The TREC Robust Retrieval Track

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD 20899
ellen.voorhees@nist.gov

Abstract

The robust retrieval track explores methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The retrieval task in the track is a traditional ad hoc retrieval task where the evaluation methodology emphasizes a system's least effective topics. The most promising approach to improving poorly performing topics is exploiting text collections other than the target collection such as the web.

The track has also investigated appropriate evaluation measures to support the focus on ineffective topics. Traditional measure are dominated by the better-performing topics, and the first two measures used in the track that do emphasize the poorly performing topics are unstable in practice. A third measure, a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results, shows promise of giving appropriate emphasis to poorly performing topics while being more stable at equal topic set sizes.

1 Introduction

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 retrieves nothing of interest 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 retrieval task in the track is a traditional ad hoc task where a system returns a ranked list of up to 1000 documents in response to a previously-unseen topic (statement of information need). The results are evaluated using traditional measures as defined by trec_eval, and also by new measures that focus more specifically on the least-well-performing topics. The first year of the track had two main technical results [9, 8]:

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. Two new measures devised to emphasize poor performers did so, 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.

A primary goal of the TREC 2004 track was to investigate the number of topics required in a test set for the new measures to be stable in practice. Stability of the measures does increase with increasing topic set size as expected [10], but the increase is is slow and the number of topics required may be impractical. A third new measure devised at the end of the track and tested with the TREC 2004 track results shows more promise as being appropriately sensitive to the poorly performing topics as well as being stable in practice.

Of course, the track also provides a forum for research on techniques to improve ad hoc retrieval, especially techniques targeting difficult topics. A subset of the test set of topics consists of topics that were difficult for automatic systems when they were used in previous TREC tasks. These topics remain difficult for the track's systems: evaluation scores for the 50 topics distinguished as difficult are approximately half as good as average scores for the remaining topics.

This paper presents an overview of the TREC robust retrieval track. The first section provides details of the task, and the following section gives the retrieval results. Section 4 examines the problem of evaluating poorly performing topics. The final section looks at the future of the track.

2 The Robust Retrieval Task

The task within the robust retrieval track is a traditional ad hoc task. The latest running of the track (TREC 2004) used a set of 250 topics as the test set, one of which was subsequently dropped due to having no relevant documents. The 250 topic set consists of 200 topics that had been used in some prior TREC plus 50 topics created for the 2004 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 new topics created for the TREC 2004 track are 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.

A subset of 50 topics from the 301–450 set were distinguished as being difficult topics for retrieval systems. This set is designated as the "hard" set in the remainder of the paper. The topics in the hard set each had low median average precision scores but at least one high outlier score in the initial TREC in which they were used. Using old topics allows the test set to contain many topics with at least some of the topics known to be difficult, but 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 runs submitted to the track. This precluded such things as true (rather than pseudo) relevance feedback and computing weights based on the known 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.

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 [9]. 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 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

3 Retrieval Results

The TREC 2004 robust track received a total of 110 runs from 14 groups. All of the runs submitted to the track were automatic runs. 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 2 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.

3.1 Retrieval methods

All of the top-performing runs used the web to expand queries [4, 5, 1]. In particular, Kwok and his colleagues had the most effective runs in both TREC 2003 and 2004 by treating the web as a large, domain-independent thesaurus and supplementing the topic statement by its terms [4]. When performed carefully, query expansion by terms in a collection other than the target collection can increase the effectiveness of many topics, including poorly performing topics. Expansion based on the target collection does not help the poor performers because pseudo-relevance feedback needs some relevant documents in the top retrieved to be effective, and that is precisely what the poorly performing topics don't have. The web is not a panacea, however, in that some approaches to exploiting the web can be more harmful than helpful [12].

Other approaches to improving the effectiveness of poor performers included selecting a query processing strategy based on a prediction of topic effectiveness[13, 7], and reodering the original ranking in a post-retrieval phase [6, 11]. Weighting functions, topic fields, and query expansion parameters were selected depending upon the prediction of topic difficulty. Documents were reordered based on trying to ensure different aspects of the topic were all represented. While each of these techniques can help some topics, the improvement was not as consistent as expanding by an external corpus.

Table 2: 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
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
JuruDesQE	0.266	0.466	11	0.010	0.295	0.398	16	0.022	0.152	0.348	14	0.008	0.272	0.452	12	0.011
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
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
polyudp2	0.185	0.364	16	0.003	0.234	0.378	6	0.025	0.083	0.240	24	0.001	0.195	0.367	14	0.004

(b) description-only runs

Table 3: 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

3.2 Difficult topics

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 [2]. He ordered the categories by the amount of natural language understanding 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 3 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.

