Score Standardization for Inter-Collection Comparison of Retrieval Systems

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

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

Justin Zobel
NICTA VRL
The University of Melbourne
Victoria 3010, Australia
jz@csse.unimelb.edu.au

ABSTRACT

The goal of system evaluation in information retrieval has always been to determine which of a set of systems is superior on a given collection. The tool used to determine system ordering is an evaluation metric such as average precision, which computes relative, collection-specific scores. We argue that a broader goal is achievable. In this paper we demonstrate that, by use of standardization, scores can be substantially independent of a particular collection, allowing systems to be compared even when they have been tested on different collections. Compared to current methods, our techniques provide richer information about system performance, improved clarity in outcome reporting, and greater simplicity in reviewing results from disparate sources.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

Keywords

Retrieval experiment, evaluation, average precision, system measurement

General Terms

Measurement, performance, experimentation

1. INTRODUCTION

A key aim of research in information retrieval (IR) is to develop search methods with improved effectiveness, but identification of improvements requires rigorous evaluation methodologies. The Cranfield evaluation methodology and its derivatives use standard test collections, consisting of documents, topics, and judgments as to which documents are relevant to which topics. The IR systems to be evaluated are used to run the topics (formulated as queries) against the document corpus to produce a ranked list of documents or *run* for each topic. The relevance judgments then show which documents are relevant to the topic, and an evaluation metric is applied to the list to provide a score for the run. Run scores are aggregated, typically by taking the arithmetic mean, to produce an overall score for the system. The effectiveness of different systems can be compared by their aggregate system scores, and the statistical significance of any score difference assessed with a hypothesis test [Smucker et al., 2007].

Several evaluation metrics are widely used in system measurement, including precision-at-d (P@d), average precision (AP), and discounted cumulative gain (DCG). However, these measures do not allow the performance of a system tested on one collection to be readily compared with that of a system tested on a different collection. Variability in topic difficulty, and hence in average topic scores, means that a good score for one collection might be poor for another, depending on the mix of topics in the collection. Normalization by the score of an ideal ranking, as employed in AP and nDCG, can be viewed as a method for correcting for such variability, but it does so with quite limited success.

In this paper, we investigate the use of *score standardization* [Webber et al., 2007] to enable inter-collection score comparisons. Under standardization, the difficulty of a query is directly estimated from the scores achieved by a sample of experimental systems, and parameters derived from these estimates are then used to normalize the scores both of the experimental systems and of future systems. Standardization gives each topic the same score mean and standard deviation for the experimental systems, and reduces the effect of topic variability in the evaluation of new systems.

Standardization makes scores interpretable in themselves, and it becomes possible to directly compare scores measured on different topics and test collections. With standardization, for example, researchers with private collections could compare their results without exchange of data sets, through the use of common standardizing systems. Within a single collection, reduction in variability means that all topics contribute equally to measured differences in effectiveness. Across multiple collections, researchers could identify how consistent a system is in different environments.

To explore the validity of score standardization, we analyze the results for runs on several key TREC collections. Our results show that standardization leads to an average two-thirds reduction in the difference in scores achieved by an IR system on different collections, enabling different systems to be evaluated against different collections and still have their performance compared. Standardization is also more robust to differences in collection formation than existing normalization schemes. Standardization has several benefits and no obvious drawbacks, and we propose that standardization parameters be published with test collections to allow richer comparison and evaluation of systems than is currently possible.

SIGIR'08, July 20–24, 2008, Singapore.

2. METRICS AND SCORE VARIABILITY

Many evaluation metrics have been described in the literature [Buckley and Voorhees, 2005, Järvelin and Kekäläinen, 2002, Moffat and Zobel, to appear]. Most of these are based on precision and recall. *Precision* is the proportion of documents up to a specified depth (that is, ordinal rank) in a run that are relevant; *recall* is the proportion of all relevant documents that are returned. Due to the number of documents in current collections, only a subset of the documents can be assessed for relevance, and recall is computed based on the set of known relevant documents $\mathcal{R}$. The degree of incompleteness of $\mathcal{R}$ is, in general, unknown.

A simple evaluation metric is precision-at-d (P@d), which is the proportion of the top d documents in a run that are relevant. There is no adjustment for the number of relevant documents $R = |\mathcal{R}|$ for each particular topic, which can vary by a factor of a hundred or more. For instance, the maximum P@10 score that any run can receive for a topic with three relevant documents is 0.3, while a ceiling effect, in which most runs have P@10 at or near 1.0, can occur when there many easy-to-find relevant documents. A richer metric, which also lacks an R adjustment, is rank-biased precision (RBP) [Moffat and Zobel, to appear], in which the effectiveness score is a biased, bounded sum of relevance values. The higher the rank of a relevant document, the greater its contribution to the score, with the bias controlled by a parameter p.

Some metrics adjust for the number of documents relevant to a topic. One of these is R-precision (RP), which modifies P@d by setting d to the number of known relevant documents R for each topic, resulting in a relatively robust metric [Buckley and Voorhees, 2005]. A more complex metric is average precision (AP), which averages the precision of a run at each relevant document returned, assigning a precision of 0 to unreturned known relevant documents. The metrics RP and AP share the characteristic that a perfect ranking (one which places all known relevant documents at the top), and only a perfect ranking, achieves a score of 1.

