Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability

Mark Sanderson
Department of Information Studies,
University of Sheffield, Sheffield, UK
m.sanderson@shef.ac.uk

Justin Zobel
School of Computer Science and Information
Technology, RMIT, Melbourne, Australia
jz@cs.rmit.edu.au

ABSTRACT

The effectiveness of information retrieval systems is measured by comparing performance on a common set of queries and documents. Significance tests are often used to evaluate the reliability of such comparisons. Previous work has examined such tests, but produced results with limited application. Other work established an alternative benchmark for significance, but the resulting test was too stringent. In this paper, we revisit the question of how such tests should be used. We find that the t-test is highly reliable (more so than the sign or Wilcoxon test), and is far more reliable than simply showing a large percentage difference in effectiveness measures between IR systems. Our results show that past empirical work on significance tests over-estimated the error of such tests. We also re-consider comparisons between the reliability of precision at rank 10 and mean average precision, arguing that past comparisons did not consider the assessor effort required to compute such measures. This investigation shows that assessor effort would be better spent building test collections with more topics, each assessed in less detail.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Systems and Software --- performance evaluation.

General Terms

Experimentation, Measurement.

Keywords

Significance tests, mean average precision, precision at 10.

1. INTRODUCTION

Test collections are the principal tool used for comparison and evaluation of retrieval systems. These collections – typically comprised of documents, queries (or topics), and relevance judgments – have been a key part of information retrieval (IR) research for decades; the use of such collections is based on research and practice in collection formation (Spärck Jones & Van Rijsbergen, 1975; Voorhees & Harman, 1999) and measurement of retrieval effectiveness (Van Rijsbergen 1979, Ch. 7; Dunlop, 1997; Järvelin, 2000; Buckley, 2004). Effectiveness is computed by measuring the ability of systems to find relevant documents. The measured score is most often used as an indicator of the performance of one system relative to another; with an assumption that similar relative performance will be observed on other test collections and in operational settings.

When researchers report results of a retrieval experiment and show, using some effectiveness measure, that one retrieval system is better than another, significance tests are often used to evaluate the result. The tests provide evidence that the observed difference in effectiveness is not due to chance. Significance tests such as the t-test or Wilcoxon are commonly used. The question of how likely it is that a significant result observed on a test collection will continue to be observed in other settings has not been as widely investigated as have other aspects of test collections.

Significance tests require that the data being tested has certain properties. Among the assumptions of the Wilcoxon signed-rank test and the t-test are that the values being tested – in this case, per-query effectiveness – are distributed, respectively, symmetrically and normally (Van Rijsbergen, 1979, Ch. 7); however, effectiveness rarely follows either distribution. The tests also assume that each set of per-query values being tested is a random sample from a broader population. What the tests determine, in comparing the *runs* of two retrieval systems, is whether the two samples are from the same population of effectiveness outcomes (the systems are equivalent) or different populations (one system gets better results than the other).

It is in this context, where significance tests are widely used in IR experiments but their impact is little understood, that we undertook the work reported in this paper. Our results suggest new procedures for evaluation of retrieval systems and show that both a relative improvement in measured effectiveness and statistical significance are required for confidence in results. If significance is omitted or the improvement is small – as is the case in many SIGIR papers – results are not reliable. We also find that assessor effort, currently expended evaluating few topics in detail, maybe better spent examining more topics in less detail.

After reviewing past work (Section 2), an existing methodology for computing the consistency of an evaluation measure is extended to assess the impact of using statistical significance when assessing IR systems (Section 3). The measures Mean Average Precision (MAP) and Precision measured at rank 10 (P@10) are compared with a view to the assessor effort required to compute such measures (Section 4). The reliability of effectiveness measures are briefly examined on test collections with large numbers of topics (Section 5), Next, a previously unreported statistical effect in the methodology is shown; means of alleviating the effect are described and its substantial impact on past and current results presented (Sections 6 and 7), before conclusions and future work are outlined (Section 8).

2. Previous work

There is limited discussion of significance tests in IR literature. Van Rijsbergen (1979) detailed the shape and form of

SIGIR '05, August 15–19, 2005, Salvador, Brazil.

data distributions to which the sign, Wilcoxon and t-tests can be applied. He showed that test collection data fails to meet all the requirements of any of the tests and warned that none can be used with confidence. Countering such caution, Hull (1993) described past (non-IR) work showing that the t-test can be reliable even when data being tested is not distributed normally. Hull described a range of significance tests, but did not empirically test them.

Savoy (1997) investigated several significance measures and echoed Van Rijsbergen's concerns. He proposed an alternative bootstrap method, based on sampling from a set of query outcomes; it is not clear whether this approach could be applied with the small sets of queries for which we have relevance judgments and to our knowledge it has not been used in practice to assess significance.

The Wilcoxon and t- are common significance tests used in IR experiments. Both take a pair of equal-sized sets of per-query effectiveness values, and assign a confidence value to the *null hypothesis*: that the values are drawn from the same population. If confidence in the hypothesis (reported as a p-value) is ≤ 0.05 ($\leq 5\%$), it is typically rejected. Although such tests only consider the null hypothesis, it is common to assume rejection implies that values are from different populations with likelihood >95%.

Apart from correctly determining significance or a lack thereof, the tests also produce *type I* and *type II* errors. A type I error is a false positive; for a p-value of 0.05 (5%), one positive test in twenty is expected to be a type I error. A type II error is a false negative; the incidence of type II errors is unknown. In the statistical and medical communities, there has been concern that the theory underestimates the rate of type I error, but that the small samples used in typical studies mean that there is insufficient data to determine significance and thus that type II errors may be common, leading to useful methods being discarded prematurely (Matthews 2003).

