Corpus Bootstrapping for Assessment of the Properties of Effectiveness Measures

Justin Zobel
The University of Melbourne, Parkville, Australia

Lida Rashidi
The University of Melbourne, Parkville, Australia

ABSTRACT

Bootstrapping is an established tool for examining the behaviour of offline information retrieval (IR) experiments, where it has primarily been used to assess statistical significance and the robustness of significance tests. In this work we consider how bootstrapping can be used to assess the reliability of effectiveness measures for experimental IR. We use bootstrapping of the corpus of documents rather than, as in most prior work, the set of queries. We demonstrate that bootstrapping can provide new insights into the behaviour of effectiveness measures: the precision of the measurement of a system for a query can be quantified; some measures are more consistent than others; rankings of systems on a test corpus likewise have a precision (or uncertainty) that can be quantified; and, in experiments with limited volumes of relevance judgements, measures can be wildly different in terms of reliability and precision. Our results show that the uncertainty in measurement and ranking of system performance can be substantial and thus our approach to corpus bootstrapping provides a key tool for helping experimenters to choose measures and understand reported outcomes.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Test collections**.

KEYWORDS

measurement, bootstrap, corpus properties, experimental design

ACM Reference Format:
Justin Zobel and Lida Rashidi. 2020. Corpus Bootstrapping for Assessment of the Properties of Effectiveness Measures. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411998

1 INTRODUCTION

Research on and development of ranking techniques for information retrieval (IR) requires reliable methods for measuring their effectiveness. Such measurement has been a key factor in the development of ranking, and is also used as a component of methods for learning to rank. There are many tools for this task, referred to here as *effectiveness measures*, including average precision (AP), reciprocal rank (RR), rank-biased precision (RBP), precision at k documents (P@k), normalised discounted cumulative gain (nDCG), inverse-square rank (INSQ), and many others [6–9, 13, 15].

A cause of measurement accuracy, or uncertainty, is that offline IR experiments are often conducted over a single document collection (or *corpus*) and set of queries, or just a few query-set–corpus combinations. Measurements depend on the composition of the corpus and set of queries, and variations in that composition leads to changes in the measured outcomes. That is, while the measurements may be accurate on a specific corpus, they are not necessarily representative of performance overall. A method that has been proposed in past work for exploring accuracy is *bootstrapping*, which smooths a limited set of observations to give estimates of uncertainty. The core idea is that the set is treated as a proxy for the population that the observations represent, with a large number of new, simulated sets created by sampling with replacement from the original. Prior work in IR has primarily used bootstrapping to examine the sensitivity of significance tests. In particular, several researchers have examined bootstrapping of the set of queries [1, 12, 17, 18, 20]. This provides estimates of the variability of the system score.

In this paper we use bootstrapping for a new purpose: examination of the behaviour of effectiveness measures. To do so we *bootstrap the document corpus*, an approach taken to date in only two papers [2, 11]. Our experiments show that corpus bootstrapping provides insight into the properties of measures:

- It provides an indication of the uncertainty (or precision) of measurements of individual systems on individual queries, as well as the uncertainty across queries or across systems.
- System scores tend to vary jointly, as removal or addition of relevant documents leads to correlated changes to measurements. However, even so system rankings are variable, with wide bounds on rank position amongst similarly scored systems, particularly for the weaker measures.
- Under bootstrapping the different effectiveness measures behave in very different ways. In particular, per-query results for AP and nDCG have a wide range of variances. While this is in principle to be expected, given the use of recall as a denominator, the extent of the uncertainty in these measures had not previously been quantified.
- The behaviour of P@10 and RR is poor.
- In a case-study, the shortcomings of AP are acute if the volume of relevance judgements is small, as can be the case in experiments designed to compare two systems, with the counter-intuitive effect of increasing overestimation of statistical confidence when the number of judgements is reduced.

Overall, our results demonstrate that bootstrapping can be an insightful tool for assessing the accuracy of IR experiments, and for assessing measurement tools themselves. One finding is that a confidence bound can be ascribed to the measurement of a system on

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411998

a query for any given effectiveness measure, as well as to system performance overall; per-query comparisons have not had significant prior investigation. Another finding is that, in a case study of small numbers of judgements, RBP was reliable, while on the same instances measurements with AP were not just inconsistent but highly misleading, showing that it should not be used unless deep assessment is undertaken.

2 BACKGROUND

Experiments in IR typically make use of a single collection and multiple queries (or of a small, non-random set of collection–query-set combinations) to provide measurements over a selection of members of the population of queries; a single query is highly unreliable as an indicator of the relative performance of two systems. The distribution of measurements can then be used to estimate whether observed differences are likely to reflect real underlying difference in behaviour or are due to chance. It is for this reason that statistical significance tests are used, to give confidence in which systems will be best on unseen data.

The measurements taken in these experiments are imprecise. First, relevance judgements are unreliable, as illustrated by the phenomenon of assessor disagreement [21]. Second, the judgements are incomplete. In the TREC experimental design, the only documents judged are those in the top 100 returned for a query by the participating systems. Pool-based judgement avoids bias amongst the systems that contributed to the pool, but may penalise new systems being examined in subsequent experiments if they retrieve relevant documents that are unjudged [25]. We do not examine these issues here; our focus is on the uncertainty that is present even when these other causes of imprecision are eliminated.

Bootstrapping. Bootstrapping is a methodology in which a set of observations is repeatedly sampled with replacement to generate an estimate of variance [3, 4, 17]. At each iteration, the bootstrap generates a sample in which some of the original values are duplicated and others are missing; the mean of the values is recorded. The result is a distribution of the means observed over a series of iterations. Bootstrapping has been formally shown to replicate aspects of repeated sampling of the true population from which the original set of measurements was drawn. The median of the bootstraps cannot significantly differ from the original mean (after a sufficient number of iterations), and thus these faux samples do not fully replace population re-sampling. However, bootstrapping relies less on assumptions such as normality about the original data and distribution than do other, more analytical tools.

In this paper, we refer to the original data set as the *root* and to the data set generated at each bootstrap iteration as an *image*. We use S to denote the number of systems participating in an experiment, Q the number of queries in the query set, and D the number of documents in a corpus.

