Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness

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

Abstract Test collections have traditionally been used by information retrieval researchers to improve their retrieval strategies. To be viable as a laboratory tool, a collection must reliably rank different retrieval variants according to their true effectiveness. In particular, the relative effectiveness of two retrieval strategies should be insensitive to modest changes in the relevant document set since individual relevance assessments are known to vary widely.

The test collections developed in the TREC workshops have become the collections of choice in the retrieval research community. To verify their reliability, NIST investigated the effect changes in the relevance assessments have on the evaluation of retrieval results. Very high correlations were found among the rankings of systems produced using different relevance judgment sets. The high correlations indicate that the comparative evaluation of retrieval performance is stable despite substantial differences in relevance judgments, and thus reaffirm the use of the TREC collections as laboratory tools.

1 Introduction

The information retrieval (IR) field has a long tradition of using retrieval experiments on test collections to advance the state of the art [2, 10, 13]. In a common experimental scenario, a particular retrieval system configuration is used to run a set of queries against a set of documents and the results are evaluated in terms of *precision* and *recall*. (The accuracy of a system is measured by precision, the proportion of retrieved documents that are relevant; the coverage of a system is measured by recall, the proportion of relevant documents that are retrieved.) A second system configuration is used to produce a second set of retrieval results that are evaluated in turn and compared to the first set. The systematic variation of key system parameters allows the researcher to isolate factors that contribute to retrieval effectiveness.

The field has almost as long a history of criticism of this experimental paradigm [15, 5, 7]. The gist of the critics' complaint is that relevance is inherently subjective. Relevance judgments are known to differ across judges and for the same judge at different times [11]. Critics question how valid conclusions can be drawn when the process is based on something as volatile as relevance.

Experimentalists have had two responses to this charge. The first response is pragmatic: experience has shown that system improvements developed on test collections prove beneficial in other environments including operational settings. The second response is a small set of studies that show that the comparative performance of two retrieval strategies is quite stable despite marked differences in the relevance assessments themselves. Lesk and Salton [9] and Burgin [1] found no differences in the relative performance of different indexing methods when evaluated using different sets of relevance judgments. Cleverdon [3] found a few differences in the ranks of 19 different indexing methods when evaluated using four independent sets of judgments, but the correlation between the methods' rankings was always very high and the absolute difference in performance of the indexing methods was quite small.

The TREC workshops continue the tradition of experimental IR [6]. An important outcome of the workshops is a set of large test collections that are now widely used by the retrieval community. A variety of different organizations—including many organizations that do not participate in the TREC workshops themselves—use the collections to develop their own retrieval strategies. In addition, the common tasks defined in TREC have increased the number of cross-system comparisons that are made. Since so much research is being conducted on the collections, it is important that the collections reliably reflect the relative merit of different retrieval strategies.

Unfortunately, the results of the studies cited above are not directly applicable to the TREC collections. Each of the studies used small test collections (fewer than 1300 documents), and therefore had a correspondingly small number of relevant documents for each query. The TREC collections are many times larger (approximately 800,000 documents each) and some queries have hundreds of relevant documents. Also, the earlier studies each compared variants of the same system, so the effect of changes in relevance assessments across system types is unknown.

To verify that the TREC collections are reliable laboratory tools, NIST investigated the effect changes in the relevance assessments have on the evaluation of retrieval results. This report is a result of that investigation. The next section provides background on the process used to create the TREC relevance assessments. Section 3 describes the effect of varying the relevance assessments for the results submitted to two TREC conferences, and analyzes why the comparative evaluation of retrieval results is as stable as it is. The final section summarizes our findings.

To be presented at the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998.

2 TREC

The TREC conference series was started in 1992 to foster research in information retrieval by offering large test collections and a forum for researchers to discuss their work on a common problem [6]. To date, TREC has issued five compact disks of English documents (roughly 5 gigabytes of text) and 350 information need statements, called *topics*. Different subsets of the topics have been evaluated against different portions of the total document set, so TREC has built a series of test collections rather than a single test collection. See `http://trec.nist.gov/data/intro_eng.html` for a description of which topics and documents comprise the different test collections.

The process of creating a test collection in TREC has been relatively unchanged since TREC-3:

- Each relevance assessor creates a set of approximately ten candidate topics, and uses the candidate topics to search the target document collection. Fifty topics total are selected from the candidate topics based on the estimated number of relevant documents and load-balancing across assessors. Figure 1 gives two example TREC topics.
- The fifty topics are used in the TREC ad hoc task for that year. Participants use the topics to make a retrieval run with their system and submit a ranking of the top 1000 documents for each of the topics to NIST for evaluation. A participant may submit two or three different runs.
- NIST forms a document pool for each topic from the top 100 documents from each run that will be judged. (Some years all submitted runs are judged; when that is not feasible, an equal number of runs per participant is judged and the participants select which of their runs will be judged.)
- The relevance assessor that created the topic makes a binary decision of 'relevant' or 'not relevant' for each of the documents in that topic's pool. Any document that is not in the pool—and therefore is not judged—is assumed to be not relevant.

NIST evaluates each of the submitted runs using the relevance assessments created by this process. The official evaluation reports several different variants of precision and recall, such as the mean precision at various cut-off levels and a recall-precision graph. One particular measure, *mean average precision*, is often used as a single summary statistic for the entire run. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved. The mean average precision for a run consisting of multiple topics is the mean of the average precision scores of each of the topics in the run. In this paper, a *system ranking* is a list of the systems under consideration sorted by decreasing mean average precision.

3 The Stability of System Rankings

The investigation into the stability of system rankings was carried out in two parts. The main study used three independent sets of assessments that NIST obtained for each of the 49 topics used in the TREC-4 evaluation (Topics 202–250). A smaller follow-up study was performed using the TREC-6 topics (Topics 301–350) and a set of alternate relevance judgments made at the University of Waterloo. Each of these studies is described in detail in the remainder of this section.