2.1 Zobel

The first IR-based work to measure the utility of significance tests was that of Zobel (1998), who split the fifty topics of TREC-5 into two disjoint sets of 25: one set holding topics 251-275, the other topics 276-300. Taking the 61 runs submitted to TREC-5, Zobel compared each run with every other, resulting in 1,830 pair-wise comparisons. If a significant difference was observed between a pair of runs measured on the first 25 topics (that is, if the null hypothesis was rejected, $p \leq 0.05$), the ordering of the runs based on an effectiveness measure was noted and the same pair of runs was compared on the second 25 topics. If the ordering of runs on both sets was the same, the significance test was judged to be correct. If the ordering was different, a type I error was recorded. Zobel examined the 1,830 pairs under four effectiveness measures, including eleven-point average precision and P@10, resulting in a total of 7,320 comparisons.

The significance tests assessed were ANOVA, Wilcoxon and the t-test. Zobel found all three to be accurate at predicting system ordering in the second set of topics: depending on the effectiveness measure used, between 97%-98% for the t-test and ANOVA, and 94%-98% for Wilcoxon. Significance via the t-test was observed in 3,810 pairs; in all but 4 of these pairs ANOVA also found significance. Significance via the Wilcoxon test was not observed in 14 of the 3,810 pairs found by t-test, but significance was observed in an additional 724 pairs.

It is to be expected that significance where $0.04 < p \leq 0.05$ will have more type I errors than significance where $0.01 < p \leq 0.05$, and even more if significance of $p \leq 0.01$ is observed. However, Zobel concatenated all such observations into a single measure of type I error for each significance test with no indication of the distribution across p values. This lack of detail in Zobel's results is addressed in the work presented here.

2.2 Voorhees and Buckley

Expanding on Zobel's topic-partitioning methodology, Voorhees & Buckley (2002) examined a simple form of significance: measuring the absolute difference in MAP between two systems. Their aim was to determine the size of difference observed for the first set of topics before it was possible to be confident that system ordering would be preserved in the second topic set. The runs used were those submitted to the ad hoc track of TRECs 3-10. The total number of runs was 476;[1] the number of pair-wise comparisons made was 16,678 (comparisons were restricted to those pairs of runs submitted to the same year of TREC). Voorhees & Buckley randomly split the 50 topics in each TREC year into two disjoint sets of 25 and computed errors rates for 20 bins of MAP differences, 0%-1%, 1%-2%, up to 19%-20%. The whole procedure was repeated 50 times to ensure that any random variation in topic selection was smoothed out.

They found that an absolute difference of between 8%-9% was required in the first set of 25 topics before the chance of the both sets of topics having the same run ordering was over 95% (that is, the error rate was less than 5%). Because the runs for each year of TREC covered only 50 topics, a 25/25 split was the largest that could be measured. Plotting the error rates by topic set size showed that as size increased, error rates reduced with a clear exponential trend. Voorhees & Buckley projected the trend lines forward to topic sizes of 50 and concluded that an absolute difference in MAP of 5%-6% would be needed between two runs measured on 50 topics before one could be 95% confident that the ordering measured on the topic set would also occur on a different set of 50 topics.

The result of Voorhees & Buckley made collections such as TREC appear less useful than they were perhaps thought to be, as few IR experiments comparing retrieval runs produce an absolute difference in mean average precision as large as 5%. However, they did not study one aspect of measuring the effectiveness of runs, namely the impact of significance tests on error rates. The use of such tests may reduce the difference in MAP required before experimenters could be confident that their result will hold when tested on other topic sets. The starting point of our work was a re-examination of these results, as we now describe.

3. Error rates reconsidered

Our first experiment was a re-run of the Voorhees & Buckley experiment with two further TREC runs added: 2 and 11. For consistency with the methodology of the earlier work, the bottom 25% of runs (as ranked by their MAP) were eliminated from the comparisons. Across the ten TRECs, 555 runs were pair-wise compared, producing 18,460 comparisons. As with Voorhees & Buckley, each comparison (of run ordering across two randomly selected topic sets) was repeated 50 times. Figure 1 shows the error rates computed for each of the bins of absolute difference in MAP (note only topics set sizes from 5 to 25 were computed). As can be seen, the larger the difference in MAP observed in the first

[1] The bottom 25% of runs submitted to each TREC were omitted due to a concern that the poorer runs might skew results.

topic set, the more likely the run ordering in the second set matches that found in the first (that is, the error rate is reduced). As with Voorhees & Buckley, to overcome the limitation of only being able to calculate error rates for sets of 25 topics, trend lines were projected[2] from the data to topics set sizes of 50, as shown in Figure 1. (As can be seen the trend lines appear to fit the data well, however, as will be seen later in the paper, trend lines provide limited accuracy.) According to the trend, when measuring the difference between two runs on TREC data, if an absolute difference of more than 5% is observed, one can be 95% confident that the ordering of the two runs will be preserved for different topic sets.

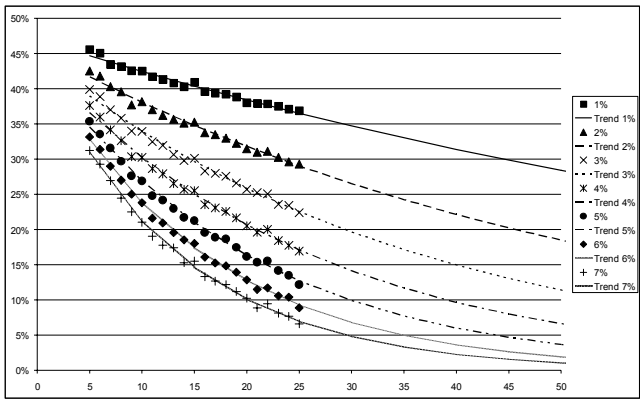


Figure 1: Projected trends of error rates for absolute differences in MAP up to topics of size 50.