Bootstrapping has been applied to the results of IR experiments in several ways. Sakai [12] and Savoy [17] used bootstrapping of query sets. In this approach, for a set of systems, a query set, and a -corpus, each query is run by each system on the corpus, giving S sets of Q runs each, that is, one set of runs per system. Here, a *run* is the list of documents returned by a system for a query; some of the documents in the run will be annotated as relevant to the query. An effectiveness measure is then applied to the run to give a score. In this approach, bootstrapping involves generating a series of images from the set of queries. The mean score for a system is the average of the per-query scores, noting again that duplicates are not removed; a query that is present multiple times in the image contributes that number of times to the average. Sakai and Savoy showed that bootstrapping can be a useful tool for system comparison as well as for understanding measurement sensitivity, but found, using AP, that a paired t-test is in principle not valid for comparing IR systems due to its underlying assumption that measurement-per-query follows a normal distribution.

Other prior work was also designed to examine sensitivity and significance. Sanderson and Zobel [16] sampled the query set without replacement. Smucker et al. [18] examined query set bootstrapping, randomization testing, and significance testing, finding that all mechanisms were similar at small p-values. Carterette [1] used bootstrapping of AP values to create normality and thus satisfy the conditions of the t-test, while examining adjustment of p-values to account for the multiple comparison problem. Urbano [20] experimentally evaluated approaches to significance and sensitivity, enumerating the assumptions made by analytical significance tests or underpinning sensitivity: normality, homoscedasticity, decorrelation, and uniformity. He found that query-set bootstrapping was similar in practice to other statistical methods such as the t-test, and more effective than ad-hoc methods such as swap rates.

Corpus bootstrapping. Cormack and Lyman [2] proposed a different approach, corpus bootstrapping, in which each image is a different set of documents. After the image has been generated, each run must be amended: documents that are duplicated in the image are correspondingly duplicated in the modified run, while missing documents are removed. That is, a root run

$$d_1, d_2, d_3, d_4, d_5, d_6, \ldots$$

might be amended to give the image run

$$d_1, d_2, d_2, d_2, d_3, d_6, d_6, \ldots$$

if d_2 was present three times in the image, d_6 was present twice, and d_4 and d_5 were absent. A relevant document will be marked as such in the modified run each time it occurs.

In this approach, for each system and query there is a set of measurements, one for each image. As for bootstrapping of query sets, for each image an average score can be computed for each system across the set of queries, and this set of averages provides an estimate of the accuracy of the original system score. A skeptic might wonder whether bootstrapping of corpuses is effective, as the number of items (documents) is so great that any image will closely resemble the original. However, the number of relevant documents per query is small, so there can be a wide range of effectiveness results per system per query across images.

Cormack and Lyman [2] used their method of corpus bootstrapping to examine how well a system's results would generalise to another corpus from the same population, showing that sampling half of a corpus could predict performance on the other half. Due to the behaviour of AP when the number of relevant documents is small, some of the analysis considers only samples with more than 5 relevant documents for a query. Their analysis is based on AP only and is focused on predictivity.

Robertson and Kanoulas [11] proposed a variation on this work, based on sampling the corpus. Again focusing on AP, and predictivity on individual queries, they found that the per-query variance is similar for different systems, and that their method is no more informative than a significance test. However, their experimental methodology appears to involve constructing an image by separately sampling the sets of relevant and non-relevant documents with respect to a specific query, with the aim of ensuring that for a system and query the number of relevant documents has a defined distribution across images, thus limiting potential variance. Forming the images in this stratified way is not obviously consistent with the principles of bootstrapping.

Effectiveness measures. Experiments in IR require tools for measurement of performance. Many have been developed, and critiqued [15]. Here, our focus is on tools that have been commonly used in IR, including in bootstrap experiments: AP, INSQ, nDCG, P@10, RBP, and RR. There have been criticisms of some of these measures [5, 8, 24].[1] For example, it has previously been observed that AP is mathematically unsuited to arithmetic averaging [10, 19, 22], and that the implicit user model is that the number of relevant documents is known in advance. However, they continue to be widely used in practice.

Note that much of the prior work on bootstrapping cited above, whether of query sets or corpuses, primarily uses only AP as a measure, so that comparison of measures is not a focus.

The question of whether observed differences in performance are robust is often addressed with significance tests. The reliability of these tests has had extensive prior investigation and is beyond the scope of our paper.

3 CORPUS BOOTSTRAPPING

Bootstrapping quantifies variation in measurements over a set of observations. In the context of IR, this helps us to interpret measurements of system behaviour, with the aim of anticipating future performance: which of two systems is expected to have greater effectiveness on a new query and new corpus.

Prior approaches are based on intuitive interpretations of the data. Sampling of queries reflects that there is a large potential population of queries for a given data set, even if the queries for which we have judgements are in some sense curated, that is, chosen by the designer of the experiment [23]. The document sets are also curated. For example, a typical collection of newspaper articles represents the material chosen by an editorial team for publication. These are not random decisions, and in practice is both selective – some material is omitted – and repetitive – a publisher with multiple outlets may or may not include multiple variants of the same article in the corpus. Bootstrap construction of images from documents corpuses could be interpreted as reflecting the ways in which a corpus might vary under different hypothetical editorial practices. It removes the assumption that either the query set or the document set is canonical. While a set of 50 queries is only small, however, and cannot be representative of user interactions, the document collections typically used in IR experiments tend to be comprehensive, such as all articles published in a particular outlet over a period.

A potential compromise is to simulate a corpus per query by sampling with replacement from the pool of judgements for that query. For a sufficiently large pool, such as those of the collections we use in this paper, this is in one respect mathematically defensible: for a given query, images will have the same distribution of numbers of relevant documents as will images of the whole corpus. However, especially for judgement pools constructed from a small number of systems, this approach is biased: omission of documents contributed by one system will mean that documents contributed by other systems must by construction be over-represented. We do not use this compromise.

We make no attempt to ensure that every query has relevant documents in every image; doing so would introduce bias into the results and violate the principles of bootstrapping. Where a query has no relevant documents, we assign an effectiveness score of 0 for each of the tools we use. We regard this decision as unproblematic. The user cannot know at query time whether a particular collection will contain useful material; and a consistent assignment of 0 does not discriminate between systems.

There is a potential confound in our experiments. The ordering of documents in a run is determined by scores computed on the root data. However, the different composition of an image could lead to different term frequency statistics, and thus evaluation of the query on the image might give a different ordering. We do not regard this as a concern in practice. First, with collections of this size the changes in term statistics are expected to be slight. Second, a strength of our experiments is that they are based on a wide range of diverse systems and, as these were the systems used to create the judgement pools, all of the documents in the first 100 positions of the root runs are judged. We cannot re-run these systems on new images, and experiments with the small set of systems to which we have access would not create the same levels of diversity. That is, any variability observed in our measurements is expected to understate the actual variability, while the high quality of data we are using provides robustness.