Assigning a score of 1 to a perfect run can be achieved for other metrics, including those supporting multi-valued relevance judgments, by a process of explicit *normalization*, where a run's raw metric score is divided by the score that an ideal ranking would achieve, based on the set of known relevant documents $\mathcal{R}$ (and, for metrics supporting multi-valued relevance, their degree of relevance). The discounted cumulative gain metric (DCG) [Järvelin and Kekäläinen, 2002], which sums the relevance contributions of each rank, discounted by a logarithmically decaying weight, is normalized by dividing by the score of an ideal ranking to produce normalized DCG (nDCG). Such normalization can be applied to essentially any metric. In fact AP (though not RP) is a metric in this category, and can be considered as a raw metric, sum of precisions (SP), normalized by the number of relevant documents R [Aslam et al., 2006], since the SP score for a ranking with all R relevant documents at the top is itself R. Thus, AP is normalized SP (nSP). We refer to normalization by ideal ranking as $\mathcal{R}$*-normalization*.

A metric gives a score for a system's run against a topic. For an evaluation experiment, the run scores for the topics in the test collection and the systems participating in the experiment can be considered as a matrix, as illustrated in Figure 1, in which systems are rows, topics are columns, and higher scores are shown by lighter-shaded cells. Note that the easy topics (white vertical lines) stand out much more clearly than the good systems, although certain poor systems are distinct as horizontal black lines. Let M be a matrix of run scores such as that in Figure 1. The score for system s achieved on topic t is denoted as m_{st}. A system's score is the mean of its per-run scores, $\overline{M}_{s*}$. It is also interesting to consider the mean score of a topic, $\overline{M}_{*t}$. Similarly, one can consider

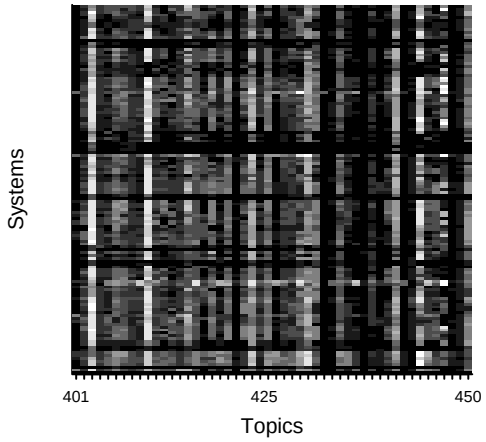

Figure 1: Per-run average precision scores for TREC 8 AdHoc Track systems. Systems are in rows, ASCII-ordered by system named; topics are in columns, ordered by topic number. Each cell represents the AP score for a system's run against a topic. Lighter shades represent higher scores.

the standard deviation of scores for a system, $\mathrm{sd}(M_{s*})$, and for a topic, $\mathrm{sd}(M_{*t})$.

Test theory, which originated in the assessment of the abilities of human subjects through examination, is built on the concept of a subject's "true" score, which the testing process is attempting to elicit [Bodoff and Li, 2007]. In information retrieval evaluation, a similar goal is attractive, that is, to be able to say that a system's true score is simply (say) 0.32, and to derive this "true" score (or a reliable estimate of it) with as little assessment effort as possible. Indeed, much contemporary IR evaluation is implicitly built on the assumption of a true score hiding behind the topic scores observed on a given collection. For instance, statistical significance tests such as the t-test implicitly ask how likely it is that the true mean metric scores of two systems are in fact the same, given the observed topic scores and assuming that they are a random sample of some larger population of topic scores.

However, such absolute interpretations of metric scores, which map from an achieved aggregate score in isolation to an evaluation of the performance of the system, are not possible with the metrics above, due to the high degree of variability in topic scores illustrated in Figure 1. The distribution of system scores depends heavily on which topics happen to be included in the collection, and is the reason why we always need to interpret scores in the context of a particular test collection. In particular, a system might achieve quite different scores on different test collections, and, even for retrieval contexts for which this collection is representative, might achieve a different score had a different set of topics been chosen. For example, Buckley [2005, p. 311] compares 8 successive versions of the SMART system on the first 8 TREC AdHoc collections, and, while the improvement in AP score of the latest over the earliest version on any single collection is of the order of 50% to 100%, the best collection AP score of the earliest (weakest) system is better than 4 of the collection scores of the latest (strongest) system. Even on the one collection, a system's score can only be recognized as good or bad by comparing it with the scores of other systems on the same collection. In fact, knowledge of a system's score is less useful than knowledge of its rank among the set of systems run against the collection.

Figure 2 illustrates the difficulty of assigning a meaning to an absolute AP score, even in the context of a single collection. The graph displays the 95% confidence intervals on mean AP scores for

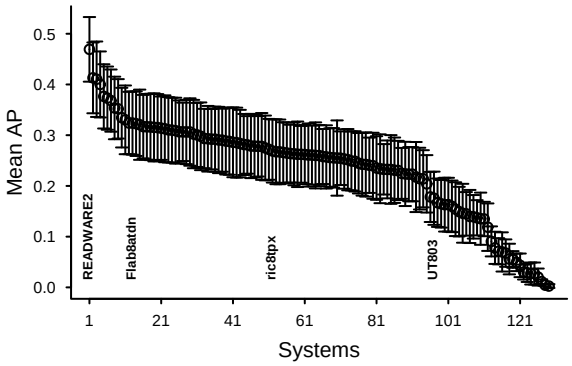

Figure 2: The 95% confidence intervals on mean AP scores for TREC 8 AdHoc Track systems, using a t distribution. Systems are ordered by their mean AP scores.