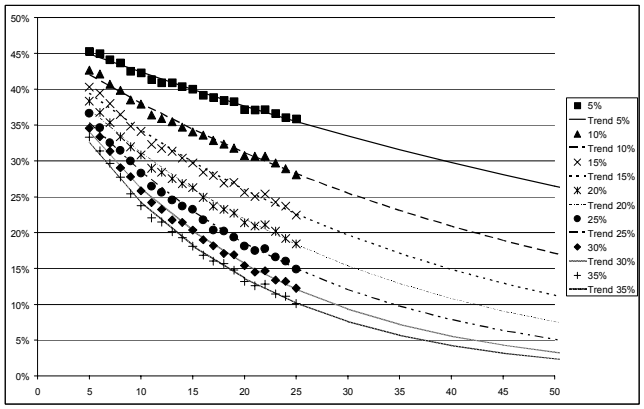


Figure 2: Error rates for relative differences in MAP with trends projected to topics of size 50.

The figures obtained in this experiment were close to those reported by Voorhees & Buckley, giving us confidence that our software was correctly implementing their methodology.

Voorhees & Buckley chose to focus on absolute differences, to test a well known rule of thumb in evaluation from Spärck Jones (1974), who stated that an absolute difference of less 5% between two runs was not meaningful. Most reported research, however uses relative percentage differences when computing the magnitude of difference between two runs. Therefore, we re-ran the experiments to compute the error rates for relative differences in MAP. As such differences tend to range more widely than absolute differences, the bins used were also wider: 0%-5%, 5%-10%, up to 95%-100%; higher relative differences were not considered. Not all bins were graphed, either to remove clutter or because insufficient data was available. The data generated from the repeated experiment, along with projection lines, is shown in Figure 2. As can be seen, according to the projection a relative difference of 25% must be observed to give confidence that the result for the first topic set is significant. Such a large difference, although sometimes observed, is much greater than differences measured in most retrieval experiments. The implication being that few observed differences between IR systems are significant.

3.1 Considering significance

We then evaluated the impact of introducing a significance test, re-running the experiment but only considering pairs of runs with statistically significant differences. The first test used was the t-test, with $0.01<p\leq0.05$ (that is, confidence in the null hypothesis was 1%-5%). Figure 3 shows the impact of significance. The relative percentage difference in effectiveness required in order to obtain significant results when measuring runs on 50 topics was projected to be less than 10%; substantially lower than the 25% required when no significance test was applied. The experiment was repeated, but only those comparisons with significance at level $0.04<p\leq0.05$ were considered (see Figure 4). Here it can be seen that the required relative difference increases, but is still much less than the 25% required without any significance test.

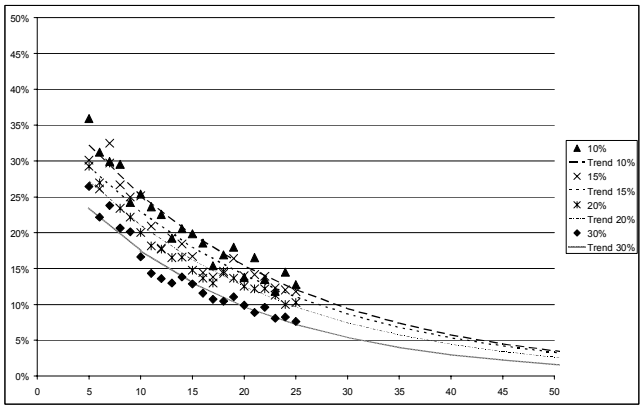


Figure 3: Error rates for those relative differences in MAP for which a t-test produced significance $0.01<p\leq0.05$.

Note that, because only pairs of runs where significance was observed were used, the quantity of data was greatly reduced. Consequently, the data points in Figure 3 and Figure 4 were more erratic and the accuracy of fit of the projections was not as high as in the previous figures. Insufficient data was available to plot a projection line for relative differences in the 0%-5% and 5%-10% bins. Note also that all the computed error rates (for topics 5-25) in Figure 3 and Figure 4 are well above the 5% error rate line. Note also, for small topic sets, even with statistical significance between two runs in the first topic set, there is no guarantee that the ordering will be preserved for other sets of topics.

Other forms of significance were also examined. For size 25 topics sets across a range of relative differences (i.e. 10%-30%) the tests – sign and Wilcoxon measured at $0.01<p\leq0.05$ – were not as accurate as the t-test: producing respectively 17%-8% and 16%-10% error rates compared to 13%-7% error for the t-test. Zobel (1998) also observed Wilcoxon producing more type I

[2] Using the "GROWTH" function in MS Excel, which computes an exponential growth from existing data.

errors than the t-test. Even though use of sign or Wilcoxon violates fewer assumptions on the nature of the data being tested, the t-test appears to be more reliable.

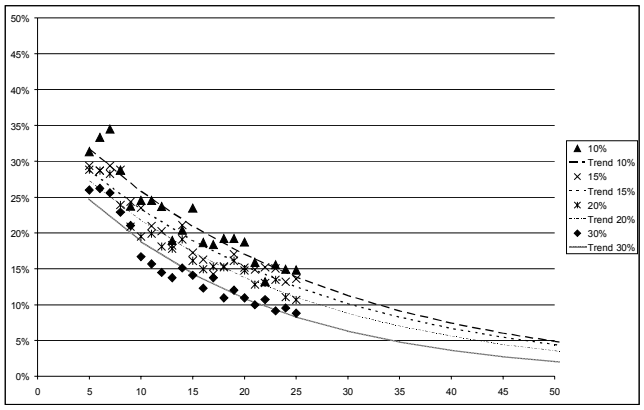


Figure 4: Same as Figure 3 but for $0.04<p\leq0.05$.

