# Efficient Construction of Large Test Collections

Gordon V. Cormack[1] Christopher R. Palmer[1] Charles L. A. Clarke[2]
MultiText Project
[1] Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
[2] Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada
mt@plg.uwaterloo.ca

**Abstract** Test collections with a million or more documents are needed for the evaluation of modern information retrieval systems. Yet their construction requires a great deal of effort. Judgements must be rendered as to whether or not documents are relevant to each of a set of queries. Exhaustive judging, in which every document is examined and a judgement rendered, is infeasible for collections of this size. Current practice is represented by the "pooling method", as used in the TREC conference series, in which only the first $k$ documents from each of a number of sources are judged. We propose two methods, *Interactive Searching and Judging* and *Move-to-Front Pooling*, that yield effective test collections while requiring many fewer judgements. Interactive Searching and Judging selects documents to be judged using an interactive search system, and may be used by a small research team to develop an effective test collection using minimal resources. Move-to-Front Pooling directly improves on the standard pooling method by using a variable number of documents from each source depending on its retrieval performance. Move-to-Front Pooling would be an appropriate replacement for the standard pooling method in future collection development efforts involving many independent groups.

## 1 Introduction

A test collection for information retrieval requires three components: 1) a set of documents, 2) a set of queries, and 3) a set of relevance judgements. Ideally, the set of relevance judgements would be complete; that is, each document would be judged as relevant or not relevant with respect to each query. For a large corpus it is infeasible to construct this ideal collection, as prohibitive effort would be required to examine and judge every document with respect to every query. A more efficient approach is to examine and judge only a subset of the documents, provided the subset can be selected to include all the relevant documents, or even a large, representative sample of the relevant documents. The unjudged documents outside of this subset are assumed to be not relevant.

Selection of a subset for judging then becomes the issue. Random selection is one possibility, but for queries with few relevant documents it is quite likely that a substantial portion of the collection will still need to be judged before any relevant documents are found. Fortunately, information retrieval itself provides an immediate solution to this problem. It is the standard practice of information retrieval systems to rank documents according to their expected probability of relevance, following the so-called probability ranking principle [15]. If a retrieval system is reasonably effective, the highest ranking documents will be excellent candidates for inclusion in the subset for judging. This idea forms the basis for the *pooling method* of collection construction, as outlined by Sparck Jones and Van Rijsbergen in early collection development proposals [17, 18], and used as the primary collection construction method in the TREC (Text Retrieval Conference) experiments [4, 12].

The pooling method examines the top-ranked $k$ documents from each of $n$ independent retrieval efforts (runs). If $k$ and $n$ are large, the set of documents judged relevant may be assumed to be representative of the ideal set and therefore suitable for evaluating retrieval results. If $k$ and $n$ are large, however, up to $kn$ documents must still be examined and judged. For example, in the TREC-6 adhoc experiments [12] values of $k = 100$ and $n = 30$ were used, requiring approximately 60,000 judgements for 50 query topics despite overlap between retrieval runs. While this number represents a large improvement over judging the entire collection of roughly 500,000 documents for all 50 queries, substantial effort was required. The aim of this paper is to examine methods for reducing this effort, while maintaining the size and effectiveness of the resulting collection.

To deal with the problem of building such sets of relevance judgements, we examine the efficiency and effectiveness of the pooling method and two new approaches: an interactive retrieval and judging method, and a variant on pooling in which documents are examined in an order determined by the judgements using a Move-To-Front (MTF) heuristic. Experimental evidence based on the TREC-6 adhoc collection indicates that either method may be used to create an effective collection more efficiently than the pooling method.

## 2 Related Work

The pooling method is outlined by Sparck-Jones and van Rijsbergen in a 1975 collection development proposal for the British Library [17]:

> "Ideally, these [relevance judgements] should be exhaustive. But if not some attempt should be made to carry out independent searches using any available information and device, to obtain a pooled output for more broadly

based relevance judgements than may be obtained only with simple user evaluation of standard search output. In this case some estimate of the recall sample should be attempted."

We consider here the pooling method as used to construct the main test collections for the six TREC conferences to date [6, 7, 8, 10, 11, 12]. These collections consist of approximately 500,000 documents, 50 queries (topics), and relevance judgements made on 25-74 result sets submitted by participants. The submitted results for each topic were ordered by likelihood of relevance, and the top $k$ documents for each topic and each submission were selected to be judged by TREC assessors. Judgements for only these documents were included in the collection; all unjudged documents were assumed not relevant. For the various TREC collections, either $k = 100$ or $k = 200$ was used to ensure that a sufficient number of relevant documents would be judged.