4 EXPERIMENTAL DESIGN

We have applied bootstrapping to several Text REtrieval Conference (TREC)[2] data sets: TREC6, TREC7, TREC8, and TREC9, and, in preliminary work, TREC3 (for which we do not show results). The experiments used for development of this work were primarily on TREC3 and TREC8. However, we observed similar trends and behaviours on all five data sets. Details of the corpora are in Table 1.

To ensure that the experiments completed in reasonable time, in most instances we restricted our attention to the top 50 systems in each collection in descending order of average RBP score on the root data (which in general was closely correlated to the order on other effectiveness measures). We refer to individual systems by their ordinal number, so for example 'system #12' refers to the 12th-best system as measured in this way.

In each experiment we generated 100 images except where indicated otherwise. In some of the preliminary experiments we used much larger numbers of iterations, as is usual in bootstrapping,

[1] Our focus here is on measures for binary relevance judgements, and thus measures for and experiments with graded judgements are out of scope.

[2] https://trec.nist.gov/

Table 1: Data used in experiments. The final column is the proportion of documents in the collection that are in the assessment pool for any query.

Identifier	Track	Systems $\|S\|$	Topics $\|Q\|$	Corpus $\|D\|$	Assessed docs (%)
TREC6	AdHoc	74	50	556,077	9.7
TREC7	AdHoc	103	50	528,155	11.5
TREC8	AdHoc	129	50	528,155	12.4
TREC9	Web	104	50	1,692,096	3.5

but this did not lead to materially changed outcomes. The wide range of experiments here and the cost of each iteration meant that consistent use of numbers larger than 100 was infeasible.

The measures used are average precision (AP), inverse-square rank (INSQ), normalized discounted cumulative gain (nDCG), precision at k documents returned (P@k), rank-biased precision (RBP), and reciprocal rank (RR). For INSQ, the anticipated number of documents is 5. For nDCG, we used a bound of 1000. For P@k, we used k of 10. For RBP, we used persistence of 0.95 and a bound of 1000. Our choice of measures was determined by their familiarity to the community and by the fact that they have different characteristics; our work is not designed to find which measure has best sensitivity and so on, but to explore and demonstrate how bootstrapping can be used to understand the properties of measures.

Validity of corpus bootstrapping

As a preliminary step, we sought to confirm that corpus bootstrapping is a robust methodology. Bootstrapping is usually described as over a set of observations, rather than over the data from which the values are computed; however, the sampling of the underlying data, if uniform, arguably meets the theoretical grounds for bootstrapping as well as does sampling of the resulting values. To confirm this, we designed a simple predictive experiment to see if bootstrapping had the expected distributive properties.

For the top 50 systems on each of our test collections we generated 100 corpus bootstraps and used the first 99 to measure the performance of each system using each effectiveness measure. For each pair of systems and each query we calculated the difference in performance on that query, and thus could estimate the 95% confidence intervals for system-system-query 3-tuples. This gives a 95% confidence interval for the remaining, 100th bootstrap sample; that is, we identified the range of values that, with 95% probability, should contain the 100th measurement under an i.i.d. sample. If the estimated intervals are unbiased, we would expect about 2.5% of the values to fall above or below the intervals, noting that this is an easy case for prediction: the same systems, with the same queries, on similarly composed corpuses.

Table 2 reports the proportion of values calculated for each of the 61,250 system-system-query triples that fall above, within, or below the 95% confidence intervals. As can be seen, the distributions are an accurate model of observed results, giving confidence in the validity of corpus bootstrapping.

5 PER-QUERY MEASUREMENT

Our first experiment uses bootstraps to examine precision of results. For each of the top 50 systems on each of the 50 queries we calculated the standard deviation in observed performance for that system-query pair. We observed that the distributions of the standard deviation are approximately normal for RBP and INSQ, but highly skew for the other measures, with many small values but some very large. Figure 1 shows the 95% percentile of the standard deviations, per query, for the top 50 systems in each TREC dataset; RR is omitted because the values were much larger than the others. INSQ and RBP are not only normal per query but also, as can be seen, consistent from query to query, while P@10 was (like RR) consistently large.

AP and nDCG, the two measures that use recall, both show wide variation in accuracy per query.[3] Arguably the unreliability of AP and nDCG at a per-query level is unsurprising, as it is caused in part by variations in the denominator, but it has not previously been quantified, and the extent to which they were variable was unanticipated. As shown in Table 3, the fact that there are many queries where the scores are in a range of $\mu \pm 0.10$ should be of concern in experiments where much smaller differences are routinely treated as a priori significant.

Table 3 shows the standard deviation of system differences per query over 100 bootstraps. We can see that AP, nDCG, and RBP are broadly consistent in precision of measurement of difference, with RBP having the most consistency in precision overall. The precision of the other measures is lower, with RR having many instances of error of 0.5 or higher, showing that it is close to meaningless. This is despite the fact that systems tend to vary in tandem, that is, a particular image happens to under-represent the root population of relevant documents for a given query, then the scores of many (though not all) of the systems on that query will fall in a similar way relative to the root.

We now explore how bootstrapping can provide a mechanism for examining how measurements of individual queries are distributed across a set of images. The usual assumption is that per-query results are uninteresting, as it is per-system averages that characterise system performance. However, these distributions do reveal properties of the tools being used for effectiveness measurement.

Results for all six measures on TREC9 are shown in Figure 2, for system #3.[4] (Plots for the other top systems were similar, though the queries with high variation were not always the same; likewise, similar behaviour was observed on the other collections.) This figure illustrates the dramatic differences in behaviour of the measures. P@10 is quantised to intervals of 0.1. RR is straightforwardly unusable. RBP and INSQ show good uniformity, with similarly scaled quantiles across all queries except those with very low median scores. However, AP and nDCG showed poor uniformity with numerous instances of high standard deviation, implying much lower

[3] In other experiments, not reported here, we found that this was a complex effect; we had anticipated a single cause such as 'small number of relevant documents' but this was not the case.

[4] As an aside, Figure 2 raises questions with the concept of an 'easy' or 'difficult' query. There are queries that are consistently low-scored under one measure that are not outliers for some other measures, and in some instances a low-scored mean can be in the presence of a reasonable level of variation. Thus the measurement tool itself can affect the variation for a given query and difficulty is not necessarily an inherent property. We have not investigated this behaviour systematically.

Table 2: Proportion (as a percentage, %) of system-system-query 3-tuples that fall above, in, or below the estimated 95% confidence interval. Note that the confidence intervals vary significantly between measures.

