Score Standardization for Robust Comparison of Retrieval Systems

William Webber *Alistair Moffat*
Department of Computer Science and Software Engineering
The University of Melbourne
Victoria, Australia 3010
wew@csse.unimelb.edu.au, alistair@csse.unimelb.edu.au

Justin Zobel
NICTA Victoria Research Laboratory
Department of Computer Science and Software Engineering
The University of Melbourne
Victoria, Australia 3010
jz@csse.unimelb.edu.au

Abstract *Information retrieval systems are evaluated by applying them to standard test collections of documents, topics, and relevance judgements. An evaluation metric is then used to score a system's output for each topic; these scores are averaged to obtain an overall measure of effectiveness. However, different topics have differing degrees of difficulty and differing variability in scores, leading to inconsistent contributions to aggregate system scores and problems in comparing scores between different test collections. In this paper, we propose that per-topic scores be standardized on the observed score distributions of the runs submitted to the original experiment from which the test collection was created. We demonstrate that standardization equalizes topic contributions to system effectiveness scores and improves inter-collection comparability.*

Keywords Retrieval system evaluation, average precision, standardization.

1 Introduction

The effectiveness of information retrieval (IR) systems is traditionally evaluated using the Cranfield methodology [Cleverdon, 1991], which involves the use of a *test collection* consisting of a *document corpus*, a set of *topics*, and, for each topic–document pair, a judgment as to whether the document is relevant to that topic; these judgments are referred to as *qrels*. To evaluate an IR system, the topics are formulated as queries, and processed against the document corpus. For each topic, the system produces a ranked list of documents or *run*, ordering the documents by decreasing estimated relevance to the topic. The position of a document in a run is known as its *rank*.

Determining the effectiveness of a run involves applying an *evaluation metric* that uses the judged relevance of the document at each rank in the run to produce a score. Run scores for the topic set are averaged to give a system score. The statistical significance of a difference in system scores is then assessed using a paired significance test [Zobel, 1998, Sanderson and Zobel, 2005, Smucker et al., 2007].

Current test collections contain millions of documents, so it is not feasible to assess every document for relevance to every topic. Instead, test collections are formed in the context of a community *retrieval experiment* [Voorhees and Harman, 2005]. The runs from each participating system are submitted to the experiment, and documents for judging are selected from some or all of the submitted runs to form a *pool* [Sparck Jones and van Rijsbergen, 1975].

Different topics in a test collection can have quite different characteristics. Some have many relevant documents, others have few. For some topics, it will be easy for retrieval algorithms to identify relevant documents, while other topics may be ambiguous, causing systems to erroneously retrieve many documents that are plausible, but irrelevant. Human assessment of relevance also varies between topics, with the assessors sometimes employing strict criteria, while at other times being more liberal.

Variations in topic characteristics result in variability in the distribution of run scores across topics. As a result, different topics make different contributions to aggregate scores, and to tests for statistical significance. Topic score variability makes it particularly difficult to compare results obtained on different topic sets.

This paper proposes a topic score adjustment technique to ameliorate the problems caused by topic variability. The technique we describe is well known in other fields of testing, but until now has not, somewhat surprisingly, been suggested for use in the evaluation of retrieval systems. The essence of the approach is that individual run scores should be *standardized* before statistical tests are applied, with the adjustment to each run

Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, December 10, 2007. **Copyright for this article remains with the authors.**

score based upon the mean and standard deviation observed for the full set of run scores for that topic across the systems participating in the retrieval experiment. Once calculated, these *standardization factors* would be published as part of the test collection, and used to standardize the scores of future runs made against that collection.

Experimental results based on the data collected during one of the TREC[1] retrieval experiments show that standardization reduces per-topic variability, and increases the ability of statistical tests to discriminate between systems. The results also show that standardization allows systems to be compared even when they have been tested on different topics. The latter observation is of particular importance for real-world web search engines, where documents, queries, and retrieval algorithms are constantly changing, and having retrieval effectiveness measures that remain comparable over time is crucial.

2 Evaluation metrics

Many evaluation metrics for information retrieval have been proposed. Probably the most widely used is *average precision* (AP). Calculation of AP involves averaging the *precision* at each rank (the proportion of documents up to that position that are relevant) at which a run has a relevant document, for every known relevant document for that topic. Unretrieved relevant documents are assigned a precision of zero. To calculate AP it is thus sufficient to sum the precisions at each known retrieved relevant document in the run, down to some limiting depth that is set as part of the experimental design, and then divide by R, the total number of relevant documents. That is, AP includes an adjustment based on the "difficulty" of the topic, and can be thought of as being an R-normalized version of a more fundamental metric, the *sum of precisions at relevance* (SP) score of the run.