Harman [5] examines the completeness of this approach and observes that doubling $k$ increases the number of relevant documents found by 11%. Harman also observes that the level of agreement between independent judges is 80%, but does not comment how this agreement affects collection effectiveness. Voorhees [20] examines the effect of judging agreement using the TREC-6 collection and the ISJ collection we built and judged independently and concludes that both are effective at ranking the TREC-6 submissions. Zobel [22] investigates several potential flaws in the pooling method of TREC and suggests a method to find a larger number of relevant documents with reasonable effort.

Sheridan et al [16], regarding the construction of a multilingual test collection, observes that there is still a substantial cost, in terms of time and money, associated with building a test collection using the pooling method. For this multilingual test collection, all data was derived from news services. Consequently, the data contained a temporal component. This temporal component was exploited to develop topic specific test collections that are excellent approximations of complete judging. Specifically, "unpredictable" events are chosen so that large segments of the base corpus, all data prior to this event, may be assumed to be not relevant. A topic-specific test collection is created by taking, as a subset of the corpus, all data from the time of the event until three days following the event. Relevance judgements are made on every documents in this three day period. It is not apparent how to generalize this approach to the construction of general test collections.

Additionally some areas of information retrieval, such as experiments using corrupted databases (simulating the problem of searching OCR generated text) and more recently spoken document retrieval and video clip retrieval. have adopted a different evaluation technique. This is referred to as either *known-item searches* [14] or as system evaluation based on a "seed" document [16]. A known-item retrieval search simulates the retrieval process undertaken by a user that is seeking a specific, half-remembered, document. Evaluation of systems using known-item searches does not require relevance judgements. Instead, a relatively distinctive document is located in the database and then a user need is associated with this document.

While known-item searching is very appealing when evaluating the retrieval of potentially corrupt information, it requires new evaluation methodologies and does not address the more general information retrieval task of categorizing large volumes of information. Test collections with relatively large numbers of relevance judgements are still required.

| Collection | Metric | P@30 | AP | RP |
|---|---|---|---|---|
| AT6 | P@30 | 1.0000 | 0.9760 | 0.9833 |
| AT6 | AP | 0.9760 | 1.0000 | 0.9918 |
| AT6 | RP | 0.9833 | 0.9918 | 1.0000 |

Table 1: AP correlation between scoring functions

## 3 Collection Effectiveness

The goal of information retrieval is, for a given query and collection, to return documents in the collection that are most likely to be relevant to the user's information need. A good result returns many documents that are relevant and few documents that are not relevant. *Precision*, the ratio of relevant documents returned to documents returned, and *recall*, the ratio of relevant documents returned to the number of relevant documents in the collection, are two standard measures used to evaluate systems. If the result documents are ordered by estimated probability of relevance, precision and recall can be computed incrementally as a function of the number of documents returned. We consider three common measures used to evaluate retrieval performance: *average precision* (AP), the integral of precision with respect to recall; *R-precision* (RP), the precision after the first $R$ results, where $R$ is the number of documents judged relevant in the collection; and *precision at 30* (P@30), the precision after the first 30 results. It is well known that there is a strong correlation among all three of these measures [19]. Table 1 show that we found comparable correlations between these metrics when they are used to evaluate the TREC-6 submissions. To be consistent with TREC practices, we have chosen to adopt AP as the standard measure for evaluating retrieval system performance for the purposes of this paper.

With this measure, we wish to compare and attempt to evaluate the "effectiveness" of collections that differ only in their relevance judgements; they contain the same documents and the same queries. One approach would be to determine how similar each collection is to the ideal collection using some metric. There appears to be no accepted metric in the literature and so we chose to use four metrics: 1) a simple count of the number of relevant documents in each collection, 2) the root mean square (RMS) of the differences in AP, 3) the linear correlation between AP values, and 4) the Kendall correlation [21] between AP values.

The number of documents judged relevant is a simple metric; it is maximized in the ideal collection. Harman [5, 9] uses this metric in evaluating the TREC pool. However, this metric ignores any bias that may exist in the subset of documents to be judged and it ignores the quality of the judgements, limiting its value for our purposes. The other metrics compare the result of computing the evaluation measure on a set of retrieval results.

The RMS difference between collections $C_1$ and $C_2$ given ranked submissions $r_1, \cdots, r_n$ is

