Forming Test Collections with No System Pooling

Mark Sanderson
Department of Information Studies
University of Sheffield
211 Portobello Street, Sheffield, S1 4DP, UK.
+44 (0) 114 22 22648
M.Sanderson@sheffield.ac.uk

Hideo Joho
Department of Information Studies
University of Sheffield
211 Portobello Street, Sheffield, S1 4DP, UK.
+44 (0) 114 22 22664
H.Joho@sheffield.ac.uk

ABSTRACT

Forming test collection relevance judgments from the pooled output of multiple retrieval systems has become the standard process for creating resources such as the TREC, CLEF, and NTCIR test collections. This paper presents a series of experiments examining three different ways of building test collections where no system pooling is used. First, a collection formation technique combining manual feedback and multiple systems is adapted to work with a single retrieval system. Second, an existing method based on pooling the output of multiple manual searches is re-examined: testing a wider range of searchers and retrieval systems than has been examined before. Third, a new approach is explored where the ranked output of a single automatic search on a single retrieval system is assessed for relevance: no pooling whatsoever. Using established techniques for evaluating the quality of relevance judgments, in all three cases, test collections are formed that are as good as TREC.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software --- *performance evaluation*.

General Terms

Experimentation, Measurement.

Keywords

Test collection formation, evaluation of qrel sets.

1. INTRODUCTION

Test collections – corpora created and shared amongst the information retrieval community to promote a common test bed for measuring the effectiveness of retrieval systems – resulted in extensive testing and comparison of retrieval algorithms. An ideal test collection is composed of a collection of documents; a set of queries; and a list of all the collection documents that are relevant to each of the queries. The document and query sets are relatively straightforward to gather, however collecting the final item, the relevance judgments – also known as *qrels* – is costly. Collections from the 1960s, '70s, and early '80s (e.g. Cranfield, NPL, CACM, etc), were small: never consisting of more than 3Mb of text. Consequently, it was possible to form qrels from an exhaustive examination of the collection, determining each document's relevance to each query.

Spärck Jones and Van Rijsbergen believed that such a strategy would not work with larger collections and a means of forming qrels without exhaustive searching was proposed. In their British library report (1975) and two follow up reports (Spärck Jones and Bates, 1977; Gilbert and Spärck Jones, 1979) the building of an ideal test collection was described. The use of *pooling* was advocated as a means of efficiently locating relevant documents within a large test collection. For each query, merging the output of diverse searches formed a pool. It was assumed that nearly all relevant documents would be found in the pool. A random sample of the document pool would be manually assessed for relevance, thereby forming the qrel set.

The pooling approach was utilized in gathering relevance judgments for the 5.5Mb Inspec collection. As described by Salton, Fox, and Wu (1983), for each query in the collection, seven different means of processing the query were run on a retrieval system, the documents retrieved by each means were merged (duplicates were removed) and the resulting pool examined by relevance assessors. It is not clear from the published work if the accuracy of the pool was tested.

TREC, the current test collection archetype, every year builds on the efforts of 50 to 100 research groups who each provide runs: the 1,000 best matching documents produced by their searching system for each of 50 topics (the name TREC give to queries). The union of the top 100 documents from each run (the pool, referred in this paper as the *system pool*) is manually assessed for relevance. TREC obtains diverse searches through the assumption that each research group used their own searching system, which, it is hoped, has a distinctive approach to retrieval. Across the first eight TRECs, the number of documents assessed per topic ranged from 1,005 to 2,310 (Voorhees, 1999). To organize groups into contributing to such a pool requires a level of organization beyond what most researchers are able to provide.

In the past, methods were presented that reduced the effort in creating test collections either by lowering the number of judgments to be assessed or cutting the size of the system pool. The initial aim of this paper is to adapt an existing method to work with no system pool and to re-examine another method

SIGIR'04, July 25-29, 2004, Sheffield, South Yorkshire, UK.

across a wider range of circumstances to better understand the applicability of such methods. To the best of our knowledge, all past work in large-scale test collection formation has assumed that a pool of some sort – either from multiple systems (a system pool) or from multiple searches on a single system (a query pool) – is necessary if a good quality test collection is to be formed. However, it is our understanding that this assumption has not been tested. Such a test is the aim of this paper.

Starting with a review of past work on reducing the number of judgments, the paper focuses on experimenting with means of reducing first, the size of the system pool to one system retaining a query pool composed of multiple manual searches per topic, but then second, reducing the queries per topic to one: i.e. no query pool. Finally conclusions are drawn and future work is proposed.