Measure	TREC6			TREC7			TREC8			TREC9		
	Below	In	Above	Below	In	Above	Below	In	Above	Below	In	Above
AP	2.6	94.2	3.2	2.5	94.7	2.8	1.8	96.3	1.9	2.4	94.5	3.1
INSQ@5	2.4	94.6	3.0	2.5	94.6	2.9	2.7	94.4	2.9	2.3	94.7	3.03
nDCG	2.7	94.6	2.7	3.2	94.1	2.7	2.4	95.1	2.5	2.3	94.9	2.8
P@10	3.0	93.9	3.1	3.0	94.0	3.0	2.1	95.1	2.7	2.0	95.2	2.8
RBP@0.95	2.4	94.2	3.4	2.7	94.8	2.5	1.4	96.9	1.7	2.3	94.5	3.2
RR	2.2	95.4	2.4	2.6	94.4	3.0	2.5	94.3	3.2	2.8	95.0	2.2

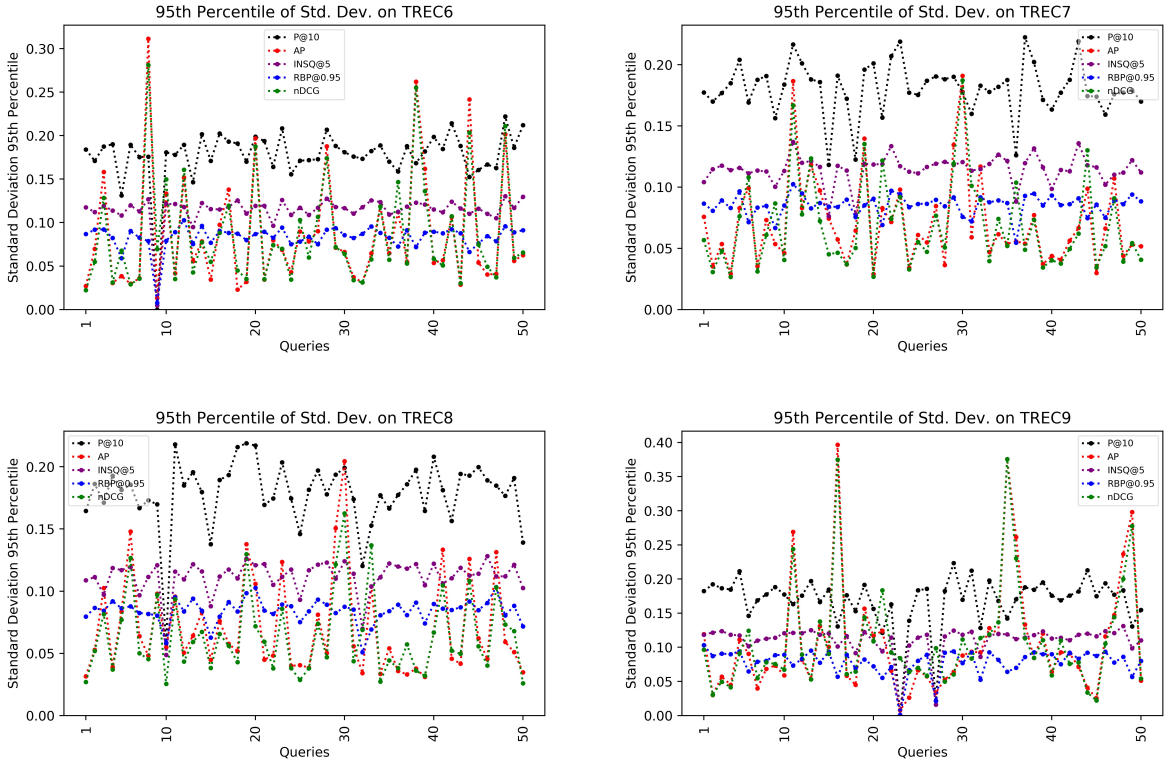

Figure 1: The 95% percentile of the standard deviations of scores for top 50 systems across the queries in TREC6, TREC7, TREC8, and TREC9, as measured by five different tools; parameters as explained in the text.

confidence in the measured result. (For nDCG, the bound used was 1000; bounds of 10, 20, and 50 gave much worse outcomes.)

Some overall results are in Table 4, in this case averages per system rather than per query, for several typical systems. These systems are all among the top 10 performers systems in their TREC, and their behaviour is like that of similarly ranked systems. As can be seen, RBP and INSQ are relatively consistent, while the other measures, including AP and nDCG, are more variable. The most stable measure in each case is highlighted in bold. RBP was the most stable measure in all cases except for the TREC8 system, where AP, nDCG, and RBP were close in terms of variability.

Note that the consistent distribution of RBP values highlights the fact that run images are reasonably stable, despite the fact that we place no constraint on the numbers of relevant documents.

These wide variations in observed behaviour mean that the per-query measured values are arguably incommensurate, implying that averages across sets of queries are not meaningful. It has been argued on mathematical grounds that AP is unsuitable for averaging [5] (though there is disagreement [14]); the same arguments apply to nDCG. The broad range of variances implies that activities such as, for example, optimisation via machine learning may

Table 3: Standard deviations of system-system-query 3-tuples.The reported values are the mean and standard deviation of the observed standard deviation across system differences per query, reported as $\mu_\sigma \pm \sigma_\sigma$.

Measure	TREC6	TREC7	TREC8	TREC9
AP	0.058 ± 0.054	0.047 ± 0.030	0.046 ± 0.035	0.067 ± 0.064
INSQ@5	0.084 ± 0.037	0.083 ± 0.034	0.078 ± 0.035	0.076 ± 0.037
nDCG	0.062 ± 0.044	0.049 ± 0.028	0.045 ± 0.028	0.071 ± 0.050
P@10	0.140 ± 0.060	0.148 ± 0.054	0.144 ± 0.056	0.131 ± 0.066
RBP@0.95	0.060 ± 0.026	0.060 ± 0.023	0.057 ± 0.024	0.051 ± 0.027
RR	0.220 ± 0.144	0.210 ± 0.142	0.201 ± 0.145	0.230 ± 0.140