Assessor Group	Overlap
Primary & A	.421
Primary & B	.494
A & B	.426
All three	.301

Table 1: Mean overlap for each assessor pair and the set of three assessors.

3.1 TREC-4 Relevance Assessments

To acquire the data necessary for the study, the TREC-4 relevance assessors were asked to judge additional topics once they had finished with the main TREC-4 assessing. The author of a topic is its primary assessor. After the primary assessor was finished with a topic, a new document pool was created for it. This new pool consisted of all of the relevant documents as judged by the primary assessor up to a maximum of 200 relevant documents (a random sample of 200 relevant documents was used if there were more than 200 relevant documents) plus 200 randomly selected documents that the primary assessor judged not relevant. The new pool was sorted by document identifier and given to two additional assessors (the secondary assessors) who each independently judged the new pool for relevance. Because of the logistics involved, a topic was given to whatever secondary assessor was available at the time, so some individual assessors judged many more topics than others. However, each topic was judged by three individuals.

Previous studies have used the *overlap* of the relevant document sets to quantify the amount of agreement among different sets of relevance assessments [9]. Overlap is defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets. Table 1 gives the mean overlap for each pair of assessors and the set of three assessors. Documents that the primary assessor judged relevant but that were not included in the secondary pool (because of the 200 document limit) were added as relevant documents to the secondary assessors' judgments for the analysis. The overlap shown in Table 1 for pairs of assessors is greater than the overlap in the earlier studies. The NIST assessors all have a similar background (retired information analyst), had the same training for the TREC task, and judged documents under identical conditions, although the primary assessor differs from the secondary assessors in that the primary assessor is the topic author. Despite the relatively high average overlap, there are significant differences in the judgments for some topics. For example, the primary assessor judged 133 documents as relevant for Topic 219, and yet no document was unanimously judged relevant. One secondary assessor judged 78 of the 133 irrelevant, and the other judged all 133 irrelevant (though judged one other document relevant). Across all topics, 30% of the documents that the primary assessor marked relevant were judged nonrelevant by both secondary assessors. In contrast, less than 3% of the documents judged nonrelevant by the primary assessor were considered relevant by both secondary assessors.

⟨ num ⟩ Number: 207

⟨ desc ⟩ What are the prospects of the Quebec separatists achieving independence from the rest of Canada?

⟨ num ⟩ Number: 312

⟨ title ⟩ Hydroponics

⟨ desc ⟩ Description:
Document will discuss the science of growing plants in water or some substance other than soil.

⟨ narr ⟩ Narrative:
A relevant document will contain specific information on the necessary nutrients, experiments, types of substrates, and/or any other pertinent facts related to the science of hydroponics. Related information includes, but is not limited to, the history of hydroponics, advantages over standard soil agricultural practices, or the approach of suspending roots in a humid enclosure and spraying them periodically with a nutrient solution to promote plant growth.

Figure 1: Sample topics from TREC-4 (207) and TREC-6 (312).

Defining different qrels sets

For a test collection, the important question is not so much how well assessors agree with one another, but how evaluation results change with the inevitable differences in assessments. To test this directly, we can evaluate the same runs using different sets of relevance judgments. A run is evaluated over a set of topics, and each topic has a set of judgments produced by each assessor. Call the concatenation of one judgment set per topic in the run a *qrels* (for query-relevance set). With three independent judgments for each of 49 topics, we can theoretically create 3^{49} different qrels by using different combinations of assessor's judgments for the topics, and evaluate the runs using each qrels. Note that each of these qrels might have been the qrels produced after the TREC conference if that set of assessors had been assigned those topics. To simplify the analysis that follows, we discarded Topic 214 since Secondary Assessor A judged no documents relevant for it[1]. That leaves 48 topics and 3^{48} possible qrels.

Three of the 3^{48} possible qrels are special cases. The original qrels set consists of the primary assessments for each topic—this is the qrels released after TREC-4 except that it lacks Topic 214. The set of judgments produced by the Secondary A judge for each topic, and the set of judgments produced by the Secondary B judge for each topic constitute the Secondary A qrels and the Secondary B qrels respectively. We created a sample of size 100,000 of the remaining qrels by randomly selecting one of the primary or secondary assessors for each of the 48 topics and combining the selected judgments into a qrels. Adding the three distinguished qrels to the sample gives a total of 100,003 qrels that were used to evaluate retrieval runs. Finally, two additional qrels, the union and intersection qrels, were created from the relevance judgments. In the union qrels a document is considered to be relevant to a topic if any assessor judged it relevant to that topic; in the intersection qrels a document is considered to be relevant to a topic if all three assessors judged it relevant to that topic. Because two topics had no documents that all assessors agreed were relevant (219 and 232), the intersection qrels contains only 46 topics.

There were 33 category A ad hoc runs submitted to TREC-4. We evaluated each of these runs against each of the qrels in the sample of 100,003 qrels and computed the sample mean of the mean average precision for each run. The means are plotted in Figure 2 where the runs are sorted by decreasing mean. The error bars in Figure 2 indicate the minimum and the maximum mean average precision obtained for that run over the sample. Also plotted in the figure are the mean average precision scores computed using the original, union, and intersection qrels. These points demonstrate how the system ranking changes for an individual qrels versus the ranking by the mean: a run with a symbol higher than the corresponding symbol of a run to its left would be ranked differently in the individual qrels ranking. For example, the `pircs2` and `uwgcl1` runs (shown in position 2 and 3 in Figure 2) would switch positions when evaluated by the Original qrels.

The plot in Figure 2 demonstrates that the mean average precision score *does* change depending on the qrels used in the evaluation. The difference between the minimum and maximum mean average precision values is greater than .05 for most runs. However, the changes are very highly correlated across runs. That is, if a particular run gets a relatively high score with a particular qrels, then it is very likely that the other runs will also get a relatively high score with that qrels. The union qrels (the triangle in Figure 2) is close to the top of the range for each run, for example.