3.2 Type I and type II errors

Another perspective on significance tests is given by consideration of the full sets of 50 queries. On the sets of runs described above, we used all 50 queries to compare every pair of runs as above, to count the proportion of runs where a significant test returned a positive result. We then binned these counts by the relative difference in MAP between the runs.

For the t-test, at a relative difference in MAP of up to 10%, only 1.3% of comparisons were significant for $p\leq0.05$. For differences of up to 30%, around 30.0% of comparisons were significant, while for differences in the band 20%-30% we observed that 57.3% of comparisons are significant. For the sign test and differences in MAP of up to 10%, the proportion of comparisons judged significant was 8.6%, a dramatic difference. For higher differences in MAP, results were similar to the t-test. The Wilcoxon test was less ready to pronounce small differences as significant, starting at 6.6% of comparisons where the relative difference in MAP was less than 10%, but reported that 78.0% of differences in the band 20%-30% were significant.

Clearly these tests are not interchangeable. Given the broad agreement between these results and those discussed above, we again conclude – tentatively, at least – that the t-test is more reliable than the alternatives, based on the expected false-positive rate of 5%. (However, the results also suggest that 5% may be an overestimate, possibly due to the strong correlation between per-query scores produced by different systems.) Both sets of results indicate that, assuming a false-positive rate of 5%, a positive t-test with a relative MAP difference of less than 10% is arguably more likely to be a type I error than a significant result.

This conclusion is consistent with the observation that, in practice, only a minimal alteration to the rankings produced by retrieval systems – such as exchange in order of a relevant and a non-relevant document in each of a fraction of the queries – is sufficient to produce a small relative MAP difference. That is, such improvements would be imperceptible to a user.

This analysis provides an upper bound for the proportion of type II errors. For example, for relative differences in MAP in the range 40%-50%, we observed that 87.7% of t-tests are positive, rising to 96.2% for MAP differences in the range 90%-100%. Only the residue – 12.3% and 3.8% respectively – are candidate type II errors. These proportions rise to 96.1% and 100.0% respectively when runs with a low MAP score (<0.1%) are discarded, and to 100% in both cases when runs with MAP <0.2% are discarded. For comparisons of runs with reasonable levels of effectiveness, we conclude that type II errors are rare.

3.3 Examining past SIGIR results

In view of these observations, some recent reported results were almost certainly invalid. Only about half of comparisons with a MAP difference of around 20% are significant, yet in recent SIGIR papers many authors claim such differences as a valid result while failing to undertake a significance test. We contend that results based on MAP differences of this order are meaningless without a significance test.

We examined a selection of papers presented at SIGIR in 2003 and 2004. We chose 26 papers that evaluated a well-defined retrieval task where MAP was a natural choice of effectiveness measure, and investigated what measures people actually used.

We found that significance was not explicitly reported in 14 of the papers. In two it was implied such tests had been tried, but outcomes were not given. In three or four of these papers, the improvements were large and arguably a significance test was unnecessary. However, in at least six papers (23% of the sample) the reported improvements were small, sometimes no more than a few percent in relative MAP. There is no reason to suppose that these are significant variations.

Some of these were questionable for other reasons. One paper reported large percentage improvements in MAP, but from a baseline of 0.03; it is unclear whether such results should be regarded as important. Another reported changes of reasonable magnitude, but on only 10 queries. In two papers no numbers were given, with all results presented graphically.

Among the 12 papers with significance tests, one used both ANOVA and the t-test, five each used either the t-test or Wilcoxon's test, and in one, the test was not identified. Most of these papers reported relative change in MAP, while a couple reported relative change in mean reciprocal rank (MRR) and P@10. Large differences in MRR and P@10 were in some cases not significant – further evidence that untested differences in these measures may not be a reliable indicator of performance. For many of the significant results, absolute MAP differences were much less than 5%, indicating that this test is too stringent.

In four papers, two using Wilcoxon and two using a t-test, authors reported small improvements in MAP (of less than 5%, and one as low as 1.6%) that were significant. On the admittedly slender basis of the summary information presented in the paper, it is our guess that these results are likely to be type I errors.

Overall, for rather less than half of the papers it was clear that the results were robust.

4. P@10 and assessor effort

The results presented so far used MAP to compute ordering of runs and statistical significance. Another common measure in IR research is P@10. Several past papers have compared these two measures. Tague-Sutcliffe & Blustein (1994) showed MAP discriminated between runs better than P@10. Buckley & Voorhees (2000) used the TREC query track, which has 21 different versions of each topic. Buckley & Voorhees compared the consistency of measures in ordering runs across the different versions of each topic. MAP was shown to be more consistent than P@10. Similar conclusions were drawn when Buckley & Voorhees (2004) examined measure stability when test collections were degraded by removing relevant documents. Again results

from their experiments show that P@10 was not as reliable as MAP. Given that our experimental setup allowed easy comparison of MAP and P@10, we undertook a fresh evaluation of the two measures.

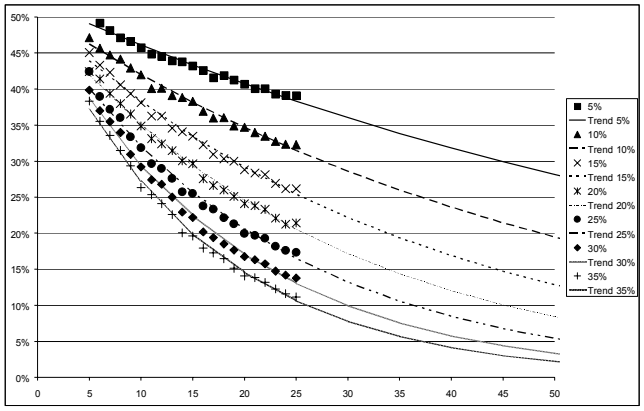


Figure 5: Error rates for relative differences in P@10 with trends projected to topics of size 50

Figure 5 shows the error rates for percentage differences between two runs against the topic set size. Compared with the graph for MAP in Figure 2, the projection out to topic sets of size 50 reveals that the two measures are almost identical: a 25%-30% relative difference is required to be confident in run ordering being preserved. Note, however, that the data points from topics sets 5-25 show P@10 with an error rate a few percent higher than the equivalent point on the MAP graph. However, the decrease in error rate for larger topic sets is greater for P@10 than it is for MAP, which results in the projection for 50 topics being very similar. Figure 6 shows error rates for P@10 for those comparisons for which a t-test at level $0.01<p\leq0.05$ is observed. In comparison with Figure 3, it can be seen that significance reduces the magnitude of relative difference required for P@10 but does not reduce it as much as for MAP.