2. PAST WORK

Means of reducing effort in building test collections have taken a number of approaches. The first and second described here use respectively, document collections and queries with certain properties that can be exploited to reduce or even eliminate assessor effort. The third approach is to propose a more efficient sampling of a document pool. The fourth approach involves using searchers to create a pool from multiple variations of the same query.

2.1 Special collections and queries

One approach to reducing human effort is to exploit collections or queries for which little assessor effort is required.

2.1.1 Document collections

There are large collections over which some form of assessment on the constituent documents' content has been conducted. News wire articles, for example, are often manually coded with broad subject categories. Text categorization systems can be trained and tested on such a collection with no additional human effort. Using the systems in evaluation of ranking algorithms and text representation methods appears to work well: such an approach was taken by Lewis who tested a form of phrase indexing using the Reuter's text categorization collection (1992).

Another approach to exploiting human assessment of content of documents was taken by Harmandas, Sanderson, and Dunlop (1997) who built a small test collection from a series of Web sites. In forming the qrels, assessors were encouraged to follow links within the sites to help locate relevant documents. The authors stated that such as approach reduced assessor effort.

2.1.2 Queries

It is also possible to create types of queries for which one can be certain only a limited part of the collection will contain relevant documents. Sheridan, Wechsler, and Schäuble (1997) built a spoken document test collection from radio news, where queries were restricted to subjects that referred to events that had a specific starting date. This allowed relevance assessors to limit their examination of the collection to only those news items broadcast on or soon after the date.

Queries can be further restricted so that only a single item in a collection will be relevant to the query. So-called known item retrieval was used in an early run of the TREC spoken document retrieval track (Garofolo, Voorhees, Stanford, and Spärck Jones, 1997). A similar form of evaluation was tried by raters of Web search engines, where the technique was referred to as perfect page searching (Sullivan, 2002).

Although a range of researchers tried the approaches described here, the majority of test collection-based evaluation is conducted on collections whose qrels are formed with a pooling approach. The next Section describes means of assessing such a pool efficiently.

2.2 Efficient pool sampling

In the pooling approach proposed by Spärck Jones and Van Rijsbergen and with the work practice of TREC, all submitting systems and queries are treated equally. In TREC, relevance assessors examine a pool that consists of the top 100 documents from each system for each query. Means of focusing effort on particular systems or particular queries have been proposed and are described here.

2.2.1 Focusing on queries

Zobel (1998) was interested in maximizing the number of relevant documents located by assessors. He recognized that the number of such documents for each of the queries of a test collection varied: some queries have many relevant, some only a few. Zobel described how the number of relevant documents found at the top of a ranking could be used to predict with some accuracy how many relevant documents would be found further down the ranking. Using this predictor, Zobel suggested that assessors could examine for each query a shallow pool formed from the top 30 documents returned from all systems. An estimator of the number of relevant documents to be found in the lower ranks would be initiated and a period of training would ensue. Assessors would continue judging documents from the lower ranks with the estimator being adjusted until it predicted expected numbers of relevant documents with sufficient accuracy. At this point, assessors would be directed to those queries that were predicted to have more relevant documents making more efficient use of their time. It would appear that this approach was not tested.

2.2.2 Focusing on systems

Cormack, Palmer, and Clarke (1998) noted that some systems contributing to a pool are more effective (i.e. find more relevant documents) than others. They presented move-to-front (MTF) pooling where documents in the pool were initially examined in rank order across all systems. As judgments of relevance were made1, systems that appeared to be locating more relevant documents for a particular query would have their un-judged documents assessed in preference to those returned by poorer performing systems. Cormack et al tested their approach by building a qrel set using MTF pooling, judging only half the number of documents TREC assessors would examine. Using the set, they measured the mean average precision (MAP) of each system that submitted a run to TREC-6 and ranked the systems by this measure. They then repeated this process using the full TREC qrels.

The *two system rankings* were correlated using Kendall's tau (Stuart, 1983). The correlation found was 0.999. Examining only

[1] Cormack et al used the recorded judgments of TREC assessors to simulate judgments being made.

a tenth of the pool using MTF, the resulting correlation reduced to 0.990. On the question of how close the correlation had to be before one was willing to use the new test collection formation method, the authors made a strong case that 0.990 was more than sufficient: in particular, citing Voorhees' study (1998), where she concluded that a Kendall's tau of approximately 0.9 between two system rankings each produced from a separate test collection indicated a sufficiently high degree of correlation between the two collections for them to be treated as effectively equivalent. In a later Voorhees paper (2001), it was stated (p. 78),

> *…evaluation schemes that produce correlations of at least .9 should be considered equivalent since it is not possible to be more precise than this. Correlations less than .8 generally reflect noticeable changes in the rankings, not simply inversions among neighbors, and suggest that the evaluation schemes have different emphases.*

As will be seen, for the experiments in this paper (Section 3), these thresholds were used to determine if one means of evaluation was equivalent to another or not.