The correlation can be quantified by using a measure of association between the different system rankings. We used a correlation based on Kendall's tau [14] as the measure of association between two rankings. Kendall's tau computes the distance between two rankings as the minimum number of pairwise adjacent swaps to turn one ranking into the other. The distance is normalized by the number of items being ranked such that two identical rankings produce a correlation of 1.0, the correlation between a ranking and its perfect inverse is -1.0, and the expected correlation of two rankings chosen at random is 0.0.

We computed the mean of the Kendall correlations in the sample of 100,003 qrels in two ways. In the first case, we took the mean of the correlations between the ranking produced by the original qrels and the rankings produced by each of the other 100,002 qrels. In the second case, we took a random subsample of 1000 qrels and computed the

[1] The means reported in Table 1 do not include Topic 214.

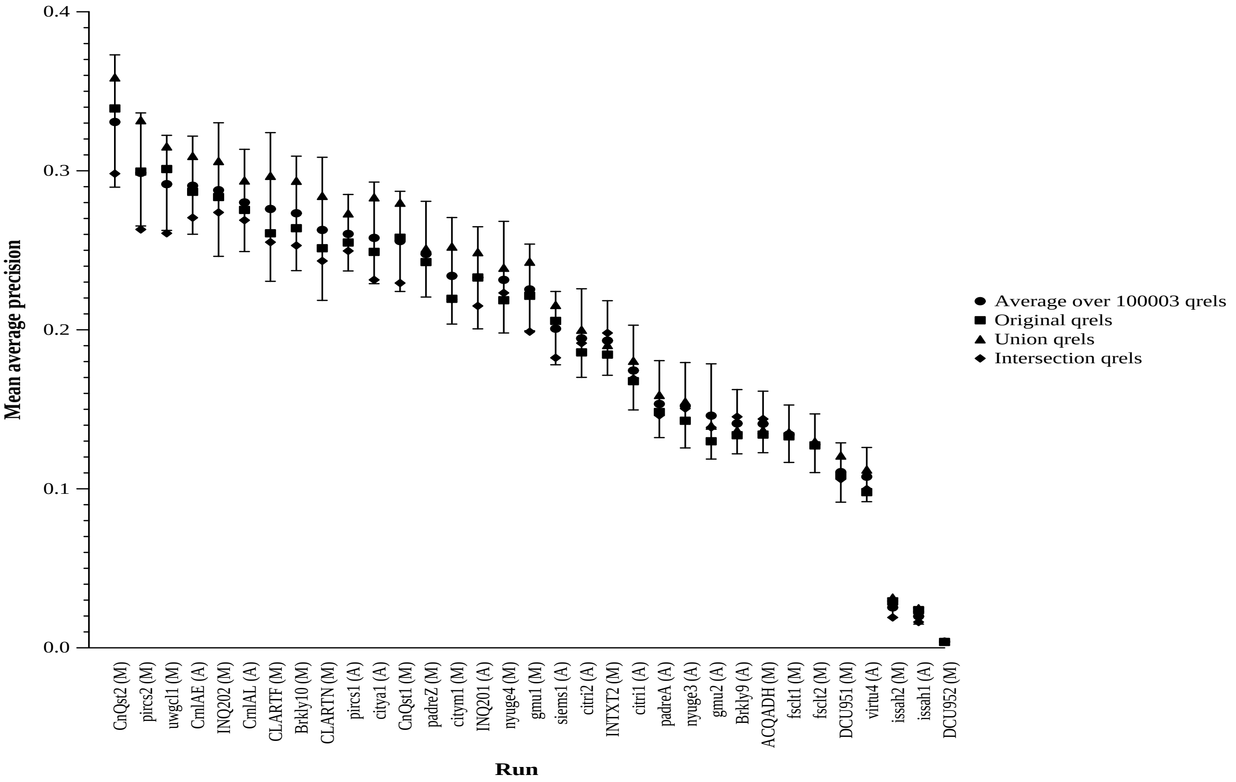

Figure 2: Sample mean, min, and max of the mean average precision computed for each of 33 TREC-4 runs over a sample of 100,003 qrels. Also plotted are the mean average precision for the original, union, and intersection qrels. Runs are labeled as either manual (M) or automatic (A).

	Mean	Min	Max
with Original	.9380 (16)	.8712 (34)	.9962 (1)
in subsample	.9382 (16)	.8409 (42)	.9962 (1)

Table 2: Kendall correlation of system rankings and corresponding number of pairwise adjacent swaps produced by different qrels. With 33 systems, there is a maximum of 528 possible pairwise adjacent swaps.

mean correlation across all pairs in the subsample. The mean, minimum, and maximum Kendall correlations for the two methods are given in Table 2. The numbers in parentheses show the number of pairwise adjacent swaps a correlation represents given that there are 33 systems in the rankings. (Note that the way in which the qrels were constructed means that any two qrels are likely to contain the same judgments for 1/3 of the topics. Since the qrels are not independent of one another, the Kendall correlation is probably slightly higher than the correlation that would result from completely independent qrels.)

On average, it takes only 16 pairwise, adjacent swaps to turn one ranking into another ranking. The vast majority of the swaps that do take place are between systems whose mean average precisions are very close (a difference of less than .01). A discussion of the probability of a swap as a function of the difference in the mean average precisions is given in the next subsection.