In agreement with previous work, therefore, we conclude that MAP is a more reliable measure than P@10. We believe the simple reason for this is that MAP takes into account the location in a ranking of all known relevant documents, whereas P@10 is influenced by at most 10. We contend that the more relevant documents (that is, data points) an effectiveness measure uses, the more accurate that measure has to the potential to be.

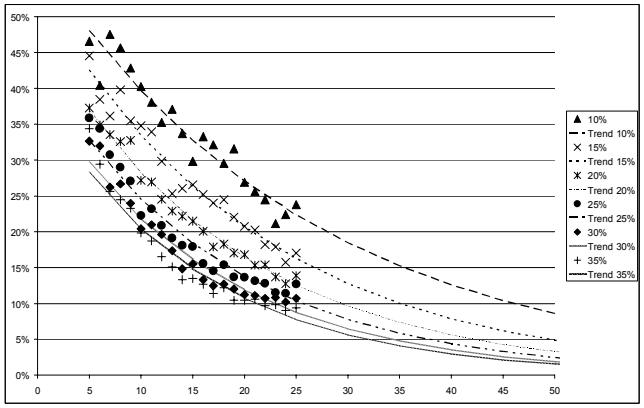
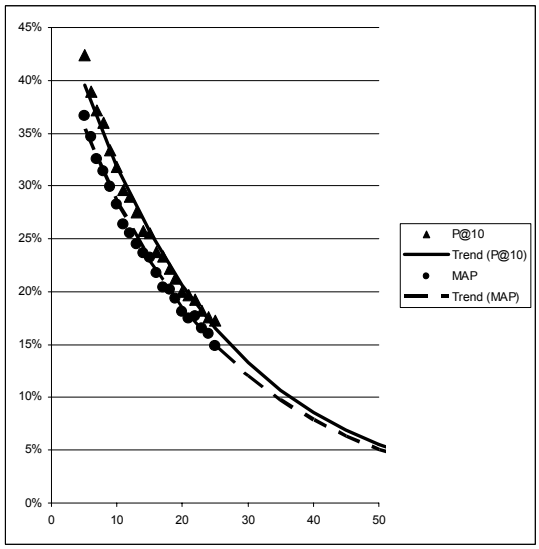


Figure 6: Error rates for those relative differences in P@10 for which a t-test produced significance $0.01<p\leq0.05$.



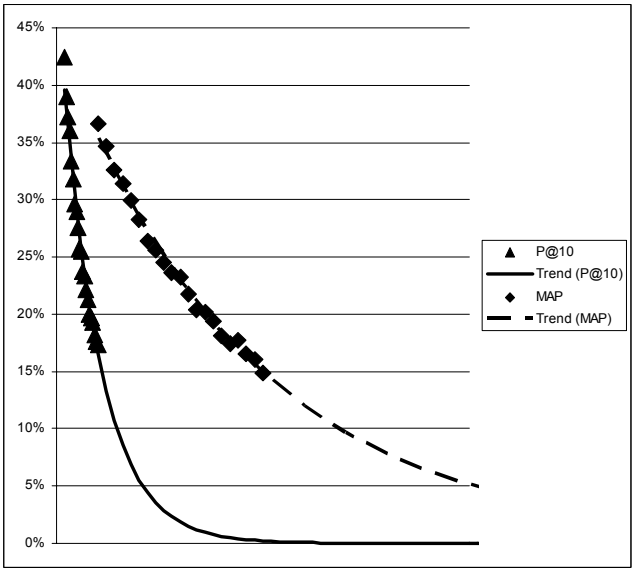


Figure 7: Top graph shows error rates for 25% relative difference in P@10 and MAP plotted against topic size; bottom graph shows the same rates and measures normalized by assessor effort. Note scale on x-axis removed.

There is, however, one aspect to the comparison of MAP and P@10 that has not previously been considered. In comparisons of the two measures in past work and in the experiments reported above, it was assumed that the test collection on which the experiments were conducted was already created; all relevance judgments were made. However, if a new collection is to be created, the distinction between the measures is not as clear. MAP as measured here uses all relevant documents identified in TREC, which were taken from a pool of documents formed from the top 100 documents ranked by each submitted run for each topic; P@10, however, only needs assessment of a pool formed from the top 10 documents of each run for each topic. Across TRECs 2-10,[3] the size of a pool formed from the top 10 is 11%-14% of the

[3] In TREC-11, pools were formed to a depth of 50, and thus we do not consider it in this part of the analysis.

size of pools formed from the top 100 documents returned by each run. That is, if one were to reduce pool size in this way, the assessment effort would be greatly reduced. Indeed, if instead of plotting error rate against topic set size, one plots against assessor effort (estimated as the number of topics multiplied by assessor effort to assess a topic to pool depth 10 or 100), the lower graph in Figure 7 is formed. P@10 appears to give a much lower error rate than MAP for a given amount of assessor effort.

