



# Spatially-Aware Information Retrieval on the Internet



SPIRIT is funded by EU IST Programme  
Contract Number: IST-2001-35047

## Test collection formation methods

<b>Deliverable number:</b>	D11 2102
<b>Deliverable type:</b>	R
<b>Deliverable nature:</b>	PU
<b>Contributing WP:</b>	WP 2
<b>Contractual date of delivery:</b>	01/01/04
<b>Actual date of delivery:</b>	31/12/03
<b>Authors:</b>	Mark Sanderson, Hideo Joho University of Sheffield

**Keywords:** Evaluation, building test collections.

**Abstract:** This document presents the methods that will be used to locate (within the SPIRIT terabyte collection) relevant documents for the queries to be used in any testing of the retrieval system. The means of locating relevant documents within test collections is commonly perceived to be a time consuming process, however the work presented here shows that a “reasonable” test collection can be formed with relatively little effort.

# Contents

1.	INTRODUCTION .....	4
2.	PAST WORK.....	5
2.1.	Special collections and queries .....	5
2.2.	Efficient pool sampling .....	6
2.3.	Interactive searching and judging .....	7
3.	THE EXPERIMENT .....	8
3.1.	Design of experiment.....	8
4.	RESULTS WITH ONE HOUR OF EFFORT .....	10
4.1.	Allowing more time per query .....	11
4.2.	More topics .....	12
5.	CONCLUSIONS .....	13
5.1.	Issues to consider for future work beyond the life of the project .....	13
6.	REFERENCES .....	14

# Executive Summary

In order for SPIRIT to conduct effective testing of its retrieval system, for the set of scenarios defined in deliverable D3 7101, it will be necessary to locate all or nearly all of the documents (within the SPIRIT terabyte collection) that are relevant to each scenario. With such a set in place, usability tests of user effectiveness in locating such documents can later be judged against a "gold standard" of relevance judgements. In the field of information retrieval, such standard test beds are common: called test collections, they are composed of a set of documents, a set of queries (the information seeking part of a scenario), and a list of the documents relevant to each query (known as relevance judgements or qrels). Creating the qrels is a time consuming resource intensive process often requiring the collaboration of a large number of research groups. So time consuming is this process that no work has been published describing the building of such a test corpus from a collection of documents as large as that held by SPIRIT.

This document describes a method for forming a test collection that requires between one and three hours of human effort per query, substantially less than that commonly perceived to be required when building a test collection. The resulting collection is shown to identify significant differences between retrieval systems with a high degree of accuracy. The collection formation method results from an analysis of where human effort is best applied: number of systems contributing to a pool; depth of ranking assessed; and human effort in re-formulating a query are all examined. Although some of these issues have been explored in the past, the work in this document presents a broader evaluation that measures the ability of new test collections to identify significant differences between retrieval systems. The presented method requires substantially less relevance assessor effort and little or no collaboration across research groups to form pools. The evaluation reveals that the accuracy is high, though lower than can be achieved with the TREC collection, which is treated here as ideal. However, as the effort required creating the new collection is considerably less than TREC, we judge our approach to be almost ideal.

# D11 2102

## ***Test collection formation methods***

### **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 - are seen by many to be one of the great innovations of IR researchers and an exemplar to other computer science fields. Their use has promoted extensive testing and comparison of retrieval algorithms. An ideal test collection is composed of the following:

- 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 judgements – 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 recognised 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 & 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, a pool was formed by merging the output of diverse searches. It was assumed that all or 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 utilised in gathering relevance judgements for the 5.5Mb Inspec collection. As described by Salton, Fox, & Wu (1983), for each query in the collection, seven different means of processing the query were run on a retrieval system and the documents retrieved by each means were merged with duplicates being removed and the resulting pool examined by relevance assessors. It is not clear from 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 each providing 1,000 top ranked documents for each of 50 queries (called *topics* by TREC) created per year. The top 100 from each ranking is added to a pool, which is exhaustively assessed. Across the first eight TRECs, the number of documents assessed per query ranged from 1,005 to 2,310 (Voorhees, 1999). In TRECs 4-8 the percentage of relevant documents located in the pool was between 5% and 8%. If 1,500 documents were being assessed at a rate of 100 documents per hour<sup>1</sup>, around 15 hours of *TREC assessor* effort per query were required to locate between 75 and 120 relevant documents. Working at 45 hours a week,  $\frac{1}{3}$  of a person year is required to assess the sampled pools for 50 queries. So great is the task taken on by TREC, there is perhaps a perception that building a test collection is beyond the resources of a single researcher or

---

<sup>1</sup> Figure taken from personal communication with Ellen Voorhees, 2002.

research group. Any “home made” collection produced would not allow accurate measurements of retrieval systems.

In the past, proposals have been made for reducing the effort in creating test collections (e.g. Zobel, 1998 and Cormack, Palmer, & Clarke, 1998), however, there would appear to be little evidence of the methods being adopted. One reason for this may be that existing methods of evaluating the new collections do not illustrate how reliable the new ones will be. This document provides an approach to evaluating such efforts, by measuring the ability of a new test collection formation method at identifying significant differences between retrieval systems. Starting with a review of past work in producing tractable test collections, the document provides a description of the evaluation method followed by the results of experiments conducted on the method. Limitations of the approach are considered and future work is proposed. Finally conclusions are drawn.

## 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.

#### 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 categorisation 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 categorisation collection (1992).

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

#### 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, & Schäuble (1997) built a spoken document test collection from radio news, where queries were restricted to subjects that referred to events, which 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* removes the need for relevance assessors. Such an approach was used in an early run of the TREC *spoken document retrieval* track (Garofolo, Voorhees, Stanford, & Spärck Jones, 1997). A similar form of evaluation was recently 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 queries 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 & Van Rijsbergen and with the work practise of TREC, all submitting systems and queries are treated equally. In TREC, relevance assessors examine the top 100 documents from each system for each query. Means of focussing effort on particular systems or particular queries have been proposed as well as a method to avoid assessor effort completely. The three approaches are described here.

### Focussing on queries

Zobel (1998) was interested in maximising the number of relevant documents located by assessors. He recognised 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.

### 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 judgements of relevance were made<sup>2</sup>, 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 rankings of systems were correlated using Kendall's Tau (Stuart, 1983). The correlation found was 0.999. Examining only 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, however they stated that

"It is difficult to determine how close to the benchmark a collection must be in order to yield reasonable judgements".

Despite their apparent worth in significantly reducing assessor effort, neither Zobel's nor Cormack et al.'s approach was adopted by TREC. Explaining why, Voorhees & Harman (1999) cited logistical problems and, with Zobel's proposal, a concern that judgements may be affected if assessors know they are being directed to examine the lower ranks of certain queries. For Cormack et al., Voorhees & Harman stated that an approach of biasing assessments to certain systems might produce a biased test collection that is more likely to rank highly certain types of retrieval system over others.

### 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.

---

<sup>2</sup> Cormack et al. used the recorded judgements of TREC assessors to simulate judgements being made.

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, relevance assessment, 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 query as he/she could think of noting all relevant documents retrieved. When no more relevant documents could be found, searchers moved onto another query. Spending on average just over 2 hours per query, 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 using the same system ranking methodology for MTF pooling: a Kendall's Tau correlation of 0.89 was obtained.

Cormack et al. stated that the lower correlation, when compared to MTF pooling, was due to both the different approach in forming the qrel set and the difference in opinion on what constitutes relevance between ISJ searchers and TREC assessors. Cormack et al. separated the two factors by identifying a set of documents that were selected by the ISJ judges, but had also been relevance assessed by TREC assessors. When comparing this set of qrels with the full TREC set, the correlation across the system rankings increased to 0.96. It appeared that somewhat more of the difference in correlation was due to differences in opinion between the TREC and ISJ judges than the clear difference in the judging process. Examining approximately 18% of the documents, in 15% of the time<sup>3</sup> and using no pooling of different system rankings, Cormack et al. produced a set of qrels that appeared to rank retrieval systems almost as well as TREC.