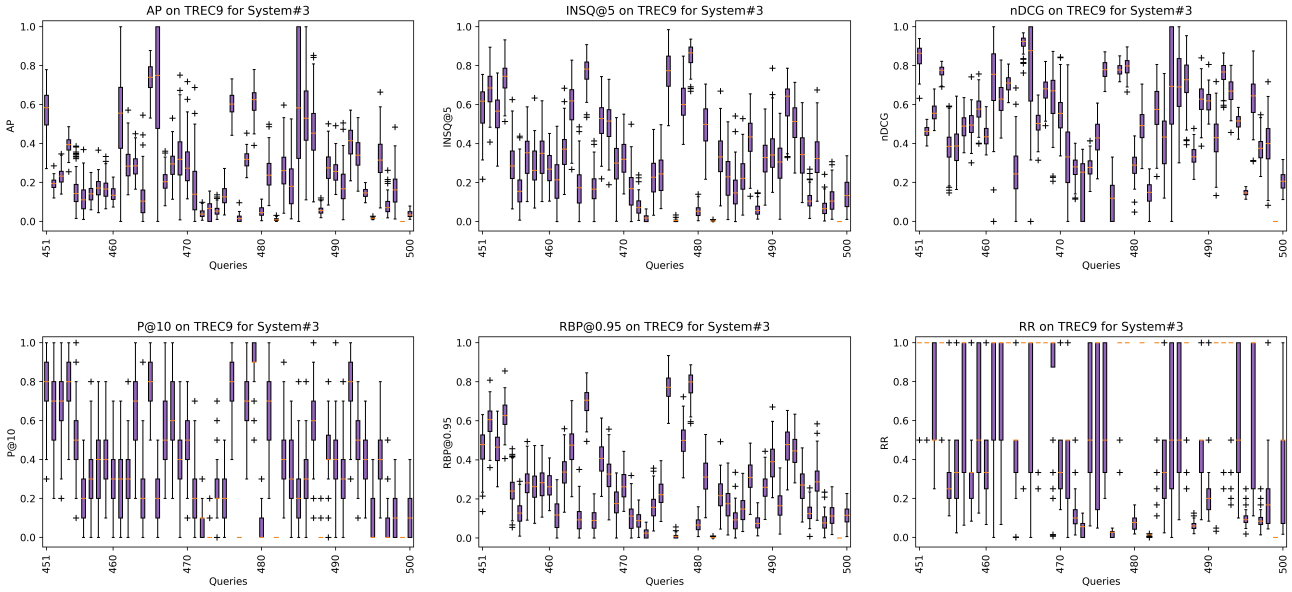

Figure 2: Distribution of scores for individual queries on TREC9 for system #3; parameters as explained in the text.

Table 4: Per-query standard deviations in query performance for typical individual systems. The reported values are, for that system, the mean and standard deviation of the observed standard deviation across the queries, reported as $\mu_\sigma \pm \sigma_\sigma$.

Measure	TREC6 System #7	TREC7 System #6	TREC8 System #2	TREC9 System #5
AP	0.07 ± 0.06	0.06 ± 0.04	0.06 ± 0.04	0.08 ± 0.07
INSQ@5	0.09 ± 0.03	0.08 ± 0.03	0.08 ± 0.03	0.08 ± 0.03
nDCG	0.07 ± 0.05	0.06 ± 0.03	**0.06 ± 0.03**	0.08 ± 0.06
P@10	0.13 ± 0.05	0.12 ± 0.06	0.12 ± 0.06	0.13 ± 0.05
RBP	**0.07 ± 0.02**	**0.07 ± 0.01**	0.07 ± 0.02	**0.06 ± 0.02**
RR	0.17 ± 0.13	0.12 ± 0.13	0.10 ± 0.12	0.18 ± 0.13

be driven by only a subset of the queries, potentially leading to undesirable outcomes.

Some of the prior work discussed earlier involved use of mechanisms to create normality; or structured the sampling to limit variance; or concluded that certain significance tests should not be used because their conditions are violated. Our results show that these concerns may be a consequence of the measures used in their experiments and are not properties of the systems or collections. With a different measure, the rationale for some of this prior work is weakened. Likewise, the relative uniformity of the distributions for RBP and INSQ, compared to the non-uniformity for other measures, illustrates that the varying discrimination power of individual queries – the rationale for other work – is at least in part due to the measure.

6 VARIABILITY IN PERFORMANCE

We now use corpus bootstrapping to examine the extent to which measured overall system results depend on the composition of the

corpus. Figure 3 shows variation in score per system on TREC9 for each measure; interestingly, these variations are much smaller than under query-set bootstrapping (results not shown), but do show that some measures operate in narrower bands than others.[5] Results were very similar on the other collections.

Figure 4 shows the standard deviation of observed average score for the top 25 systems in each collection. These results show the variation in precision of the different measures, with, broadly, AP, INSQ, nDCG, and RBP all in a similar band but with some variation between collections.

The distribution of per-query scores shown earlier, coupled with the broad score-confidence bands implied here for whole systems, shows that reporting of more than 2 decimal places of precision cannot be justified, for experiments on small numbers of queries such as 50. Most of these scores are only known to within a range of around 0.05 to 0.10. In the IR literature it is common to see tables of results showing system differences of 0.01 or less. While it is usually acknowledged that such small differences may not be significant, our results strongly indicate that they are not meaningful, and they should only be reported if contextualised by estimates of error.

Another perspective on relative performance is shown in Figure 5, which depicts variation in system ranks as boxplots of quartiles. On each bootstrap, the average system scores give an ordering and thus each system has a rank: first, second, and so on. As these orderings vary between bootstraps, there is a distribution of rank positions. (Note that the apparent stability of RBP is to a small extent an artefact of the fact that it was used to determine system order.) The different measures have different variabilities, and thus the median rank or original rank can only be regarded as indicative. RR, P@10, and to some extent AP show a relatively wide range of variation in rank, while RBP, INSQ, and nDCG relatively stable.

7 BEHAVIOUR WITH SHALLOW POOLS

We examine the behaviour of AP and RBP when the volume of relevance judgements is small. In the well-resourced TREC experiments, and specifically in the collections we have used, for each query all documents ranked in the top 100 for any system are placed in a *pool*, and all documents in the pool are judged.

However, for a novel collection being used in an individual experiment, it is prohibitive to do so many judgements, and indeed unnecessary if the aim is just to compare two systems. Instead, the researcher might use just the two systems being compared to create the pool, and to contain costs might also limit the pool depth, on the assumption that scores calculated by effectiveness measures are largely determined by the documents early in the ranking.

As a case study, consider the top two systems in TREC9. When they are compared using RBP and the full judgements, there is a score difference of about 0.07 and with a t-test a very small p-value, of about 0.001. AP likewise gives a score difference of about 0.07, and a p-value of about 0.028.

We used bootstrapping to generate 100 images and for each query and each of the two systems took the top-ranked 10 and 25 documents respectively to create small pools, then re-computed the score differences between the systems and the p-value; and likewise for each bootstrap computed the score and the p-value with the full judgements (notated as 'depth 100' in the figures). Results are shown in Figure 6, in which the red lines are the original score differences and p-values. As can be seen in the rightmost graphs, for the full collection the bootstrapped scores and p-values are somewhat clustered around the original, though AP shows much wider spread in p-value than does RBP.