Spending substantially less assessor effort building a "P@10 only test collection" is certainly attractive; however, there is one unknown about such a collection. The reason that TREC organizers form pools from the top 100 results of each run (instead of the top 10) is to try to ensure the collection will be effective at assessing retrieval systems that did not contribute to the pool. A new type of retrieval system tested on a collection built from a shallow pool may retrieve relevant documents that were not assessed when the collection was formed. Such a system is likely to be better assessed using a collection where more documents were judged. It is important to point out however, that the degree to which this is a problem in practice is unknown. The stability of test collections – formed from shallow pools – at assessing new retrieval systems is as yet untested. If a test collection is not going to be re-used in later assessment, forming collections from shallow pools appears to be an efficient and accurate approach.

One can consider using the time saved by reducing pool depth to form a test collection with many more topics. Dividing the number of topics per year at TREC (50) by the reduction in pool size (11%-14%), one can estimate that test collections with topic set sizes in the range 357-454[4] could be formed if pool depths were kept to 10. The density of relevant documents at the top of a ranking is higher than the density lower down, so an additional benefit of examining more topics to a shallow depth is that more relevant documents will be found for the same amount of assessor effort as is expended examining 50 topics to a depth of 100. Counting the number of relevant documents typically found in a pool formed from the top 10 of runs submitted to TRECs 2-10, one can compute that 1.7-3.6 times more relevant documents would be found by using shallow pools. Intuitively, it is clear that the more data points – relevant documents – available, the more reliable a run comparison will be. We speculate that such a collection would be better at assessing new retrieval systems, than a collection formed from deep pools, such as TREC. These considerations and the results above suggest that, in contrast to the current TREC methodology, it is better to have larger numbers of topics (perhaps 400) and shallower pools (perhaps depth 10).

5. Examining large topic sets

To examine some aspects of the reliability of measuring runs on large topic sets, we ran 10 variations of the Zettair IR system (configured with 5 different ranking functions each using either title or title+description of topics) across approximately 400 topics produced from 10 years of TREC. Consider any run and a sample of 5 queries; these queries have an average effectiveness. Over a series of samples (say ten) the standard deviation in the averages can be determined. We can then average the standard deviation across the 10 samples. (That is, we determine the average standard deviation of the MAP.) As the total size of the samples is small compared to the number of topics, the samples are more or less independent; by the central limit theorem, the averages follow a normal distribution regardless of the distribution of the per-query results.

We observed that, as sample size was increased, the standard deviation fell. At a sample size of 5, it was 0.056; to take one run, where the observed average was 0.283, we expected 95% of averages of size-5 samples to be between 0.171 and 0.395. At a sample size of 25 the average standard deviation was 0.025, so that 95% of averages were between 0.233 and 0.333. Observing this trend, the standard deviation continued to decline with increasing sample size, dropping below 0.01 at about 150-200 queries, thus suggesting that effectiveness over a large number of queries – when determined in a consistent way – approaches an absolute value for a given run.

6. Selection without replacement

At the core of the methodology for measuring the accuracy of significance tests used here (and in the previous work of Zobel, 1998 and Voorhees & Buckley, 2002) is an assumption that the two disjoint sets of topics selected from the original 50 were formed *independently* of each other. Independence is critical, as, if an ordering of runs and some measure of significance was observed on one set, what was tested in the methodology was the reliability of that significance measure to predict that the ordering was preserved on another wholly independent topic set. If the two sets were randomly selected from the large spread of topics that are typically submitted to an operational search engine, the chance of there being overlap of topics is low. Therefore, the two sets in the methodology were chosen to be disjoint.

However, ensuring the sets were disjoint removes independence from the selection process. To illustrate, when forming two topics sets of size 5, the first is randomly chosen from the available 50 topics. In order to ensure the second set is disjoint from the first, selection of topics is performed *without replacement*; random selection of topics for the second set is from the remaining 45. The choice of the topics in the first set has influenced the choice in the second, but choosing 5 topics from 45 is not that different from choosing 5 from 50. When building two topics sets of 25, however, after randomly selecting one set, the 25 for the second are already chosen. Here, the selection of the second set was wholly *dependent* on the selection of the first.

As the formation of the two subsets was not an independent process we examined the impact of dependent formation of disjoint topic sets from a relatively small topic pool of 50. When measuring the difference between two runs, *a* & *b*, over a set of topics, one will typically find that for some topics, *a* is better than *b* (*a*>*b*), for others, *b* is better than *a* (*b*>*a*), and for some topics, there is little difference between the runs (*a*≈*b*). Generally, if more topics show *a*>*b*, then, averaging across all topics, *a* will be measured to be better than *b*. When forming two large disjoint topic subsets, one set could hold nearly all the *a*>*b* topics. The second set would therefore, be composed of what remains: mainly *a*≈*b* and *b*>*a* topics. With such sets, it is likely that the first set will show *a* is better than *b* and the second set will show the

[4] Note that the calculation ignores two factors. If one gave a group of assessors more topics to assess, the overall time taken to become familiar with the subject matter of the topics would increase. However, Zobel (1998) showed that there is a greater density of relevant documents near the top of a ranking. There is anecdotal evidence from TREC and other environments that assessors can judge relevant documents more quickly than non-relevant, which may reduce assessment time.

opposite. The difference in MAP in the first set is likely to be large, as that set is composed mainly of $a>b$ topics. If the difference in the first set is larger than a certain measure of significance (such as 25% relative difference in MAP), because the second set shows an opposite run order, the measure of significance would be shown to predict run ordering wrongly.

Such an improbable selection of topics is unlikely when sampling from a large number of topics; the probability of such a selection occurring when selecting from 50, however, is relatively high. We concluded that the methodology, used here and in previous work, may be identifying artificially high error rates for significance measures.