Other metrics also incorporate some form of score *normalization*, so as to reduce scores for "easy" topics and increase them for "hard" ones. For example, the *normalized discounted cumulative gain* (NDCG) method of Järvelin and Kekäläinen [2002] is derived from the *discounted cumulative gain* of a run (the sum of the relevance contributions of documents in a run, each discounted by the logarithm of its rank). In this case the normalization factor is the maximum possible DCG score that could have been achieved to that evaluation depth. On the other hand, some metrics make no attempt to adjust for topic variability. In precision-at-depth-d, for example, a run is scored on the proportion of top-d documents that are relevant, no matter what the total number of relevant documents for the topic are. Topics with many, easily-identified relevant documents get higher precision-at-d scores than topics with few, hard-to-find relevant documents.

[1] `http://trec.nist.gov`

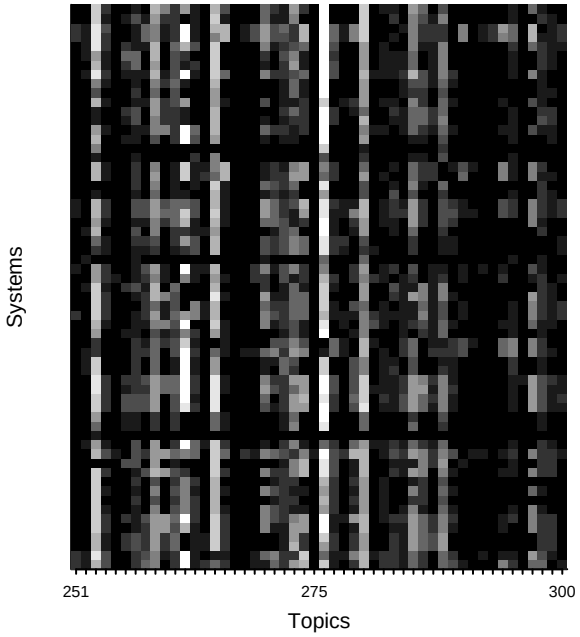
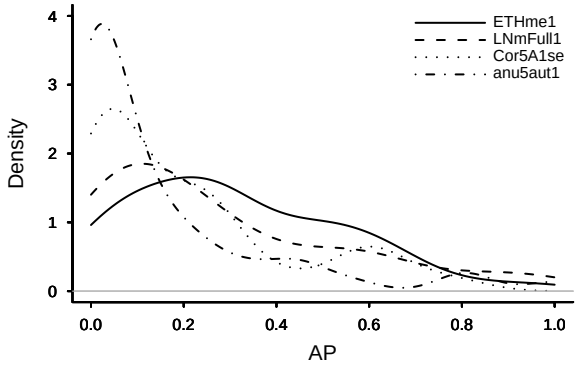


Figure 1: *Intensity visualization of run AP scores from the TREC5 Adhoc Track. The columns represent topics, ordered by topic number, and the rows represent systems, in ASCII order of system name. Each cell represents the AP score of a single run, with white cells indicating scores above* 0.9*; black cells scores below* 0.1*; and shades of grey indicating intermediate run scores. Easy topics stand out more clearly than good systems.*

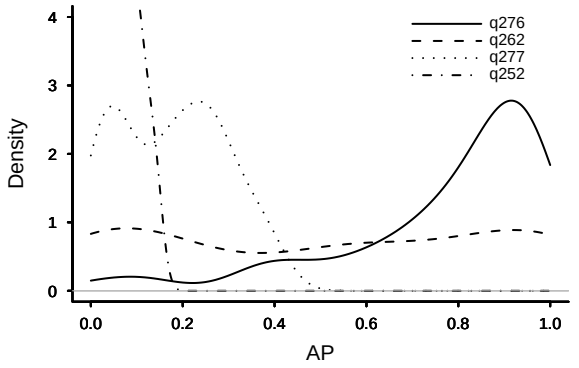
The normalization methods of AP and NDCG can be considered separately from the metrics themselves. Any metric could be normalized by dividing by the total number of relevant documents (as in AP), or by the maximum possible score that could be achieved under that metric (as in NDCG). But note that both of these normalization methods require that the size of the set of relevant documents be known or estimated. Where qrels have been formed by pooling, the actual number of relevant documents is not known, and instead the number of pooled relevant documents is used. This makes the possibility of performing new relevance judgments for a new system problematic. If the normalization factor is updated when new relevant documents are found, then all previously reported scores have to be modified downwards; and if the normalization factor is not updated, then it is possible for a new system to achieve a score higher than the nominal maximum.