Note, it would be wrong to conclude from such a result that the pooling method for forming qrels, proposed by Spärck Jones and Van Rijsbergen, is unnecessary. A pool is still being formed with ISJ, however it is a pool formed from a single retrieval system retrieving on many variations of the same query (as occurred with the Inspec collection, see above). The success of a single system approach does indicate, however, that the pooling of multiple systems (the approach used in TREC) *may* be unnecessary. Before one could reach such a conclusion, however, a more detailed exploration of differences in the ISJ and TREC qrels must be made. Voorhees (1998) conducted such a comparison. She examined the ranking of systems in more detail by focussing on pairs of systems. She measured the probability of a pair, ranked in order by one set of qrels (those from ISJ), having their order swapped under the other set (from TREC). Voorhees measured the probability of a system rank order swap in relation to the difference in mean average precision measured between the system pairs. She showed that the probability of a swap varied inversely with the magnitude of difference in MAP: the greater the difference between two systems measured on one set of qrels, the lower the probability of a swap measured on the other. For differences greater than 0.05 the probability of a swap on the other collection was negligible. Voorhees used this and other evidence to conclude, "Different relevance assessments, created under disparate conditions, produce essentially the same comparative evaluation results." (Page 322).

---

<sup>3</sup> It is hard to make comparisons on time between ISJ and TREC as the ISJ process covers both the document retrieval and assessment phase. Estimates of time, taken by TREC assessors, ignore the set up and processing time of all the retrieval systems used to contribute to the TREC pool.

### 3. The experiment

The ISJ method appeared to be well suited to building a “home made” test collection, one built by a single researcher or research group: it required access to one retrieval system and approximately 2.5 weeks of searcher time to process 50 queries. However, it was not clear from the work of Cormack et al. how successfully the ISJ-based approach would work on other retrieval systems; it may only work well with the retrieval system used by Cormack et al. due to some feature of a retrieval algorithm, user interface, or query re-formulation method. The aim of the work presented in this document was to examine a range of other retrieval systems and manual searching strategies to understand the effectiveness of an ISJ-based approach to obtaining relevance judgements. In order to do this, it was necessary to design an experimental procedure that would allow extensive testing that could be conducted in a tractable amount of time.

#### 3.1. Design of experiment

Each year that TREC ran the so-called *ad hoc run*, the classic test collection experiment, both automatic and manual processing of TREC topics was allowed. Although to many, the manual runs held little research interest, TREC organisers encouraged them as manual runs were almost always more effective than the automatic: adding many relevant documents to the pool. In TREC-7, for example, Voorhees and Harman (1998) stated that although documents from the manual runs constituted only 21% of the assessed pool, 90% of the relevant documents located that year were found in manual runs, 24% were unique to them. The eight most effective such runs were described in brief by Voorhees and Harman. Most were 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 not entirely dissimilar to ISJ. Given the stated success of the manual runs in locating relevant documents, it was decided to form qrels from the rankings returned from each TREC-7 manual run as a means of simulating ISJ. The top N documents submitted from each run judged relevant by TREC assessors formed a *manual qrel set* for that run. The Mean Average Precision (MAP) for each of the *ad hoc* runs submitted to TREC-7 was then computed from the set. The MAP computed for the runs from the full TREC qrels was also computed and comparisons between the two sets made. Apart from measuring qrel effectiveness, it was also necessary to assess the human effort in building the qrel sets as full TREC, ISJ, and manual runs all took different amounts of time to form. To this end, the following assumptions were made:

- Manual run searchers were reported in “TREC run report documents” to take “minutes” to complete their query; this was rarely specified more accurately. Therefore, for the experiments in this document, it was assumed that minutes meant 30 minutes, probably an over estimate. Ideally, one would have manual run searchers judge their own documents, but this information was not present in the TREC result archive. The judgements of the TREC assessors were used instead: it was assumed documents were assessed at the rate of 100 per hour.
- As stated above, the full TREC qrels required 15 hours of assessor effort to build per query, this estimate ignored the efforts of those who submitted runs to TREC for the reason that it was very hard to estimate. It is likely to be, however, a substantial amount of time.
- ISJ searchers took just over 2 hours per query to both build the query and judge all documents returned in the submitted TREC run.