2.2.3 No manual assessment

Given that the document pool produced by multiple systems is a rich source of relevant documents, Soboroff, Nicholas, and Cahan (2001) examined the possibility of using just the raw pool as the qrel set with no manual assessor effort. Working with TREC data and using the same assessment procedure as Cormack et al, Soboroff et al ranked TREC submissions using the pool qrels and compared the ranking with one formed from the standard TREC qrels. Although the judgements were successful in determining poorly performing systems as poorly performing, and medium performing systems as being better than the poor, the best performing systems were measured to be no better than the poor. Soboroff et al tried a number of refinements to their technique, but were unable to build a pool that could distinguish the best performing retrieval systems from the worst.

It would appear that some level of human assessment is needed to provide effective measurement of retrieval systems. Given the work of Zobel and Cormack et al, the question is how little human effort is required to produce a reasonable test collection?

2.3 Interactive searching and judging

In addition to proposing move-to-front pooling, Cormack et al also proposed a means of forming qrels using a combination of interactive searching, judging, and query re-formulation referred to as ISJ (Interactive Searching and Judging). For each query in TREC-6, Cormack et al instructed a searcher to search as many variations and refinements of the topic as he/she could think of noting all relevant documents retrieved. When no more relevant could be found, searchers moved onto another. Spending on average just over two hours per topic, the searchers assessed, on average, 260 documents identifying 78 (30%) as being relevant. Cormack et al compared the ISJ qrels with the full TREC-6 qrels. Although almost exactly the same number of relevant documents were identified (3,900 for ISJ, 3,923 for TREC-6) only 40% were common to the two qrel sets. Cormack et al measured the Kendall's tau correlation of system rankings measured on the two sets: a value of 0.89 was obtained; just below Voorhees's upper threshold of 0.9.

However, Cormack et al stated that the lower correlation, when compared to TREC, was due to both the different approach in forming the qrel set and the difference in opinion on what constitutes relevance between ISJ searcher/assessors and TREC's assessors. Cormack et al separated the two factors by identifying a subset of the documents that were selected by the ISJ judges that had also been relevance assessed by TREC assessors. When comparing this set of qrels (1,568) with the full TREC set (3,923), the correlation across the system rankings increased to 0.96 despite the ISJ set being just under 40% of the TREC qrel set. It appeared that more of the difference in correlation was due to difference in opinion between the TREC and ISJ judges than the difference in the judging process.

Taking into account Voorhees's view of what is a sufficient correlation between system rankings,, one can conclude that Cormack et al produced a set of qrels that rank retrieval systems as well as TREC, but with no system pooling.

2.4 Interactive relevance feedback and judging

Given the striking success of ISJ, one may wonder why the method was not adopted by TREC to save time in assessor effort. Soboroff and Robertson (2003) explained that TREC's assessors were judged to be far better assessors of documents than generators of queries to locate them. Therefore, they adapted the ISJ approach to work with the assessors' strengths: maintaining an iterative search for relevant documents, but using relevance feedback to generate the query at each iteration.

Soboroff and Robertson generated fifty topics for the filtering track of TREC 2002. After the topics and their statement were given, seven runs from four retrieved systems were prepared and used to make a small pool composed of the top-ranked 100 documents from each run. The CombMNZ fusion algorithm (Fox and Shaw, 1993) was then used to select the top-ranked 100 documents for the pool, which were passed to the assessors for relevance judgment. The documents judged as relevant were passed to each system to be processed for relevance feedback, which generated a new query, which generated a new set of documents to be assessed. The process was run for five iterations or until no relevant document was found in a previous iteration. Overall, the assessors made 21,000 relevance judgments.

The set of relevant documents identified at this stage was called the first round qrel. When all filtering track submissions were completed, they were examined for relevance documents. These qrels were known as the second round qrels. The second round generated another 42,000 documents to judge, and seven topics had more than fifty new relevant documents, four topics had more than twenty new, the overall median of fifty topics was 8.5. Despite these additional efforts, the Kendall's tau correlations between the qrel from the first round only and the accumulated one from both rounds were between 0.912 and 0.996 depending on the tasks. This led the authors to conclude (pp. 248-249),

> *…the* [system] *rankings are virtually identical.*

3. EXPERIMENTS

As described in the introduction, the assumption in all the work presented here is that some form of pooling, either system or

query, is required in order for a good test collection to be formed. The experiments in this Section will test that assumption.

