DRAFT:
Overview of the TREC 2005 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 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 2005 edition of the track used 50 topics that had been demonstrated to be difficult on one document collection, and ran those topics on a different document collection. Relevance information from the first collection could be exploited in producing a query for the second collection, if desired. The main measure for evaluating system effectiveness is "gmap", a variant of the traditional MAP measure that uses a geometric mean rather than an arithmetic mean to average individual topic results. This measure emphasizes the poorly-performing topics while being stable with as few as 50 topics.

Systems were also required to rank the topics by predicted difficulty. This task is motivated by the hope that systems will eventually be able to use such predictions to do topic-specific processing. Prediction quality is measured by the area between two curves of MAP scores, each curve plotting the change in MAP scores as the set of topics the average is computed over decreases by removing the current least effective topic. For one curve, the least effective topic is defined as the topic with the actual worst average precision score, while for the other curve the least effective topic is defined as the predicted worst topic.

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 request. The previous two editions of the robust track have demonstrated that average effectiveness masks individual topic effectiveness, and that optimizing standard average effectiveness measures usually harms the already ineffective topics.

A focus of the robust track since its inception has been developing the evaluation methodology for measuring how well systems avoid abysmal results for individual topics. Two measures introduced in the initial track were subsequently shown to be relatively unstable even for as many as 100 topics in the test set [3]. Those measures have been dropped from this year's results and have been replaced by the geometric MAP, or "gmap", measure. Gmap is computed as a geometric mean of the average precision scores of the test set of topics, as opposed to the arithmetic mean used to compute the standard MAP measure. Experiments using the TREC 2004 robust track results suggest that the measure gives appropriate emphasis to poorly performing topics while being stable with as few as 50 topics.

In addition to producing a ranked list of documents for each topic, systems were required to rank the *topics* by predicted difficulty. The motivation for this task is the hope that systems will eventually be 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. The final section looks at the future of the track.

1 The Robust Retrieval Task

The task within the robust retrieval track is a traditional ad hoc task. The document set used in this year's track was the AQUAINT Corpus of English News Text (LDC catalog number LDC2002T31). This collection consists of documents from three different sources: the AP newswire from 1998–2000, the New York Times newswire from 1998–2000, and the (English portion of the) Xinhua News Agency from 1996–2000. There are approximately 1,033,000 documents and 3 gigabytes of text in the collection.

The topic set consisted of 50 topics that had been used in ad hoc and robust tracks in previous years where they were run against the document set comprised of the documents on TREC disks 4&5 (minus the *Congressional Record*). These topics each had low median average precision scores in both the initial TREC in which they were used and in previous robust tracks, and were chosen for the track precisely because they are assumed to be difficult topics.

The 50 test topics were selected from a somewhat larger set based on having at least three relevant documents in the AQUAINT collection. NIST assessors were given a set of of topic statements and asked to search the AQUAINT collection looking for at least three relevant documents. Assessors were given the general guideline that they should spend no more than about 30 minutes searching for any one topic. The assessor stopped searching for relevant documents as soon as he or she found three relevant documents or when they felt they had exhausted the collection without finding three relevant documents. The topics for which fewer than three relevant documents were retrieved were discarded. The entire process was terminated as soon as a total of 50 topics with a minimum of three relevant documents was found.

The assessor who judged a topic on the AQUAINT data set was in general different from the assessor who originally judged the topic on the disks 4&5 collection. Thus, both the document set and the assessor differed between original runs using the topics and the robust 2005 runs. Nonetheless, systems were allowed to exploit the existing judgments in creating their queries for the track if the chose to do so. (Such runs were labeled as manual or "human-assisted" runs since the previous judgments were manually created. Runs that used other types of manual processing are also labeled as human-assisted.) Using the existing judgments in this manner is equivalent to the routing task performed in early TRECs.

The TREC 2005 HARD track used the same test collections as the robust track. Pools for document judging were created from one baseline and one final run for each HARD track participant, and one run per robust track participant. Because there were limited assessing resources, relatively shallow pools were created. The top 55 documents per topic for each pool run were added to the pools, producing pools that had a mean size of 756 documents (minimum 350, maximum 1390). While the pools are shallow, we expect the diversity of the runs added to the pools to make the pools sufficiently comprehensive. This hypothesis will be explored later in section 2.2. Documents in the pools were judged not relevant, relevant, or highly relevant, with both highly relevant and relevant judgments used as the relevant set for evaluation.

