How Reliable are the Results of Large-Scale Information Retrieval Experiments?

Justin Zobel
Department of Computer Science, RMIT, GPO Box, 2476V, Melbourne 3001, Australia
jz@cs.rmit.edu.au

Abstract

Two stages in measurement of techniques for information retrieval are gathering of documents for relevance assessment and use of the assessments to numerically evaluate effectiveness. We consider both of these stages in the context of the TREC experiments, to determine whether they lead to measurements that are trustworthy and fair. Our detailed empirical investigation of the TREC results shows that the measured relative performance of systems appears to be reliable, but that recall is overestimated: it is likely that many relevant documents have not been found. We propose a new pooling strategy that can significantly increase the number of relevant documents found for given effort, without compromising fairness.

1 Introduction

The assumptions underlying research in information retrieval are straightforward. A user poses a query that represents an information need; the retrieval system uses a matching algorithm to identify documents that are likely to satisfy this need; and the user reads the returned documents to find answers to the query. Based on these assumptions, an information retrieval system can be measured with respect to a test collection, which is comprised of a set of documents, a set of queries, and relevance information about each document with respect to each query. The system resolves each query and is scored according to its ability to fetch the relevant documents.

It is well known that the reliability of measurement of a system depends on the quality of the relevance judgements, and that relevance assessors are rarely in exact agreement [4, 5, 7, 11, 14, 15]. Such "human factors" problems can introduce error into information retrieval experiments, but, assuming the assessment is sufficiently careful (that the assessor, for example, has not simply checked whether the query terms occur in each document), they should not in the general case introduce bias into measurement of the relative performance of systems. Similarly, given assessments there are many techniques for assigning a score to a system [8, 9, 10, 12], which can be based on theoretical considerations or pragmatic assumptions concerning the purposes of a system.

Other aspects of the experimental methodology are, however, also open to question. In particular, both the method for choosing which documents to assess and the method for measuring system performance can be unreliable. Until this decade most information retrieval experiments used fairly small databases of a few thousand records only, which, it could be argued, can trivialise the retrieval problem and are not always rich enough to distinguish between retrieval methods of different power. However, their size did allow complete relevance judgements to be formed. The 1990s has seen widespread use of much larger experimental databases, in particular the TREC collection [3], which provides a more realistic test environment but prohibits comprehensive relevance assessment. With such collections it is necessary to use techniques such as pooling to identify documents to be considered for relevance assessment, but it is possible for pooling to introduce bias. For example, use of a fixed-depth pool can potentially favour the numerical performance of a "new" system that is a simple combination of two other successful methods, and, if recall is measured, can discriminate against a method that is based on novel principles. Another problem is that unreliable estimates of recall can bias results.

With regard to measurement of system performance, absolute figures only concern effectiveness for particular queries on a particular data set. Of more interest are relative measures of whether one system is better than another, or whether modifying a similarity measure leads to improved performance. Relative performance can be established by comparing absolute effectiveness over a test collection, but such comparison is not meaningful unless the figures include some form of confidence or error bar [10]. An alternative to an error bar is to use a statistical test to evaluate confidence in the hypothesis that any difference in performance is significant. There are several such significance tests, some of which have been applied to information retrieval experiments.

In this paper we use the data generated in the TREC experiments to investigate these issues. Our aim is to determine the degree to which the use of pooling produces reliable results, and whether the relative measures of system performance are fair. We first investigate significance, primarily so that we can be confident that our later comparisons show valid variation, but we do show that ensuring that results are significant can be more important than choice of measure of effectiveness.

We then investigate pooling, and show that the TREC results are indeed reasonably reliable: pooling bias does not appear to have a significant impact and the available relevance judgements provide a fair basis for measurement of new systems. However, we show that it is likely that at best 50%–70% of the relevant documents have been discovered, in particular because of the queries that have large numbers of answers; and we show that the measurement strategy of assuming unjudged documents to be irrelevant is questionable.

We also show that it is possible to obtain useful estimates of the likely numbers of new relevant documents that can be discovered for each query if pool depth is increased. These results suggest a variation on standard pooling strategies that can increase the number of relevant documents discovered for given judgement effort, without introducing bias.

2 Test Data

For the results in this paper we have used the data generated by the TREC project [3], managed by NIST. In TREC, each participating group is given the same data and queries, and returns to NIST their *runs*, a listing of the identifiers of the top-ranked documents for each query. Each run contains up to 1000 identifiers. For each query, the top 100 identifiers from each run are *pooled*, that is, merged to eliminate duplicates and to remove any association between document and retrieval method. The number of identifiers taken from each run is the *pool depth*. The documents in each pool are then manually assessed for relevance; unjudged documents are assumed to be not relevant. Each year NIST then compiles all of this data: documents, queries, every run, and every relevance assessment. We primarily use information from TREC 5 (held in 1996, using 61 runs) and TREC 3 (1994, using 32 runs), and also consider TREC 4 (1995, using 33 runs).

The TREC relevance judgements include documents identified by a variety of experiments, not just the main runs. In our analysis we have considered the main runs only, and have been careful to prune the judgements to documents identified by these runs.