4 Evaluating Ineffectiveness

Most TREC topic sets contain 50 topics. The TREC 2003 robust track 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.

Table 4: 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

4.1 Stability of %no and area measure

The motivation for using 250 topics in the TREC 2004 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 [9] 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 4 shows the error rate and proportion of ties computed for the four different measures measures used in Table 2 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 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.

The error rates shown in the table for for topic set size 50 are somewhat higher than those computed for the TREC 2003 track, probably reflecting the greater variety of topics the error rate was computed from. 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 4 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) [10]. Table 5 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.

Table 5: 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 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. Note that the same argument applies to the "Success@10" measure (i.e., the number of topics that retrieve a relevant document in the top 10 retrieved) that is being used to evaluate tasks such as home page finding and the document retrieval phase of question answering.

The sensitivity of the area measure is more reasonable. The area measure appears to be an acceptably stable measure for topic set sizes of at least 100 topics, though as mentioned above, its emphasis on the worst performing topics lessens as topic size grows.

4.2 Geometric MAP

The problem with using MAP as a measure for poorly performing topics is that changes in the scores of better-performing topics mask changes in the scores of poorly performing topics. For example, the MAP of a run in which the effectiveness of topic A doubles from 0.02 to 0.04 while the effectiveness of topic B decreases 5% from 0.4 to 0.38 is identical to the baseline run's MAP. This suggests using a nonlinear rescaling of the individual topics' average precision scores before averaging over the topic set as a way of emphasizing the poorly performing topics.

The geometric mean of the individual topics' average precision scores has the desired effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between larger scores. The geometric mean is equivalent to taking the log of the the individual topics' average precision scores, computing the arithmetic mean of the logs, and exponentiating back for the final geometric MAP score. Since the average precision score for a single topic can be 0.0—and trec_eval reports scores to 4 significant digits—we take the expedient of adding 0.00001 to all scores before taking the log (and then subtracting 0.00001 from the result after exponentiating).

To understand the effect of the various measures, Figure 1 shows a plot of the individual topic average precision scores for three runs from the TREC 2004 robust track. For each run, the average precision scores are sorted by increasing score and plotted in that order. Thus the x-axis in the figure represents a topic rank and the y-axis is the average precision score obtained by the topic at that rank. The three runs were selected to illustrate the differences in the measures. The `pircRB04td2` run was the most effective run as measured by both standard MAP over all 249 topics and geometric MAP over all 249 topics. The `NLPR04clus10` run has relatively few abysmal topics and also relatively few excellent topics, while the `uogRobLWR10` run has relatively many of both abysmal and excellent topics. The evaluation scores for these three runs are given in Table 6. The `uogRobLWR10` run has a better standard MAP score than the `NLPR04clus10` run, and a worse area and geometric MAP score. The P(10) score for the two runs are essentially identical.

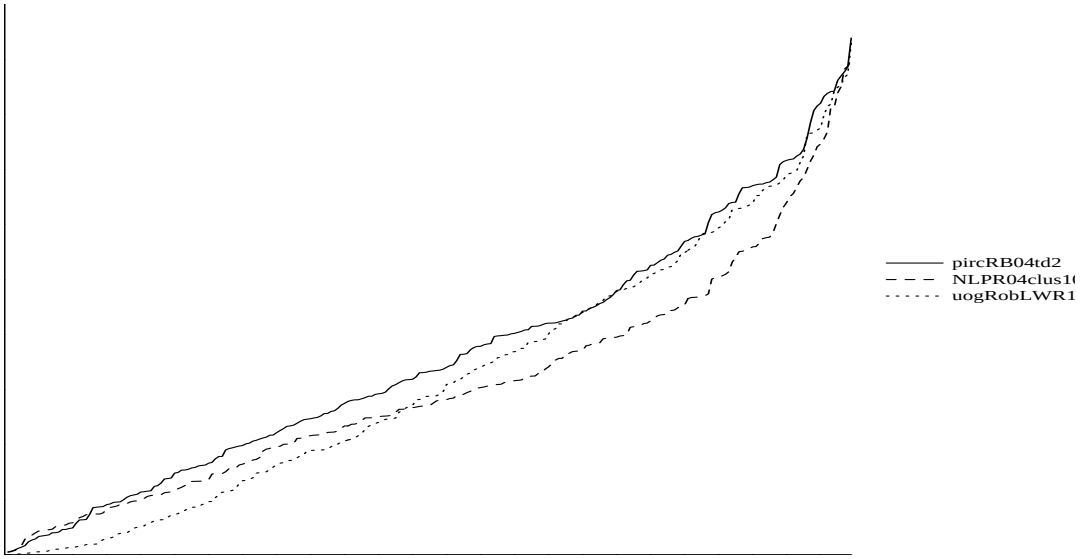

Figure 1: Individual topic average precision scores for three TREC 2004 runs.