3 Topic variability

A community evaluation experiment involves running T topics against S systems. Each system thus produces T runs, one for each topic; and each topic receives S runs, one from each system. The runs are scored using an IR evaluation metric, such as AP. The resulting set of scores can be thought of as forming an $S \times T$ matrix. Figure 1 visualizes this matrix for the AP scores of



(a) System AP score density



(b) Topic AP score density

Figure 2: *Kernel density estimates of AP scores for the (a) systems, and (b) topics, at the first, tenth, twentieth, and seventy-fifth percentile (top to bottom in the legend) when ordered by average (a) system, or (b) topic, AP scores. Systems behave more like each other than do topics. All data is from the the TREC5 Adhoc Track.*

the TREC5 Adhoc Track as an intensity image, with lighter shades being higher scores. Topics are columns; systems are rows. The easy topics give rise to vertical white lines in the plot, and are easy to spot. Good systems should similarly give rise to horizontal white lines, but are much harder to pick out. Only the very worst systems, those with some programming bug or serious algorithmic misapplication, stand out as dark horizontal lines. This simple visualization makes it clear that the score matrix holds more information about the topics than it does about the systems.

The scores of the T topic runs for each system (in a row of the visualization matrix) form a distribution, as do the scores of the S system runs against each topic (in a column of the matrix). Figure 2 displays topic and system score distribution kernel density estimates (a form of smoothed histogram) from the TREC5 Adhoc Track for the first, tenth, twentieth, and seventy-fifth percentile topic and system, as ordered (in part (a)) by row averages, and (in part (b)) by column averages. The homogeneity of the system scores, and the heterogeneity of the topic scores, is immediately obvious. The system distributions all have the same, right-skewed unimodal shape, similar dispersions, and even relatively similar localities. In contrast, the topic score distributions vary greatly.

	System AP			
	ETHme1	LNmFull1	Cor5A1se	anu5aut1
mean	0.317	0.282	0.206	0.154
st.dev	0.231	0.271	0.220	0.232

	Topic AP			
	q276	q262	q277	q252
mean	0.771	0.506	0.175	0.056
st.dev	0.235	0.383	0.118	0.039

Table 1: *Mean and standard deviation of AP scores for sample systems and topics from the TREC5 Adhoc Track. The kernel density estimates for the same data sets are plotted in Figure 2.*

Systems	Significance test	
	Paired	Two-sample
All	0.636	0.364
Auto	0.495	0.130

Table 2: *Proportion of system pairs from the TREC5 Adhoc Track found to have significantly different average precision at* $p = 0.05$ *in a two-tailed* t*-test, either paired or two-sample, and including either all* 61 *systems or only the* 30 *automatic systems minus the four faulty ones with* MAP < 0.050.

Table 1 summarizes the locality and dispersions of the system and topic distributions plotted in Figure 2, in terms of their means and standard deviations. All systems have similar standard deviations, and the mean of the best system is only twice that of the seventy-fifth percentile. In contrast, topic means and standard deviations each vary tenfold from lowest to highest value.

Clearly, the mean AP scores for a test collection depend greatly on which topics happen to be selected, making it difficult to compare AP scores from different test collections. This is reflected in the lower discrimination resulting from using two-sample rather than paired significance tests to compare systems. In a two sample test, the null hypothesis is that the two samples are independent random samples from the same population. In a paired test, each item in one sample is assumed to have a shared dependency with a corresponding item in the other sample. Where the same test collection is used in evaluating two IR systems, the runs from each system for the same topic are paired, and the values fed into the hypothesis test are the deltas between the paired run scores. Pairing helps to control the effect of topic variability on scores, whereas reverting to a two-sample test gives an indication of the ability to find statistical significance when using different test collections sampled from the same (conceptual) population.