3 Standards for Significance

Much of the research in information retrieval is concerned with measurement of retrieval systems: examining performance on test collections and investigating the effect of changes to retrieval techniques. A question that has often arisen is how a system should be measured. Here we are concerned with a variant of that question: given a measurement technique and measurements of two systems A and B (where, say, A's measured effectiveness is greater than B's), how to decide whether the difference in the measurements is significant.

In the context of information retrieval, significance concerns whether (to a certain likelihood) the difference in the mean performance of the two systems for a given set of queries—a sample of a large underlying query population—is likely to represent a difference in the mean performance of the systems on the population as a whole. For information retrieval experiments we are interested in paired or correlated tests. Two well-known paired tests are "Student's" t-test and Wilcoxon's signed rank test; and analysis of variance, or ANOVA, can also be used for this task. Choice of test relies on somewhat informal principles [6]; in this context the main criteria are whether the sample set is sufficiently large, and, if not sufficiently large, whether the distribution of per-query results of A and B (or the per-query differences between A and B) is likely to be normal. Wilcoxon's test is said to have more power in the absence of a normal distribution; otherwise the t-test or ANOVA are likely to be more discriminating.

Tests for significance are to a confidence level, typically 95%, so that a report that the difference between A and B is significant has a 5% probability of being a false positive. Thus, if a significance test is reliable, then in 95% of choices of A and B, and for sufficiently large further query sets (that is, further samples from the population of queries), the performance of A will exceed that of B other than for a tiny proportion of cases that are flukes; in the remaining 5% the performance of A will exceed that of B approximately half the time.[1] On this basis it is possible to validate significance tests empirically.

To investigate these issues we have used the t-test, ANOVA, and Wilcoxon's test to examine all 1830 distinct A–B pairs on TREC 5, for query sets 251–275 and 276–300. We first used Lilliefors test for normality on the per-query results and per-query differences; for 11-point average (denoted 11pt), precision at 10 documents retrieved (p@10), and precision at 100 documents retrieved (p@100) results on a sample of systems we found that the results were unlikely to be normally distributed, suggesting that Wilcoxon's test is the more appropriate. However, the sample size of 25 is close to the lower bound of 30 suggested in texts as "sufficiently large".

Per-query results are highly correlated between systems, in typical cases giving a Pearson score of close to 1, because some queries are easier to resolve or have more answers than others; this correlation can affect assessment of significance. To address this problem we also considered normalised 11pt (denoted n11pt) results, where for each query the score of each system was divided by the score of the highest score obtained by any system for that query. This transformation eliminates the "ease of query" correlation, and moreover yields results that are apparently normal. (More sophisticated tests for normality, such as those used by Savoy [10], might reveal otherwise, but whether these values are normal is incidental to our aims and we have not investigated the properties of n11pt further.)

Across the total of 7320 A–B comparisons, comprised of 1830 on each of four methods for measuring performance, results were as follows.

- According to the t-test there were 3810 instances of significant difference. This result is consistent in form with that of Tague-Sutcliffe and Blustein [13]; it implies that the systems form about 12 non-disjoint groups where the members of each group have not been shown to have different means.
- ANOVA and t-test results were remarkably consistent. There were no cases of ANOVA identifying significance when the t-test did not, and only 4 cases (less than 0.1%) of t-test significance not confirmed by ANOVA.
- Wilcoxon's test was not consistent with the others. In 14 cases (0.2%) Wilcoxon's test failed to find significance when the others detected it. In 724 cases (9.9%) Wilcoxon's test found significance when the others did not, and was similarly inconsistent for 8.9% of the comparisons based on the normally distributed n11pt results.

We then investigated whether each significance result determined for one query set and pair of systems was validated on the other query set by a difference in means of the appropriate sign. Results were as follows.

[1] It is because of this uncertainty that researchers are often advised against re-sampling their data to explore for significance—if sampled sufficiently often a false positive or false negative will eventually arise. However, such sampling does allow exploration of the properties of the significance tests themselves.

Note that these tests are one-sided; failure does not mean that there is 95% probability (or any particular probability) of the means of the underlying populations being equal.