With such an experimental set up and an understanding of time taken, the following questions were examinable.

- How consistent is the manual qrel approach when applied on other retrieval systems with different retrieval features using different relevance assessors?
- Is there a better way of assessing the effectiveness of a set of qrels at ranking systems than Kendall's Tau or probability of a swap?
- How much of the ranking from a single system should be examined?

- How much better a pool is formed if two or three manual systems are combined?
- What improvement does the increase of topic size produce?

**Consistency of approach**

A concern of someone considering using an ISJ-like approach might be that its reported success was due to some special quality of the retrieval system used by Cormack et al. and that the approach would not work on other systems. In order to test this, manual run qrel sets were formed from the 50 top ranked documents from each of the 17 manual runs of TREC-7. MAP calculated on the qrels ranked the TREC-7 adhoc runs. For each run, the ranking was correlated (using Kendall's Tau) with the MAP ranking formed from the full TREC qrels. The results are shown in Table 1.

*Table 1 Kendall's Tau of manual runs*

System	Tau	System	Tau	System	Tau
nthu1	0.57	brkly26	0.82	clarit98comb	0.85
lanl981	0.70	uoftimgu	0.82	acsys7mi	0.85
uoftimgu	0.78	clarit98rank	0.83	iit98ma1	0.87
gersh1	0.79	harris1	0.84	uwmt7a2	0.88
lnmanual7	0.81	uwmt7a1	0.84	t7miti1	0.89
fsclt7m	0.81	clarit98clus	0.85		
		<b>Std. Dev.</b>	<i>0.08</i>	<b>Average</b>	<i>0.81</i>

As might expected, with less time taken than ISJ (1 hour<sup>4</sup> as opposed to just over 2), the correlation from manual qrels was lower than that reported by Cormack et al. (0.96). What is striking about the results, however, is the consistency of the Kendall's Tau measure. The standard deviation of the average across the 17 runs was under 10% indicating a strong uniformity of measure. The only exceptions were the NTHU1 and LANL981 runs. As no papers describing these runs appeared in the TREC-7 proceedings, it was hard to speculate what caused the lower Tau. Assuming the two runs to be outliers, it was concluded that using a form of ISJ to build qrel sets will work consistently across different types of retrieval system and across different users conducting the searching.

**Other means of measuring a test collection**

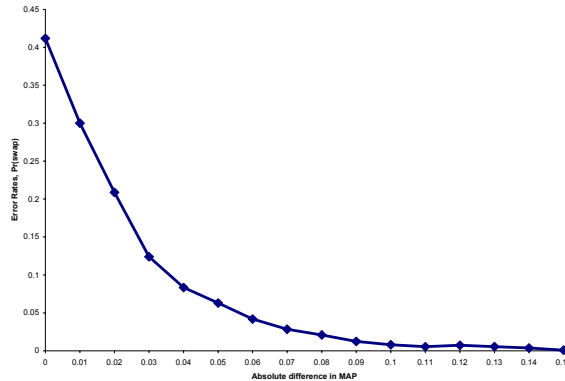
Although the Tau measure was lower for the manual qrels than those from ISJ, it was not possible to determine from the single figure if the qrels formed were good enough for someone wishing to use them for a test collection: what level of Kendall's Tau was sufficient? In order to understand better, the swap analysis used by Voorhees was conducted, as shown in Figure 1. Here, qrels formed from each of the manual runs, was examined. The probability of rank order swap between a pair of submitted runs was measured in relation to the absolute difference in MAP between each pair. The probability of a swap for differences greater than 6% was just under 0.05. Using this analysis, it would appear that for such differences, the collections were essentially performing as well as TREC, but with only one hour of human effort per query to form them.