Runs were evaluated using `trec_eval`, and the standard measures are included in the evaluation report for robust runs. The primary measure for the track is the geometric MAP (gmap) score computed over the 50 test topics. Gmap was introduced in the TREC 2004 robust track [3] as a measure that emphasizes poorly performing topics while remaining stable with as few as 50 topics. Gmap takes a geometric mean of the individual topics' average precision scores, which has the 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 gmap score. The gmap value reported for robust track runs was computed using the current version of `trec_eval` (invoked with the -a option). In this implementation, all individual topic average precision scores that are less than 0.00001 are set to 0.00001 to avoid taking logs of 0.0.

2 Retrieval Results

2.1 Participant results

The robust track received a total of 74 runs from the 17 groups listed in Table 1. Participants were allowed to submit up to five runs. To have comparable runs across participating sites, if the participant submitted any automatic runs, one run was required to use just the description field of the topic statements, and one run was required to use just the title field of the topic statements. Four of the runs submitted to the track were human-assisted runs; the remaining seventy were completely automatic runs. Of the automatic runs, 24 runs were description-only runs, 34 were title-only runs, and 12 used various combinations of the topic statement.

Table 1: Groups participating in the robust track.

Arizona State University (Roussinov)	Chinese Academy of Sciences (ICT)
Ecole des Mines de Saint-Etienne	The Hong Kong Polytechnic University
Hummingbird	IBM Research, Haifa
Indiana University	IRIT/SIG
Johns Hopkins University/APL	Meiji University
Queens College, CUNY	Queensland University of Technology
RMIT University	Sabir Research, Inc.
University of Illinois at Chicago	University of Illinois at Urbana-Champaign
University of Massachusetts	

Table 2: Evaluation results for the best title-only and description-only runs for the top eight groups as measured by gmap.

Title-only Runs				Description-only Runs			
Run	gmap	MAP	P10	Run	gmap	MAP	P10
uic0501	0.233	0.310	0.592	ASUDE	0.178	0.289	0.536
indri05RdmnT	0.206	0.332	0.524	indri05RdmeD	0.161	0.282	0.498
pircRB05t2	0.196	0.280	0.542	ICT05qerfD	0.155	0.259	0.446
ICT05qerfTg	0.189	0.271	0.444	JuruDWE	0.129	0.230	0.472
UIUCrAt1	0.189	0.268	0.498	pircRB05d1	0.125	0.230	0.466
JuruTiWE	0.157	0.239	0.496	sab05rod1	0.114	0.184	0.404
humR05txle	0.150	0.242	0.490	humR05dle	0.114	0.201	0.432
wdf1t3qs0	0.149	0.235	0.456	wdf1t3qd	0.110	0.187	0.376

Table 2 gives the evaluation scores for the best run for the top eight groups who submitted either a title-only run or a description-only run. The table gives the gmap, MAP, and average P(10) scores over the 50 topics. The run shown in the table is the run with the highest gmap; the table is sorted by this same value.

2.2 Test collection results

There are some differences in the ranking of systems by different measures, though this is to be expected since the measures were designed to emphasize different aspects of retrieval. For all measures, title-only runs are more effective than description-only runs, the opposite of the effect found in earlier robust tracks when the topics were run on the TREC disks 4&5 document set. Retrieval effectiveness in general is better on the AQUAINT collection than the disks 4&5 collection as illustrated in figure 1. The figure shows box-and-whisker plots of the average precision scores for each of the topics across the set of description-only runs submitted to TREC 2004 (top plot) and TREC 2005 (bottom plot). The line in the middle of a box indicates the median average precision score for that topic. The plots are computed over different numbers of runs (24 description-only runs in TREC 2005 vs. 30 description-only runs in TREC 2004) and in general involve different systems, but aggregate scores should be valid to compare. The majority of topics have higher medians for TREC 2005 than for TREC 2004.

It is extremely unlikely that the entire set of systems that submitted description-only runs to TREC 2005 are significantly improved over TREC 2004 systems. Instead, these results remind us that topics are not inherently easy or difficult in isolation—the difficulty depends on the interaction between the information need and information source. If we accept that the topic set is easier on the AQUAINT corpus than on the disks 4&5 corpus, are there characteristics of the collections that would explain the difference?