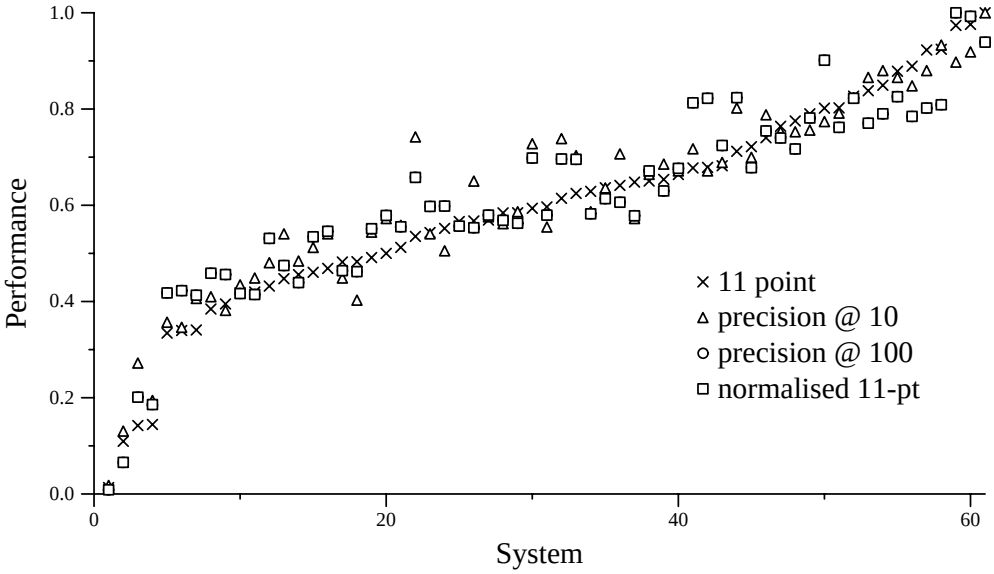

Figure 1: *Comparison of four measures of effectiveness. Each of the 61 columns represents one system in TREC 5; each mark in each column represents a different measure of effectiveness. For each measure the scores have been scaled so that the best score is 1.0, thus allowing direct comparison of the measures.*

- For ANOVA and the t-test, 97%–98% of significance results were confirmed. Perhaps surprisingly, a positive significance result for (say) 11pt results on queries 251–275 was almost as likely to confirmed on queries 276–300 by n11pt, p@10, or p@100 results as by 11pt results. We conclude that the type of measurement is relatively unimportant, so long as the difference between the systems is significant. (This is despite the fact that the ordering of systems by performance according to these measures is not consistent, as illustrated in Figure 1.) However, by a small margin p@10 and p@100 tended to give worse confirmation than the other two measures, which were indistinguishable.
- For Wilcoxon's test, 94%–98% of results were confirmed. For 11pt and n11pt around 97% of results were confirmed, close to the expected figure of "just under 97.5%" discussed above. We conclude that, given its reliability and greater power, the Wilcoxon test should be used for determining significance.

As noted by Savoy [10], a mistake that is sometimes made in interpretation of retrieval results is to assume that if two systems have similar average effectiveness then the difference in effectiveness is not significant, and that a difference of more than a few percent is probably significant. That this supposition is false is confirmed by the experiments above; for example, in one case a difference in 11-point effectiveness of 0.002 (from 0.149 to 0.151) is significant, presumably because in virtually every query the first system has slightly outperformed the second; while in another case a difference of 0.118 (from 0.210 to 0.328) is not significant. Overall around 20% of differences that were smaller than 0.05 were significant. In the results described below there are instances of tiny differences being both significant and interesting.

This aspect of significance is important in the context of measurement of retrieval systems: often, such measurement is not used for the broad task of comparing two disparate approaches to retrieval, but to identify whether a particular technique under study leads to performance gains. Such gains are usually small increments; if the gains can be reliably identified then the study of information retrieval is advanced. While small gains do little for the ultimate goal of improving user satisfaction—no user would be aware of the difference between 43% and 47% recall-precision on a certain query—over the course of a research program a series of small gains can cumulate into large improvements. Thus a difference that is imperceptible to a user may well be significant, that is, a consequence of a valuable change in the underlying retrieval method.

4 Difficulties with Depth

For measurement of effectiveness on a large database, some mechanism is needed to limit the number of documents that must be judged for each query. Pooling is used because it is perceived to be fair: each system contributes the same number of documents for assessment. With enough systems and a sufficiently deep pool, there is some likelihood that most of the relevant documents will be found.

Even with pooling, however, the cost of manual judgements can be large—it is typically the dominant cost of an information retrieval experiment—mandating some compromises. In TREC the pool depth is 100, but the run depth (or *measurement depth*) is 1000; that is, all of the 1000 documents in each run are considered when measuring system performance. We believe that the rationale for this approach is as follows: systems that are good at recall may continue to fetch relevant documents at depths greater than 100, which are therefore not added to the pool, but most of these documents should have been brought into the pool by other systems. If this rationale is correct then overall measurements of system effectiveness would be little changed if the pool depth is increased, because few new relevant documents would be identified; thus the labour of making further judgements is unnecessary, while the greater measurement depth should give better discrimination between systems.

System reinforcement

A potential disadvantage of having measurement depth m exceed pool depth p is that similar systems can reinforce each other. Consider a pool of three systems, A, B, and C. Suppose A and B use similar retrieval mechanisms such that some of the documents retrieved by A at a depth between p and m are retrieved by B at depth less than p, and vice versa; and suppose C is based on different principles to A and B. Then the performance of C can be underestimated. That is, this methodology may misjudge the performance of novel retrieval techniques.[2] Conversely, the effectiveness of techniques such as combination of evidence may be overestimated if the techniques being combined all contributed to the pool.

It can be argued that these effects are likely to be unimportant, assuming that the pool is sufficiently deep. The argument is as follows: since the performance in the first 100 documents is the main determinant of score, a few more judged documents at greater depth can only make a slight difference. However, although this argument may be correct in the typical case, there are circumstances in which it is misleading: for queries where there are many relevant documents but only a few in the top 100 for each query; for queries with a large number of relevant documents; and for queries where the total pool size is large, because each system found different documents. As we show below, it is likely that at best 50%–70% of the relevant documents have been identified, so distortions can arise where only some techniques are fetching relevant documents that have not been judged. Moreover, as discussed above improvements in retrieval techniques often consist of a series of small increments, the evidence for which may be individually masked by such effects.