The plots of the union and intersection evaluations in Figure 2 show that the evaluation using those qrels sets is not different from the evaluation using the other qrels sets. The intersection and union qrels differ from the other sets in that each topics' judgments are a combination of judges' opinions in the union and intersection qrels. The intersection qrels set represents a particularly stringent definition of relevance, and the union qrels a very weak definition of relevance. Nonetheless, in all but two cases (intersection qrels for runs `pircs2` and `uwgcl1`), the mean average precision as computed by the union and intersection qrels falls within the range of values observed in the sample of 100,003 qrels. The corresponding rankings are also similar: the Kendall correlation between the original qrels ranking and the union qrels ranking is .9508, and between the original and intersection rankings is .9015. Thus despite the intrinsic difference in the composition of the union and intersection qrels on the one hand and the one-judge-per-topic qrels on the other, the evaluation behavior remains the same.

Estimating the probability of a swap

The purpose of a test collection is to enable developers of IR systems to isolate factors that contribute to retrieval effectiveness. To this end, statements about the average stability of the ranking of a set of runs are not as interesting as statements regarding the probability that any two particular runs will be ranked in a different order if the relevance judgments change. Define a *swap* to be the situation in which one qrels ranks run i before run j and a second qrels ranks run j before run i. This subsection uses the sample of 100,003 qrels to estimate the probability that a swap exists for each pair of TREC-4 runs.

Let $B[i,j]$ be the number of qrels that cause system i to evaluate as better than system j. Then the esti-

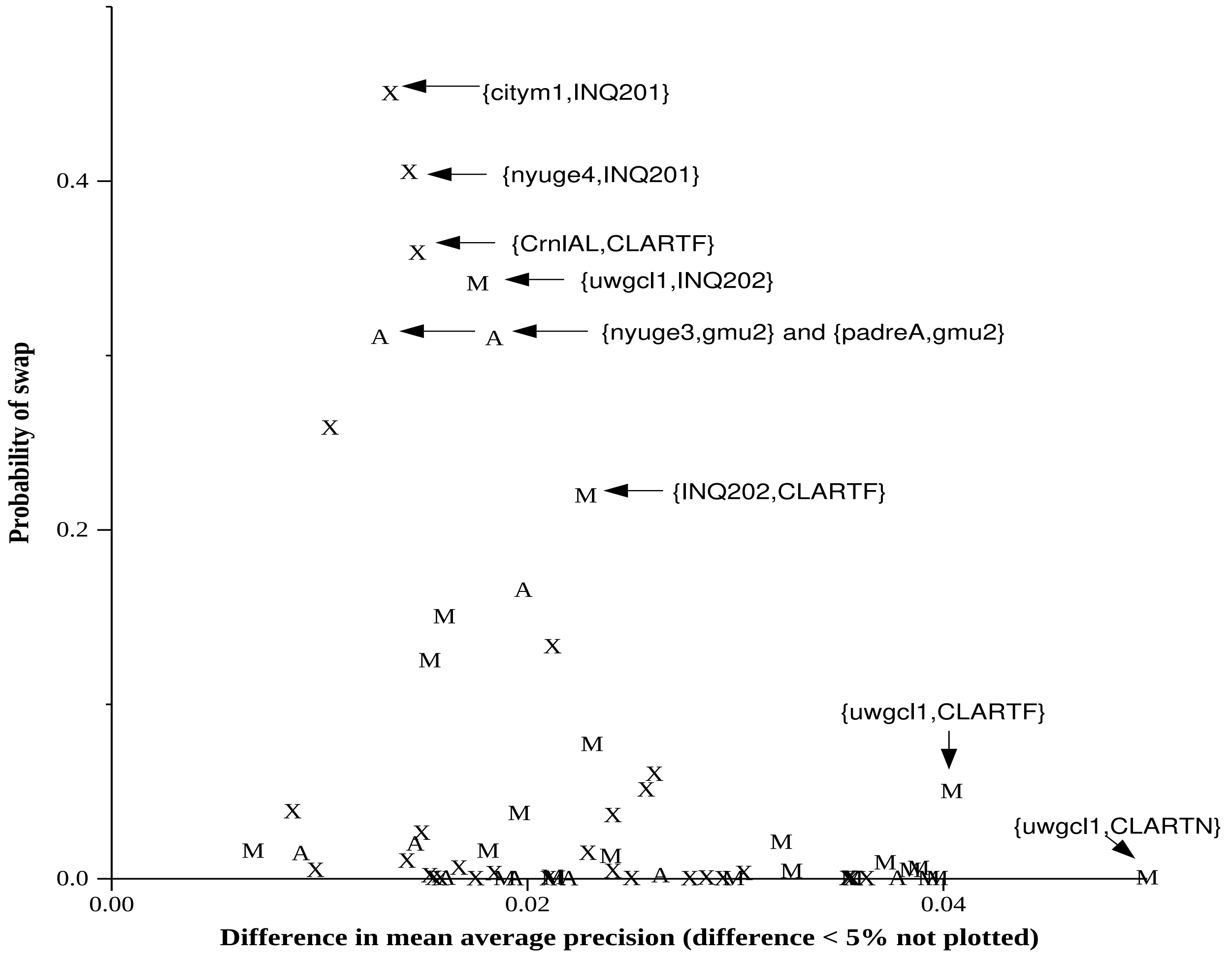

Figure 3: Probability of a swap versus difference in mean average precision for TREC-4 runs. Only pairs of systems that have a difference greater than 5% and a non-zero probability of a swap are plotted. The label of a point indicates whether the runs in the pair are both automatic (A), both manual (M), or one of each (X).

mated probability of a swap between systems i and j is $\frac{\min(B[i,j],B[j,i])}{100{,}003}$. Note that since we treat i and j symmetrically in this definition, the maximum probability of a swap is .5. Figure 3 plots the probability of a swap against the difference in the mean average precision of the two runs as evaluated using the original qrels. In the plot, each point is labeled with whether the two systems are manual systems (M), automatic systems (A), or one of each (X). Only pairs of systems that actually swapped and have a difference of at least 5% in mean average precision are plotted. Of the 528 pairs of systems, 427 have no swaps for all 100,003 qrels. Of the remaining 101 pairs, 63 have a difference in mean average precision of at least 5%, and 27 have a difference of at least 10%.