There are a number of differences between the two test collections. The AQUAINT document set is much larger than the disks 4&5 document set: AQUAINT has more than one million documents and 3 gigabytes of text while

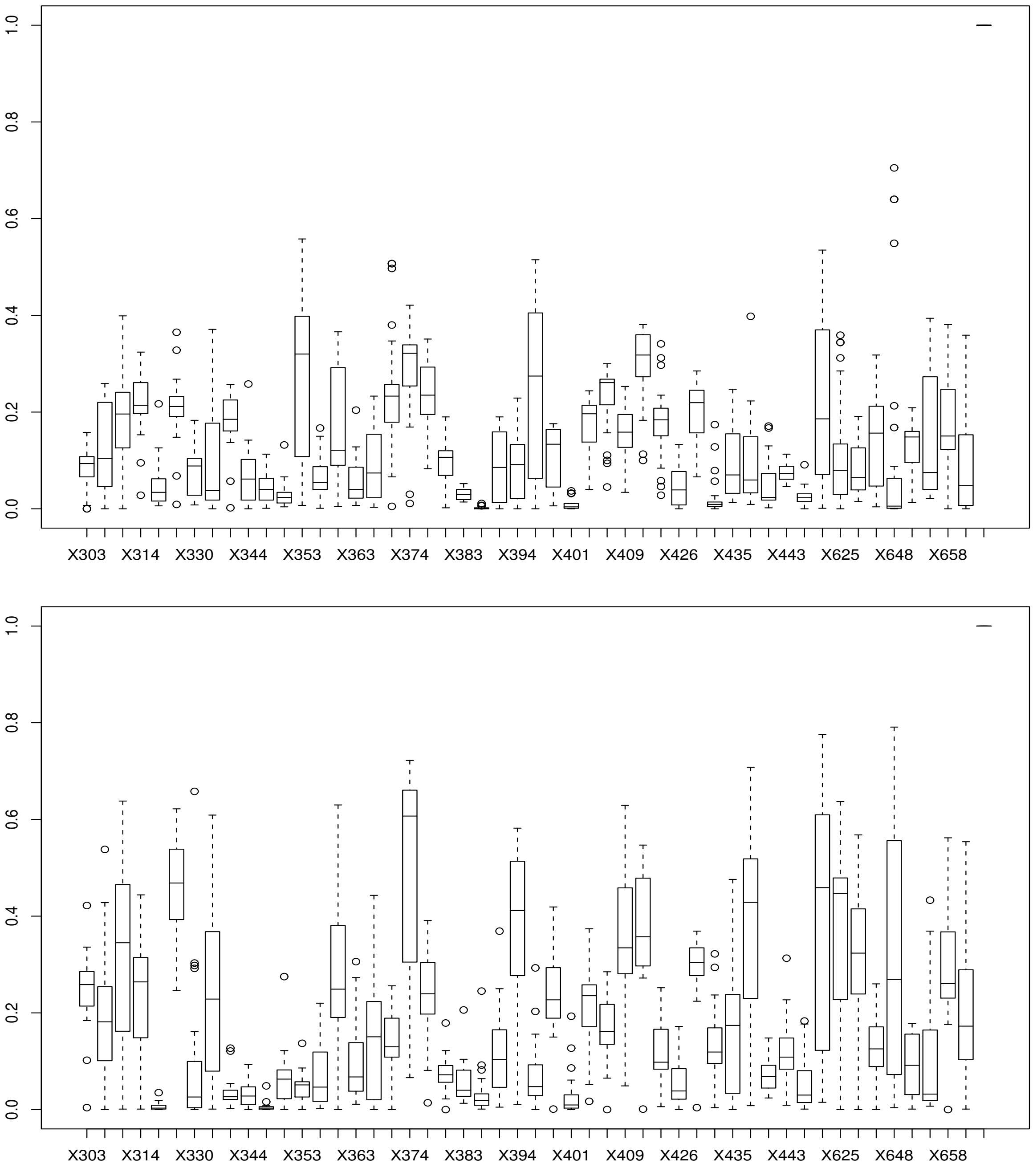

Figure 1: Box-and-whiskers plot of average precision scores for each of the 50 TREC 2005 test topics across description-only runs submitted to TREC 2004 (top) and TREC 2005 (bottom)

