. The idea is essentially the inverse of the area measure: instead of computing MAP over the X worst topics, compute it over the best Y topics where $Y = 249 \ldots 199$ and the best topics are defined as the first Y topics in either the predicted or actual ranking. The difference between the two curves produced using the actual ranking on the one hand and the predicted ranking on the other is the measure of how accurate the predictions are. Figure 1 shows these curves plotted for the `uic0401` run, the run with the highest Kendall correlation, on the left and the `humR04d5` run, the run with the (second[1]) smallest difference between curves, on the right. In the figure, the MAP scores computed when eliminating topics from the actual ranking are plotted with circles and scores using the predicted ranking are plotted with triangles.

Figure 2 shows a scatter plot of the area between the MAP curves versus the Kendall τ between the rankings for each of the 110 runs submitted to the track. If the τ and area-between-MAP-curves agreed as to which runs made good predictions, the points would lie on a line from the upper left to the lower right. While the general tendency is roughly in that direction, there are enough outliers to argue against using Kendall's τ over the entire topic ranking for this purpose.

Figure 2 also shows that there is quite a range in systems' abilities to predict which topics will be poor performers for them. Twenty-two of the 110 runs representing 5 of the 14 groups had area-between-MAP-curves scores of 0.5 or less. Thirty runs representing six groups (all distinct from the first group) had area-between-MAP-curves scores of greater than 1.0 How much accuracy is required—including whether accurate predictions can be exploited at all—remains to be seen.

[1] The run with the smallest difference was an ineffective run where almost all topics had very small average precision scores.

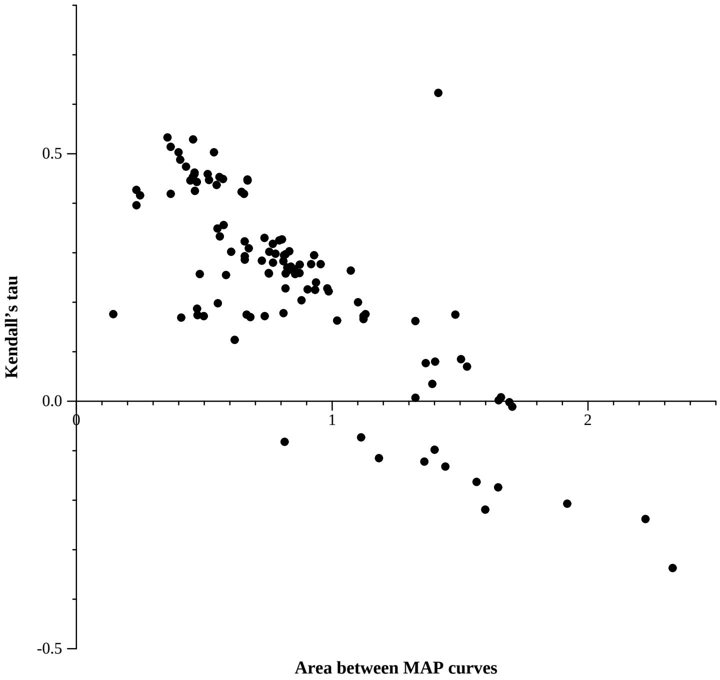


Figure 2: Scatter plot of area-between-MAP-curves vs. Kendall's τ for robust track runs.

Table 5: Error rate and proportion of ties for different measures and topic set sizes.

	50 Topics		75 Topics		100 Topics		124 Topics	
	Error Rate (%)	Proportion of Ties	Error Rate (%)	Proportion of Ties	Error Rate (%)	Proportion of Ties	Error Rate (%)	Proportion of Ties
MAP	2.4	0.144	1.3	0.146	0.7	0.146	0.3	0.145
P10	4.0	0.215	2.1	0.223	1.1	0.226	0.6	0.228
%no	14.1	0.107	11.8	0.146	9.6	0.064	7.6	0.065
area	10.6	0.040	7.9	0.041	5.9	0.042	4.7	0.042

4 Measure stability

Most TREC topic sets contain 50 topics. In the TREC 2003 robust track we showed that the %no and area measures that emphasize poorly performing topics are unstable when used with topic sets as small as 50 topics. The problem is that the measures are defined over a subset of the topics in the set causing them to be much less stable than traditional measures for a given topic set size. In turn, the instability causes the margin of error associated with the measures to be large relative to the difference in scores observed in practice.