Points plotted furthest from the origin of the graph in Figure 3 are of interest because they represent pairs of systems that either swap often or have large differences in the mean average precision. For example, each of the `citym1` and `INQ201` runs had almost an equal number of qrels for which it was ranked higher than the other, though the pair just barely made the 5% difference in mean average precision cut-off. At the other extreme, the `uwgcl1` and `CLARTN` runs swapped only 50 times, but had a difference in mean average precision of approximately 12%. Four runs that participate in most of the outlying points—`CLARTF`, `CLRTN`, `INQ202`, and `gmu2`—are the four runs that have the largest sample variance of the mean average precision as computed over the 100,003 qrels.

Figure 3 shows that automatic systems are more stable than manual systems. (The definition of a "manual" run is very broad in TREC: a run is labeled as a manual run if *any* human intervention occurred during the processing of the run. Creating the original query by hand from the topic, manually tweaking automatically assigned weights, and doing relevance feedback all receive the "manual" label.) Ten of the 63 points plotted in the graph are of pairs of automatic systems, and for all but three of those pairs the probability of a swap is less than .02. The two automatic pairs with a probability of a swap close to .3 involve the `gmu2` run. The relatively large variation in the `gmu2` run is completely attributable to Topic 232. The Secondary B assessor judged exactly one document relevant for Topic 232, and the `gmu2` run (only) ranked that document first. Retrieving the single relevant document first makes the average precision for that topic 1.0 as compared to an average precision of .1152 or .0021 for the other assessor's judgments. The mean average precision score for run `gmu2` thus changes significantly depending on whether or not Secondary B assessments are used for Topic 232.

TREC-4 participants were allowed to submit up to two ad hoc runs, and 14 participants submitted two. Thirteen of the 14 pairs of runs in which both runs were submitted by the same organization had no swaps at all; the other pair (the two Cornell runs, `CrnlAE` and `CrnlAL`) ranked in the same order for all but 5 of the 100,003 qrels. The two Cornell runs had a difference of .0114 or 4.1% in the mean average precisions as evaluated by the original qrels. The difference for other same-organization pairs was usually larger, though the difference between the two CLARITECH runs (`CLARTF` and `CLARTN`) was 3.7%, and the difference between the two runs from the National University of Singapore (`issah2` and `issah1`) was 2.3%. The fact that the two CLARITECH runs never swapped is particularly striking. These two runs had the largest variances in the set of 33 runs, and between the two of them are involved in 14 of the 63 points plotted in Figure 3. Yet even with the very small absolute difference in mean average precision, the two runs were ranked in the same order by all qrels. The evidence of the same-organization pairs demonstrates that the TREC test collections are viable laboratory tools for same-system experimentation.

Cross-system comparisons need to be made with some more care, especially when user interaction is involved in producing the retrieval results. A common rule of thumb is that differences of less than 5% are not meaningful, differences between 5 and 10% are noticeable, and differences of greater than 10% are material [12]. However Figure 3 suggests that it is possible—though quite unlikely—that runs from different systems that have greater than a 10% difference in their mean average precision scores when evaluated by one qrels may sort differently if another set of relevance judgments is used.

3.2 TREC-6 Relevance Assessments

In the discussion of the overlap of the TREC-4 relevance assessments, it was noted that the assessments were all made by NIST assessors who have similar backgrounds and training. Perhaps the stability of the evaluation results reflects only the common background of the NIST assessors. We can test this hypothesis by comparing the evaluation results using qrels produced by NIST versus qrels created outside of NIST.

In the course of their participation in TREC-6, the University of Waterloo judged over 13,000 documents for relevance across the 50 TREC-6 ad hoc topics [4], averaging about 2 hours per topic making judgments. The circumstances of the Waterloo judgments differed dramatically from those of the NIST TREC-6 judgments: the Waterloo assessors have very different backgrounds from the NIST assessors; the pool of documents Waterloo used to make the judgments was smaller and the result of a single system's searches, but also contained a higher percentage of relevant documents than the NIST pool; the Waterloo assessors were not the topic authors. Both sets of assessors had the common goal of finding the complete relevant document set.

The Waterloo assessors used a three-point relevance scale—relevant, not relevant, and 'iffy'—but binary assessments are needed to be compatible with the NIST assessments and the evaluation software. For this analysis, we forced all iffy judgments to not relevant. The mean overlap between the two assessments sets for the 50 TREC-6 topics is .328. One topic, Topic 309, has an overlap of 0.0 since no documents are in the intersection of the relevant sets. Topic 320 has an overlap of 1.0 since the relevant sets for that topic are identical.

qrels pair	correlation	swaps
NIST & Waterloo	.8956	141
NIST & union	.9556	60
NIST & intersection	.8904	148
Waterloo & union	.8956	141
Waterloo & intersection	.9237	103
union & intersection	.8652	182
mean	.9044	129

Table 3: Kendall correlation between rankings produced by different qrels sets on the TREC-6 data. Also shown is the number of swaps the correlation represents given that there are 74 systems represented in the rankings (and thus 2701 possible swaps).

There were 74 category A, ad hoc runs submitted to TREC-6. Each of these runs was evaluated using four different qrels: the original NIST assessments, the Waterloo assessments, the union of the two assessment sets, and the intersection of the two sets. Figure 4 shows the mean average precision for the runs as evaluated using each of the four qrels where the runs are ordered by the mean average precision using the original qrels. (The Waterloo run `uwmt6a0` is not plotted because of the scale of the plot. The ranking submitted by Waterloo for that run was constructed such that each of the documents they judged relevant were ranked first; this gives the run a mean average precision of 1.0 when using the Waterloo qrels. The run is ranked first using the NIST qrels as well, with a mean average precision of .4631.)