For the small pools, however, very different behaviour emerges. RBP scores and p-values are much the same as on the full judgements. The p-values for RBP in Figure 6 range between 0.00 and 0.03 for all pool depths, excluding a few outliers. For AP, however, the score differences are exaggerated as the number of judgements is reduced, and worse, the p-values fall, showing greater confidence in the result. The p-values for the pool depths 10, 25, and 100 range from 0.00 to 0.03, 0.00 to 0.06, and 0.00 to 0.30 respectively, showing a 10-times decrease in p-values on the small pool in comparison to the full judgement pool. This is only a single instance, but we have observed similar behaviour for AP in a good number of other system pairs on TREC9 and the other collections; this example was chosen as typical amongst those we inspected. These instances demonstrate that results with AP with small numbers of judgements can be highly unreliable.

Figure 7 shows the distribution of differences in scores between the two systems, per query. This shows that the distribution of differences for RBP is more or less independent of the depth of the pool, with only minimal change as more information is available. For AP, as the volume of judgements is reduced the distribution of scores dramatically widens, so that even with pools of depth 25 the spread of score differences is many times greater than on the full judgements. Again, it is clear that AP should not be used unless deep judgements have been made.

8 CONCLUSIONS

We have shown that corpus bootstrapping can be used to compare the behaviour of measurement tools and to produce error bounds on scores, even to the level of individual queries. This is the first use of bootstrapping to comparatively assess the accuracy and variability of effectiveness measures. The previous use of bootstrapping in IR research has been to establish statistical significance, assess sensitivity, and interrogate other significance tests.

Most of this previous work used bootstrapping of the set of queries; we have used the less common approach of bootstrapping the corpus. In contrast to the limitations in some prior work, including sampling bias and selection of data to be bootstrapped, our sampling is oblivious to pools, queries, and judgements.

We examined multiple measurement tools with the aim of comparing their efficacy.[6] Past criticism has focused on principles [5] and the lack of independence in the measures [24]. Our results further underline their imprecision and the risks inherent when interpreting results without estimation of confidence; even if other sources of error are removed, such as judgement inconsistencies, only two digits of precision are justified for these collections.

[5] Recall that, as a consequence of the properties of bootstrapping (which is not as powerful as true sampling of the underlying population), the means of the observed average system scores from the bootstraps, per system, tend to the average system score on the root data – with the exception of the poorly formed RR and to a lesser extent nDCG. It does not imply that the original scores were precise.

[6] Which is distinct from the common, dubious practice of 'metric shopping', in which the results of several tools are reported as if they were independent.

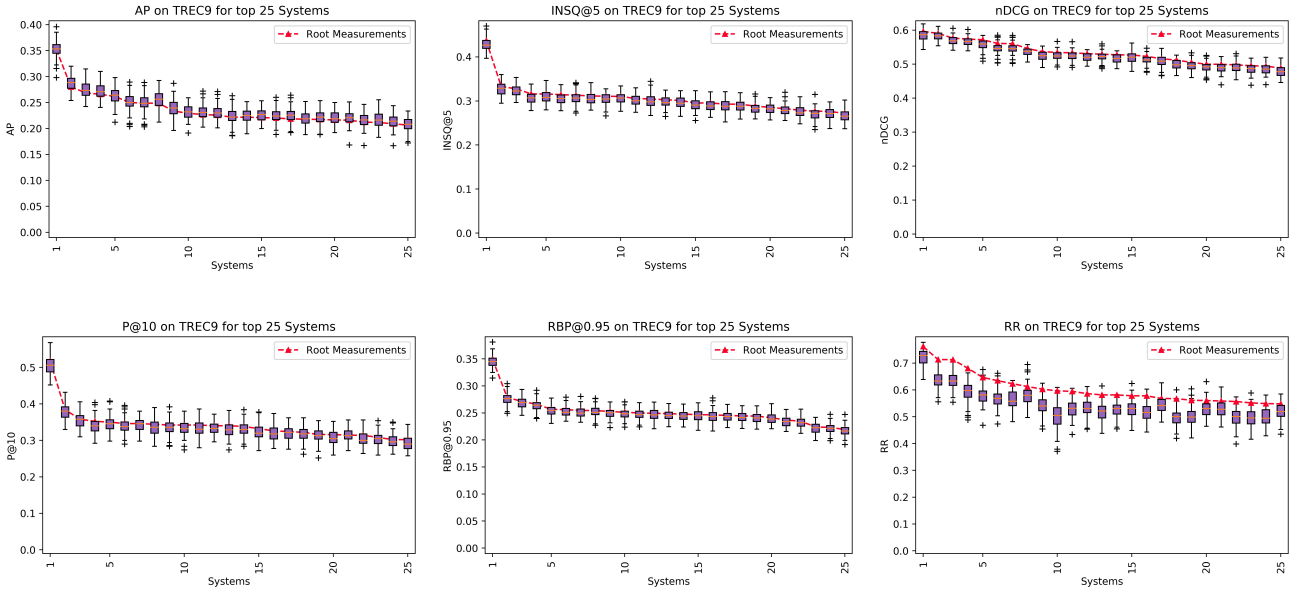

Figure 3: Quartiles of RBP, nDCG, AP, RR, INSQ, and P@10 scores for the top 25 systems on TREC9.