	System AP (M_{s*})			
	READWARE2	Flab8atdn	ric8tpx	UT803
mean	0.469	0.324	0.269	0.176
st.dev	0.224	0.220	0.219	0.185

	Topic AP (M_{*t})			
	q403	q430	q414	q401
mean	0.708	0.486	0.203	0.048
st.dev	0.226	0.217	0.108	0.090

Table 1: Mean and standard deviation of AP scores for the first, tenth, fortieth, and seventy-fifth percentile systems and topics by mean AP from TREC 8 AdHoc Track.

the 129 systems participating in the AdHoc Track of TREC 8. The fundamental assumption here, as with statistical significance testing, is that the observed scores have been randomly sampled from an underlying population of scores. Thus, for instance, the system Flab8atdn achieved an observed mean AP score of 0.324; but the best that can be said under the random sampling hypothesis is that (with 95% probability) the system's true mean AP score is between 0.262 and 0.386. Despite being the thirteenth-ranked system by observed score, its score range overlaps with those of the second-ranked and ninety-fourth-ranked systems. So, even assuming that the collection is perfectly representative of what it is intended to test, a mean AP score is, in isolation, not very informative.

The reason for the wide confidence interval on AP scores is the variability of per-topic scores. Table 1 lists AP score means and standard deviations for the first, tenth, fortieth, and seventy-fifth percentile systems and topics as ordered by mean AP. System standard deviations are remarkably similar, and the four system means range by a factor of only 2.5. Topics, on the other hand, are far more variable; the four means range by a factor of almost 15, and standard deviations by 2.5; ordered by standard deviation, the ratio from the first to the seventy-fifth percentile is almost 6. And this variability is despite AP's use of $\mathcal{R}$-normalization.

Even if system comparisons are on a single collection, and paired hypothesis tests are being used to help control the difference in topic score means, the wide variability in topic score standard deviations means that the comparisons may be less reliable than the test results suggest. In a paired hypothesis test, the computation is based on the score deltas between the two systems; but if the score standard deviation of one topic is 6 times that of another topic, then the average score delta will also be 6 times larger, meaning the higher-variance topic will have 6 times as much influence in system score deltas and paired hypothesis testing as the lower-variance topic. Nor are high-variance topics more reliable indicators of performance than low-variance topics. The Pearson's correlation between topic reliability, as measured by item-total correlation [Bodoff and Li, 2007], and topic AP standard deviation, considering the best 75% of TREC 8 AdHoc Track systems by mean AP, is only 0.005, indicating no meaningful correlation. Measured differences between systems are disproportionately due to a subset of the topics rather than to the topic set as a whole.

3. STANDARDIZATION

We propose a direct form of normalization, namely *standardization*. Score standardization is a well-known technique in tests applied to human subjects [Hays, 1991, chapter 4], but it has not to our knowledge been applied to IR evaluation. In standardization, topic scores are directly adjusted by the observed mean score and standard deviation for that topic on a sample of systems. If a topic t has a score mean of $\mu_t = \overline{M}_{*t}$ and a score standard deviation of $\sigma_t = \mathrm{sd}(M_{*t})$, and if a system s receives a score for that topic of m_{st}, then the standardized score m'_{st} for that run is:

$$m'_{st} = \frac{m_{st} - \mu_t}{\sigma_t} \tag{1}$$

The values μ_t and σ_t are the *standardization factors* for topic t. Such a score is known as a z *score*, and expresses how many standard deviations m_{st} is from the sample mean. As such, a standardized score is immediately informative in a way that an unstandardized one is not: one can tell directly from a run's score whether the system has performed well for the topic.

In a recent workshop paper [Webber et al., 2007], we explored the impact of standardization on an historical TREC collection. We found that, within the one collection and set of experimental systems, standardization evens out topic score variances, making individual run scores more meaningful. In this paper we build on and extend those results, and apply the techniques to the problem of practical inter-collection system comparisons.

Standardized z scores are centered on zero and unbounded, while most IR metrics are bounded in the range $[0, 1]$. To follow this practice, z-scores can to be mapped into the $[0, 1]$ range, with one attractive candidate being the cumulative density function of the standard normal distribution:

$$F_X(m') = \int_{-\infty}^{m'} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x \tag{2}$$

Normal-CDF-converted standardization is used throughout this paper, and is referred to simply as "standardization" from here on. Observe that, by design, a standardized score of 0.5 means "average", and 0.84 and 0.16 represent one standard deviation above and below average. Conversion to the range $[0, 1]$ also has the desirable property of reducing the influence of outlier data points, for instance when only one system finds relevant documents for a topic.

Figure 3 shows the distribution of per-run unstandardized and standardized AP scores for TREC 8 AdHoc Track systems. The raw AP scores are heavily skewed towards lower values, with almost half of the per-run scores being below 0.2. In contrast, the standardized scores are evenly distributed across the $[0, 1]$ range, with the 25th, 50th, and 75th percentiles being 0.26, 0.50, and 0.74.

Standardization of raw scores produces precisely the same score values as standardization of $\mathcal{R}$-normalized scores, since $\mathcal{R}$-normalization involves division by a per-topic constant factor, which identically scales topic mean and standard deviation. So, standardized

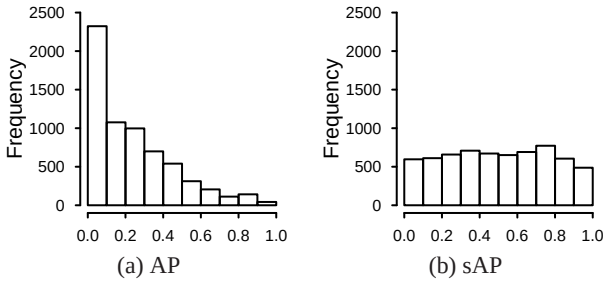

Figure 3: Distribution of the 6,450 per-run AP and standardized AP scores for TREC 8 AdHoc Track systems.