The Kendall correlations between the rankings produced by each pair of qrels is given in Table 3. Once again the correlations among the rankings are quite high, with a mean correlation of .9044. The NIST qrels is more highly correlated with the union qrels, and the Waterloo qrels with the intersection qrels, reflecting the diversity of the respective pools. The smallest correlation is between the union and intersection qrels.

From the set of 74 runs, 66 pairs of runs are such that the same participant submitted both runs. From these 66 pairs, only 10 pairs were ranked in different orders by some pair of qrels. Four of the pairs had a difference in mean average precision as measured by the NIST qrels of less than 5%, another four had a difference of between 5 and 10%, and the remaining two pairs had a difference of greater than 10%. This differs somewhat from the TREC-4 runs in which no pair of runs submitted by the same organization had any swaps to speak of. However, the evidence suggests that this difference is more a consequence of a different set of topics than a different source for the relevance judgments.

The definition of mean average precision of a run gives equal weight to each topic in the run regardless of the number of relevant documents a topic has. Usually this is desirable because it emphasizes a user's view of the system. However, when there are very few relevant documents (say, 5 or fewer) the average precision measure itself is unstable in that small perturbations in the document ranking can cause large differences in the average precision. The extreme case of this is demonstrated in the TREC-4 `gmu2` run. With one relevant document, ranking that document second instead of first halves the average precision of the topic: from 1.0 when the relevant document is ranked first to 0.5 when it is ranked second. The TREC-6 collection appears to be affected by this instability more so than the other TREC collections. For

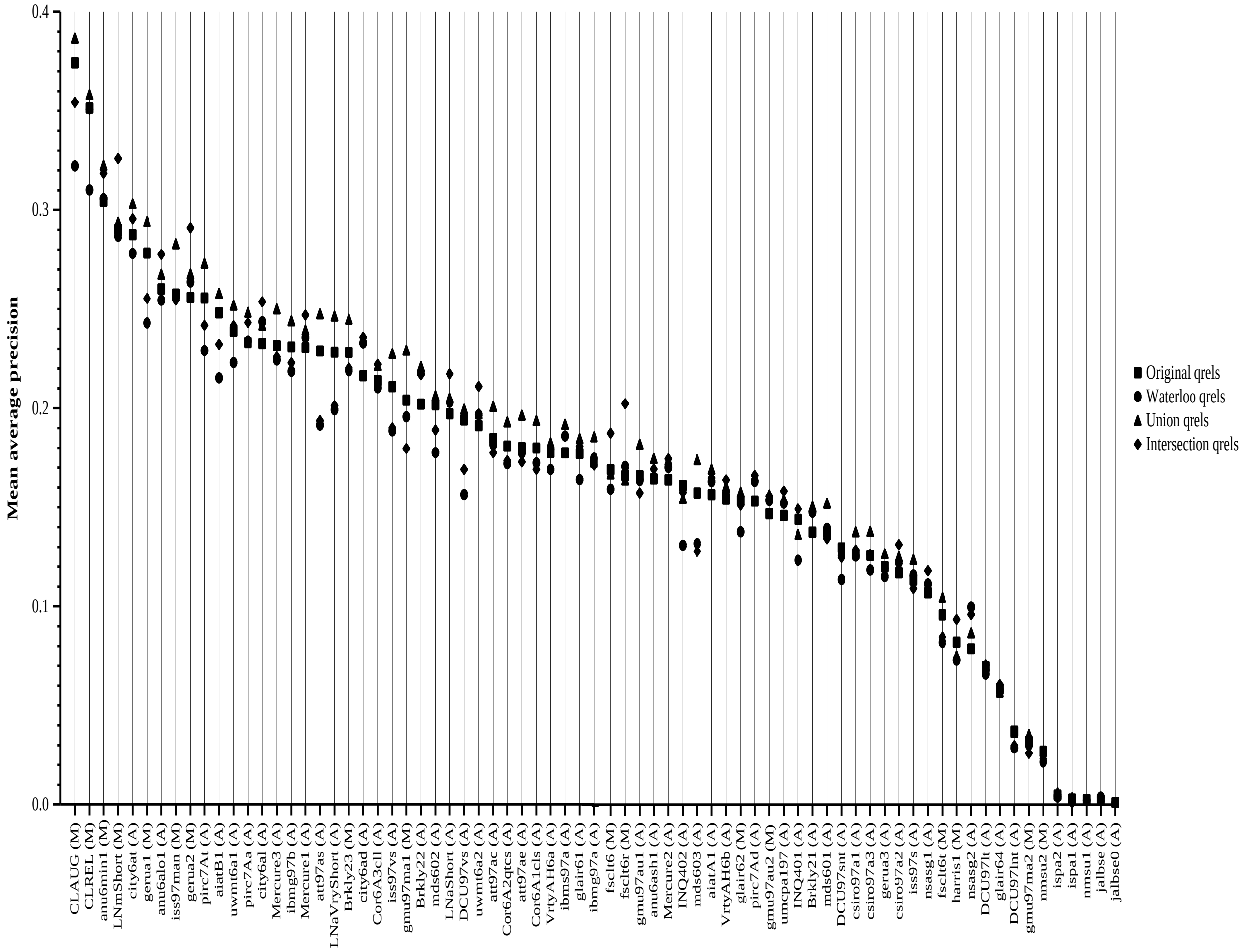

Figure 4: Mean average precision for TREC-6 runs as evaluated using four different qrels.

example, in the official TREC-6 evaluation, CUNY run `pirc7At` had a mean average precision of .2556 while run `pirc7Aa` had a mean average precision of .2332 (a 10% difference). Yet by most other evaluation measures the `pirc7Aa` was more effective: it retrieved more relevant documents in total, it had higher precision at ranks 10 and 20, and the average precision was greater than the median average precision for more topics [8].