Table 2 contrasts the discriminative power of paired and two-sample significance tests on the the TREC5 Adhoc Track systems. The paired t-test finds a signif-

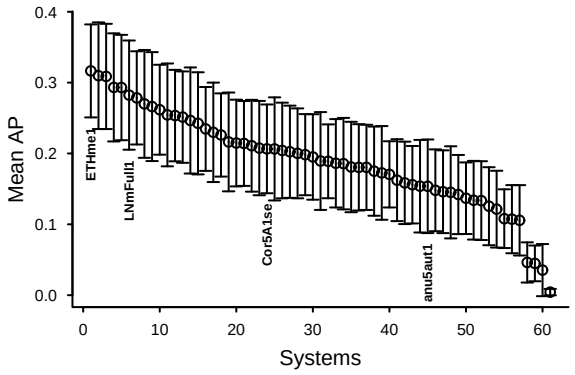


Figure 3: *The 95% confidence intervals on mean AP scores for the TREC5 Adhoc Track systems, using a t distribution. Systems are ordered by their official MAP score.*

icant difference between almost two-thirds of system pairs, whereas the two-sample test only finds significance for slightly over a third. If only the automatic runs are considered, and the four apparently faulty systems are excluded (all with a mean system AP score less than 0.050), then the outcome is even more stark. Half of the system pairs are found significantly different at the $p = 0.05$ level by the paired test, but only one eighth by the two-sample test. These results highlight the difficulty caused by inter-topic score variability when making comparisons between systems tested on different but similarly constituted test collections.

Figure 3 displays another effect of high inter-topic score variability. Here, 95% confidence intervals have been plotted on the "true" mean system AP (MAP) scores of the TREC5 Adhoc Track systems. These confidence intervals are statistically related to (though not identical with) the results of a two-sample significance test. The intervals on the MAP scores are wide: the best system could have a true MAP score of anywhere between 0.25 and 0.38; the median system's confidence interval is from 0.12 to 0.26; and the worst system (excluding the four faulty ones) sits in the range 0.06 to 0.16. With such wide confidence intervals, the strength of any conclusions based on mean AP score comparisons is extremely limited.

The marked difference in mean topic scores constrains experimenters to evaluate systems on the same test collection, so that they can employ paired significance tests. However, the high degree of difference in intra-topic AP standard deviations for different topics causes problems that even the one test collection and paired hypothesis testing cannot control. Consider once again the topic score standard deviations displayed in Table 1. Topic q262, the tenth-percentile topic by mean AP in the collection, has roughly ten times the AP score standard deviation of topic q252, the seventy-fifth-percentile topic. As a result, topic q262 will on average have ten times as much influence on the difference between the mean AP scores of any two systems as topic q252, and will have a similar impact in paired significance tests. Now, it happens that topic q262 is the highest-variance topic in the TREC5 Adhoc Track. However, even the twenty-fifth percentile topic when ordered by AP standard deviation has two and a half times the standard deviation of the seventy-fifth percentile topic (0.170 against 0.067), and therefore two and a half times the influence.

It might be supposed that topics with higher score variance are better at discriminating good from bad systems than are topics with lower variance, based on the intuition that the low-variance topics suffer more from "random noise" than the high-variance ones, or have too few relevant documents, or are too hard. The problem with testing such a conjecture is, of course, its circularity: we must first determine which are the good and which the bad systems before the reliable discrimination of a topic's scores can be determined. One way of approaching the issue is to consider how well the system scores for a topic correlate with the mean system scores for the rest of the topic set excluding that topic. This *item-total correlation* is used in Test Theory to indicate the *reliability* of a test component [Bodoff and Li, 2007]. Topic reliability can then be correlated with topic score standard deviation. Doing so for the TREC5 Adhoc Track systems (after excluding the four faulty runs, which score at or near zero for every topic and thus artificially inflate both the standard deviation and the reliability of topics with high mean scores) gives a Pearson correlation of 0.091 – that is, there is essentially no correlation. And even with this figure, we are not controlling for the fact that high-variance topics as a class still have more influence in determining the total scores even when individual high-variance topics are excluded. Thus, although high variance topics have greater influence on system mean AP scores and on paired significance tests, they are not inherently more reliable, and so their greater influence is neither deserved nor desirable.

4 Standardization

Topic score variability and the problems it causes can be addressed using score adjustment, and as was noted above, many effectiveness metrics have some kind of normalization built in. They are, however, rather indirect in their application to the problem of score variability. As the preceding discussion has demonstrated, AP's embedded normalization by R, the number of relevant documents, is not particularly effective.