$$\sqrt{\frac{\sum_{i=1}^{n}(AP[C_1](r_i) - AP[C_2](r_i))^2}{n}}$$

where $AP[C_i](r_j)$ is the average precision of submission

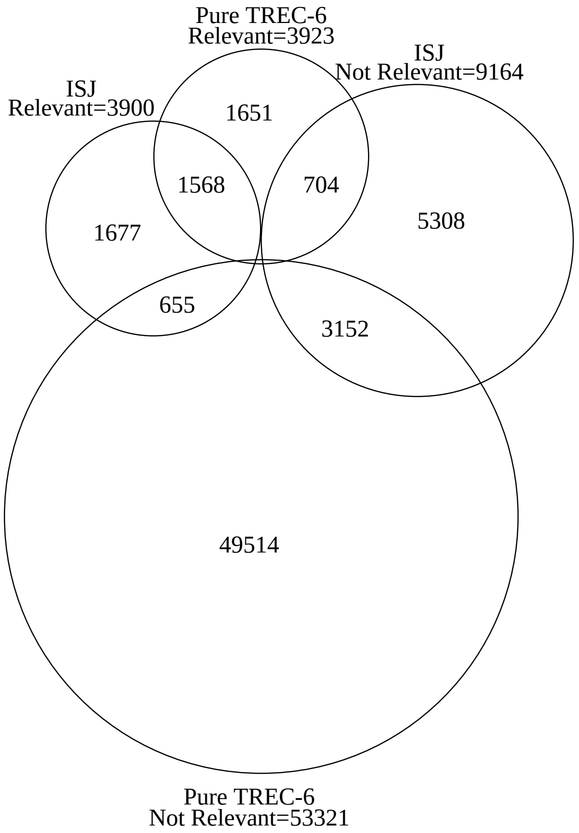


Figure 1: ISJ and PT6 Judging Agreement.

$r_j$ evaluated using collection $C_i$. RMS difference measures the absolute difference in average precision values between two collections when used to evaluate the same set of retrieval results. RMS difference is influenced by both magnitude and correlation and so provides an indicator of the direct comparability of inter-collection average precision values.

The Kendall correlation [21] is used by Voorhees [20] to compare the effectiveness of collections in ranking the submissions to TREC-4 and TREC-6, and is included here for comparison. It is a measure of the number of transpositions in the permutation between rankings normalized such that identical rankings have a Kendall correlation of 1.

Of course, it is infeasible to construct the ideal collection with which to compare. Instead, two independently derived collections will be used as benchmarks.

## 4 TREC-6 Collections

In the experiments that follow, we use for baseline purposes two collections derived from the TREC-6 collection: the actual TREC-6 collection used for TREC-6 judging (AT6), which includes the ancillary judgements, and a "purified" TREC-6 collection (PT6) which excludes the judgements contributed by the ancillary experiments and the judgements contributed by the MultiText manual adhoc submission on which the ISJ experiments reported below are based. The ancillary experiments are excluded as they were not contributed by the standard pooling method and the MultiText submission is excluded to maintain independence between the construction strategies being compared in this paper.

For the actual TREC collection (AT6), about 70,000 documents were judged. Some 60,000 came from the pool of submitted results, and the rest from ancillary experiments. In total, 4,611 documents were judged relevant. The "pure" TREC collection (PT6) is the result of 15,000 fewer judgements due to the exclusion of the MultiText manual adhoc submission and the ancillary experiments. In addition, due to an I/O error when assembling the TREC-6 data, one other submission was inadvertently omitted: Brkly21 (submitted by a group at the University of California, Berkeley). The judgements in PT6 are those that would be generated by applying the standard pooling method to the remaining 28 TREC-6 submissions evaluated by the official TREC-6 judges. The PT6 collection contains 3,923 documents judged relevant.

## 5 Interactive Searching and Judging

Extensive interactive searching by multiple searchers can be used to produce a large set of relevance judgements. By combining the efforts of multiple searchers, we may expect to identify a representative subset of the relevant documents. As part of our TREC-6 experimental work [3] we used this *Interactive Searching and Judging* (ISJ) strategy to create a a set of relevance judgements for the TREC-6 adhoc queries and documents.

Working together, four searchers created a set of judgements for the TREC-6 adhoc queries and documents. Our effort was prior to and independent of the TREC-6 judging effort, except for the fact that the documents we found were used in our submission to TREC-6 and therefore contributed to the pool used in creating the official collection (AT6).

The interactive search system used was that of our MultiText project, which performed well in TREC-4 and TREC-5 [1, 2]. The system uses manual boolean query construction and ranks documents based on the length and number of passages that satisfy the query. The system displays passages satisfying the query with the search terms highlighted and allows assessors to record their judgements. A complete description of the retrieval system and the ISJ interface appears elsewhere [1, 3].

The searchers used no formal strategy for searching other than to try to find as many relevant documents as possible for each topic with reasonable effort. No specific limit was placed on time spent per topic. Informally, the strategy used was to formulate a query and to judge the results of the query until the frequency of relevant judgements dropped to a level where continuing seemed fruitless. At this point, another query was formulated or the topic was abandoned. The searchers spent 105 hours, 2.1 hours per topic on average, formulating queries and judging documents. The searchers judged 13,000 documents, of which 3900 were judged relevant.