Three experiments are presented: the first, Soboroff's iterative relevance feedback technique with no system pooling; second, a re-examination of Cormack et al's ISJ method; and third, examining how good a test collection can be formed when no pooling whatsoever is used.

3.1 Relevance Feedback

The experimental procedure to adjust Soboroff and Robertson's method was as follows. First an initial query was composed by the title and description in the topic statement of TREC-7. The top-ranked 100 documents were recorded from the result. Rather than use a new set of assessors to judge the documents for relevance, existing TREC judgments were used to determine which documents in the ranking were relevant. All such documents found in the recorded set were fed to relevance feedback. The next 100 unseen documents in the result of the first iteration were then recorded; the relevant documents were again extracted, and fed to relevance feedback. This process was repeated as many as five times. Accumulating the distinct relevant documents found in the previous iterations and the result of the current retrieval generated qrels.

Three retrieval models available in the Lemur Toolkit2 were used in this experiment both for indexing and retrieving. The models were a Vector Space Model (annotated as TFIDF), Okapi BM25 Probabilistic Model (Okapi), and KL-divergence Language Model (KL-Div.). The default values given by the toolkit were used in most model-specific parameters. However, the following changes were made to optimize the models to the TREC-7 collection to our knowledge3. For TFIDF, both doc.tfMethod and query.tfMethod (TF weighting method) were set to log-TF. For Okapi, K1=1.4, b=0.6, K3=1000 (taken from Robertson, et al., 1998) were used. For KL-divergence, Dirichlet Prior value (a smoothing parameter) was estimated based on the TREC-7 collection using the toolkit command EstimateDirPrior, and set to 331. queryUpdateMethod was set to divergence minimization (see the Lemur manual for the details of these parameters).

Other variable parameters for relevance feedback are summarized in Table 1. The main variables are the retrieval models, number of expansion terms, and selection of relevant documents. The retrieval models were described above. The number of expansion terms was either fixed to 30 (i.e. every iteration adds 30 terms), or incremental 30 (i.e. 30 for the first feedback, 60 for the second, etc). The selection of relevant documents was either to use all relevant documents found in the previous iterations (annotated as Accum.), or use only those found in the last iteration (New). The last two variables were tested because we were interested in them as a factor of retrieving new relevant documents over a relatively large number of iterations.

Table 2 shows the Kendall's tau correlation of the relevance feedback runs with the official TREC-7 system rankings. The 8th column can be viewed as the baseline of each run, where the qrel was generated from all relevant documents found in the top-ranked 1000 documents of initial queries. The 2nd to 7th columns are the correlations of each iteration where the qrel was generated from all relevant documents found in the previous iterations as described above. In our scenario, the assessors were supposed to judge 100 documents per iteration. Therefore, by the end of the 5th iteration for each topic, 600 judgments were made.

Table 1: Summary of relevance feedback parameters

Run ID	RetModel	TermCount	RelDocs
RF_01	TFIDF	Fixed 30	Accum.
RF_02	TFIDF	Fixed 30	New
RF_03	TFIDF	Incre 30	Accum
RF_04	TFIDF	Incre 30	New
RF_05	Okapi	Fixed 30	Accum.
RF_06	Okapi	Fixed 30	New
RF_07	Okapi	Incre 30	Accum
RF_08	Okapi	Incre 30	New
RF_09	KL-Div.	Fixed 30	Accum.
RF_10	KL-Div.	Fixed 30	New
RF_11	KL-Div.	Incre 30	Accum
RF_12	KL-Div.	Incre 30	New

Table 2: Kendall's tau of relevance feedback runs: the 2nd to 6th columns indicate the number of iterations, and the 7th column is the correlation of the qrel that consists of the relevant documents in the top-ranked 1000 documents of the initial search. All figures but averages are statistically significant at .01 level