SP gives precisely the same values as standardized AP (nSP), and standardized DCG as standardized nDCG. Thus, metrics can be considered as raw, $\mathcal{R}$-normalized, and standardized.

The production of standardization factors requires the existence of experimental runs from which they can be calculated. However, since the current test collection creation methodology also requires such experimental systems to form the judgment pool, in practice the requirements for producing standardized scores are no higher than for producing unstandardized scores. Once standardization factors for each topic and metric have been determined from the experimental systems, they can be published along with the test collection, and used to standardize the scores of new systems run against the collection. How many experimental systems are required to derive reliable standardization factors, and how long these factors remain reliable in the face of changing (and hopefully improving) systems, is considered in the next section.

4. ROBUSTNESS OF STANDARDIZATION

Experimental data

The data used in this paper is the TREC 2004 Robust Track test collection and runs from the participating systems. The Robust Track of TREC is designed to examine and improve the consistency of information retrieval systems by attempting to predict difficult topics and emphasizing them in evaluation metrics. The document set is the AdHoc Track corpus, namely TREC disks 4 and 5, minus the *Congressional Record*. The TREC 2004 Robust Track topic set consists not only of 49 topics newly created for the task (a 50th was dropped when no relevant documents were found), but also the 50 topics from the TREC 2003 Robust task, and 150 topics from the AdHoc tracks of TREC 6 through TREC 8 (1997, 1998, and 1999). Relevance judgments for the earlier topic sets are reused, with new judgments being made only for the new topics. A total of 110 systems from 14 different groups participated, with participating systems submitting runs against both the 49 new topics and the 200 old ones. The runs submitted to the original experiments in which the older experimental sub-collections were created are also available. This data set therefore is well-suited for exploring questions of inter-collection comparability and the durability of standardization factors. However, we exclude the TREC 6 AdHoc sub-collection as the lack of topic title keywords from many topic descriptions, an issue unique to this sub-collection, causes anomalous performance from description-only runs.

Longevity of standardization factors

Standardization factors for a collection are calculated based on the results of the *standardizing systems*, that is, the systems that contributed to the original experiment. These standardization factors are then used to standardize the scores of new systems being evaluated against the collection. Over time, as systems improve, the standardization factors may become out of date. This can have two effects. First, comparisons within a collection can become inaccurate, as the relative difficulty of topics may change. Second, standardizations on different collections may be based on experimental systems of different intrinsic quality. The second issue is considered in Section 5, the first here.

	Collection		
	T7.adh	T8.adh	T03.rob
Pearson	0.998	0.998	0.998
Kendall	0.969	0.962	0.970

Table 2: Pearson's correlation of scores and Kendall's τ rank correlation using standardized AP for TREC 2004 Robust Track systems on each of the earlier sub-collections, comparing the effect of standardizing based on the original experimental systems and standardizing based on the TREC 2004 Robust Track systems.

Table 2 gives the Pearson's correlation for system scores and the Kendall's τ correlation for system rankings for the TREC 2004 Robust systems on each of the earlier sub-collections, comparing in each case the results obtained by standardizing using the original experimental systems and standardizing using the TREC 2004 Robust systems. Note that the Pearson and Kendall's τ correlation coefficients work on different scales and so cannot be directly compared to each other. The Kendall's τ should be compared with the 0.742 correlation for ranking the TREC 2004 systems based on the TREC 2003 versus the TREC 2004 topics; the Pearson's coefficients should be compared with the 0.943 correlation on scores between the two topic sets. Clearly, standardization using relatively outdated systems is much less distorting than comparisons between different test collections.

Estimation of standardization factors

The standardization factors derived from the standardizing systems can be thought of as estimates of the true factors for the topic. A key question is how many systems are required in order to get reliable standardization factors, and, in particular, what effect reducing the number of standardizing systems has upon system ranking. Here, the benchmark is the ranking obtained from standardization factors derived from the full experimental set. The following experiments use the relevance judgments from the full judgment pool, limiting the number of participating systems only when calculating standardization factors.

The test uses as standardizing systems the TREC 2003 Robust Track systems, and the test collection is the topics created for that track. The evaluated systems are the TREC 2004 Robust Track systems as run against the TREC 2003 topics. The procedure is to sample from the standardizing systems, derive standardization factors from the sample, use these to standardize the scores of the evaluated systems, and calculate the Kendall's τ between the system ranking from the sampled standardization and from the full standardization. This is repeated multiple times for each sample size. The 50th (median), 95th and 99th percentile lowest Kendall's τ figures are recorded. The whole process is then repeated for other sample sizes.