### 5.1 Comparison

The ISJ and PT6 collections contain respectively 3900 and 3923 documents judged relevant, with an intersection of 1568 documents. Of those not in the intersection, 655 were judged relevant in ISJ and not relevant in PT6, while 704 were judged relevant in PT6 and not relevant in ISJ. This agreement is comparable to the agreement reported between TREC judges [10]. The remaining documents judged relevant were unjudged in the other collection. The total number of documents examined in

| Collection | ISJ | PT6/ISJ | PT6 | ISJ/PT6 | AT6 |
|---|---|---|---|---|---|
| ISJ | 1.0000 | 0.9990 | 0.9859 | 0.9870 | 0.9849 |
| PT6/ISJ | 0.9990 | 1.0000 | 0.9866 | 0.9874 | 0.9850 |
| PT6 | 0.9859 | 0.9866 | 1.0000 | 0.9973 | 0.9996 |
| ISJ/PT6 | 0.9870 | 0.9874 | 0.9973 | 1.0000 | 0.9975 |
| AT6 | 0.9849 | 0.9850 | 0.9996 | 0.9975 | 1.0000 |

Table 2: Effect of judging variations between ISJ and PT6, AP correlation

| Collection | ISJ | PT6/ISJ | PT6 | ISJ/PT6 | AT6 |
|---|---|---|---|---|---|
| ISJ | 1.00000 | 0.97965 | 0.89749 | 0.89454 | 0.89201 |
| PT6/ISJ | 0.97965 | 1.00000 | 0.90375 | 0.90097 | 0.89671 |
| PT6 | 0.89749 | 0.90375 | 1.00000 | 0.96106 | 0.98669 |
| ISJ/PT6 | 0.89514 | 0.90140 | 0.96087 | 1.00039 | 0.95539 |
| AT6 | 0.89201 | 0.89671 | 0.98669 | 0.95558 | 1.00000 |

Table 3: Effect of judging variations between ISJ and PT6, Kendall correlation

constructing ISJ and PT6 were 13064 and 57244 respectively. Figure 1 show these sets as a Venn diagram.

Two factors account for the difference between the ISJ and PT6 collections: a different set of documents were judged, and different judgements were made on some of the same documents. To separate these two factors, we created two additional collections: ISJ/PT6 and PT6/ISJ. ISJ/PT6 uses the judgements made by NIST assessors for the documents in the ISJ pool. Similarly, PT6/ISJ is constructed by using the judgements made in ISJ for the documents in the pool constructed for PT6. That is, ISJ and ISJ/PT6 judge the same set of documents, but use different judging, as do PT6 and PT6/ISJ. On the other hand, ISJ and PT6/ISJ use the same judging on different sets of documents, as do PT6 and ISJ/PT6.

Table 2 gives the correlation among summary average precision values between all pairs of the four collections (as well as AT6 for comparison). From this matrix we can conclude that there is a strong correlation among the results from all four pools. The correlation between ISJ and PT6 is 0.986. The strongest correlations (0.999 and 0.997) occur for ISJ vs. PT6/ISJ and PT6 vs. ISJ/PT6 — collections with the same judging and on different sets of documents. The weaker correlations (0.987 and 0.987) occur for ISJ vs. ISJ/PT6, PT6 vs. PT6/ISJ — collections with different judgements on the same set of documents. The Kendall correlations given in table 3 strongly shows the same effect. Our conclusion is that the difference in effectiveness between ISJ and PT6 is small and that the dominating factor in this difference is the difference in judging rather than the difference in the set of judged documents. Had ISJ been judged by the same assessors as PT6, we suggest that much of the difference would vanish. In any event, Salton and Lesk [13] argue that much larger differences in agreement in judges produce no substantial difference in effectiveness. Voorhees [20] independently compared our ISJ collection with AT6 and also compared variants of the TREC-4 collection made with different judgements, concluding that there was no substantive difference in effectiveness in either case.

To place the magnitude of the difference between ISJ and PT6 in context, these correlations may be compared to the correlations between different measures on the full TREC-6 collection given in table 1. These correlation values are comparable to the correlation values between ISJ and PT6, and between their variants. The magnitude of the difference in effectiveness between these two independently derived collections can be accounted for almost entirely by judging differences, and is comparable to the difference between different measures on the same collection.

## 6 Move-To-Front Pooling

In contrast to the pooling method of TREC, which examines documents in arbitrary order, Move-To-Front Pooling examines documents in order of their estimated likelihood of relevance. Within a submission, documents are assumed to be ordered by likelihood of relevance; among submissions, likelihood of relevance is estimated using the judgements rendered on documents examined so far. A submission that has more recently yielded a relevant document is assumed to be more likely to yield another, and is examined first. The net effect is that more documents are judged from submissions that are observed to perform well, while fewer documents are judged from the others.

The pooling method is justified by the assumption that the documents in each submission are ranked by their probability of relevance. That is, if we denote by $P_s(r)$ the probability that the document with rank $r$ in submission $s$ is relevant, $P_s$ is assumed to be a monotone decreasing function. It follows from this assumption that if $k$ documents are to be judged from submission $s$ that the top-ranked $k$ documents can be expected to yield the most relevant documents and hence the most effective collection. We evaluated the effectiveness of the pooling method for $k$ from 1 to 100.