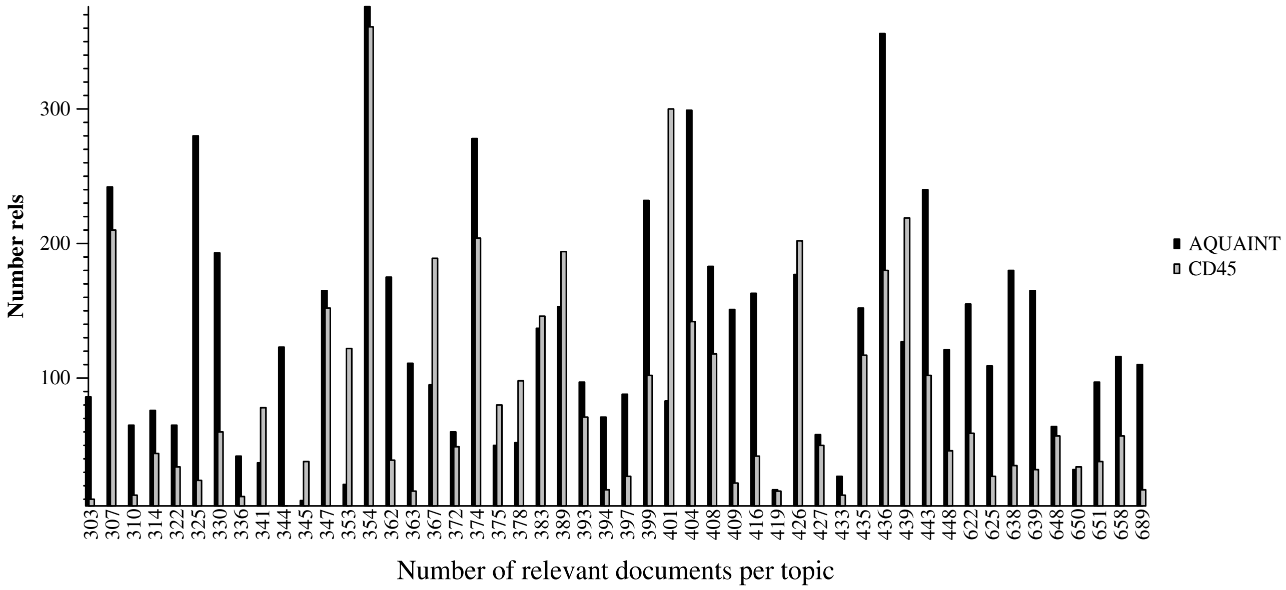

Figure 2: number of relevant documents per topic in the TREC 2005 test set in both the AQUAINT document set and the disks 4&5 document set

the disks 4&5 collection has 528,000 documents and 1904 MB of text. The AQUAINT collection contains newswire data only while the disks 4&5 collection contains the 1994 *Federal Register* and FBIS documents. The AQUAINT collection covers a later time period.

Earlier work in the TREC VLC track demonstrated that things being otherwise the same, P(10) scores increase when the size of the document set increases [1]. Yet the distribution of relevant documents across topics for the two test collections is not obviously different. For the AQUAINT test set, there is a mean of 131.2 relevant documents per topic, with a minimum number of relevant of 9 and a maximum number of relevant of 376. For the disks 4&5 collection the corresponding statistics are a mean of 86.4, minimum 5, and maximum 361. Figure 2 shows the number of relevant documents per topic for each of the two collections. Figure 3 shows a scatter plot of the number of relevant documents for the topic vs. the median average precision score across description-only runs. The topic number is used as the marker in the plot, with bold numbers representing TREC 2005 and lighter numbers TREC 2004. The plot confirms early results that topic difficulty (represented by median score) is not correlated with the number of relevant documents for the topic. Thus, simple counts of the number of relevant documents for the collections do not explain the different behavior.

Since absolute evaluation scores are known to vary when different relevance assessors are used [2], and there were different relevance assessors for the different document sets, perhaps the absolute scores for the topics are greater for the AQUAINT set, but the relative difficulty of the topics is the same for the two document sets. To test this hypothesis, we computed the Kendall τ score between topics ranked by median average precision score as computed over description-only runs submitted to TREC 2004 and TREC 2005. Figure 4 shows the two topic rankings where each of the 50 topics was assigned a different character code from a–X. The τ score between these rankings is only 0.326, demonstrating that the topics have different relative difficulty on the two document sets.

The pools from which the TREC 2005 test collection was created were more shallow than previous pools. Topics first used in the ad hoc tasks for TRECs 6–8 (topics 301–450) in particular had pools that were deeper and were comprised from more groups' runs than this year's pools. Perhaps the difference in retrieval behavior is an artifact of the pooling process. For example, title-only runs being more effective than runs that used more of the topic statement could be consistent with the AQUAINT collection's shallow pools containing only easy-to-retrieve relevant documents.