Figure 4 reports the Kendall's τ for varying sample sets, giving median values and lower-end percentiles, using standardized AP. In comparison, the Kendall's τ on system rankings on the TREC 2004 Robust systems between the TREC 2003 and the TREC 2004 topics using unstandardized AP is 0.742, and between the unstandardized and standardized AP scores for the TREC 2003

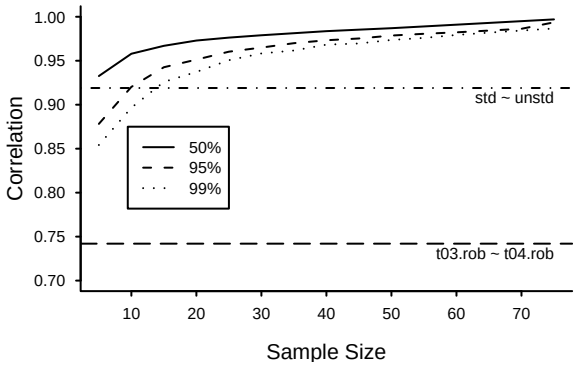

Figure 4: Percentiles of Kendall's τ between rankings on partial and full standardization system sets, using standardized AP. The test collection and standardization systems are from the Robust Track of TREC 2003; the evaluated systems are those submitted to the Robust Track of TREC 2004, considering only their runs against the TREC 2003 topics. There were 2,000 random samples made for each sample size. The full standardization set has 78 systems. The Kendall's τ between the TREC 2003 and TREC 2004 sub-collections on unstandardized AP, and between unstandardized and standardized AP on TREC 2003, are also shown.

sub-collection is 0.919. Even taking the 99th lowest percentile, as few as 5 of the 78 systems need to be sampled for standardization factors that give more consistent results than inter-collection comparisons, while 10 to 15 systems are sufficient to better the correlation with unstandardized scores at the 95th and 99th percentiles. A small set of systems, therefore, is sufficient to provide standardization factors that give reliable system rankings, far smaller than is needed to provide the relevance judgments.

5. CROSS-COLLECTION COMPARISONS

Comparability of identically sampled collections

In investigating the question of cross-collection comparability, two kinds of collections need to be considered. The first is collections that we know to be drawn from the same population under the random sampling hypothesis. By definition, significance tests between two such collections are statistically valid, if it is understood that their results are being extended only to other samples of this population. The second is collections where it cannot be assumed that they have been randomly sampled from the same underlying population, that is, where there may be factors that cause one collection to be significantly different from another.

If we use random sampling, then the sampled values will behave as an independent and identically distributed variable, and the theoretical basis of hypothesis testing will be met. Such randomly-sampled collection pairs can be approximated by randomly sampling from the topics of an existing collection, or set of collections. Any set of collections can be used and still, via random sampling, be considered identically sampled, but it is preferable to choose collections that are relatively homogeneous. Here, the 100 topics from the AdHoc tracks of TREC 7 and TREC 8, Topics 351–450, are used, and the runs are those made by the TREC 2004 Robust Track systems. The topics (and associated runs) are randomly partitioned into two halves to form two randomly sampled collections. (The fact that we are sampling from such a small population, without replacement, means that the assumption of independence is violated, but the results are adequate for our current purposes.) The random partitioning is repeated multiple times to generate a set of identically-sampled collections.

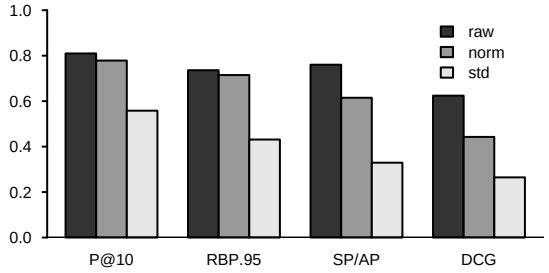

Figure 5: Mean standard-deviation normalized root mean square error (dRMSE) for TREC 2004 Robust Track systems on 1,000 random partitionings of Topics 351–450 from the TREC 7 and TREC 8 AdHoc tracks, for different metrics, with and without normalization and standardization. Standardization factors are derived from the original TREC 7 and TREC 8 systems.

Score comparability between collections means that, if we run the same system against two collections, it should receive similar scores for each collection. Here, "similar" can be understood in the loose sense of producing aggregate system scores that are not too different; or, more narrowly, as producing sets of topic scores that are not found to be significantly different under statistical testing.

System score comparability can be measured using root mean squared error (RMSE). Continuing the notation of Equation 1, let S be our set of evaluated systems. Consider two collections, C and D. Let $\overline{M^C}_{s*}$ be the score under some metric that system $s \in S$ achieves on collection C (that is, the mean of the scores that s achieved on the topics making up C), and similarly for $\overline{M^D}_{s*}$. Then the root mean squared error between C and D is:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{s \in S}(\overline{M^C}_{s*} - \overline{M^D}_{s*})^2}{|S|}} \tag{3}$$

The RMSE is dependent upon the magnitude of the score values for a metric; if scores for one metric are precisely ten times the scores for another, then the RMSE will be ten times greater, even though comparability is effectively the same. To facilitate comparisons between different metrics we normalize by dividing by the average standard deviation of system scores for each collection, to derive standard-deviation normalized root mean square error or dRMSE:

$$\mathrm{dRMSE} = \frac{2 \cdot \mathrm{RMSE}}{\mathrm{sd}(\{\overline{M^C}_{s*} : s \in S\}) + \mathrm{sd}(\{\overline{M^D}_{s*} : s \in S\})} \tag{4}$$

Note that normalizing by the geometric rather than the arithmetic mean of the two standard deviations produces almost identical results in practice.

Randomized topic set re-sampling can be used to derive distributions of dRMSE figures for different metrics. Figure 5 gives the results of multiple random partitionings of the TREC 7 and TREC 8 AdHoc topics. The metrics P@10, RBP with persistence $p = 0.95$, SP (unnormalized AP), and DCG are compared, together with their $\mathcal{R}$-normalized and standardized versions. The results show that every metric with standardization is more stable than all metrics in their raw form. And standardization leads to significantly greater stability than $\mathcal{R}$-normalization, even on identically-sampled collections. (As will be seen later, normalization is far less robust to differently-sampled collections.)