The pooling method, as implemented for TREC, judges $k$ documents from each submission. This approach maximizes the number of relevant documents found only if we assume that $P_s(r)$ is identical for every $s$ in the pool. This assumption is not realistic as it is obvious that some submissions have better performance than others. That is, documents at the same rank in different submissions have different probabilities of relevance. If $P_s(r)$ were known exactly, the strategy to find the maximal number of relevant documents would be to judge every document from run $s$ such that $P_s(r) > p$ for some probability threshold $p$. This would examine more documents from submissions with better performance. Alternatively, we could build a priority queue consisting of all documents, and repeatedly select and judge the document with highest $P_s(r)$. Such a priority queue is approximated using the *Move-To-Front* heuristic (MTF).

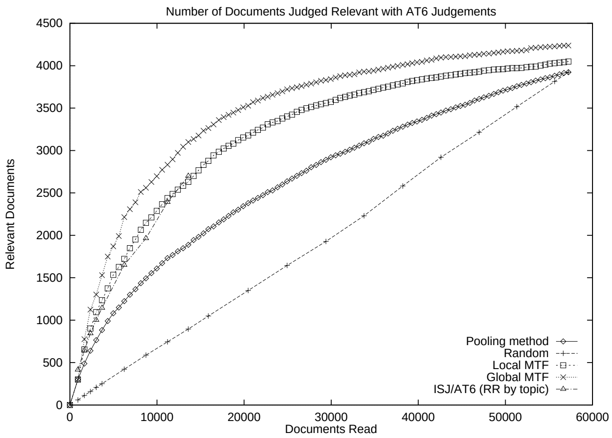


Figure 2: Relevant documents found using the AT6 benchmark

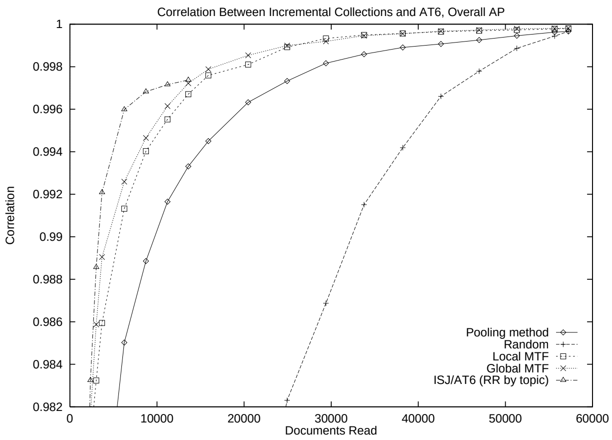


Figure 3: Incremental correlation using the AT6 benchmark

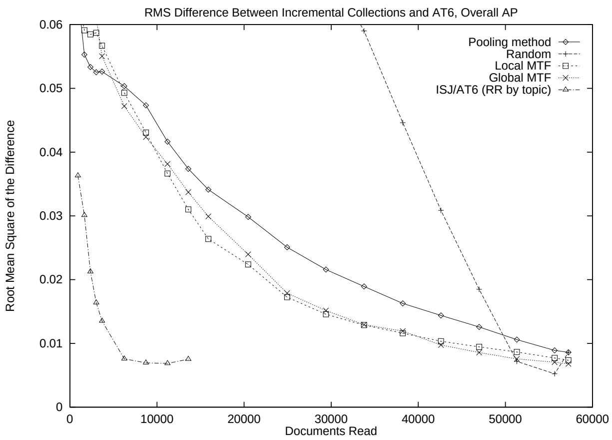


Figure 4: RMS difference in AP using the AT6 benchmark

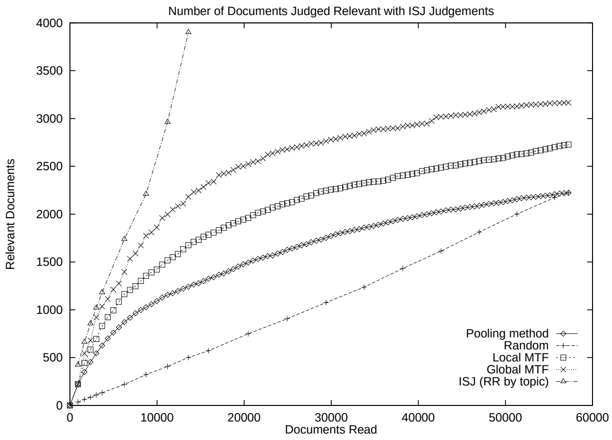


Figure 5: Relevant documents found using the ISJ benchmark

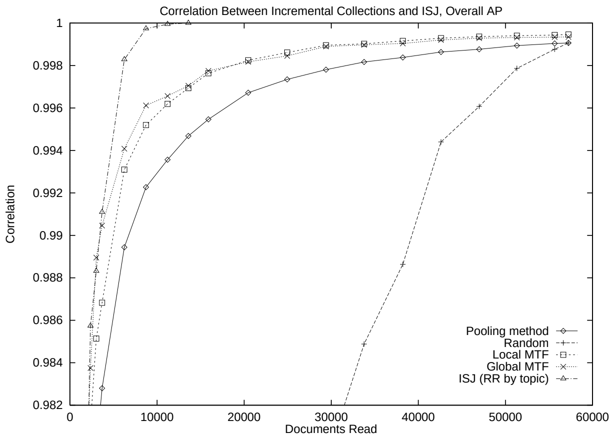


Figure 6: Incremental correlation using the ISJ benchmark

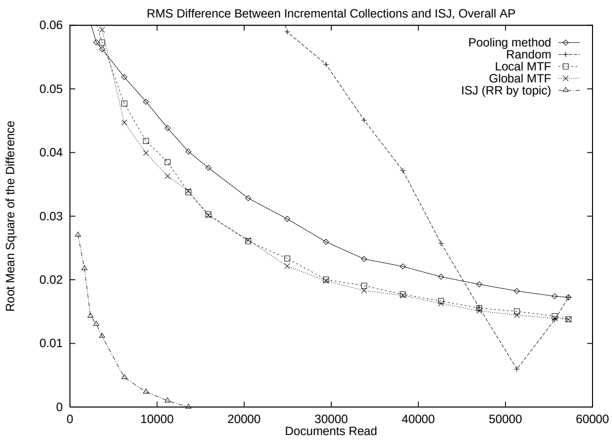