We sought to understand if such topic set selections occurred. It was assumed that any improbable topic selections would be identifiable by a large difference in MAP (between the two runs) in the first topic set and a high error rate in predicting run order. The selections would be more prevalent in larger topic subsets (e.g. 15-25) than in the smaller subsets (5-10). As stated earlier, each pair-wise comparison between runs was repeated 50 times, each time with a different random selection of topic sets. For this analysis, the 50 pairs of topics – of sizes 5, 10, 15, 20 and 25 – selected for each pair-wise comparison were sorted by the difference in MAP measured in the first topic set. The error rate in predicting run ordering across the two sets was measured and the average rate across all comparisons was plotted on a graph as shown in Figure 8. Each data point in the graph was an average of error rates measured across 18,460 pair-wise comparisons. The points on the left are the average error rate for the topic sets that had the smallest difference in MAP across the 50 trials of each pair-wise comparison. The points on the right are the average for sets producing the largest MAP difference. Scanning the graph left to right for topics of size 5, the prediction of run ordering improves as differences of MAP in the first set grow, though at the far right of the graph a slight increase in error rate appears. As can be seen, for larger topic sets, the slight increase in error on the right grows with topic set size as the dependency between the first and second topic sets grows. This we took to be evidence that the dependency between large topic sets has a substantial influence on the error rate measured by the methodology used by Zobel (1998), by Voorhees and Buckley (2002), and used so far in this paper. The rising level of error on the right side of the graphs is not a true measure of error, but an artifact of the methodology.

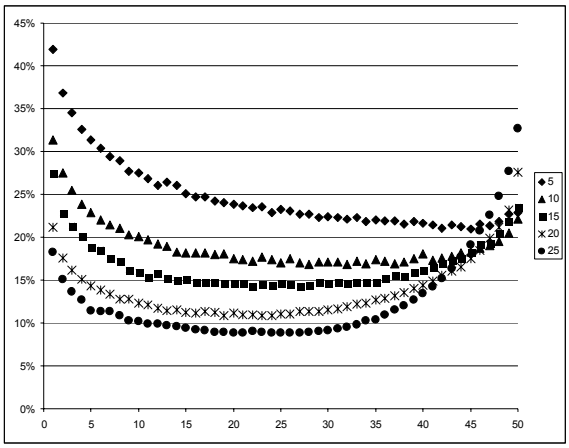


Figure 8: Error rates of different topic set sizes for topic sets selected without replacement.

Since error rates appear to be artificially high, we view all work conducted here and in past works as indicating an *upper bound* on differences required between systems, which was encouraging as actual differences required for 95% confidence were probably lower. Attempts to determine how much lower were the next stage of work.

We attempted to determine a means of removing the improbable topic set selections from the data. However, this was soon abandoned as each attempt resulted in error rate graphs with unexpected properties that questioned the validity of each attempt. An alternative approach was explored.

7. Selection with replacement

If one were forming two topic sets sampling from a large population to test a pair of runs a and b, across the types of topics ($a>b$, $a\approx b$, and $a<b$) it would be expected that the distribution of types in the two sets would be similar. As was shown in the Section above, when selecting without replacement, such expectations are not always fulfilled. A solution to ensure distributions are the same is to select *with replacement*: to pick the first topic set from the population of 50 queries and do the same with the second. The disadvantage to such a strategy is that some topics will be common to both sets, which may cause error rates to be underestimated. However, it was decided to examine such an approach as it would establish a *lower bound* on the error rates experimenters could expect when using TREC collections, thereby complementing the upper bound determined by the earlier work of this paper and by Voorhees and Buckley (2002).

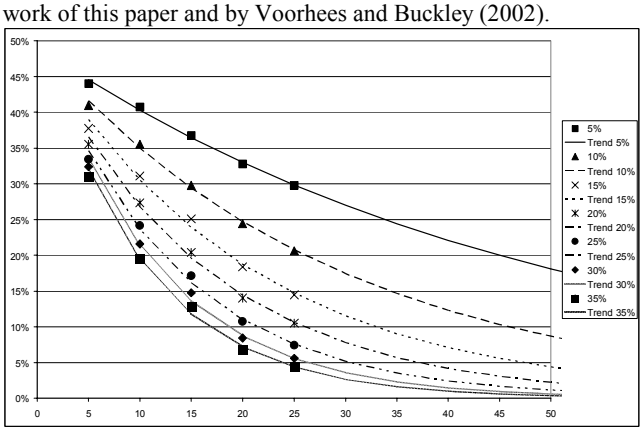


Figure 9: Error rates for relative differences in MAP, trends projected to topics of size 50; topics selected with replacement.

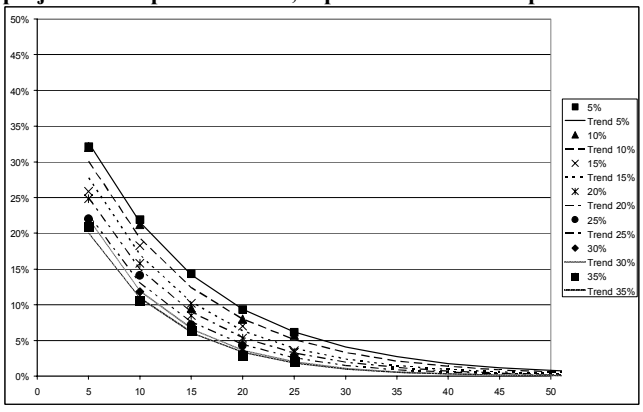


Figure 10: Error rates for those relative differences in MAP for which a t-test produced significance $0.01<p\leq0.05$. Topic sets selected with replacement.