Run ID	0	1	2	3	4	5	1000
RF_01	0.79	0.87	0.89	0.90	0.90	0.91	0.87
RF_02	0.79	0.87	0.89	0.89	0.89	0.90	0.87
RF_03	0.79	0.87	0.88	0.89	0.90	0.90	0.87
RF_04	0.79	0.87	0.89	0.89	0.89	0.89	0.87
RF_05	0.83	0.93	0.95	0.96	0.96	0.97	0.90
RF_06	0.83	0.93	0.95	0.96	0.96	0.97	0.90
RF_07	0.83	0.93	0.95	0.96	0.97	0.97	0.90
RF_08	0.83	0.93	0.95	0.96	0.96	0.97	0.90
RF_09	0.82	0.88	0.90	0.91	0.91	0.92	0.89
RF_10	0.82	0.88	0.90	0.92	0.92	0.93	0.89
RF_11	0.82	0.88	0.90	0.90	0.91	0.92	0.89
RF_12	0.82	0.88	0.90	0.92	0.93	0.93	0.89
Average	0.82	0.89	0.91	0.92	0.93	0.93	0.89

Several points can be emphasized from the result. First, in all runs, the baseline was outperformed by the qrel of the second iteration, which would reduce the number of judgments to 30% of 1000 judges of the baseline. Second, although TFIDF runs (RF_01 to RF_04) seem to require a careful parameter setting, the correlation above 0.9 (i.e. Voorhees's threshold) was consistently obtained as early as by the end of the third iteration in all Okapi and KL-Divergence runs (RF_05 to RF_12). It was also found

[2] http://www-2.cs.cmu.edu/~lemur/

[3] which we believe it fair to do as our aim is to build a usable set of qrels as opposed to evaluate the retrieval effectiveness.

that it is possible to get the correlation 0.97 by the end of the fifth iteration (600 judges per topic). These findings highlight the advantage of using relevance feedback to general qrels efficiently.

As for the number of expansion terms and selection of relevant documents, the former does not seem to have a major impact on the correlation in Okapi and KL-Divergence runs (e.g. RF_05 vs. RF_07 or RF_09 vs. RF_11) while TFIDF preferred the fixed 30 terms to the iterations. On the other hand, the selection of relevant documents showed a more noticeable difference in KL-divergence runs (e.g. RF_09 vs. RF_10). This suggests that new relevant documents found in the last iteration are also a good resource to keep gaining new relevant documents over the iterations.

Overall, from this first experiment, it is drawn that a single system can generate a usable set of qrels, and that the process building qrels using a single system can be facilitated by the relevance feedback, thus, suggested. Finally, in our environment, increasing the size of expansion terms (to 100), using alternative queryUpdateMethod such as Mixture model, changing feedbackCoefficient (the balance between the query and relevant documents) over the iteration did not make a noteworthy difference in the correlation.

3.1.1 Conclusion

Remembering that different topics and documents were examined in these experiments compared to Soboroff et al's work, it is nevertheless striking that the Kendall's tau correlations presented in Table 2 are similar to those reported by Soboroff; this despite Soboroff's use of system pooling. We conclude from this experiment that system pooling in this relevance feedback based approach is not necessary as was previously thought.

3.2 Interactive Searching and Judging

Since being published, by Cormack, Palmer and Clarke, Interactive Searching and Judging was further validated and used in forming qrels for the NTCIR evaluation exercise (Kuriyama et al, 2002). In addition, the method was adopted in more recent TDT work (Cieri et al, 2002). As ISJ involved a particular searching system and a particular set of searchers, the system or searchers may have influenced the success measured in the two validations of the approach. Given the variability of IR systems and searchers, it was decided to re-run the experiments using more searching systems and searchers to better understand the breadth of applicability of the technique. In order to do this in a tractable amount of time, it was desirable to design an experimental procedure that would allow extensive testing.

TREC allows groups to submit both *automatic runs* (where the searching system processes the TREC topic and returns retrieved documents with no manual interference allowed) and *manual runs*, where any amount of human intervention in the process of searching for relevant documents for a particular topic is permitted. It was realized that the manual runs submitted to TREC were similar to the ISJ process. For some manual runs intervention was minimal, for others, it involved searchers spending a great deal of time issuing many queries for each topic locating as many relevant documents as possible. For example, Voorhees and Harman (1998) described the building of the "t7miti1" run (the most effective run of TREC-7): where upwards of eighteen manually chosen queries were submitted per topic and the searchers only put into the run submitted to TREC, documents they judged to be relevant. The next seven most effective runs were mostly variants of users issuing queries, examining results, reformulating queries using some form of relevance feedback and eventually returning to TREC the ranking from the final query, with perhaps earlier identified relevant documents inserted at the top: processes similar to ISJ. Many of the most effective manual runs in TREC-8 also took a similar approach (e.g. CL99XTopt, iit99ma1 and orcl99man, see Table 4).