We investigated the existence of such effects as follows. We created a pool of depth 100, and measured each run to a depth of 100, recording 11-point recall-precision and precision at depth 100. We then repeated the measurements, but using a pool depth of 10. For many systems, the two measurements are consistent—the relative performance is unaltered.[3] However, for a few systems the impact is marked; for example, two systems that scored close to 0.20 with a pool of depth 100 have scored 0.25 and 0.30 respectively with a pool of depth 10. In part the latter system may have done well because other runs have identified relevant documents on its behalf, while the former system had no such reinforcement; and in part the latter system may be achieving its score through high precision, while the former used high recall.

Comparing a pool depth of 50 to a pool depth of 100, however, the inconsistencies are small. In a further experiment we built relevance judgements for a pool of depth 50, but for each system in turn allowed that system alone to contribute a further 50 documents, thus mimicking perfect reinforcement for that system. In each case we measured each system on both the fair pool of depth 50 and the biased pool to which one system had contributed twice. Overall, with a biased pool 11-point effectiveness decreased for the non-cheating systems by a mere 0.5%, and the cheating systems increased by 1%.[4] However, these increases are in general significant, since they represent a small increase in effectiveness for every query, not the usual mix of increases and decreases in typical system-to-system comparisons. The largest increase observed was 3.5%, and cheating tended to be more effective for systems that scored highly.

Another perspective on this issue is as follows. For each system we assumed a pool depth of 100, then computed recall-precision using measurement depths of 100 and (the TREC convention) 1000; we denote these measurements 11pt100 and 11pt1000 respectively. We then used Wilcoxon's test to compare every pair A–B of systems, over (for consistency with our earlier results) query sets 251-275 and 276-300. There were only seven instances, of approximately 4500, where A was significantly greater than B according to one measure and $A \leq B$ according to the other—in other words, confirmation was almost 100%.

Increasing measurement depth has another effect: the number of instances in which one system is shown to be significantly better than another increases, so that discrimination between systems is improved. However, the change does affect the relative ordering of systems, varying position in the system ranking by as much as six places. Moreover, some systems are fetching a great many more unjudged documents than others, and, given our estimates of recall below, there is a high likelihood that many of these documents are relevant.

Surprisingly, neglecting the two or three lowest-scoring systems, there is little correlation between score and the number of judged documents to depth 1000. There is, however, a correlation between the number of relevant documents fetched by one system only (to depth 100) and the number of unjudged documents fetched by the system between depths 101 and 1000. That is, systems that identify more new relevant documents than others also get less benefit from the other contributors to the pool, and measurement to depth 1000 of these systems is likely to underestimate performance.

Given that most of the unjudged relevant documents are answers to around only 10 of the queries, it is reasonable to consider omitting these queries from the system-to-system comparisions. This would have several effects: the number of instances of measured significance would fall; in practice the reliability of the measurements would increase, because of the elimination of a cause of uncertainty; and the possibility of measurement bias is introduced, because there is potentially a causal relationship between query type, number of answers, and type of system. Selection of queries according to the number of answers is dubious at best.

We conclude that the phenomenon of reinforcement is identifiable in practice and introduces small but generally unimportant distortions in relative performance for systems that have contributed to the pool. The practice of measurement depth exceeding pool depth improves measured discrimination between systems, but introduces further uncertainty into the results; on balance we believe that this practice is not justified.

System omission

Another potential disadvantage of pooling is that, if only a fraction of the relevant documents are identified, a technique that did not have an opportunity to contribute to the pool may have its effectiveness underestimated. Such

[2] One can speculate that a process such as that used by TREC may be self-reinforcing—by tuning on previous data researchers are, potentially, exploring only one or two families of effective techniques, neglecting other approaches because they do not score well on past data.

[3] But the measurements are not identical. Increasing the pool depth reduces recall-precision scores, since each system identifies a smaller proportion of the relevant documents.

[4] These are percentage changes, not absolute improvements in effectiveness. As 11-point recall-precision for these systems is typically in the range 0.10–0.30, these percentages represent changes in effectiveness of 0.0005–0.0030.

behaviour was observed by Zobel et al. [16] in experiments during the early years of the TREC project, when testing passage retrieval on the Federal Register subcollection and evaluating the results with the previous year's TREC data: for some queries the retrieval mechanism highly ranked documents that appeared to be relevant, but had not been fetched by any of the official TREC runs; and overall the mechanism fetched many unjudged documents. (The separate question of the number of unjudged relevant documents is considered in Sections 5 and 6.) However, the number of contributing systems in TREC has since increased, and the problem may not be important in practice.

To investigate this aspect of pooling, we selected a run, formed a pool using all runs, then removed from the pool those documents contributed only by the selected run. By comparing performance on the original pool and the modified pool we can measure the degree to which contributing to the pool improves perceived effectiveness. By repeating this experiment for each system, we can get an average improvement over all systems.