However, the swap analysis provided a limited view of the utility of the manual run qrels. If two methods were measured to be very different using manual qrels but the TREC qrel set recorded only a small insignificant difference between them, the TREC qrels would be trusted more and the large difference in manual would be seen as an error of the set. However, the swap analysis would not reveal such an error. It was therefore necessary to devise another form of measuring the effectiveness of qrel sets.

When comparing two retrieval methods, the role of a test collection is to inform an experimenter which method across the collection topics is more effective at retrieving relevant documents. A measure of significance is typically used to indicate if the ordering of the methods is expected to be preserved for topics out with those of the collection. When comparing qrels, especially when one set (manual run) is a subset of another (TREC), it is important to understand if rank order is preserved across qrel sets (i.e. the Voorhees swap analysis) *and* if measures of significance computed on the smaller qrel set are consistent with significance measures determined from the larger. One can view measuring the effectiveness

<sup>4</sup> 30 searcher minutes forming the query, plus 30 TREC assessor minutes to judge the top 50.

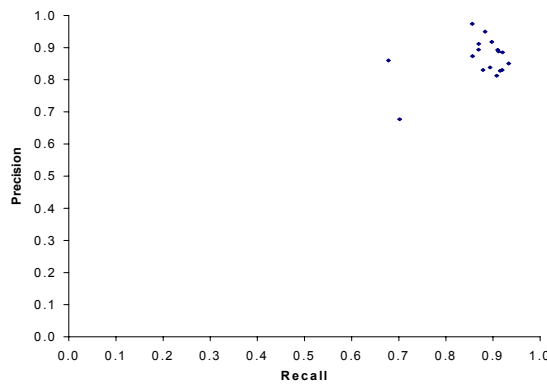
of the smaller qrel set as being similar to an evaluation of categorisation: full TREC qrels are used to categorise the pairs of runs that have a significant difference between them<sup>5</sup>. The precision and recall of the manual qrels in locating the TREC significant differences can be measured. Precision is the fraction of significant differences identified on manual qrels that were also identified on TREC. Similarly, recall is the fraction of TREC identified differences found with the manual qrels. All that remained was to define what was meant by significant. For these initial experiments, a simple test was chosen: 5% or greater absolute difference in MAP indicated significance (as suggested by Spärck Jones, 1974). With such a test in place, it was possible to examine variations in forming manual qrel sets.



**Figure 1 Error Rate (# of swap)**

#### 4. Results with one hour of effort

A series of experiments were conducted assessing the effectiveness of the manual run qrels. To start, the 17 runs were assessed measuring the precision and recall of each run at locating significant differences (see Figure 2). As can be seen (apart from the two outliers), the runs are similarly effective. They identify significant differences with between 81% and 97% precision and range in recall from 85% to 93%. Averaged across all 17 runs, recall was 87%, precision 86%. Ignoring the two outliers, recall rose to 89%, precision to 88%. One can vary the test of significance in the manual qrels to other values of absolute difference in MAP: Figure 3 shows the results of this variation averaged across all manual runs. For differences of 10% or higher, the average precision of the runs was 98%, however, recall dropped to 57%. Conversely, MAP differences of 1% or higher located 99% of significant differences, but they were identified with a precision of 66%.



*Figure 2 Precision/Recall of manual runs*

<sup>5</sup> It is assumed here that judgments of significance on TREC are correct. This is not strictly true, however, as any significance measure has a level of error associated with it.