The motivation for using 250 topics in the this year's track was to test the stability of the measures on larger topic set sizes. The empirical procedures to compute the error rates and error margins are the same as were used in the 2003 track [6] except the topic set size is varied. Since the combined topic set contained 249 topics, topic set sizes up to 124 (half 249) can be tested.

Table 5 shows the error rate and proportion of ties computed for the four different measures measures used in table 3 and four different topic set sizes: 50, 75, 100, and 124. The error rate shows how likely it is that a single comparison of two systems using the given topic set size and evaluation measure will rank the systems in the wrong order. For example, an error rate of 3% says that in 3 out of 100 cases the comparison will be wrong. Larger error rates imply a less stable measure. The proportion of ties 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 6: Sensitivity of measures: given is the critical value required to have an error rate no greater than 5% plus the percentage of comparisons over track run pairs that exceeded the critical value.

	50 Topics		75 Topics		100 Topics		124 Topics	
	Critical Value	% Significant	Critical Value	% Significant	Critical Value	% Significant	Critical Value	% Significant
%no	11 (22%)	3.8	16 (21%)	3.9	11 (10%)	15.7	13 (10%)	16.3
area	0.025	16.5	0.020	38.6	0.015	62.4	0.015	68.8

The error rates computed for topic set size 50 are somewhat higher than those computed for the TREC 2003 track, probably reflecting the greater variety of topics from which to choose. The general trends in the error rates are strong and consistent: error rate decreases as topic set size increases, and the %no and area measures have a significantly higher error rate than MAP or P(10) at equal topic set sizes.

Using the standard of no larger than a 5% error rate, the area measure can be used with test sets of at least 124 topics, while the %no measure requires still larger topics sets. Note that since the area measure is defined using the worst quarter topics, a 124 topic set size implies the measure is using 31 topics in its computation. While this is good for stability, it is no longer as focused on the very poor topics.

The error rates shown in table 5 assumed two runs whose difference in score was less than 5% of the larger score were equally as effective. By using a larger value for the difference before deciding two runs are different, we can decrease the error rate for a given topic set size (because the discrimination power is reduced) [7]. Table 6 gives the critical value required to to obtain no more than a 5% error rate for a given topic set size. For the area measure, the critical value is the minimum difference in area scores needed. For the %no measure, the critical value is the number of additional questions that must have no relevant in the top ten, also expressed as a percentage of the total topic set size. Also given in the table is the percentage of the comparisons that exceeded the critical value when comparing all pairs of runs submitted to the track over all 1000 topic sets used to estimate the error rates. This percentage demonstrates how sensitive the measure is to score differences encountered in practice.

The sensitivity of the %no measure does increase with topic set size, but the sensitivity is still very poor even at 124 topics. While intuitively appealing, this measure is just too coarse to be useful unless there are massive numbers of topics. The sensitivity of the area measure is more reasonable. The area measure appears to be an acceptable measure for topic set sizes of at least 100 topics.

5 Conclusion

Track future to be discussed at the conference.

References

[1] Chris Buckley. Why current IR engines fail. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Reserach and Development in Information Retrieval*, pages 584–585, 2004.

[2] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.

[3] Donna Harman and Chris Buckley. The NRRC Reliable Information Access (RIA) Workshop. In *Proceedings of the Twenty-Seventh Annual International ACM SIGIR Conference on Reserach and Development in Information Retrieval*, pages 528–529, 2004.

[4] 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.

[5] Ellen M. Voorhees. Overview of the TREC 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*, pages 57–68, 2003. NIST Special Publication 500-251.

[6] Ellen M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 69–77, 2004.

[7] 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.