For TREC 5, a measurement depth of 100, and a pool depth of 100, the average improvement was a tiny 0.5% and the maximum (neglecting the poorest-performing systems) was 3.5%, thus suggesting that the effect is unimportant. However, over the 10 queries with the most answers the average improvement was 7%. For TREC 3, the effect was more acute, with an overall average improvement of 2.2% and a startling 19% for the 10 queries with the most answers. Not surprisingly, these problems are more serious with a small pool; using a depth of 10 there is a 2.3% overall improvement for TREC 5 and 14% overall for TREC 3.

These results show that use of adequate pool depth is essential. However, they confirm that the TREC methodology is in this respect reasonably reliable, particularly for queries with smaller numbers of answers, and that the existing judgements can be used to evaluate new retrieval methods. However, these fresh evaluations should consider the number of unjudged documents being fetched, and experimenters should be aware that performance is probably being underestimated.

5 Reliability of Recall

A particular failing of the pooled method for identifying relevant documents is that it is impossible to be sure that most of the relevant documents have been located. Thus pooling cannot be used for the (arguably less important case of) measuring systems designed to maximise recall. Moreover, if it is possible that many of the unjudged documents are relevant, then an existing set of judgements may seriously underestimate the performance of a method that is good at finding "difficult" relevant documents, as well as underestimate the performance of systems that did not contribute to the pool.

Extrapolation from pool depth

We estimate total recall in the context of a pool as follows. Consider two pools, of depths $p-1$ and p, constructed from a set of runs. The set of relevant documents from the second pool contains that from the first; the n relevant documents in the second but not in the first are new arrivals. We can plot n against p to observe the rate at which new relevant documents continue to appear as pool depth is increased, as in the irregular line in Figure 2. As can be seen, the rate of new arrivals (totalled over the 50 queries and 61 runs from TREC 5) does diminish as pool size is increased, but is still around 20 per depth as the limit of 100 is reached, when 5040 relevant documents have been observed.

A straightforward method for using this data to estimate the total number of relevant documents is to fit some curve and extrapolate it, then compute the total area under the curve. The function used was

$$n = Cp^s - 1$$

where C and s are constants; fitting this function to the data requires only a straightforward linear regression on $\log_e p$ and $\log_e(n+1)$, where the addition of 1 allows $n = 0$. For TREC 5 we find $C = 396.3$ and $s = -0.6304$ using depths 1–100, and $C = 382.5$ and $s = -0.6182$ using depths 1–50. In both cases the standard error in $\log_e C$ is approximately 0.065 and in s is approximately 0.017. These fits are plotted for depths 3–100 in Figure 2. As can be seen the fit is extremely good, and is also excellent for depths 1 and 2, omitted to allow larger scale on the vertical axis. Use of a limited depth of 50 also gives an excellent fit, as to a lesser extent does use of a depth of 20, thus showing that for the data we have the fit provides good prediction. (We subsequently tested other functions but did not obtain a better fit.)

The fit on depths 1–100 can be used to estimate the total number of relevant documents for queries 251–300, which is simply the total area under the curve, but this estimate is at best only accurate to within an order of magnitude. Of more interest are smaller extrapolations—to say pool depth of 200 or 500. At these depths the estimated totals of relevant documents are 6707 and 9358 respectively. That is, increasing pool depth to 200 would be expected to identify a further 35 relevant documents per query, an increase of 32%. Similar measurements on subsets of the query set show that around 85% of these further relevant documents are due to the 10 queries with the most answers in a pool of depth 100.

These numbers are likely to be fairly accurate, because the errors in the fit are small. Moreover, there is a good fit for depths 51–100 based on depths 1–50; the fit predicted 1296 new relevant documents (in the range 1104 to 1519, allowing plus or minus one standard error in each parameter), while 1350 actual new relevant documents were observed.

We observed a similarly good fit for TREC 3, which has 7278 known relevant documents derived from the 32 runs. This fit gives an estimate of 10,138 relevant documents by depth 200 and 15,347 relevant documents by depth 500, large increases on the number found at depth 100. The numbers for TREC 4 are 5487, 7472, and 10,909 respectively, again based on an excellent fit.

Extrapolation from system count

Increasing the number of systems contributing to the pool can be used in a similar way. Each system brings in new relevant documents, but the number brought in by each successive system should drop as, gradually, all relevant documents are identified. This method, however, may underestimate the total number of relevant documents, as some will not be brought into the top 100 of the ranking of any likely retrieval system.