An alternative method for reducing topic score variability – and one widely used in other experimental settings – is to normalize scores for a topic by the observed score mean and standard deviation of that topic. If a topic t has an unnormalized score mean of μ_t and an unnormalized score standard deviation of σ_t, and if a run for that topic receives the unnormalized score of x, then the normalized score x' for that run is:

$$x' = \frac{x - \mu_t}{\sigma_t}$$

Topic	Unstandardized AP				Standardized AP			
	ETHme1	LNmFull1	Cor5A1se	anu5aut1	ETHme1	LNmFull1	Cor5A1se	anu5aut1
q276	0.968	1.000	0.615	0.814	0.840	0.975	−0.667	0.180
q262	0.500	0.950	0.017	1.000	−0.015	1.161	−1.277	1.291
q277	0.344	0.301	0.256	0.059	1.434	1.067	0.689	−0.985
q252	0.045	0.058	0.030	0.109	−0.275	0.059	−0.665	1.340

Table 3: *Selected topics from the TREC5 Adhoc Track, and selected system AP scores, before and after standardization. The parameters μ_t and σ_t for these four topics are as listed in the bottom section of Table 1.*

Such a value is known as a z *score*, and the process of deriving it is commonly called *standardization* [Hays, 1991, chapter 4].

Note that a standardized distribution has the same shape as the unstandardized one, and that standardization only affects locality and dispersion. After standardization the mean score for each topic is zero, and the standard deviation is one. There are no fixed upper or lower bounds on a topic's scores. Chebyshev's inequality, however, guarantees that at least 75% of standardized scores for a topic will be between -2.0 and 2.0, and in practice the proportion will generally be much higher; for the TREC5 Adhoc Track standardized AP scores, it is 96%.

Unfortunately, the true mean and true standard deviation of each topic – the values that would be obtained from the conceptual population of runs of which the actually observed runs are only considered to be a sample – is not known. However, it can be estimated in the usual way from the statistics of the observed sample. That is, the standardization factors μ_t and σ_t for a topic t are estimated as the mean and standard deviation of the runs made against t in the original retrieval experiment. These estimates rely on there being an original retrieval experiment, of course, whereas normalization by total number of relevant documents or maximum achievable score could theoretically be performed in the absence of an experiment, by assessing every document for relevance. In practice, though, full assessment is impractical, and qrel sets are derived from pooling experimental systems. Therefore, the requirements for standardization are in practice no greater than for other forms of normalization.

It has already been mentioned that AP incorporates a crude form of normalization, and can be thought of as an unnormalized metric, sum of precisions at relevance, or SP, which is then normalized by the total number of known relevant documents, R. It is also possible to directly standardize the SP run scores rather than the AP scores, and in fact *exactly the same values* result from standardizing the SP scores as the AP scores. In other words, standardized SP is the same metric as standardized AP, freeing us from the need to know or estimate R. This observation alone would be sufficient to justify our interest in standardization of effectiveness scores. Similarly, standardized DCG and standardized NDCG are numerically identical measures.

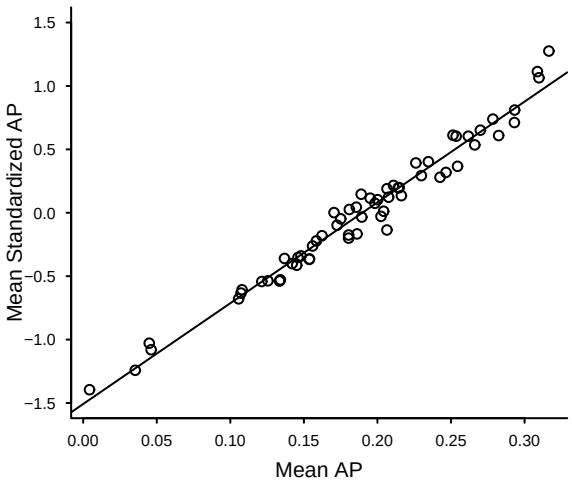


Figure 4: *Mean unstandardized AP and mean standardized AP scores for the TREC5 Adhoc Track systems, with the line of best fit.*

Table 3 shows the unstandardized and standardized AP scores for the previously examined topics and systems. The standardization factors for the topics are the sample mean and standard deviations previously reported in Table 1. The unstandardized figures are difficult to interpret and compare. For example, ETHme1 scored 0.344 for topic q277 and 0.500 for topic q262; but does the latter score represent a better result than the former, or was it simply an easier topic? Similarly, Cor5A1se scored 0.2 higher than anu5aut1 on topic q277, but 0.08 lower on q252; is the former result more significant than the latter? In contrast, the standardized results are directly informative. A positive score indicates the run outperformed the community mean for that topic, a negative score that it underperformed it; a score of 1.0 means the run is one standard deviation above the mean, and so on. So, without examining any other figures, we can immediately see that anu5aut1 has done well on topic q252, and Cor5A1se poorly on topic q262.