Figure 7: RMS difference in AP using the ISJ benchmark

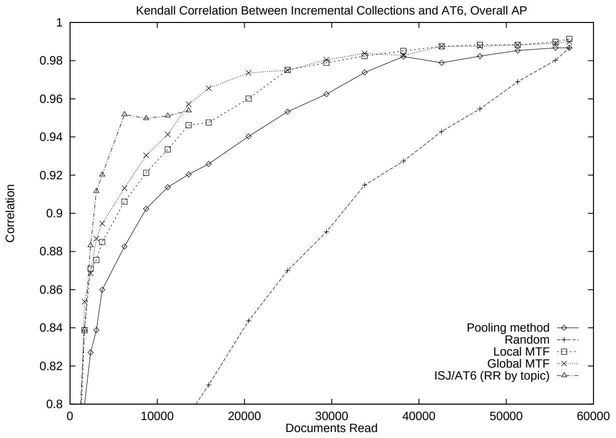


Figure 8: Kendall correlation using the AT6 benchmark

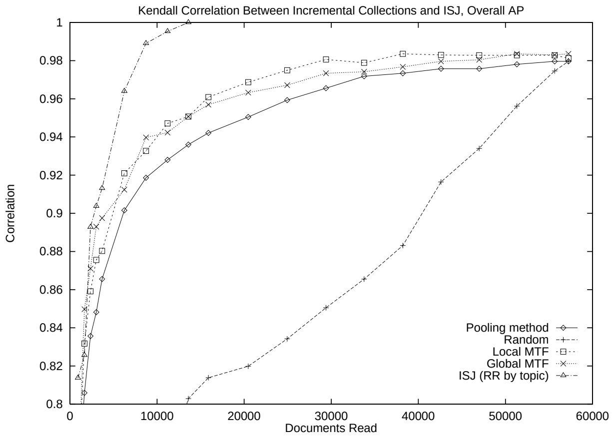


Figure 9: Kendall correlation using the ISJ benchmark

Under MTF, the documents from each submission are judged in order of rank. The submissions themselves are prioritized, and the top-ranked document from the submission with the top priority is judged. If it is judged relevant (or has been previously judged relevant because it appeared in some other submission) its priority is set to the maximum. Otherwise, its priority is reduced. MTF approximates a probability-ordered priority queue under the assumption that the difference in rank between consecutive relevant documents in the submission is a good indicator of the reciprocal of the probability of relevance.

We examine two variants of MTF. The first, *global MTF*, considers all submissions for all topics together: the queue selects the next document to be judged using the MTF heuristic, independent of the topic for which the document has been submitted as relevant. This approach can be expected to judge more documents for "easy" topics than for hard ones. The second, *local MTF*, considers the same number of documents for each topic, to a maximum of the number of documents judged by the pooling method for $k = 100$. Local MTF ensures that each topic receives a comparable number of judgements.