A problem with this approach is to choose an ordering of the systems. The 61! possible orders have widely varying characteristics, with in some cases all the more successful systems early in the ordering and in other cases the more successful systems towards the end; in the worst case the

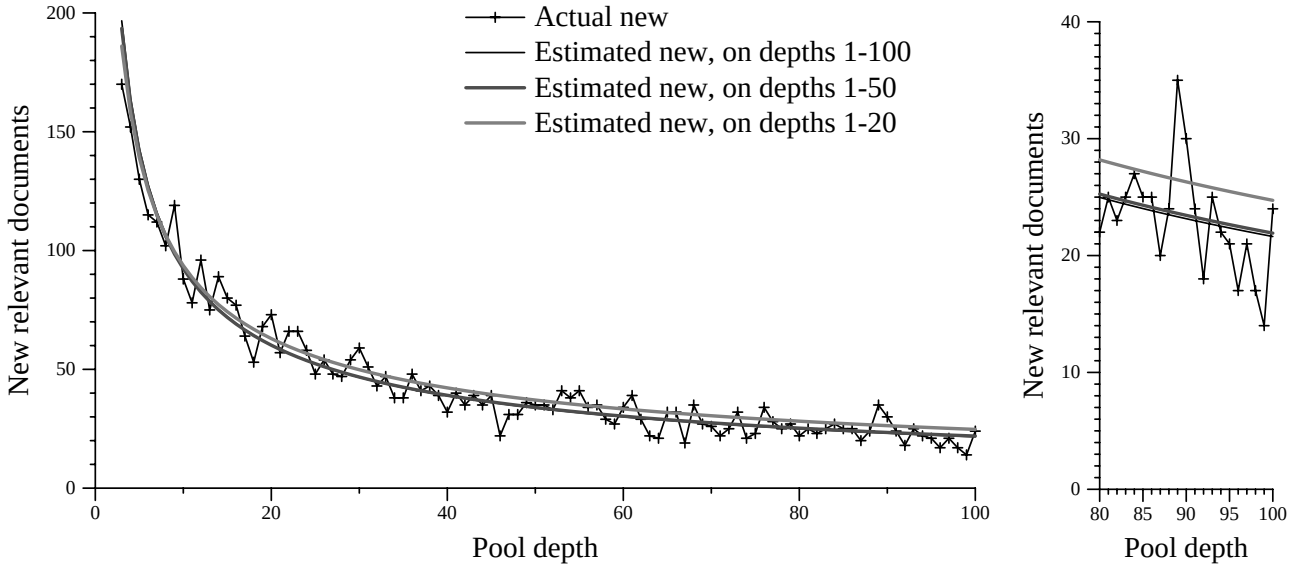

Figure 2: *Total number of new relevant documents at each pool depth, actual and estimated, for queries 251–300 from TREC 5. On left, depths 3-100. On right, depths 80-100, expanded to show detail.*

last system identifies over 200 new relevant documents [14]. To address this difficulty we generated a series of over 1000 random permutations of the list of systems and averaged the number of new documents introduced by the kth system over all the permutations. This average behaviour is plotted for TREC 5 in Figure 3, together with a curve fitted as in Figure 2; the standard errors of the parameters of this curve are 0.02 of $\log_e C$ and 0.007 of s. This curve predicts a total of 6054 relevant documents after 122 systems have contributed—that is, doubling the number of systems—or 7375 by 305 systems. The accuracy of this fit can be judged from extrapolation from 30 systems to 61, where the prediction for new relevant documents is 916, in the range 869 to 967; the actual number observed is 870. For TREC 3 the predictions are 8658 after 64 systems and 10323 after 160 systems, again with very small error. These results are consistent with those derived by increasing pool depth.

Overall, we conclude that the TREC experiments have been reasonably successful at identifying relevant documents, but that many relevant documents remain unidentified—only 5040 by the main runs and 5524 by all contributors for TREC 5, well short of our predictions. The assumption that unjudged documents are irrelevant is not well founded.

These results are confirmed by observed numbers of new relevant documents. As reported by Harman [2], subsequent to the main TREC 3 event pools were formed from depths 101-200 and assessed in a search for further relevant documents. This experiment found around 35 more relevant documents per query, or around 9000 relevant documents in total. Our results show that, in addition to the documents found to pool depth 200, many more relevant documents remain unidentified.

6 Principles for Pooling

Given that it is possible to obtain good estimates of the overall number of relevant documents, it is interesting to investigate whether similar estimates can be obtained for individual queries. (We consider only the case of increasing pool depth, and have not investigated increasing system number.) Not surprisingly, the fits obtained by per-query extrapolation are highly approximate.

Consider the example of using regression on depths 1–50 for some query to obtain a prediction for the number of new relevant documents at depths 51–60 for that query, on TREC 5. Noting that negative numbers of relevant documents can be predicted because the function $Cp^s - 1$ can take negative values, in a typical case (query 271, with 7 new relevant documents at these depths) our method predicts 2 more relevant documents, but allowing one standard error in each parameter gives the range -2 to 10. An extreme case is a prediction of 31 to 71 (query 269, 45 actual new relevant documents), while predictions of ranges such as -3 to 1 are common.

However, for the 7×50 cases of using depths 30 to 90 (counting in tens) to predict the next 10, there are only six instances of the predicted range not bounding the actual number of relevant documents. Thus, for example, when using extrapolation over depths 1–70 to predict the number of new relevant documents per query at depths 71–80, the prediction of -3 to 1 for query 260 underestimates the actual number, which is 2, but for the other 49 queries the number is within the estimated range. It follows that this extrapolation technique can be used to give a broad indication of the number of relevant documents that can be identified by deepening the pool for each query.