Table 6: Evaluation scores for the runs of Figure 1.

	MAP	geometric MAP	P10	area	%no
pircRB04td2	0.359	0.263	0.541	0.047	4
NLPR04clus10	0.306	0.230	0.449	0.048	8
uogRobLWR10	0.320	0.176	0.448	0.015	11

Table 7: Error rate and proportion of ties computed over different topic set sizes for the geometric MAP measure.

Topic Set Size	Error Rate (%)	Proportion of Ties
25	9.1	0.081
50	5.2	0.086
63	4.1	0.088
75	3.4	0.090
100	2.3	0.092
124	1.5	0.094

Table 7 shows that the geometric mean measure is also a stable measure. The table gives the error rate and proportion of ties for geometric MAP for various topic set sizes. As in Table 4, the geometric MAP's error rates are computed assuming a difference in scores less than 5% of the larger score is a tie. Compared to the error rates for the measures given in Table 4, geometric MAP's error rate is larger than both standard MAP and P(10) for equal topic set sizes, but much reduced compared to the area and %no measures. The geometric MAP measure has the additional benefit over the area measure of being less complex. Given just the geometric MAP scores for a run over two sets of topics, the geometric MAP score for that run on the combined set of topics can be computed, which is not the case for the area measure.

5 Conclusion

The first two years of the TREC robust retrieval track have focused on trying to ensure that all topics obtain minimum effectiveness levels. The most promising approach to accomplishing this feat is exploiting text collections other than the target collection, usually the web. Believing that you cannot improve that which you cannot measure, the track has also examined evaluation measures that emphasize poorly performing topics. The geometric MAP measure is the most stable measure with a suitable emphasis.

The robust retrieval track is scheduled to run again in TREC 2005, though the focus of the track is expected to change. The current thinking is that the track will test the robustness of ad hoc retrieval technology by examining how stable it is in face of changes to the retrieval environment. To accomplish this, participants in the robust track will be asked to use their system for the ad hoc task in at least two of the other TREC tracks (for example, genomics and terabyte or terabyte and HARD). Within the robust track, same-system runs will be contrasted to see how differences in the tasks affect performance. Runs will also be evaluated using existing robust track measures, particularly geometric MAP.

Acknowledgements

Steve Robertson and Chris Buckley were instrumental in the development of the geometric MAP measure.

References

[1] Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Fondazione Ugo Bordoni at TREC 2004. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

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

[3] Chris Buckley and Donna Harman. Reliable information access final workshop report. ARDA Northeast Regional Research Center Technical Report, 2004.

[4] K.L. Kwok, L. Grunfeld, H.L. Sun, and P. Deng. TREC2004 robust track experiments using PIRCS. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[5] Shuang Liu, Chaojing Sun, and Clement Yu. UIC at TREC-2004: Robust track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[6] Christine Piatko, James Mayfield, Paul McNamee, and Scott Cost. JHU/APL at TREC 2004: Robust and terabyte tracks. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[7] Vassilis Plachouras, Ben He, and Iadh Ounis. University of Glasgow at TREC2004: Experiments in web, robust and terabyte tracks with Terrier. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[8] Ellen M. Voorhees. Measuring ineffectiveness. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 562–563, 2004.

[9] 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.

[10] 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.

[11] Jin Xu, Jun Zhao, and Bo Xu. NLPR at TREC 2004: Robust experiments. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[12] Kiduk Yang, Ning Yu, Adam Wead, Gavin La Rowe, Yu-Hsiu Li, Christopher Friend, and Yoon Lee. WIDIT in TREC-2004 genomics, HARD, robust, and web tracks. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.

[13] Elad Yom-Tov, Shai Fine, David Carmel, Adam Darlow, and Einat Amitay. Juru at TREC 2004: Experiments with prediction of query difficulty. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2005.