However, the tests that NIST runs to gauge the quality of pools do not indicate anything out of the ordinary with the TREC 2005 HARD-robust pools. One of these tests looks at the number of relevant documents that were contributed to the pools by exactly one group. Such a "unique relevant" document would not have been in the pool, and therefore

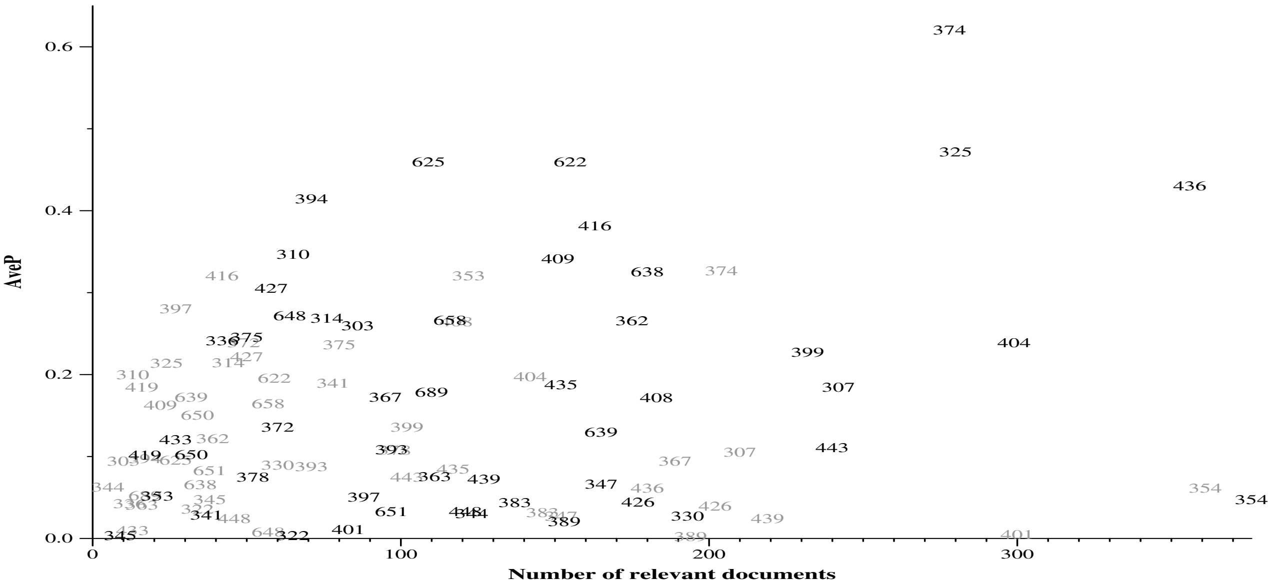

Figure 3: Scatter plot of median average precision score vs. number of relevant documents for that topic. Bold numbers represent TREC 2005 values and lighter numbers TREC 2004 values.

TREC 2004 topic ranking:
`smFzDrtIdfcCPiGSWEUAoubyQaqgxKVNRjnLXkhpHevlMOJTBw`

TREC 2005 topic ranking:
`sfPQLyFcERITdWoathCAKbXqDrSJNxUGpuMlmznHvVOjigwBke`

Figure 4: Ranking of TREC 2005 test topics by median average precision score across description-only runs.

would have been assumed to be irrelevant, if the one group that retrieved it had not participated. For the AQUAINT collection, there are topics for which a large percentage (slightly more than half) of the known relevant documents are unique relevants. But that has been true with past pools, including TREC 6–8 ad hoc pools, as well. Consistent with earlier pools, the groups that contributed many unique relevant documents are groups that did manual runs such as the human-assisted robust run `sab05ror1` from Sabir Research (405 unique relevant over 50 topics) and the manual HARD track runs `MARYB1` and `MARY05C1` from the University of Maryland (326 unique relevant over 50 topics). When runs were evaluated using the standard qrels and a qrels that eliminated the unique relevant for that group, the Sabir run suffered a 23% degradation in MAP, and the Maryland runs about a 12% degradation in MAP, but the mean change (counting these manual runs) was only a 3.2% degradation. For TREC-8 ad hoc, the manual runs from READWARE contributed 478 unique relevant over the 50 TREC-8 ad hoc topics and suffered a 10% degradation in MAP when their unique relevants were removed. The mean change in MAP across all TREC-8 pool runs with and without unique relevant documents was a 0.8% degradation.