These predictions can be used to guide construction of variable-depth pools, in which the final depth used can vary from query to query. The rationale for constructing such pools is as follows. The aim of pooling should be to construct a set of judgements that is not biased towards any one system—that is, each system should be equally able to contribute to the pool—and to identify as many relevant documents as possible. In particular, if it is has become likely that for a certain query no more relevant documents will be identified, then continuing to judge documents for that query is a waste of resources. A potential objection to use of variable-depth pools is that documents must be judged in pool-depth order, conceivably suggesting to the assessors that later documents are from deeper in the pool and are thus less likely to be relevant. However, since pool depth is increased only if there is a reasonable likelihood of

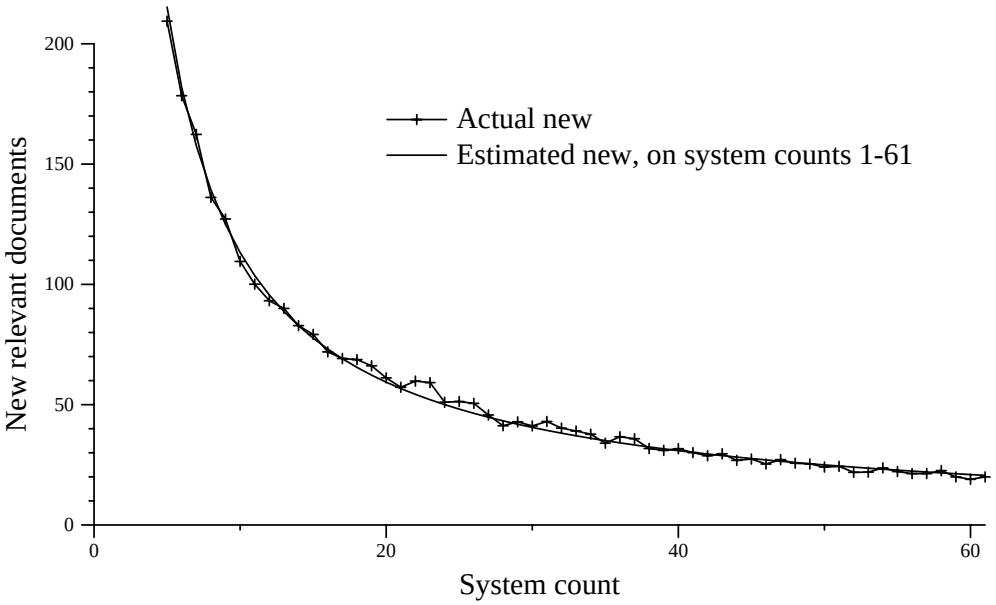

Figure 3: *Total number of new relevant documents at each system count, actual and estimated, for queries 251–300 from TREC 5.*

further relevant documents being found, and at each stage can be increased by a substantial increment, we believe that this problem should not be an obstacle in practice.

In this approach, pools can be constructed as follows.

1. Fully judge each query to some initial depth—say a pool formed from the top 30 documents in each run.
2. For each query, use extrapolation to predict the likely number of relevant documents to be found when the per-query pool depth is increased, say by 10. Compare this prediction to the number of new documents in the increased pool for each query, to produce a predicted cost of finding each new relevant document.
3. Then either
   - identify the most promising queries and make judgements on their extended pools; or
   - remove the least promising queries so that they are not considered for further judgements, and judge the pools of all the remaining queries.

   If assessors are available to make further judgements return to step 2, otherwise stop.

There are many possible variants of this scheme, based on decisions on how to assign costs to an increase in pool depths, which could conceivably involve factors such as length of documents to be assessed and the number of documents judged for that query so far. A Boolean alternative to our pooled approach is suggested by Blair [1], but this alternative does not quantify the likely numbers of relevant documents to be found.

As an instance of this process, consider increasing the pool depth from 50 to 60 in TREC 5. If increased uniformly for all queries, 7689 documents must be judged, of which 347 are relevant. Now suppose instead that, based on extrapolation from pool depths 1–50, the 25 most promising queries (by predicted number of answers) have their pool depth increased to 70. Then 7777 documents must be judged, of which 577 are relevant; thus 230 more relevant documents (an increase of 66%) can be found for the same total effort.

The effect of selectively increasing pool size is shown in detail for TREC 5 in Table 1. Each line shows the behaviour when all queries have been fully judged to some depth (for example, to depth 30 in the first line) and pool depth is to be increased by 10 (to 40 in the first line). If all queries are considered, in each case the pool size is between 7000 and 8000 and for depth 30 approximately 5% are relevant; this percentage drops as pool depth increases. If only the best 25 queries are considered, 8.9% are relevant, and only 49 relevant documents are not discovered; the 10 least promising queries contribute only a few percent of the relevant documents.

Moreover, for these queries with small numbers of answers most of the relevant documents are found by pool depth 30—indeed, the percentage found almost certainly exceeds the percentage of relevant documents found by depth 100 for the 10 queries with the most answers. It is clear that varying pool size for each query can dramatically increase the number of relevant documents that are found; and that, both overall and for the queries with the most answers, doing so should improve the reliability of the measured results.