Although the average recall and precision figures were not as high as one may like, it must be remembered that the research groups who constructed the retrieval systems and experimental approaches behind most of the manual runs were not motivated to build test collections, neither were the searchers who conducted the runs. The runs were used as a convenient way of measuring the consistency of the ISJ approach when using different retrieval systems. Examining standard deviation of average recall and precision reveals it was less than 10% of the two averages for the 17 runs, around 5% of the average without the two outliers. The low deviation emphasized the strong consistency of the manual runs in forming qrel sets. ISJ is a reliable approach to building qrel sets.

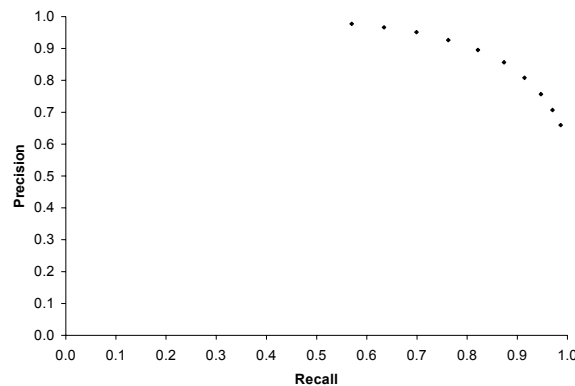


Figure 3 Precision/Recall with different MAP cut-off

Examining individual systems, one run was worthy of greater study. The building of the t7miti1 run was described by Voorhees and Harman as being conducted in a manner apparently similar to ISJ: upwards of 18 queries were submitted per topic and searchers only put into the run documents they judged to be relevant. The recall and precision of the qrels formed from the top 50 documents for this run was 86% and 97% respectively. Roughly 3 in 20 TREC identified significant differences were missed, but of those that were identified, only 3 in 100 were incorrectly identified. For most people's purposes, a test collection with this level of accuracy and coverage would be more than sufficient. That it was formed in a  $\frac{1}{15}$  of the time taken by TREC assessors (just over one week of effort) and did not require multiple systems contributing to a pool, is a significant bonus.

#### 4.1. Allowing more time per query

Given that one hour of searcher/assessor effort produced good quality test collections, it was decided to explore how more searcher effort could be best applied. The only way to improve the quality of a qrel set was to add relevant documents to it. Given that time was limited, it was important to maximize the number of relevant documents encountered by assessors. Three means of expending more effort were explored: looking further down a ranking; allowing some amount of system pooling; and judging more topics.

##### Examining depth of ranking

The results shown so far were based on an estimated one hour of assessor effort, examining the top 50 documents returned per query. An investigation of the benefits of examining more of the ranking returned from each manual run was conducted. Table 2 shows the results averaged across all 17 runs: as stated above, the top 50 documents (one hour of effort) obtained on average 87% recall and 86% precision. Using the same searcher effort, but requiring an assessor to examine more of the returned ranking (i.e. the top 150: 2 hours in total) improved recall/precision to 91%/89%. Increasing effort again by examining the top 250 (3 hours) raised average recall/precision to 92%/90%. Both Zobel (1998) and later Manmatha (2001) showed that the frequency of occurrence of relevant documents decreased as rank increased. Expending additional assessor effort examining more of the ranking of a particular retrieval system appeared to produce diminishing returns.

**More systems**

Another way of expending more effort in building a test collection is to allow an amount of system pooling. Although one of the aims of the work reported here was to investigate building a “home made” collection where it was assumed that access to retrieval resources was limited, it was judged reasonable to expect a research group to have access to two or three different systems. A series of experiments was conducted where all possible pairings of the 17 manual runs (136 pairings) were examined. The one hundred unique top ranked documents returned from the two systems were examined. It was assumed that 30 minutes of searching time per system was expended and that one hour of assessor effort was required to examine 100 documents. A repeat experiment where all system triples were examined (680) and the top 150 unique documents were examined.