## 6.1 Comparison

For comparison, we evaluated two other methods: "random", and ISJ. For random, we selected documents to be judged at random from the 100 top-ranked documents of each submission, and evaluated the effectiveness as a function of the number of documents judged. For ISJ, we use the documents judged in creating the ISJ collection, but using the same judgements as the remaining runs. To form a collection using $k_d$ judgements, we select an equal number of documents for each topic, in the order they were found by ISJ.

The effectiveness of each strategy was evaluated as a function of the number of documents judged using the relevance judgements of the benchmark collection. For each strategy, collections were built by judging the same number of documents as the pooling method for each $k$ from 1 to 100. The resulting collections were used to compute summary average precision for each of the 72 TREC-6 submissions, excluding our submission derived from ISJ. These values were compared to the corresponding values computed using a benchmark collection. Judgements for all collections were taken from the benchmark; documents unjudged in the benchmark were deemed not relevant.

We chose two benchmark collections: AT6, containing all judgements rendered for TREC-6, and ISJ, containing our independent judgements. AT6 is biased toward the pooling method, as pooling yielded the vast majority of documents that were judged in the collection. The ISJ collection is similarly biased toward the ISJ strategy used to create it, but independent of the incremental pooling strategies being compared. We considered combining AT6 and ISJ in various ways to create a pool with more judgements that would therefore be closer to ideal. But we rejected these combinations as benchmarks because they were too similar in effect to AT6 with its intendent bias.

Figures 2, 3, 4 and 8 show the effectiveness of each of the judging strategies as a function of the number of documents judged, for each of the metrics, using AT6 as the benchmark. Global MTF finds more relevant documents more efficiently than local MTF, which is in turn more efficient than pooling using a fixed number of documents per submission. Using correlation or RMS difference as a metric, the advantage of global over local MTF is substantially reduced, but both are more efficient and effective than pooling. As predicted, the random strategy reaches the same end point as pooling, but reaches this point much less efficiently. The RMS difference curve for the random strategy dips below the other curves due to the influence of the magnitude of the average precision values, which increase with the number of documents judged. At about 56,000 documents judged, the mean value reaches and later exceeds that of the benchmark, resulting in an increase in RMS difference beyond this point.

ISJ achieves good effectiveness very efficiently, but stops before achieving the overall effectiveness of the other methods because it finds only about two-thirds of the documents judged relevant in the AT6 benchmark. The RMS difference in average precision relative to the benchmark is slightly smaller than achieved by the other methods - less than 0.01. We are lead to conclude that the average precision values between ISJ and the benchmark compare well in magnitude over and above being well correlated.

Figures 5, 6, 7 and 9 give the same metrics relative to ISJ as the benchmark. The results for all metrics are similar, with the relative performance for all approaches unchanged for all metrics. The performance of ISJ is exaggerated because it was used to determine the documents to be judged in the benchmark; nevertheless the curves suggest that it would have been possible to build an effective pool with fewer judgements. The results from both the ISJ and the AT6 benchmarks are remarkably similar in spite of the inherent bias toward the pooling method of the AT6 benchmark.

It is difficult to determine how close to the benchmark a collection must be in order to yield reasonable judgements. Judging a very small number of documents yields a collection whose performance is well correlated with the benchmark. For example, the pooling method with $k = 1$ judges 922 documents, finds 300 relevant and has a correlation of 0.92099 with respect to the AT6 benchmark. By judging half as many documents as pooling with $k = 100$, either MTF strategy creates a collection that correlates as well with the benchmark (0.999). Correlations of 0.990 can be achieved with one tenth as many judgements. For comparison, recall that the correlations among average precision, R-precision and precision @30 for the TREC-6 collection range from 0.976 to 0.992. Recall also that different judging strategies yield a correlation of only 0.987, a level that is achieved very early in the effectiveness curves. We must ask the question, "How close is close enough?" because the efficiency of building a collection depends heavily on the answer. We leave the task of answering this question as a topic for future research.

## 7 Conclusion

The manual effort associated with creating a set of relevance judgements for an information retrieval test collection can be reduced considerably without compromising the quality of the collection. Using interactive search we independently created a set of relevance judgements for the queries and corpus used in the TREC-6 retrieval experiments. Although the number of relevant documents contained in this set is approximately the same as in the official set, its creation required less than one-quarter as many judgements and took only 105 person-hours, a level of effort well within the capabilities of a small research team. Despite the difference in effort, the suitability of the resulting collection as a tool for evaluating retrieval effectiveness, the collection effectiveness, was not compromised. We found that the collections have similar effect, and that the difference between the collections appears to be dominated by disagreement between judges rather than by the differences in the set of documents selected for judging.

The selection of documents for judging may be based on a set of independent retrieval runs. In this circumstance, it is standard practice to pool the top $k$ documents from each run and then judge all the documents in this pool. The efficiency of this pooling method may be improved by using the effectiveness of the runs themselves as a guide to the judging process. Using a Move-To-Front heuristic, runs are evaluated in rank order with the next target for judging coming from the run in which a relevant document was most recently found. We examined both a global and a local version of this heuristic. While the global version produced relevant documents more quickly, neither was an obvious favorite in terms of the quality of collection produced for a given number of judgements made. Both versions produced an effective collection with considerably fewer judgements than would be required under the pooling method.