5 Experiments

Figure 4 plots the mean unstandardized AP and mean standardized AP scores for the TREC5 Adhoc Track systems against each other. The two metrics correlate closely. The Pearson correlation on the scores is 0.985, and the Kendall's τ correlation on the system ranks is

Systems	Significance test	
	Paired	Two-sample
All	0.683 *+0.047*	0.683 *+0.319*
Auto	0.561 *+0.066*	0.542 *+0.412*

Table 4: *Proportion of system pairs from the TREC5 Adhoc Track found to be significantly different using standardized AP at $p = 0.05$ in a two-tailed t-test, either paired or two-sample, and including either all 61 systems or only the 30 automatic systems minus the 4 faulty ones with* MAP < 0.05*. The improvement over the results with unstandardized AP reported in Table 2 is shown in italics.*

0.903. By way of comparison, the respective correlations of NDCG with mean (unstandardized) AP are 0.973 and 0.915. There are some local perturbations of ordering, however, which may represent improvements if we accept the thesis that topics should have similar impacts. The standardized AP scores separate out the three best systems, in the top right-hand corner, better than the unstandardized AP scores do. Note also that standardization shows the three best systems to perform almost as well as the four faulty systems do poorly, the former being on average one standard deviation better than the mean, the latter one standard deviation worse. It is not possible to make such a judgment simply by looking at the unstandardized MAP scores.

Table 4 compares the discriminative power of two-sample and paired significance tests on the standardized AP scores. In contrast to the unstandardized scores (see Table 2), with standardized scores, the two-sample significance test approaches the discriminative power of the paired test. The level of agreement between the two is high, the overlap (size of intersection divided by size of union) being 0.9 for the full system set, which is acceptable given the $p = 0.05$ significance level. Note also that the paired test itself is more discriminative for standardized than for unstandardized AP scores, rising for the full system set from 0.636 to 0.683 of pairs. This is the result of standardizing intra-topic score variance: deltas become comparable between topics, the variance of deltas falls on average, and the paired test is more confident about differences.

Figure 5 displays the 95% confidence intervals on the mean standardized AP scores for the TREC5 Adhoc Track systems, which are considerably narrower than for the mean unstandardized AP scores (Figure 3). Before standardization, the top system's confidence interval overlapped with the median's; after standardization it is well clear by the end of the first quartile.

One of the stated goals of score standardization is to improve the comparability of scores between different topic sets, provided that they are (actually or conceptually) randomly sampled from the same population of topics. As mentioned, when comparing two systems run against different topic sets, a two-sample significance test must be used. For significance testing, comparability can be considered in two aspects: the rate of false positives, and the rate of false negatives.

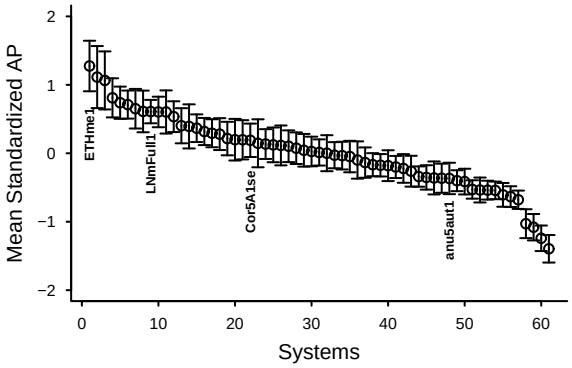


Figure 5: *The 95% confidence intervals on mean standardized AP scores for the TREC5 Adhoc Track systems, using a t distribution.*

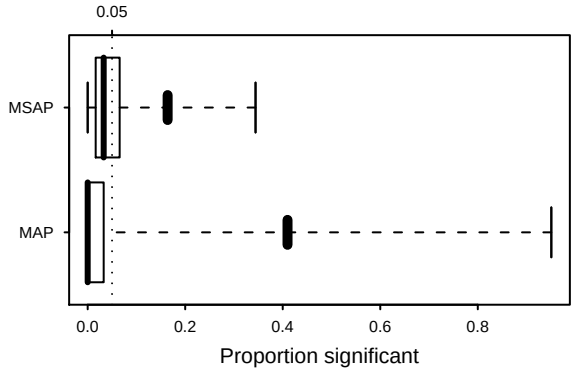


Figure 6: *False positives on two-tailed two-sample t-tests at $p = 0.05$ for the TREC5 Adhoc Track systems using 25-topic randomly-sampled subsets, repeated $5{,}000$ times. The line within the box is the median; the left and right box edges are the 25th and 75th percentiles, respectively; the dotted whiskers extend to the extreme values; and the thick line on the right whiskers marks the upper bound of the 95% confidence interval. The dotted vertical line is the expected and approximate mean of both distributions.*