The entire procedure for computing error rates described in Section 3 was repeated, but using selection with replacement for topic selection. To save time in computing all the 18,460 pair-wise comparisons for all the topic sets (repeated 50 times), only sets of size 5, 10, 15, 20 and 25 were processed. The graph in Figure 9 shows the resulting error rates. Here, it can be seen that for a topic set size of 50, a relative difference of between 15%-20% is required in order to be 95% confident that the order of runs will be preserved across different topic sets, less than the 25%-30% percentage difference computed using the previous methodology. When only considering those comparisons where a percentage difference was recorded and statistical significance ($0.01<p\leq0.05$) was observed, error rates dropped substantially (see Figure 10). Here it can be seen that, if statistical significance is observed for any percentage difference between two runs, one can be highly confident that the ordering of the runs will be observed on different topic sets. The drop in error rate when significance was applied to the with replacement topic sets was much more substantial than to the sets formed without replacement. This effect was repeated when P@10 was examined as shown in Figure 11 and Figure 12. We speculate that the improbable topic selections observed in the without replacement methodology caused particularly high error rates for significance tests.

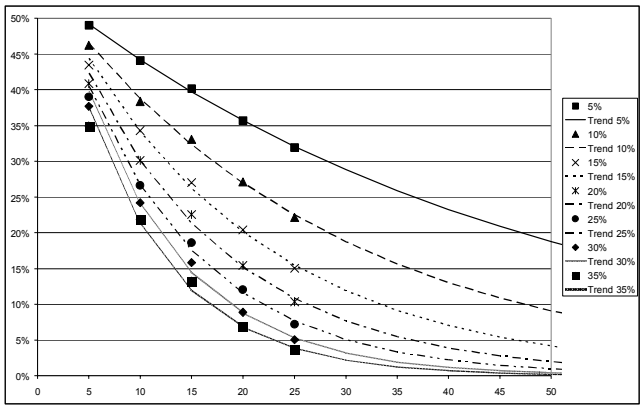


Figure 11: Same as Figure 9 but for P@10.

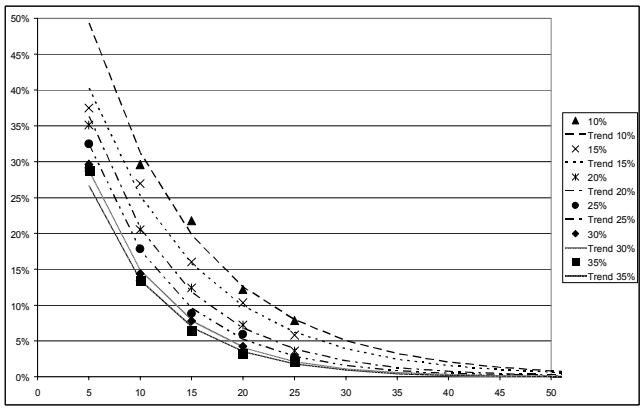


Figure 12: Same as Figure 10 but for P@10.

8. Conclusions and future work

We established upper and lower bounds on the error rates of significance applied to the TREC test collections. From the results, it was clear that significance substantially increases the reliability of retrieval effectiveness measures. When considering upper and lower bounds for 50 topics, at worst almost any measure of t-test significance coupled with at least a 10% relative difference in MAP between two runs is significant; at best, once significance is observed, experimenters can view the result as significant. However for small topic set sizes (≤25) observing statistical significance does not guarantee that a result will be repeatable on other sets of topics.

In comparisons of significance tests, the t-test was shown to produce lower error rates than sign and Wilcoxon. It was shown that, if a set of relevance judgments have already been created, MAP is a more reliable effectiveness measure than P@10.

If a collection is to be created, once assessor effort is taken into account, P@10 produces lower error rates for a given unit of assessor effort than MAP. We also showed that building test collections with shallow pools locates more relevant documents, which we believe results in more accurate measurement.

Future work will examine how to better sample topic sets to more reliably estimate error rates: stratified sampling will be explored. We will also examine the error rates of relevance judgments built from shallow pools, particularly comparing error rates of P@10 to MAP measured with small numbers of relevant documents.

9. Acknowledgements

Thanks to Jamie Callan for sparking much of the original interest in this work and to Alistair Moffat with whom discussions while struggling with data helped resolve matters tremendously. Financial support for the work was provided by the University of Melbourne, the Australian Research Council, and the EU projects SPIRIT and BRICKS: contract numbers IST-2001-35047 & IST-2002-2.3.1.12.

10. Bibliography

[1] Buckley, C., Voorhees, E.M. (2000) Evaluating evaluation measure stability, *Proc. ACM SIGIR*, 33-40.

[2] Buckley, C., Voorhees, E.M. (2004) Retrieval evaluation with incomplete information, in *Proc. ACM SIGIR*, 25-32

[3] Dunlop, M.D. (1997) Time Relevance and Interaction Modeling for Information Retrieval, in *Proc. ACM SIGIR*, 206-213.

[4] Hull, D. (1993) Using statistical testing in the evaluation of retrieval experiments, in *Proc. of ACM SIGIR*, 329-338.

[5] Järvelin, K. & Kekäläinen, J. (2000) IR evaluation methods for retrieving highly relevant documents, in *Proc. ACM SIGIR*, 41-48.

[6] Matthews, R. (2003) The numbers don't add up, *New Scientist*, March, p. 28, issue 2385.

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

[8] Spärck Jones, K. (1974) Automatic indexing. *Journal of Documentation*, 30:393-432, 1974.

[9] 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.

[10] Tague-Sutcliffe, J., Blustein (1994) A Statistical Analysis of the TREC-3 Data, in *Proc. TREC-3*, 385-398.

[11] Van Rijsbergen, C.J. (1979) *Information Retrieval*, London: Butterworths.

[12] Voorhees, E.M., Buckley, C. (2002) The effect of topic set size on retrieval experiment error, in *Proc. ACM SIGIR*, 316-323.

[13] Voorhees, E.M., Harman, D. (1999) Overview of the 8th Text REtrieval Conference (TREC-8), in *Proc. 8th Text REtrieval Conf.*

[14] Zobel, J. (1998) How reliable are the results of large-scale information retrieval experiments?, in *Proc. ACM SIGIR*, 307-31