Given the stated success of the manual runs in locating relevant documents, it was decided to treat each of the manual runs (submitted to TREC) as simulations of the ISJ process: forming a qrel set from the ranking returned from each run. As with the later part of Cormack et al's experiments with ISJ, the formed set would come from the intersection of the official TREC qrels with the top 1,000 retrieved documents from the manual run. The assessments are kept constant, only the means of obtaining documents to be judged was varied. The Mean Average Precision (MAP) for each of the ad hoc runs submitted to the same year the TREC manual run was submitted to was computed from the new qrel set and compared to the MAP computed for the runs from the full TREC qrels: i.e. the two system rankings from the two qrels sets were compared.

The key question addressed by the experiments was how consistent is the ISJ manual qrel approach when applied on other retrieval systems with different retrieval features using different relevance assessors? In other words, if one were to adopt the ISJ approach to build a test collection, could one be confident that the approach would work for the sets of searchers and search system used.

3.2.1 The experiments

In order to test a re-running of the ISJ approach on a wide range of manual runs, the TREC qrels and all the runs submitted to the ad hoc task of TRECs 5, 6, 7, and 8 were downloaded from the TREC web site. The number of runs per TREC year is shown in Table 3.

Table 3: No. of manual and automatic runs across four TRECs

TREC	Manual runs	Automatic runs
5	31	30
6	17	57
7	17	86
8	13	116

For each manual run, a qrel set was formed from the documents in the full run (1,000 documents per topic) that intersect with the qrel set produced from TREC. The automatic ad hoc runs were evaluated (using trec_eval), ranked by mean average precision, and the resulting system ranking correlated (using Kendall's tau) to the ranking obtained through the official TREC evaluation (see Table 4).

As can be seen in the TREC-8 experiment, three runs in thirteen have a Kendall's tau less than Voorhees's stated threshold of 0.9 (indicating a qrel set indistinguishable from TREC); of those, two are below the 0.8 threshold (suggesting that the evaluation schemes have different emphases). Information on the results for the other three TRECS are shown in Table 5.

Table 4: Kendall's tau correlations for the qrels formed from each manual run in TREC-8.

TREC-8 run	Kendall's tau
CL99XTopt	0.968
CL99XT	0.967
CL99SD	0.960
CL99SDopt1	0.953
iit99ma1	0.947
orcl99man	0.943
CL99SDopt2	0.940
GE8MTD2	0.923
READWARE2	0.917
8manexT3D1N0	0.904
READWARE	0.897
cirtrc82	0.722
disco1	0.679

Table 5: number of manual runs achieving Kendall's tau correlations; percentages are shown in brackets.

TREC	tau<.8	.8≥tau<.9	tau≥.9
5	3 (10)	6 (19)	22 (71)
6	4 (24)	3 (18)	10 (59)
7	0 (00)	5 (29)	12 (71)
8	2 (15)	1 (08)	10 (77)

Across the four TRECs considered, 69% of manual runs provided qrel sets that formed a viable test collection, a further 19% formed a collection that was somewhat different from TREC, and the remaining 12% formed collections that were noticeably different from TREC. Given that the purpose of almost every manual run submitted to the four TRECs was not to create a definitive qrel set, but to showcase or experiment with a searching method or interface, it is perhaps striking how few of the runs (just over 1 in 10) produce poor quality qrels.

3.2.2 Poor qrels

An experimenter wishing to create a test collection using ISJ would want to be assured that they are not going to be unfortunate enough to create such a poor qrel set. Therefore, the nine runs that produced such sets were examined in more detail, examining how the runs rated in comparison to all other submissions (manual and automatic) to TREC ad hoc.

Table 6: upper and lower rank positions of the least effective manual runs. Note there were no "least effective" runs in TREC-7.

TREC	upper	lower	size of rank
5	43	60	61
6	48	69	74
7	-	-	-
8	121	126	129

As can be seen in Table 6, results show that the runs were ineffective or very ineffective at retrieving relevant documents in comparison to other manual or automatic runs: appearing far down the overall system ranking. It was judged that it would be most unlikely that someone using the Cormack et al ISJ method would create such runs, as consistent poor performance in retrieving relevant documents would be noticed by the experimenter.

3.2.3 Conclusion

From these results, we concluded that the Interactive Searching and Judging method is broadly applicable regardless of retrieval system used or people employed to conduct the searching process.