Another hypothesis as to how pooling could affect evaluation scores is that the TREC 2004 runs may have had more unjudged documents in top ranks since no new pools for old topics were created. Since unjudged documents are considered to be not relevant, this could explain lower scores for TREC 2004 as compared to TREC 2005. Yet this hypothesis is not borne out, either. The 24 TREC 2005 description-only runs had a mean of 0.69 unjudged documents in the top 10 retrieved, and a mean of 25.63 unjudged documents in the top 100 retrieved. In contrast, the 30 TREC 2004 description-only runs had a mean of 0.24 unjudged in the top 10 retrieved and a mean of 11.9 unjudged in the top 100 retrieved (these means are computed over the 200 old topics in the TREC 2004 test set, which include the 50 TREC 2005 topics).

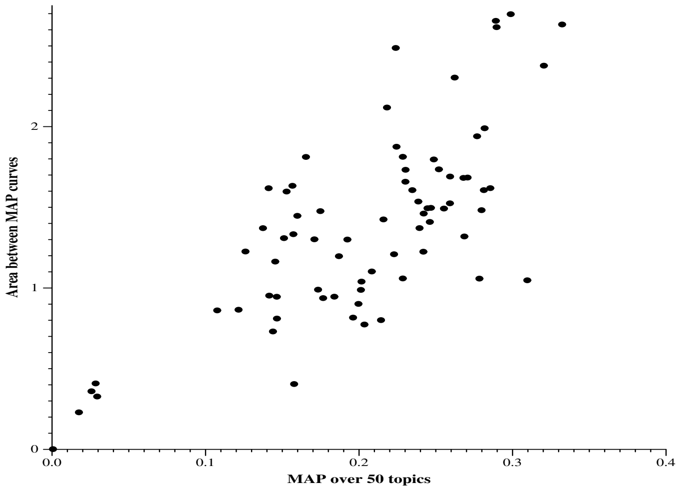

Figure 5: Scatter plot of area prediction measure vs. MAP for TREC 2005 robust track runs illustrating positive correlation of the scores.

In summary, the retrieval results from the robust track support the contention that aggregate retrieval behavior differs between the AQUAINT and disks 4&5 test collections for the same topic set, though to date we have not been able to determine any characteristics of the collections that explain the difference.

3 Predicting difficulty

Having a system predict whether it can effectively answer a topic is a necessary precursor to having that system modify its behavior to avoid poor performers. The difficulty prediction task was introduced into the robust track in TREC 2004. The task requires systems to rank the test set topics in strict order from 1 to 50 such that the topic at rank 1 is the topic the system predicted it had done best on, the topic at rank 2 is the topic the system predicted it had done next best on, etc.

The ranking submitted by a system is called its *predicted* ranking. The topics ranked by the average precision scores obtained by the system is called its *actual* ranking. The quality of a system's prediction is a function of how different the predicted ranking is from the actual ranking. The original measure used in 2004 for how the rankings differed was the Kendall τ measure between the two rankings, though it quickly became obvious that this is not a good measure for the intended goal of the predictions. The Kendall τ measure is sensitive to any change in the ranking across the entire set of topics, while the task is focused on the poor performers. A second way to measure the difference in the rankings is to look at how MAP scores change when successively greater numbers of topics are eliminated from the evaluation. In particular, compute the MAP score for a run over the best X topics where $X = 50 \dots 25$ and the best topics are defined as the first X 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.

While the area between the two curves is a better match than Kendall τ as a quality measure of predictions for our task, it has its own faults. The biggest fault is that the area between the MAP curves is dependent on the quality of the run itself, making the area measure alone unreliable as a gauge of how good the prediction was. For example, poorly performing runs will have a small area (implying good prediction) simply because there is no room for the graphs to differ. Figure 3 shows a scatter plot of the area measure vs. the MAP score over all 50 topics for each of the runs submitted to the TREC 2005 robust track. A perfect submission would have a MAP of 1.0 and an area score of 0.0, making the lower right corner of the graph the target. Unfortunately, the strong bottom-left to top-right orientation of the plot illustrates the dependency between the two measures. Some form of normalization of the area score by the full-set MAP score may render the measure more usable.

4 Future of the Track

To be discussed at the conference.

References

[1] David Hawking and Stephen E. Robertson. On collection size and retrieval effectiveness. *Information Retrieval*, 6(1):99–105, 2003.

[2] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management*, 36:697–716, 2000.

[3] Ellen M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*, pages 70–79, 2005.