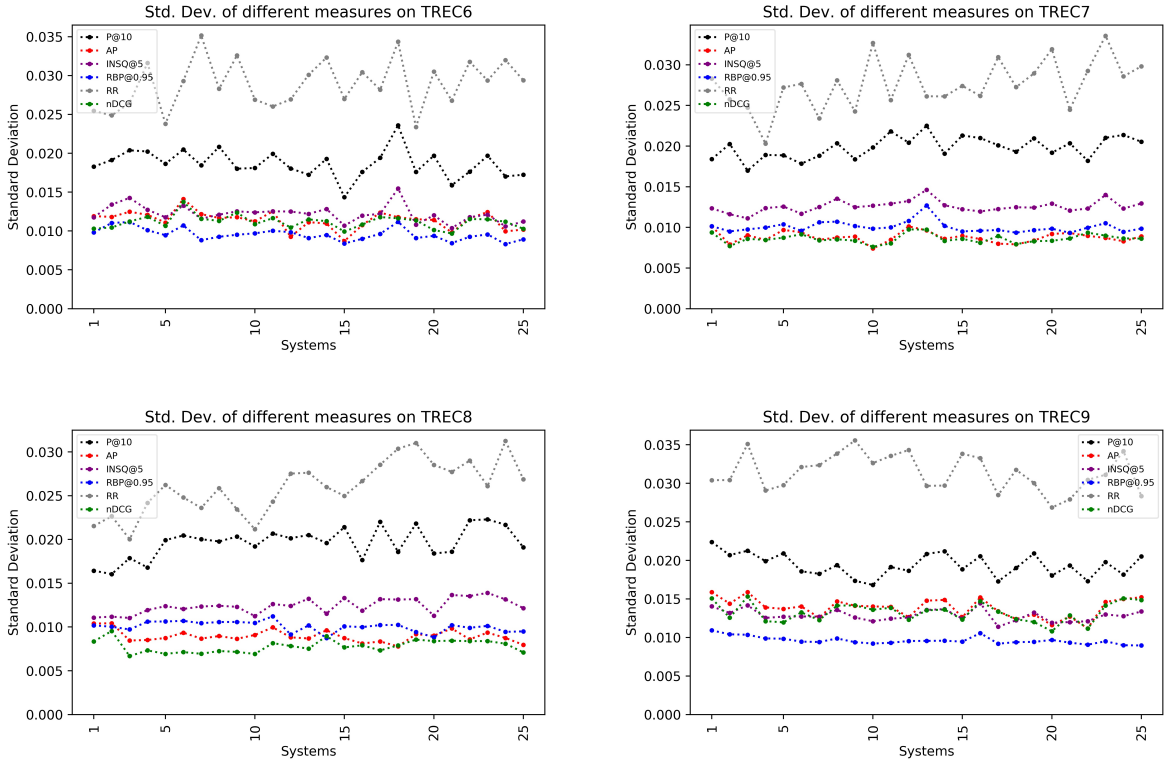

Figure 4: Standard deviations of measured scores for the top 25 systems on TREC6, TREC7, TREC8, and TREC9.

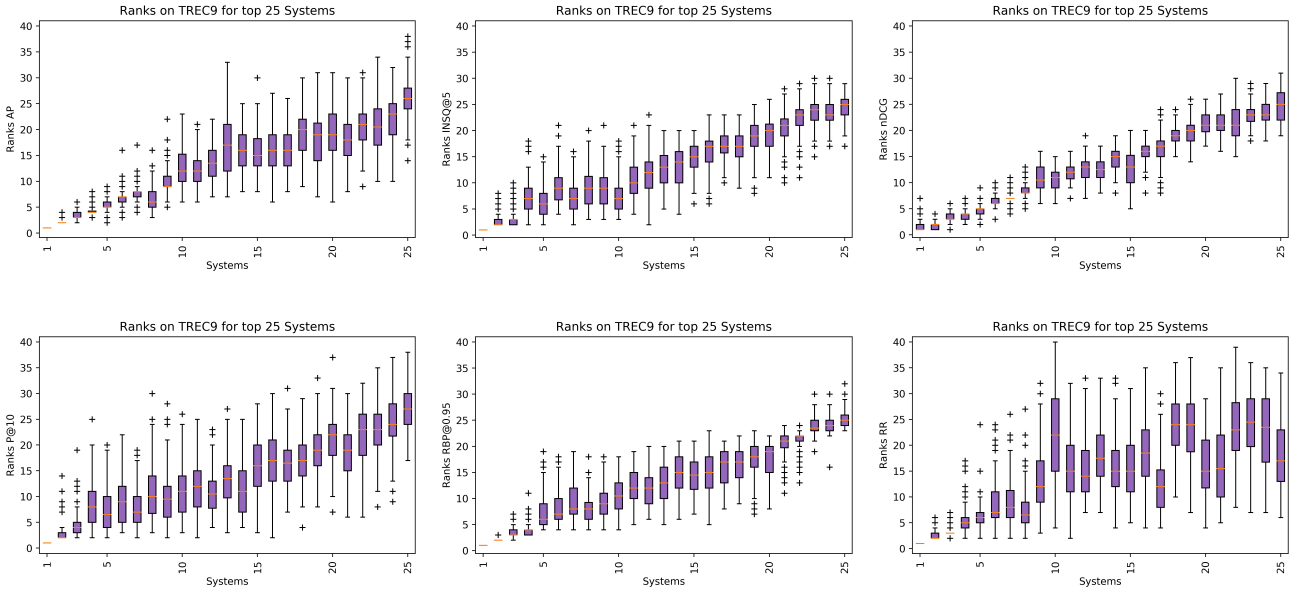

Figure 5: Variations in ranks for the top 25 systems on TREC9.

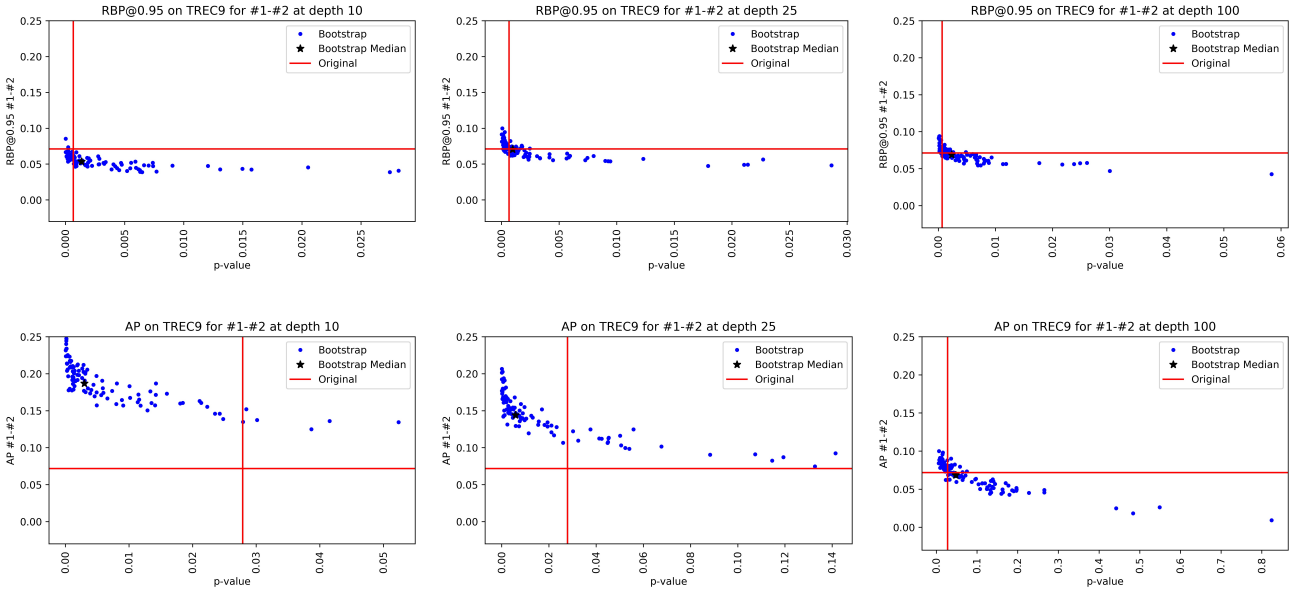

Figure 6: Distribution of p-values vs the score differences between the top two systems for the bootstraps on TREC9.