A second form of collection comparability is finding statistically significant differences. If the same system is tested on two different collections, then the results on the two collections should

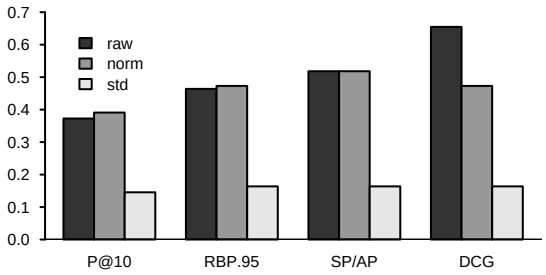

Figure 6: The 97.5th highest percentile false positive rates for various metrics, with different forms of normalization. A false positive is a finding that a system is significantly different from itself using a two-tailed, two-sample t-test at level $\alpha = 0.05$. False positive rates are calculated for the 110 TREC 2004 Robust Track systems, by 2,000 random re-partitionings of Topics 351–450 from the TREC 7 and TREC 8 AdHoc tracks. Standardization factors are derived from the original TREC 7 and TREC 8 systems.

not be found to be significantly different; if they are, then that is a false positive, or at least the collections are not comparable, since obviously a system is not significantly different from itself. The false positive rate for a metric on two collections therefore is taken from the number of systems found to be significantly different from themselves on the two collections. The significance test employed here is a two-tailed, two-sample t-test, at significance level $\alpha = 0.05$.

Randomized topic set re-sampling can also be used to test the false positive rate. Figure 6 gives the upper end of the 95% confidence interval on false positive rates for the TREC 2004 Robust Track systems over the TREC 7 and TREC 8 AdHoc Track topics. Due to random sampling, the mean false positive rates for every metric and form of normalization are close to the significance level of 0.05 (they range from 0.042 to 0.052). By looking at the upper end of the confidence interval, we are instead examining a reasonable upper bound on how high the false positive rate can go when comparing two (identically sampled) test collections. SP and DCG have higher discriminative power than RBP or P@10, so the fact that they have higher potential false positive rates is not surprising. However, standardization enormously decreases the upper-end false positive rates, from around 50% to just over 15%. This is achieved without harming discriminative power. For instance, for the TREC 8 sub-collection, the proportion of system pairs found significantly different on a two-tailed, paired t-test at level $\alpha = 0.05$ is 68.7% for DCG, 69.3% for nDCG, and 68.8% for sDCG. Normalization by $\mathcal{R}$, in contrast, does little to improve false positive rates. That is to say, even where the hypothesis of random sampling from an underlying population is observed (as is the case here), use of standardized metrics rather than $\mathcal{R}$-normalized metrics leads to far more reliable inter-collection comparisons.

Comparability between distinct collections

Examination of inter-collection metric comparability between two identically-sampled collections is a best-case situation, where the statistical equivalence of the collections is artificially created. In practice, different collections are not identically sampled. However, the AdHoc and Robust TREC collections use the same document corpus and were built with similar methodologies, so comparability between them would be desirable. We now explore the comparability of metrics in these circumstances, and the effect of $\mathcal{R}$-normalization and standardization on this comparability.

Table 3 shows the dRMSE of system AP scores for each pair of collections used in the TREC 2004 Robust Track. The two AdHoc collections are relatively close to each other, as are the two Robust collections. For instance, the observed dRMSE of 0.63 between the TREC 7 and TREC 8 AdHoc collections is close to the mean randomized dRMSE over these topics of 0.60 reported in Figure 5, indicating that from the perspective of this statistic the two collections are not significantly different. However, comparisons between any of the AdHoc and any of the Robust collections are problematic. The observed dRMSE of 1.857 between the TREC 7 and TREC 2003 collections, for example, compares with the mean randomized dRMSE across those two collections of 0.596, and in fact falls beyond the 99th percentile of randomized values, meaning that the two collections are highly significantly different for this statistic when using AP. Table 4, which gives false positive rates, also indicates severe problems. Almost all systems seem significantly better than themselves when evaluated using AP against the TREC 2003 collection than when evaluated against the TREC 7 collection, and again this false positive rate is beyond the 99th percentile of randomized values.

	T8.adh	T03.rob	T04.rob
T7.adh	0.627	1.857	1.285
T8.adh		1.387	0.859
T03.rob			0.583

Table 3: Standard-deviation normalized root mean square error for system AP scores between each pair of collections in the TREC 2004 Robust set for all systems participating in the track.

	T7.adh	T8.adh	T03.rob	T04.rob
T7.adh		0	0	0
T8.adh	2		0	0
T03.rob	103	57		2
T04.rob	61	8	0	

Table 4: Number of the 110 TREC 2004 Robust Track systems that were found to be significantly better when tested on the sub-collection in the row than on the sub-collection in the column, using unstandardized mean AP. Significance is determined by a two-sample, one-tailed t-test, at level $\alpha = 0.025$.

Table 5 gives the inter-collection dRMSE of standardized SP/AP scores. As anticipated from Figure 5, the standardized scores have a much lower dRMSE for every collection pair than do the $\mathcal{R}$-normalized AP scores in Table 3. More particularly, the dRMSE figures are similar for every collection pair. The observed dRMSE figures for standardized AP are well within the 95% confidence interval found by randomization, and in fact sit quite close to the respective means, indicating that, for dRMSE with standardized AP, the collections are not significantly different. The false positive rates (not tabulated for space reasons) are also much improved, averaging 5% and not exceeding 11% for any collection pair, with no strong effect between AdHoc and Robust collections.