Table 3 displays the results of the experiment showing minimum, maximum, and average recall & precision. As the number of systems was increased, the average rose. It was notable also that the minimum recorded value of recall and precision across all experiments rose greatly. By averaging two or three systems, the poor quality rankings from weaker manual runs, such as NTHU1 and LANL981, were alleviated by rankings from better runs: an advantage from some degree of system pooling. It is notable also that the maximum recall recorded across the experiments also increased although precision stayed the same.

*Table 2 Precision/Recall with depth of pool*

Hrs.	Top N	Recall	Precision
1	50	0.87	0.86
2	150	0.91	0.89
3	250	0.92	0.90

*Table 3 Precision/Recall with No. of systems*

# of sys	Hrs	Top N	Mn. R	Mn. P	Mx. R	Mx. P	Av. R	Av. P
1	1	50	0.72	0.70	0.93	0.97	0.87	0.86
2	2	100	0.81	0.77	0.97	0.97	0.93	0.89
3	3	150	0.88	0.82	0.98	0.97	0.96	0.91

Comparing precision and recall across tables 2 & 3, it can be seen that for both two and three hours of effort, working with multiple systems produced higher precision and recall than examining more of the ranking of a single system even though fewer documents were examined in the system pooling.

**4.2. More topics**

Lastly, an alternate use of assessor/searcher effort was to explore any possible improvements to a qrel set by examining more topics. With more, tests of significance were expected to be more reliable. Unfortunately, the experimental methodology of working with the manual runs of past TREC results did not allow a full wide ranging experiment as no run covered more than fifty topics. We attempted to examine the effect of topic size in the following manner.

The research question examined was: given a limited fixed amount of searcher/assessor time, is it better to assess the rankings from many topics to only a shallow depth; or is it better to examine the rankings of a few topics deeply?

To answer the question, the following experiment was conducted. The allowable searcher/assessor time was fixed to fifty hours (as was assumed in the one hour per topic experiments at the start of this Section). Between ten and fifty topics were examined and the depth of pool per topic was accordingly adjusted based on the time left after generating each query (i.e. thirty minutes per topic). Topics were randomly sampled from TREC-7; each experiment was repeated ten times. The results in Table 4 show that the accuracy increased as the topic size grew, but at thirty to forty topics the effectiveness of the qrels formed stopped increasing and for fifty topics, effectiveness was slight worse. It would appear there is an optimum depth of ranking some where between 75 and 116. If one is to examine more topics, it can not be at the expense of examining less than this depth of ranking.

Table 4 Precision/Recall with topic size

Topics	Top N	Recall	Precision
50	50	0.87	0.86
40	75	0.88	0.86
30	116	0.89	0.85
20	200	0.86	0.82
10	450	0.83	0.80

## 5. Conclusions

The work in this document described an analysis of TREC run data to explore more efficient means of creating a test collection. Assessing the rankings returned from 17 manual runs, qrels sets were formed. Means of testing the effectiveness of the sets was established by measuring the precision and recall of the sets at determining significant differences between pairs of retrieval systems. The assessment method provided a clearer understanding of the utility of a qrel set than previous measures. It was shown that despite variation in the retrieval systems and in the searchers forming the manual runs, interactive use of a single retrieval system was consistently effective at forming a test collection.

With a motivated searcher formulating the multiple queries for each topic, expending approximately one hour of effort per query, a test collection was produced that was accurate at identifying significant differences between retrieval systems. The level of accuracy was approaching that expected from a TREC collection. With additional effort and access to two or three retrieval systems, use of limited system pooling was found to be an effective way of increasing accuracy of the qrels even further.

It is to be concluded from such work that building a “home made” collection - using interactive searching followed by assessment to produce a qrel set - *is* something that researchers and research groups can undertake in a tractable amount of time.

The interactive searching and judgement process will be used in SPIRIT to produce the test collection upon which usability experiments can be based.

### 5.1. Issues to consider for future work beyond the life of the project

We consider that a number of further topics are worthy of investigation, both within the experimental framework described and beyond.