The rate of false positives can be investigated by testing a system against itself. Obviously, a system is not significantly different from itself, so if a significance test finds that it is, it must be a false positive. To explore this aspect, we took all 50 of the TREC5 Adhoc Track topics and randomly sampled two subsets of 25 topics each. The two subsets were not required to be disjoint; a disjoint partitioning would distort the extreme results, since if one subset happened to get all the hardest topics, then necessarily the other subset would get all the easiest ones [Sanderson and Zobel, 2005]. Then a two-tailed two-sample t-test was performed for each of the TREC5 Adhoc Track systems using the two topic subsets, and the proportion of false positives was recorded, using both the unstandardized and the standardized AP scores. This process was then repeated $5{,}000$ times, to provide a distribution of false positives. Figure 6 displays the results of this experiment. The mean rate of false positives is 0.043 and 0.049 for the unstandardized and standard-

ized AP scores respectively, as is to be expected with $p = 0.05$. However, the variability of false positive rates for the unstandardized AP scores is considerably greater than for the standardized ones. Most topic partitionings show no false positives on unstandardized AP, indicating an insufficiently sensitive test (a rate of 0.05 is expected), but at the other extreme, there is a partitioning in which 95% of the systems are found to be better than themselves – a real boon for IR researchers wishing to publish papers. These results also demonstrate an important point about significance tests based around fixed test collections, which is that the chance of error (false positives or false negatives) is not independent between different pairs of systems; that is, if a collection makes an error on one pair, then it is more likely to make an error on the others. The seriousness of an error bias in a test collection is, however, greatly reduced by standardization.

Determining the rate of false negatives is more difficult, in that we do not know what the true positives are – that is, which systems truly are better than which. The approach taken here is to consider systems pairs that a paired, two-tailed t-test finds significant at $p = 0.001$ for both unstandardized and standardized AP – that is, pairs in which one system is almost certainly better than the other. One of these system pairs was chosen at random, along with a topic partitioning; then, a two-tailed, two-sample t-test was performed between the two systems, the first system using one topic subset, the second using the other, with both standardized and unstandardized AP scores used to evaluate effectiveness. Repeating this above experiment $5{,}000$ times, no significant difference at $p = 0.05$ was found in 41.6% of the pairs using unstandardized AP, whereas with standardized AP, only 2.6% of pairs were not found to be significantly different. Moreover, in only 0.2% of cases was significance found with unstandardized scores but not with standardized ones. These results provide compelling evidence that standardization enormously boosts the power of two-sample significance tests.

6 Related Work

Bodoff and Li [2007] introduce Test Theory concepts such as topic and test reliability into the evaluation of IR test collections. They find the TREC collections they investigate to be reliable in a test-theoretic sense, when AP is the evaluation metric. However, the standard of reliability is a rule-of-thumb, and what may be acceptable reliability in a test applied to human subjects may not be equally acceptable in IR evaluation. Reliability is assessed in terms of average correlation between topic scores, and inter-topic variation is not addressed.

Determining whether one metric is superior to another is problematic in the absence of a gold standard. Aslam et al. [2005] suggest the use of a maximum entropy method, finding AP to be superior to R-precision and precision-at-depth-d. Aslam et al. do not address the question of comparability of scores between systems and topics. Sakai [2006] proposes that metrics should be assessed based on the proportion of all system pairs in an experiment whose means are found to be significantly different in a paired bootstrap significance test. Smucker et al. [2007] compare the bootstrap, randomization, and t-tests, and find them to give similar results. Earlier, Voorhees and Buckley [2002] proposed an ad-hoc variant.

A popular alternative to AP is NDCG, proposed by Järvelin and Kekäläinen [2002], which incorporates its own normalization method, as described in Section 2. Another alternative is rank-biased precision or RBP [Moffat et al., 2007], which is bounded to the range $[0, 1]$, but is not normalized in the sense used here.