7 Conclusions

We have had long-standing concerns about some aspects of the methodology used for large-scale information retrieval experiments such as TREC: whether the measured results are trustworthy; whether the use of a limited-depth pool significantly distorts results for "new" systems that did not contribute to the pool; and whether the pooling strategy does indeed discover most of the relevant documents. Despite commencing this work expecting to discover that the TREC strategies might be seriously flawed, we have found that overall they do indeed lead to reliable results.

In particular, our empirical investigation has shown that results based on the relevance judgements formed from a limited depth pool are reliable—if the pool is sufficiently deep—both for systems that contributed to the pool and for "new" systems. In this respect, the TREC limit of 100 appears to be adequate. These results answer misgiv-

Pool depth	All queries	10 best	25 best	10 worst
30 → 40	7892/420 (5.3%)	1890/266 (14.1%)	4265/371 (8.9%)	1496/12 (0.8%)
40 → 50	7688/343 (4.5%)	1730/221 (12.8%)	4096/304 (7.4%)	1483/11 (0.7%)
50 → 60	7689/347 (4.5%)	1861/233 (12.5%)	3940/316 (8.0%)	1424/ 5 (0.4%)
60 → 70	7544/282 (3.7%)	1676/196 (11.7%)	3837/261 (6.8%)	1513/ 3 (0.2%)
70 → 80	7518/259 (3.4%)	1768/196 (11.1%)	3788/239 (6.3%)	1386/11 (0.8%)
80 → 90	7327/259 (3.5%)	1671/173 (10.4%)	3861/230 (6.0%)	1391/12 (0.9%)
90 → 100	7349/203 (2.8%)	1739/150 (8.6%)	3861/180 (4.7%)	1422/ 3 (0.2%)

Table 1: *Numbers of retrieved documents and of relevant documents found at a range of pool depths for TREC 5, for different strategies for choosing queries whose pools are increased. Each pair* X/Y *is the number* X *of new documents in the pool and the number* Y *of new relevant documents in the pool, followed in parentheses by the percentage of new relevant documents in the pool of new documents.*

ings raised in this paper and elsewhere [15, 16]. We have also argued that significance results are trustworthy, and indeed that showing that the difference in performance between two methods is significant is probably of more value than precise choice of performance measure, particularly as there are many instances in the TREC experiments of large yet insignificant differences and of much smaller differences that are significant. Our results indicate that Wilcoxon's signed-rank test is reliable and, in contrast to ANOVA and "Student's" t-test, provides greater discrimination between systems.

However, we have identified some limitations of the pooling strategy. The practice of using the top 1000 documents to measure systems when only the top 100 have contributed to the pool allows greater discrimination between systems, but introduces uncertainty. Also, our estimates of the number of unjudged relevant documents show that it is likely that at best 50%–70% of the relevant documents have been found; most of these unjudged relevant documents are for the 10 or so queries that already have the most known answers. For this reason measures based on recall are highly uncertain.

These results have allowed us to propose a new method for pooling that increases the number of relevant documents found for given judgement effort. In this method the pool size of each query should be increased by small increments, say 10, and the pool judged only where there is reasonable likelihood of it containing relevant documents. Simple regression on the per-query number of new relevant documents found at each pool depth, although highly approximate, is a good basis for choice of queries for further judgement effort. Overall this technique allows many more relevant documents to be found for given effort, and should increase the reliability of measured results in large-scale information retrieval experiments.

Acknowledgements

I am grateful to Alistair Moffat for his help with this work. I am also grateful to Ross Wilkinson and Hugh Williams for their comments, and to Donna Harman and NIST for making the TREC data available. This work was supported by the Australian Research Council.

References

[1] D.C. Blair. STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1):4–22, 1996.

[2] D. Harman. Overview of the fourth text retrieval conference (TREC-4). In D. Harman, editor, *Proc. Text Retrieval Conference (TREC)*, October 1995.

[3] D. Harman. Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, 31(3):271–289, May 1995.

[4] S.P. Harter. The Cranfield II relevance assessments: A critical evaluation. *Library Quarterly*, 41:229–243, 1971.

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

[6] C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach, second edition*. Open Court, Chicago Illinois, 1993.

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

[8] V.V. Raghavan, G.S. Jung, and P. Bollman. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, 1989.

[9] G. Salton. The state of retrieval system evaluation. *Information Processing & Management*, 28(4):441–449, 1992.

[10] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4):495–512, 1997.

[11] D.R. Swanson. Some unexplained aspects of the Cranfield tests of indexing performance factors. *Library Quarterly*, 41:223–228, 1971.

[12] J. Tague-Sutcliffe. The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, 28(4):467–490, 1992.

[13] J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the TREC-3 data. In D. Harman, editor, *Proc. Text Retrieval Conference (TREC)*, pages 385–398, 1994.

[14] E. Voorhees and D. Harman. Overview of the fifth text retrieval conference (TREC-5). In E. Voorhees and D. Harman, editors, *Proc. Text Retrieval Conference (TREC)*, November 1996.

[15] P. Wallis and J.A. Thom. Relevance judgements for assessing recall. *Information Processing & Management*, 32(3):273–286, 1996.

[16] J. Zobel, A. Moffat, R. Wilkinson, and R. Sacks-Davis. Efficient retrieval of partial documents. *Information Processing & Management*, 31(3):361–377, 1995.