- Repeating these experiments on other years of TREC is an obvious extension to our work.
- A re-running of the number of topics experiment (Section 4.2) where limited system pooling is factored in as an additional variable is an area of investigation worthy of future work.
- Use of 5% absolute difference in MAP is a simple test of significance, but it is probably not the best. Zobel (1998) pointed out that small MAP differences between retrieval systems can be measured to be significant by tests like the sign test, t-test or Wilcoxon. Initial experiments repeating some of the experiments using these tests have been conducted producing largely similar results, however further investigation is warranted.
- When working with qrels from a single system, there is the important question of whether those qrels would effectively prefer certain types of document and of retrieval approaches to others. While the measures of recall and precision indicate that if there is bias, it is not wide spread, examining this issue in depth is worthy of exploration.

It must be remembered that the experiments described in this document were simulations of people forming test collections. The simulations inevitably had limitations: estimations of time taken were hard to make and were probably over estimated; searchers undoubtedly made relevance judgements, but these judgements were ignored; TREC assessor judgements were used where searcher judgements would have been preferred. An experiment testing the

simulations by getting users to search for relevant documents over a period of time will be conducted.

## 6. REFERENCES

- Cormack, G.V., Palmer, C.R. & 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.
- 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 6<sup>th</sup> Text REtrieval Conference (TREC 6)*, NIST Special Publication 500-240, 83-92.
- Gilbert, H. & 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.
- Harman, D (1996), Panel: building and using test collections, in Proceedings of the 19<sup>th</sup> annual international ACM SIGIR conference on research and development in information retrieval, 335-337.
- Harmandas, V., Sanderson, M., Dunlop, M.D. (1997), Image retrieval by hypertext links, in Proceedings of the 20<sup>th</sup> annual international ACM SIGIR conference on research and development in information retrieval, 296-303.
- Lewis, D.D. (1992), An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task in Proceedings of the 15<sup>th</sup> annual international ACM SIGIR conference on Research and development in information retrieval, 37-46
- Manmatha, R., Rath, T., Feng, F. (2001): Modeling Score Distributions for Combining the Outputs of Search Engines, in *Proceedings of the 24<sup>th</sup> annual international ACM SIGIR conference on Research and development in information retrieval*
- Salton, G., Fox, E.A., Wu, H. (1983): Extended Boolean Information Retrieval, in *Communications of the ACM*, 26(11): 1022-1036
- Sheridan, P., Wechsler, M., & Schäuble, P. (1997), Cross-Language Speech Retrieval: Establishing a Baseline Performance, in *Proceedings of the 20<sup>th</sup> annual international ACM SIGIR conference on research and development in information retrieval*, 99-108.
- Spärck Jones, K., (1974), Progress in Documentation: Automatic Indexing, *Journal of Documentation*, 30(4), 393-432.
- 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.
- 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.
- Stuart, A. (1983), Kendall's tau. In Kotz, S and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*, vol. 4, 367-369. John Wiley & Sons.
- Sullivan, D. (2002), The Search Engine "Perfect Page", in *Search Engine Watch* accessed from <http://searchenginewatch.com/searchday/02/sd1104-pptest.html>.
- Voorhees, E.M. (1998) Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness, in Proceedings of the 21<sup>st</sup> annual international ACM SIGIR conference on research and development in information retrieval, 315-323.
- Voorhees, E.M., Harman, D. (1998) Overview of the 7<sup>th</sup> Text REtrieval Conference (TREC-7), in *Proceedings of the 7<sup>th</sup> Text REtrieval Conference (TREC-7) NIST Special Publication 500-242*, 1-24.
- Voorhees, E.M., Harman, D. (1999) Overview of the 8<sup>th</sup> Text REtrieval Conference (TREC-8), in *Proceedings of the 8<sup>th</sup> Text REtrieval Conference (TREC-8) NIST Special Publication 500-246*, 1-24.
- Voorhees, E. (2002), Personal Communication.

**SPiRiT project**

Test collection formation methods

IST-2001-35047

*D11 2102*

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.