3.3 Automatic runs

In the reexamination of ISJ, so effective was the strategy of using a single manual run to form a qrel set, it was decided to extend the experiment to all automatic runs as well. Again, forming a qrel set by intersecting the 1,000 documents per topic returned by each automatic run with the official TREC qrels; then ranking all ad hoc runs by the mean average precision measured from the newly formed qrel set and correlating (using Kendall's tau) the ranking with that produced from TREC.

Table 7: number of automatic runs achieving Kendall's tau correlations; percentages are shown in brackets.

TREC	tau<.8	.8≥tau<.9	tau≥.9
5	11 (37)	10 (33)	9 (30)
6	22 (39)	27 (47)	8 (14)
7	17 (20)	29 (34)	40 (46)
8	17 (15)	42 (36)	57 (49)

As can be seen in Table 7, in all four TRECs, the runs that produced a tau of less than 0.8 never constituted more than 39% of ad hoc runs examined. In TRECs 7 and 8, respectively 46% and 49%, of the submitted ad hoc runs produced qrels with a correlation above 0.9. Given the wide range of automatic runs submitted to TREC over the years, it striking that such a large number of effective qrels can be derived from the 1,000 documents of an automatic run where no form of pooling, system or query, is utilized.

Quite why TRECs 7 and 8 are better than TRECs 6 or 5 is not clear. One speculation is that it may reflect an improvement in the overall quality of ad hoc systems submitted to TREC in later years.

3.3.1 Poor qrels

As with the manual runs, the automatic runs that produced poor qrels sets were examined: as can be seen in Table 8, they were found to occur towards the bottom of a full (manual and automatic) ranking from TREC. With the exception of one run in TREC-6, all such poor runs occurred in the bottom half of each TREC system ranking.

Table 8: upper and lower rank positions of the least effective automatic runs.

TREC	Upper	lower	size of rank
5	41	61	61
6	33	74	74
7	63	103	103
8	107	129	129

3.3.2 Conclusion

It would appear from the results presented here that it is possible to create a set of qrels from the run of a single effective retrieval system. While results indicate that the methods presented in Sections 3.1 and 3.2 provide more effective, more efficient approaches to forming qrel sets, the result presented in this Section, runs counter to a largely held assumption (based on extensive past work) that pooling of some type is required to form test collections. The results in this experiment show that this is not the case, large test collections can be formed with no pooling.

4. CONCLUSIONS

While there is no dispute that with sufficient collaborating groups and person power, a combination of extensive system pooling and query pooling (as the automatic and manual runs in TREC provides) produces a high quality set of qrels. However, the result of the experiments presented in this paper showed that it is quite possible to create a usable set of qrels avoiding either one or both forms of pooling. Situations will arise where experimenters need to build a new test collection quickly and with limited resources. Through the adaptation of an existing relevance feedback-based method; the validation of an interactive searching and judging method; and the presentation of a new approach to building qrels using the output of a single automatic run, three methods experimenters can use were presented and shown to be effective.

5. FUTURE WORK

The evaluation of qrels presented here used the de-facto standard of measuring the Kendall's tau correlation between two system rankings, each produced by a different qrel set. The task the qrels are being used for in this standard is to rank a large number of retrieval systems (between 30 and 116 in TRECs 5-8), some very effective, some not. However, such a task is not the one most users of test collections employ, most experimenters are more likely to be comparing a small number of runs produced from variations of the same retrieval system. As pointed out by Bland and Altman (1986, p.308),

> *…Correlation depends on the range of the true quantity in the sample. If this is wide, the correlation will be greater than if it is narrow.*

Testing qrel sets on a large number of runs is likely to produce high correlations. The range of values when experimenting with a few run variants from a single system is likely to be smaller. A next step in the work presented here will be to examine how well all the qrel formation methods presented here rank smaller sets of runs.

It is also our intention to examine the *bpref* measure recently introduced (Buckley and Voorhees, 2004), which is specifically designed to be used in situations where qrels are not complete. Tests so far appear to indicate that the measure provides a more reliable means of ranking systems than mean average precision. Tests, however, have so far been on qrels formed from degraded system pools, our intention is to test it with the efficient qrel formation methods described here.

A further consideration is to examine if the accuracy of qrel sets in determining rank order of retrieval systems varies depending on the effectiveness of the retrieval system: in other words can a test collection distinguish between two ineffective retrieval systems better, as well as, or worse than it can between two highly effective retrieval systems? Such an important consideration does not appear to have been examined in the past and is work we shall be addressing next.