The NIST qrels for TREC-6 contain 5 topics (308, 309, 338, 344, 348) that have 5 or fewer relevant documents. The Waterloo qrels have three such topics (308, 312, 348), two of which are the same as the NIST topics and one additional topic. To remove some of the confounding effects of the inherent instability of the average precision measure from the variability caused by different assessors, we repeated the TREC-6 study using only the 44 topics that had more than 5 relevant documents in both assessment sets.

The mean overlap between the NIST and Waterloo qrels increases slightly from .328 to .337 when the 6 topics are discarded, mostly because the one topic with an overlap of 0.0 is among the discarded. Table 4 shows the Kendall correlations between each of the pairs of rankings for both 50 topics and 44 topics. The mean correlation and the correlations for all but one pair of qrels also increase. Finally, seven[2] of the 66 pairs of single-participant runs swapped, one of which has a difference in mean average precision of greater than 10%. This pair, `gerua1` and `gerua2`, consists of manual runs in which users judged documents and performed relevance feedback. Such feedback runs essentially add a third relevance judge into the mix, and are consequently less stable.

qrels pair	50	44
NIST & Waterloo	.8956	.9074
NIST & union	.9556	.9637
NIST & intersection	.8904	.8882
Waterloo & union	.8956	.9126
Waterloo & intersection	.9237	.9511
union & intersection	.8652	.8800
mean	.9044	.9172

Table 4: Kendall correlation between rankings produced by different qrels sets evaluated using the entire set of 50 topics and the set of 44 topics that have more than 5 relevant documents.

The increased stability in the system rankings when topics with small numbers of relevant documents are removed supports the hypothesis that these topics are in-

[2]These seven pairs are not a proper subset of the 10 pairs that inverted when using 50 topics.

herently less stable. However even with all 50 topics included in the comparison, the results of the TREC-6 experiment are the same as those of the TREC-4 experiment. Different relevance assessments, created under disparate conditions, produce essentially the same comparative evaluation results.

3.3 Analysis

Lesk and Salton concluded that one of the reasons system rankings remain stable despite differences in the relevance assessments is that evaluation results are reported as averages over many queries [9]. While retrieval results for individual queries can vary widely, an average over a sufficient number of queries is more stable. The problem, of course, is defining "sufficient". We empirically investigated the effect of averaging by looking at the correlation of rankings produced by different same-size sets of topics. The results are shown in Figure 5.

The figure shows a histogram of Kendall correlations: the x-axis plots the number of topics used in the topic sets (5, 10, 15, 20, or 25); the y-axis plots the value of the correlation between the system rankings produced by pairs of topic sets; and the z-axis shows the number of times a correlation was obtained for topic sets of a given size. The histogram was produced by evaluating the TREC-6 systems on two, equal-sized, randomly selected, disjoint subsets of topics (using the NIST qrels). One thousand trials were run for each topic set size. A trial consisted of choosing two sets of topics of the required size, forming the system rankings based on the evaluation results from the two different sets of topics, and computing the correlation between the two rankings. As an example, the figure shows that there were approximately 10 topic set pairs of size 5 that produced a Kendall correlation of .80, and 2 pairs of size 25 that produced a Kendall correlation of .80.

A significant majority of the 1000 trials show a correlation of greater than .8 when using topic set sizes of at least 15. When using topic sets of size 25, all but 2 pairs have a correlation of at least .8, and the vast majority of pairs have a correlation greater than .9. In contrast, the range of correlations for topic sets of size 5 is from slightly less then .5 to greater than .9, with most correlations less than .9. Thus, at least for the TREC-6 environment, as few as 25 topics can be used to compare the relative effectiveness of different retrieval runs with great confidence.

We can extend this finding to the case of different relevance assessments by viewing different assessment sets for one query as "gold standard" assessments for slightly different queries. Even if the assessments change radically for half the queries, the system ranking derived from 50 queries is likely to remain stable. This average-of-averages technique typically used in IR evaluation has been criticized for hiding significant topic-to-topic variation. While there is no doubt that averaging does mask individual topic differences, it also increases the reliability of overall effectiveness comparisons.

4 Conclusion

Test collections are research tools that provide a means for researchers to explore the relative benefits of different retrieval strategies in a laboratory setting. As such they are abstractions of an operational retrieval environment: they represent a user's interest as a static set of (usually binary) decisions regarding the relevance of each document, making no provision for the fact that a real user's perception of relevance will change as he or she interacts with the retrieved documents, or for the fact that 'relevance' is idiosyncratic. Despite the abstraction, test collections have proven useful in developing retrieval techniques for operational settings.

Because a test collection abstracts away from the vagaries of users' views of relevance, the relative performance of two retrieval techniques evaluated on the collection must be stable despite small changes in the set of relevance judgments included in the collection. This paper investigated the effects changes in the relevance judgments of the TREC collections had on the evaluation of TREC retrieval runs. The TREC collections are many times larger and more diverse than any other collection for which a similar analysis has been done. Nonetheless, the major conclusion of those studies holds for the TREC collections as well: the relative effectiveness of different retrieval strategies is stable despite marked differences in the relevance judgments used to define perfect retrieval.

A variety of different conditions were tested in the experiments, including judgments made by query authors vs. judgments by non-authors; judgments made by different non-authors; judgments made by a single judge vs. group judgments; and judgments made by different people in the same environment vs. judgments made in very different environments. The actual value of the effectiveness measure was affected by the different conditions, but in each case the relative performance of the retrieval runs was almost always the same. These results validate the use of the TREC test collections in retrieval experiments.