Mean AP scores a system by the arithmetic mean of the run AP scores. An alternative is to take the geometric mean, resulting in GMAP. Robertson [2006] discusses GMAP, observing that it is equivalent to the exponentiated average of the log of the per-topic AP values. Robertson further observes that taking the mean of a set of scores implies the assumption that score intervals, rather than score ratios, are the important unit of comparison; taking the log of the scores before averaging converts ratios to intervals. The GMAP metric is used in the Robust track of TREC, as it gives more emphasis to topics with low mean AP, which are regarded as "hard" [Voorhees, 2004]. We believe that score standardization may be a better solution.

Mizzaro and Robertson [2007] investigate normalizing per-run AP (or log AP) scores either by topic or by system, by subtracting the mean score for that topic or system (but not adjusting for the standard deviation). The adjusted AP scores are used as edge weights in a system-topic graph. Mizzaro and Robertson find considerable variation in topic difficulty and system performance on topics; that on the whole effective systems are better at distinguishing easy topics from hard ones; and that easy topics are better at distinguishing between more or less effective systems. The last finding they regard as undesirable. Their finding that easy topics distinguish good from bad systems may be a consequence of easy topics having larger variances under AP, and therefore more influence on final scores. Finally, Tague-Sutcliffe and Blustein [1994] undertake an analysis-of-variance on the TREC3 experimental data, and find that there is a stronger topic than system effect; Banks et al. [1999] discuss the analysis in more detail. Our experiments confirm their observations.

7 Conclusions and future work

We have shown that the power of a text retrieval experiment to discriminate between systems can be improved by standardizing run scores across a set of systems. Our application of standardization to AP, chosen because it is the commonest metric in current work, shows that not only does it allow better discrimination, but should allow comparison of results between dif-

ferent test collections. We plan to undertake similar experiments with other metrics, including discounted cumulative gain and rank biased precision.

Standardization removes what on reflection is a surprising limitation on the way TREC test collections have been used: although the runs submitted by these systems provide a wealth of potential information about the test topics and document corpus, the submitted runs are used solely to determine a pool of documents for relevance assessment, and are then ignored. Using these runs to estimate topic difficulty, and then standardizing topic scores based on these estimates, is one way in which the information in the submitted runs can be used to improve the reliability of the test collection. There are doubtless others.

There are also issues with some of the underlying assumptions in standard approaches to measurement of statistical significance. First, the assumption of random sampling from an underlying population is problematic. Second, significance testing is undertaken between pairs of systems in isolation, ignoring the wealth of data that is available from the other runs that are part of the experimental collection. Score standardization perhaps helps to resolve some of this oddity, but it still seems a strange procedure.

Several authors have, using AP and other measures, pursued investigations such as sampling the topic sets to see how robust the original pooling experiments were [Zobel, 1998, Sanderson and Zobel, 2005, Buckley and Voorhees, 2000, Voorhees and Buckley, 2002]. We expect that standardization would lead to substantial improvements, an avenue we also plan to explore. Finally, we will investigate the comparability of standardized results between different collections.

Acknowledgements This work was supported by the Australian Research Council.

References

J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 27–34, Salvador, Brazil, 2005.

D. Banks, P. Over, and N.-F. Zhang. Blind men and elephants: six approaches to TREC data. *Information Retrieval*, 1(1): 7–34, April 1999.

D. Bodoff and P. Li. Test theory for assessing IR test collections. In *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 367–374, Amsterdam, The Netherlands, July 2007.

C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proc. 23rd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 33–40, Athens, Greece, 2000.

C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proc. 14th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 3–12, Chicago, IL, USA, 1991.

W. L. Hays. *Statistics*. Harcourt Brace, Fort Worth, 4th edition, 1991.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

S. Mizzaro and S. Robertson. Hits hits TREC: exploring IR evaluation results with network analysis. In *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 479–486, Amsterdam, The Netherlands, July 2007.

A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 375–382, Amsterdam, The Netherlands, July 2007.

S. Robertson. On GMAP: and other transformations. In *Proc. 15th ACM Int. Conf. on Information and Knowledge Management*, pages 78–83, Arlington, VA, USA, November 2006.

T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proc. 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 525–532, Seattle, WA, USA, 2006.

M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 162–169, Salvador, Brazil, 2005.

M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. 16th ACM Int. Conf. on Information and Knowledge Management*, Lisboa, Portugal, 2007.

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

J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the TREC-3 data. In *Proc. TREC-3*, November 1994. NIST Special Publication 500-225.

E. Voorhees and D. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proc. TREC-13*, November 2004. NIST Special Publication 500-261.

E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proc. 25th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 316–323, Tampere, Finland, 2002.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998.