Our experiments show that, unsurprisingly, P@10 and RR are poor. Relatively, RBP (strictly, RBP@0.95) is accurate and consistent across all the datasets we considered, as to a lesser extent is INSQ. (We used INSQ@5 throughout, but note that INSQ@10, which we did not evaluate, should be similar in behaviour to RBP@0.95.

AP and nDCG (strictly, nDCG@1000), although widely relied on, do not have these properties; of particular concern is that AP can produce results that are strongly misleading if the volume of relevance judgements is small. We have further observed that some of the concerns in prior work about non-normality of scores potentially invalidating significance tests is in fact (arguably, at least) due to poor choice of effectiveness measure, thus removing the rationale for that work.

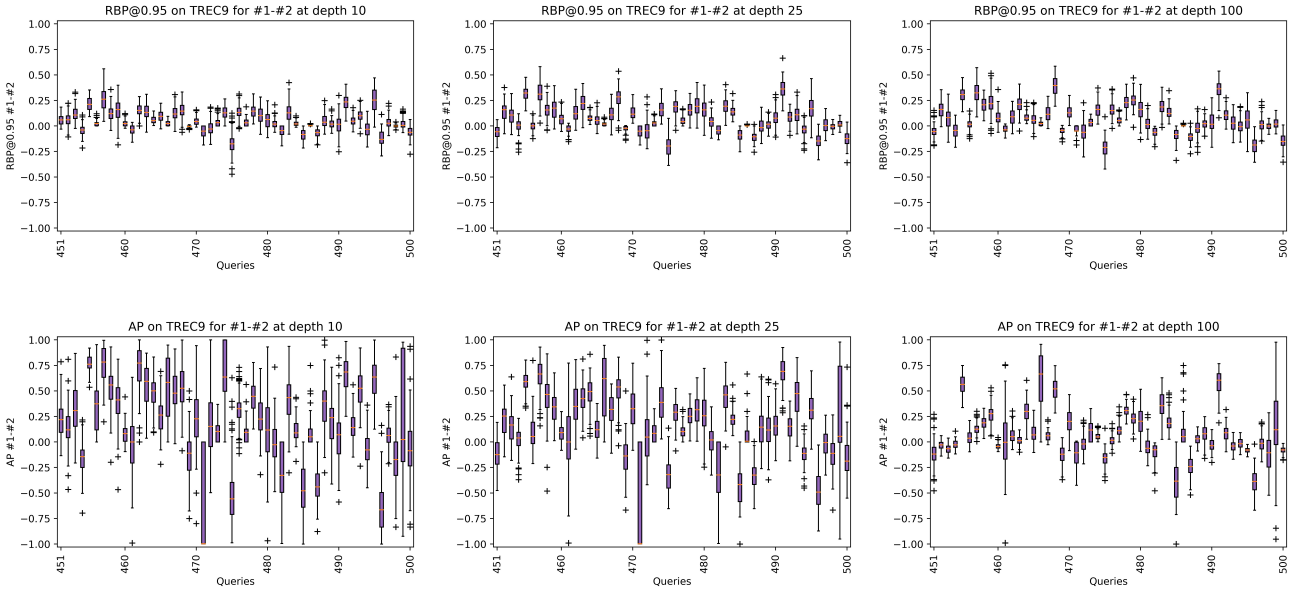

Figure 7: Distribution of score differences between the top two systems per query for the bootstraps on TREC9.

Considering further work, we have observed that the properties of the measurement tools remain in some respects poorly understood; for example, AP is imprecise and not well founded, yet in other work and our own preliminary experiments does perform well in predicting future performance, a combination that seems paradoxical. Also, in initial experiments we have observed that pairs of systems whose scores lie outside each other's confidence bounds are consistently found to be significantly different by a t-test. We plan to explore the robustness of this relationship, as it would allow significance to be determined even by researchers who do not have the access to the experimental runs or per-query measures of performance.

REFERENCES

[1] B. A. Carterette. 2012. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. on Information Systems (TOIS)* 30, 1 (2012), 4.

[2] G. V. Cormack and T. R. Lynam. 2006. Statistical precision of information retrieval evaluation. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 533–540.

[3] B. Efron and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* (1986), 54–75.

[4] B. Efron and R. J. Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

[5] N. Fuhr. 2018. Some common mistakes in IR evaluation, and how they can be avoided. In *ACM SIGIR Forum*, Vol. 51. ACM, 32–41.

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

[7] A. Moffat, F. Scholer, and P. Thomas. 2012. Models and metrics: IR evaluation as a user process. In *Proceedings of the 17th Australasian Document Computing Symposium*. 47–54.

[8] A. Moffat and J. Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. on Information Systems (TOIS)* 27, 1 (2008).

[9] C. J. Van Rijsbergen. 1979. *Information Retrieval*.

[10] S. Robertson. 2006. On GMAP: and other transformations. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM, 78–83.

[11] S. E. Robertson and E. Kanoulas. 2012. On per-topic variance in IR evaluation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 891–900.

[12] T. Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 525–532.

[13] T. Sakai. 2018. *Laboratory experiments in information retrieval: Sample sizes, effect sizes, and statistical power*. Vol. 40. Springer.

[14] T. Sakai. 2020. On FuhrâĂŹs Guideline for IR Evaluation. *ACM SIGIR Forum* 54, 1 (2020).

[15] M. Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval* 4, 4 (2010), 247–375.

[16] M. Sanderson and J. Zobel. 2005. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 162–169.

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

[18] M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. ACM, 623–632.

[19] I. Soboroff. 2004. On evaluating web search with very few relevant documents. In *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 530–531.

[20] J. Urbano. 2016. Test collection reliability: a study of bias and robustness to statistical assumptions via stochastic simulation. *Information Retrieval Journal* 19, 3 (2016), 313–350.

[21] E. M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* 36, 5 (2000), 697–716.

[22] E. M. Voorhees and C. Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 316–323.

[23] E. M. Voorhees, D. K. Harman, et al. 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge.

[24] W. Webber, A. Moffat, J. Zobel, and T. Sakai. 2008. Precision-at-ten considered redundant. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 695–696.

[25] J. Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 307–314.