The study did detect circumstances in which run comparisons need to be done with more caution. When queries have very few relevant documents (fewer than 5 or so), summary evaluation measures such as average precision are themselves unstable; runs that include many such queries are more variable. Runs that involve significant amounts of user interaction, especially in the context of relevance feedback, also vary more, reflecting the amount of agreement between the user doing the feedback and the official relevance judge. As expected, cross-system comparisons were found to need a larger difference in mean average precision than intra-system comparisons for the difference to be meaningful. The flip side of these cautions is that for the prototypical use of a test collection—comparing algorithmic variants of the same retrieval system—the TREC test collections are extremely reliable. And given appropriate regard for the cautions listed here, the collections are useful tools for a wide range of retrieval research.

Acknowledgements Special thanks to Gord Cormack of the University of Waterloo for providing the Waterloo TREC-6 relevance assessments. The analysis of the effect of averaging, including creating Figure 5, was performed by Paul Over of NIST. The content of this paper was shaped in large part by discussions with Chris Buckley, Donna Harman, and Paul Over. David Banks and other members of the Statistical Engineering Division at NIST made helpful suggestions regarding the analysis of the data.

References

[1] Robert Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Infor-*

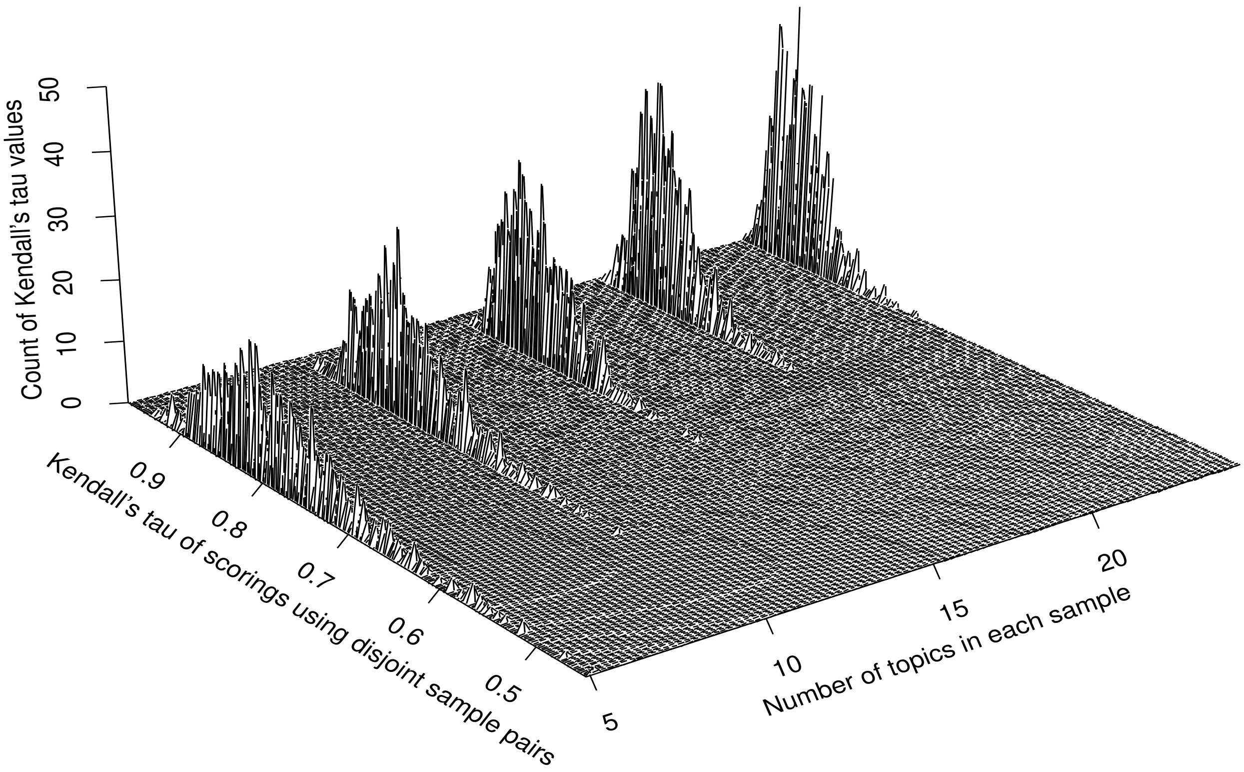

Figure 5: Histogram of correlation among rankings produced by different equal-sized subsets of the TREC-6 topics.

mation Processing and Management, 28(5):619–627, 1992.

[2] C. W. Cleverdon, J. Mills, and E. M. Keen. Factors determining the performance of indexing systems. Two volumes, Cranfield, England, 1968.

[3] Cyril W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical Report Cranfield Library Report No. 3, Cranfield Institute of Technology, Cranfield, UK, October 1970.

[4] Gordon V. Cormack, Charles L.A. Clarke, Christoher R. Palmer, and Samuel S.L. To. Passage-based refinement (MultiText experiments for TREC-6). In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 171–186, November 1997. Draft for conference notebook.

[5] C. A. Cuadra and R. V. Katter. Opening the black box of relevance. *Journal of Documentation*, 23(4):291–303, 1967.

[6] Donna K. Harman. The first Text REtrieval Conference (TREC-1), Rockville, MD, U.S.A, 4–6 November, 1992. *Information Processing and Management*, 29(4):411–414, 1993.

[7] Stephen P. Harter. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, 47(1):37–49, 1996.

[8] K.L. Kwok, L. Grunfeld, and J.H. Xu. TREC-6 english and chinese retrieval experiments using PIRCS. In *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 96–102, November 1997. Draft for conference notebook.

[9] M.E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4:343–359, 1969.

[10] G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.

[11] Linda Schamber. Relevance and information behavior. *Annual Review of Information Science and Technology*, 29:3–48, 1994.

[12] Karen Sparck Jones. Automatic indexing. *Journal of Documentation*, 30:393–432, 1974.

[13] Karen Sparck Jones. *Information Retrieval Experiment*. Butterworths, London, 1981.

[14] 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.

[15] M. Taube. A note on the pseudomathematics of relevance. *American Documentation*, 16(2):69–72, April 1965.