## 8 Acknowledgements

The authors thank Ilana Rosenshein for her assistance in rendering the ISJ judgements. This research was supported by Communications and Information Technology of Ontario (CITO) and by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

[1] C. L. A. Clarke and G. V. Cormack. Interactive substring retrival. In D. K. Harman and E. M. Voorhees, editors, *Information Technology: The Fifth Text REtrieval Conference (TREC-5)*, Gaithersburg, Maryland, November 1996. National Institute of Standards and Technology (NIST), United States Department of Commerce. Available electronically at `http://trec.nist.gov`.

[2] C. L. A. Clarke, G. V. Cormack, and F. J. Burkowski. Shortest substring ranking. In D. K. Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, pages 295–304, Gaithersburg, Maryland, November 1995. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-238. Available electronically at `http://trec.nist.gov`.

[3] G. V. Cormack, C. L. A. Clarke, C. R. Palmer, and S. S.-L. To. Passage based refinement. In *Sixth Text REtrieval Conference (TREC-6)*, Gaithersburg, Maryland, November 1997. National Institute of Standards and Technology (NIST), United States Department of Commerce. Available electronically at `http://trec.nist.gov`.

[4] D. Harman. Overview of the first TREC conference. In *16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 36–47, Pittsburgh, PA, June 1993.

[5] D. Harman. Overview of the fourth Text REtrieval Conference (TREC-4). In D. K. Harman, editor, *The Fourth Text REtrieval Conference (TREC-4)*, pages 1–23, Gaithersburg, Maryland, November 1995. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-236. Available electronically at `http://trec.nist.gov`.

[6] D. K. Harman, editor. *The First Text REtrieval Conference (TREC-1)*, Gaithersburg, Maryland, November 1992. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-207.

[7] D. K. Harman, editor. *The Second Text REtrieval Conference (TREC-2)*, Gaithersburg, Maryland, November 1993. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-215.

[8] D. K. Harman, editor. *Overview of the Third Text REtrieval Conference (TREC-3)*, Gaithersburg, Maryland, November 1994. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-225. Available electronically at `http://trec.nist.gov`.

[9] D. K. Harman. Overview of the third Text REtrieval Conference (TREC-3). In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 1–19, Gaithersburg, Maryland, November 1994. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-225. Available electronically at `http://trec.nist.gov`.

[10] D. K. Harman, editor. *The Fourth Text REtrieval Conference (TREC-4)*, Gaithersburg, Maryland, November 1995. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-236. Available electronically at `http://trec.nist.gov`.

[11] D. K. Harman and E. M. Voorhees, editors. *Information Technology: The Fifth Text REtrieval Conference (TREC-5)*, Gaithersburg, Maryland, November 1996. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-238. Available electronically at `http://trec.nist.gov`.

[12] D. K. Harman and E. M. Voorhees, editors. *The Sixth Text REtrieval Conference (TREC-5)*, Gaithersburg, Maryland, November 1997. National Institute of Standards and Technology (NIST), United States Department of Commerce. Available electronically at `http://trec.nist.gov`.

[13] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Management*, 4:343–359, 1966.

[14] E. V. Paul B. Kantor. Report on the TREC-5 confusion track. In D. K. Harman, editor, *Information Technology: The Fifth Text REtrieval Conference (TREC-5)*, pages 65–74, Gaithersburg, Maryland, November 1996. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-236. Available electronically at `http://trec.nist.gov`.

[15] S. E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33:294–304, 1977.

[16] P. Sheridan, J. P. Ballerini, and P. Schäuble. Building a large multilingual test collection from comparable news documents. In G. Grefenstette, A. Smeaton, and P. Sheridan, editors, *Workshop on Cross-Linguistic Information Retrieval*, pages 56–65. ACM SIGIR, Aug. 1996.

[17] K. Sparck Jones and C. J. Van Rijsbergen. Report on the need for and provision of an 'ideal' test collection. Technical report, University Computer Laboratory, Cambridge, 1975.

[18] K. Sparck Jones and C. J. Van Rijsbergen. Information retrieval test collections. *Journal of Documentation*, 32(1):59–72, March 1976.

[19] J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the TREC-3 data. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 385–398, Gaithersburg, Maryland, November 1994. National Institute of Standards and Technology (NIST), United States Department of Commerce. NIST Special Publication 500-225. Available electronically at `http://trec.nist.gov`.

[20] E. M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, August 1998.

[21] N. West. *Applied Statistics for Marine Affairs Professionals*. Praeger, Westport, CT, 1996.

[22] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, August 1998.