6. ACKNOWLEDGMENTS

The authors wish to thank Keith van Rijsbergen and Jamie Callan for valuable inputs at a number of stages of the work. Financial support for the work was provided by the EU 5th Framework RTD project SPIRIT: contract number IST-2001-35047.

7. REFERENCES

[1] Bland, J.M., Altman, D.G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307-310.

[2] Buckley, C., Voorhees, E.M. (2004), Retrieval Evaluation with Incomplete Information, in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*

[3] Cieri, C., Strassel, S., Graff, D., Martey, N., Rennert, K. and Liberman, M. (2002), Corpora for Topic Detection and Tracking, In: Allan, J. (ed.), *Topic Detection and Tracking: Event-based Information Organization*, 33-66, Kluwer.

[4] Cormack, G.V., Palmer, C.R. and Clarke, C.L.A. (1998), Efficient Construction of Large Test Collections, in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 282-289.

[5] Fox, E.A. and Shaw, J.A. (1993), Combination of Multiple Searches, in NIST Special Publication 500-215: *The 2nd Text REtrieval Conference (TREC-2)*, Gaithersburg, MD, 243-252.

[6] Garofolo, J.S., Voorhees, E.M., Stanford, V.M., Spärck Jones, K. (1997), TREC-6 1997 Spoken Document Retrieval Track Overview and Results, in *Proceedings of the 6th Text REtrieval Conference (TREC 6)*, NIST Special Publication 500-240, 83-92.

[7] Gilbert, H. and Spärck Jones, K. (1979), Statistical bases of relevance assessment for the 'ideal' information retrieval test collection, *British Library Research and Development Report 5481*, Computer Laboratory, University of Cambridge.

[8] Harman, D (1996), Panel: building and using test collections, in *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, 335-337.

[9] Harmandas, V., Sanderson, M., Dunlop, M.D. (1997), Image retrieval by hypertext links, in *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval*, 296-303.

[10] Kuriyama, K., Kando, N., Nozue, T. and Eguchi, K. (2002), Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop, *Information Retrieval*, 5 (1), 41-59.

[11] Lewis, D.D. (1992), An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task in

Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, 37-46

[12] Manmatha, R., Rath, T., Feng, F. (2001): Modeling Score Distributions for Combining the Outputs of Search Engines, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*

[13] Salton, G., Fox, E.A., Wu, H. (1983): Extended Boolean Information Retrieval, in *Communications of the ACM*, 26(11): 1022-1036

[14] Sheridan, P., Wechsler, M., and Schäuble, P. (1997), Cross-Language Speech Retrieval: Establishing a Baseline Performance, in *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval*, 99-108.

[15] Soboroff, I., Nicholas, C., and Cahan, P. (2001), Ranking retrieval systems without relevance judgments, in *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, 66-73.

[16] Soboroff, I. and Robertson, S. (2003), Building a filtering test collection for TREC 2002, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 243-250.

[17] Spärck Jones, K., (1974), Progress in Documentation: Automatic Indexing, *Journal of Documentation*, 30(4), 393-432.

[18] Spärck Jones, K., Van Rijsbergen, C.J. (1975), Report on the need for and provision of an 'ideal' information retrieval test collection, *British Library Research and Development Report 5266*, University Computer Laboratory, Cambridge.

[19] Spärck Jones, K., Bates, R.G. (1977), Report on a design study for the 'ideal' information retrieval test collection, *British Library Research and Development Report 5428*, Computer Laboratory, University of Cambridge.

[20] Stuart, A. (1983), Kendall's tau. In Kotz, S and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*, vol. 4, 367-369. John Wiley and Sons.

[21] Sullivan, D. (2002), The Search Engine "Perfect Page", in *Search Engine Watch* accessed from http://searchenginewatch.com/searchday/02/sd1104-pptest.html.

[22] Voorhees, E.M. (1998) Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, in *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, 315-323.

[23] Voorhees, E.M., Harman, D. (1998) Overview of the 7th Text REtrieval Conference (TREC-7), in *Proceedings of the 7th Text REtrieval Conference (TREC-7)* NIST Special Publication 500-242, 1-24.

[24] Voorhees, E.M., Harman, D. (1999) Overview of the 8th Text REtrieval Conference (TREC-8), in *Proceedings of the 8th Text REtrieval Conference (TREC-8)* NIST Special Publication 500-246, 1-24.

[25] Voorhees, E. (2001) Evaluation by Highly Relevant Documents, in *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, 74-82.

[26] Voorhees, E. (2002), Personal Communication.

[27] Zobel, J. (1998), How reliable are the results of large-scale information retrieval experiments? in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 307-314.