Figure 7 gives the mean dRMSE scores for various metrics, in their raw, $\mathcal{R}$-normalized, and standardized forms. The value of 1.1 in the middle bar of the SP/AP group, for instance, is the mean of the six values reported in Table 3. Note that these means include both the two same-track pairs and the four different-track (Robust-to-AdHoc) pairs; if only the latter were included, the results would be even less flattering to $\mathcal{R}$-normalization. Standardization moderately improves RBP's observed cross-collection comparability, and, unexpectedly, marginally worsens that for P@10. However, the improvements for SP/AP and DCG are dramatic, even from their $\mathcal{R}$-normalized forms.

	T8.adh	T03.rob	T04.rob
T7.adh	0.320	0.342	0.373
T8.adh		0.406	0.397
T03.rob			0.398

Table 5: Standard-deviation normalized root mean square error for system standardized AP scores between each pair of collections in the TREC 2004 Robust set for all systems participating in the track. Standardization factors are derived from the original experiments.

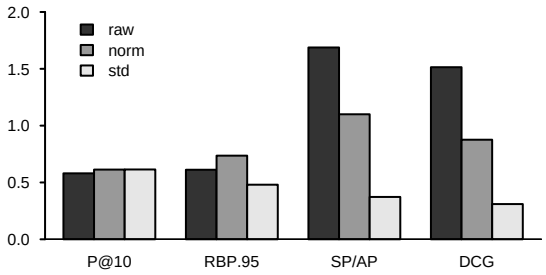

Figure 7: Mean standard-deviation normalized root mean square error (dRMSE) for TREC 2004 Robust Track systems between each pair of the TREC 7 AdHoc, TREC 8 AdHoc, TREC 2003 Robust, and TREC 2004 Robust collections, for various metrics, without and with standardization. Standardization is performed based on the original experimental systems.

The $\mathcal{R}$-normalized metrics are even less comparable between the Robust and AdHoc collections than for identically sampled collections because of differences in the constitution of the set of known relevant documents $\mathcal{R}$. Both Robust and AdHoc judgment pools were formed by pooling to depth 100 (depth 125 for TREC 2003), but the number of participant groups and therefore pooled systems was quite different, with 42 and 41 systems pooled for the two AdHoc collections and only 16 and 14 for the Robust ones. Moreover, the AdHoc tracks included a large number of manual runs, identifying around 25% of the known relevant documents, whereas the Robust tracks had none. The consequences can be seen in Table 6. The average number of known relevant documents per topic is greater for the AdHoc than for the Robust collections. The Robust topics are not harder than the AdHoc ones, with the TREC 2004 Robust systems receiving very similar average system P@10 (and also RBP, not shown) scores in each of the four test environments. However, the $\mathcal{R}$-normalized metrics such as AP (and nDCG, not shown) are misled by the smaller values of $\mathcal{R}$ in the two Robust test environments into thinking their topics are harder, and the corresponding normalized scores are higher than for the AdHoc test environments. Conversely, SP (and DCG, not shown), being non-convergent metrics that evaluate deep in the runs, give higher average scores to the sub-collections with more known relevant documents. Standardization, shown in the last row, is not affected by the changes in $\mathcal{R}$. Note that, as one would hope, slightly improved sAP scores are calculated for the TREC 2004 Robust systems when they are standardized using the original systems' scores.

The conclusion of these experiments is clear: although (or perhaps because) it sets out to adjust scores to reflect the weight of relevance for a topic, $\mathcal{R}$-normalization is in fact very sensitive to variability in the way in which the set of known relevant documents is determined. In contrast, standardization is robust to such differences, making collections with significantly different $\mathcal{R}$ formations comparable in the same way that identically sampled ones are. And even where $\mathcal{R}$ estimates are compatible, standardization offers far greater comparability, as the randomized tests predicted.

	T7.adh	T8.adh	T03.rob	T04.rob
Judged	1606.9	1736.6	958.7	710.0
Relevant	93.5	94.6	33.2	42.1
P@10	0.452	0.450	0.466	0.434
AP	0.212	0.244	0.327	0.293
SP	17.88	20.29	9.71	11.61
sAP	0.516	0.503	0.517	0.500

Table 6: Mean number of documents judged and mean number of documents found to be relevant for the different sub-collections of the TREC 2004 Robust collection, and mean P@10, AP, SP, and sAP scores for the TREC 2004 Robust Track systems run against each sub-collection.

6. PREVIOUS WORK

Average precision was developed in the context of TREC [Buckley and Voorhees, 2005]. Although it has been widely used for over a decade, there is no definitive paper describing the metric, and it has only recently been analyzed in the literature. Discounted cumulative gain and its variants are described in Järvelin and Kekäläinen [2002]. Rank-biased precision is described by Moffat and Zobel [to appear].

Determining the quality of a metric can easily become a circular problem: a good metric is one that highly ranks good systems, but how do we know what the good systems are without first using a metric to judge them? A common approach is to examine the statistical features of metrics. Buckley and Voorhees [2000] and Sanderson and Zobel [2005] calculate the error rate of a metric by randomly partitioning a topic set and counting the number of times the resulting subsets order system pairs differently; metrics with lower error rates are regarded as more stable and therefore better. Similarly, Sakai [2006] suggests that the sensitivity of a metric be determined by the proportion of system pairs found to be significantly different under an hypothesis test; he proposes the bootstrap test for this purpose. Aslam et al. [2005] propose that the quality of a metric can be determined by using a maximum entropy analysis: the more constraints that a given metric score places upon the possible rankings it could have been derived from, the more information that metric provides, and hence the better it is.

An alternative approach to assessing evaluation metrics is to examine how well they correlate with user experience. Huffman and Hochster [2007] found that reported satisfaction of assessors correlates fairly strongly with relevance among the top three documents or even simply the very top-ranked document; however, their experiments used professional assessors attempting to interpret the information needs and satisfaction of the users who submitted the sampled queries. In contrast, Al-Maskari et al. [2007], working with users judging their own satisfaction, found only weak correlation between most metrics and user satisfaction. Rather than self-satisfaction, Turpin and Scholer [2006] gave users two specific tasks: find a single relevant document in the least time; and find as many relevant documents as possible in five minutes. Turpin and Scholer found no significant correlation between the average AP score of a system and user performance on the first (precision) task, and only a weak correlation on the second (recall) task.

It is one thing to determine that system A has scored higher than system B on a given collection and metric; it is another to confirm that this difference in scores is significant. Zobel [1998] examines the use of the t-test, ANOVA, and Wilcoxon test, and finds that the t-test and Wilcoxon diverge. Savoy [1997] examines the theoretical basis of hypothesis testing in the IR environment, and proposes the use of the bootstrap hypothesis test. Smucker et al. [2007] propose the randomized permutation test as requiring less assumptions

about data distribution and sampling. They demonstrate that the t-test and Bootstrap tests give almost identical results, with the randomization test being similar, but that the Wilcoxon test diverges.

Bodoff and Li [2007] suggest that collections be viewed less as random samples from an underlying population, and more as purposefully created tests, similar to tests that might be applied to students. They then introduce ideas from test theory such as the reliability of individual test components, including individual topics.

Zobel [1998] normalizes metric scores by dividing a run's score by the highest score achieved by any run for that topic; this is done primarily to improve the comparability of scores achieved by different topics. Järvelin and Kekäläinen [2002] propose that scores should be normalized, not by the highest scores achieved, but by the highest score achievable, given the known distribution of relevance. Mizzaro and Robertson [2007] normalize per-run scores, either by topic or system, by subtracting the mean observed score for that topic or run; they do not, however, adjust for variance.

The high degree of variance in topic score distribution and by implication topic difficulty has been widely commented on. Using ANOVA techniques, Tague-Sutcliffe and Blustein [1994] observe that the topic effect is much stronger than the system effect; that is, there is more variation between topic scores than between system scores.

To our knowledge, comparing systems on disparate collections has not been systematically explored, although the practical results of Buckley [2005] indicate the difficulty of doing this with AP.

7. CONCLUSION

Accurate measurement is integral to improvement in all fields of science. Having measures that are reproducible, comparable, and immediately interpretable would enormously facilitate the identification and acceptance of advancements in the discipline. The evaluation metrics currently in use, however, do not provide these characteristics. Instead, experimental results for one system can only be interpreted by explicit comparison with other systems, and system comparison can only meaningfully be pursued by testing all systems on the one collection, something that is always inconvenient and often impossible. Worse, the existing normalization methods, reliant as they are upon an inevitably incomplete sample of the set of relevant documents for each topic, can exacerbate the problem of non-comparability between different collections, if the different collections have had different relevance assessment inputs. In contrast, standardization greatly increases the ability to compare system results within and between test collections, and allows for wide differences in performance to be immediately detected from aggregate scores, without the need to exhaustively test all systems on the one collection.

Acknowledgment. This work was supported by the Australian Research Council.

Standardization factors for common TREC collections and metrics can be found at:

`http://www.csse.unimelb.edu.au/~alistair/ir_eval/`

References

A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In Clarke et al. [2007], pages 773–774.

J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In Marchionini et al. [2005], pages 27–34.

J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In Dumais et al. [2006], pages 541–548.

D. Bodoff and P. Li. Test theory for assessing IR test collections. In Clarke et al. [2007], pages 367–374.

C. Buckley. The SMART project at TREC. In Voorhees and Harman [2005], chapter 13.

C. Buckley and E. Voorhees. Retrieval system evaluation. In Voorhees and Harman [2005], chapter 3.

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

C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors. *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Amsterdam, July 2007.

S. Dumais, E. N. Efthimiadis, D. Hawking, and K. Järvelin, editors. *Proc. 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Seattle, Washington, August 2006.

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

S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In Clarke et al. [2007], pages 567–574.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors. *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Salvador, Brazil, August 2005.

S. Mizzaro and S. Robertson. HITS hits TREC: exploring IR evaluation results with network analysis. In Clarke et al. [2007], pages 479–486.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, to appear.

T. Sakai. Evaluating evaluation metrics based on the bootstrap. In Dumais et al. [2006], pages 525–532.

M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In Marchionini et al. [2005], pages 162–169.

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

M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In M. J. Silva, A. A. F. Laender, R. Baeza-Yates, D. L. McGuinness, B. Olstad, Ø. H. Olsen, and A. O. Falcão, editors, *Proc. 16th ACM Int. Conf. on Information and Knowledge Management*, pages 623–632, Lisboa, Portugal, November 2007.

J. Tague-Sutcliffe and J. Blustein. A statistical analysis of the TREC-3 data. In D. K. Harman, editor, *Proc. TREC-3*, pages 385–398, November 1994. NIST Special Publication 500-225.

A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In Dumais et al. [2006], pages 11–18.

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

W. Webber, A. Moffat, and J. Zobel. Score standardization for robust comparison of retrieval systems. In M. Wu, A. Turpin, and A. Spink, editors, *Proc. 12th Australasian Document Computing Symposium*, pages 1–8, Melbourne, December 2007.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998.