Statistical Precision of Information Retrieval Evaluation

Gordon V. Cormack and Thomas R. Lynam
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
gvcormac@uwaterloo.ca, trlynam@uwaterloo.ca

ABSTRACT

We introduce and validate bootstrap techniques to compute confidence intervals that quantify the effect of test-collection variability on average precision (AP) and mean average precision (MAP) IR effectiveness measures. We consider the test collection in IR evaluation to be a representative of a population of materially similar collections, whose documents are drawn from an infinite pool with similar characteristics. Our model accurately predicts the degree of concordance between system results on randomly selected halves of the TREC-6 ad hoc corpus. We advance a framework for statistical evaluation that uses the same general framework to model other sources of chance variation as a source of input for meta-analysis techniques.

Categories and Subject Descriptors

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

General Terms

Experimentation, Measurement

Keywords

bootstrap, confidence interval, precision

1. INTRODUCTION

The purpose of IR evaluation is to measure the effectiveness, or relative effectiveness, of information retrieval systems. *Statistical precision*[1] is the degree to which the measurement is free from random error; *validity* is the degree to which the measurement truly reflects retrieval effectiveness. Validity may be further qualified as *internal validity*, the aptness of the measure under test conditions, or *external validity*, the generalizability of the results to other situations ([8], pp. 115-134).

Our primary concern is the statistical precision of information retrieval experiments. If an experiment were to be repeated with different but materially similar data, how similar would the results be? Is it possible, when the test is conducted, to predict accurately this degree of similarity? While these questions are largely amenable to statistical inference, they may be understood only in the context of a general investigative framework that includes questions of validity which are not necessarily statistical. Instead they are addressed by the tools of scientific inquiry – observation, induction, deduction and experiment.

We argue here that one source of random error – that associated with the test corpus – should be considered in assessing IR evaluation methods. We develop and evaluate bootstrap methods that estimate this source of statistical imprecision.

We further argue that another source of random error – that associated with the topics – is poorly modelled by regarding the set of test topics as a random sample of some "true" population; experiments assuming such a model are uncompelling to establish either statistical precision or validity. Instead, we propose, each topic should be regarded as a separate test, and the results of these tests should be combined using meta-analysis techniques ([8], pp. 643-673).

2. TREC AD HOC RETRIEVAL

The object of our study is the TREC ad hoc retrieval evaluation technique[19]. Given a topic T and a set of documents D, each tested IR system returns an ordered subset $S = s_1 s_2 ... s_n$ of D, ranked by the system's estimate of the likelihood that each document is relevant to T. Several effectiveness measures are computed, including average precision (AP), precision at k returned documents ($P@k$), and R-precision ($P@R$) defined as

$$AP = \sum_{k=1}^{|S|} rel(s_k) * P@k/R$$

$$P@k = \sum_{i=1}^{k} rel(s_i)/k$$

$$R = \sum_{d_i \in D} rel(d)$$

[1]Known simply as *precision* in the statistics literature; denoted *statistical precision* here to distinguish it from the IR effectiveness measure of the same name.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.

$$rel(d) = \left\{ \begin{array}{l} 1 \text{ if document } d \text{ is relevant to } T \\ 0 \text{ otherwise} \end{array} \right\}.$$

The test is repeated for several topics; the effectiveness measures from these tests are reported separately and also averaged across topics. In particular, mean average precision (MAP) is typically used and accepted as a valid effectiveness measure.

The evaluation measures depend on the truth value of *if document d is relevant to T,* which must be adjudicated. TREC uses the pooling method in which the top-ranked t documents from a set of systems (perhaps the systems under test) are combined, eliminating duplicates, and presented in random order to a human assessor. The assessor records a judgement of *relevant* or *not relevant* for each document in the pool. Documents not in the pool are assumed to be *not relevant*.

TREC evaluations have typically tested about 50 systems using 50 topics, $n = 1000$, $|D| = 500\,000$, $t = 100$, and a pool size of 40 000 (for all 50 topics).[19]

3. PHILOSOPHICAL FRAMEWORK

The notion of *population* has been the subject of historical and current philosophical debate [7]. We adopt Fisher's view [4] of an infinite hypothetical population:

> If, in a Mendelian experiment, we say that the probability is one half that a mouse born of a certain mating shall be white, we must conceive of our mouse as one of an infinite population of mice which might have been produced by that mating. The population must be infinite for in sampling from a finite population the fact of one mouse being white would affect the probability of others being white, and this is not the hypothesis which we wish to consider; moreover, the probability may not always be a rational number. Being infinite the population is clearly hypothetical, for not only must the actual number produced by any parents be finite, but we might wish to consider the possibility that the probability should depend on the age of the parents, or their nutritional conditions. We can, however, imagine an unlimited number of mice produced upon the conditions of our experiment, that is, by similar parents, in the same age, in the same environment. The proportion of white mice in this imaginary population appears to be the actual meaning to be assigned to our statement of probability. Briefly, the hypothetical population is the conceptual resultant of the conditions which we are studying. The probability, like other statistical parameters, is a numerical characteristic of that population.

A *model* is a characterization with a few parameters that *abstracts* the hypothetical population, removing irrelevant information while preserving that which reflects measures of interest; in this instance, measures of retrieval effectiveness. As such a model is a scientific theory that is posed to explain known facts; its worth is judged by its simplicity, its ability to explain existing observations and its ability to predict new ones. A theory is never "proved" as it is simply a model, but our confidence in it builds as these criteria (degree of abstraction, explanatory ability, predictive ability) are demonstrated.

With respect to IR evaluation, it is possible to identify several hypothetical *target* populations: the topics that might be presented to a system, the corpora from which the system may be expected to retrieve documents relevant to the topic, the set of relevance assessments for the topics; even the set of systems that might be subject to test may be considered to be a hypothetical population of interest. However, these populations are exceedingly difficult to specify, let alone model with a small number of parameters. And sampling them would be a hopeless task as many members of the population exist only in the future. Instead, we select readily available data and observations, and treat them as representing the hypothetical population of all data like that which we collected – the *source* population. The meaning of the word "like" must be considered carefully in modelling such a population; a narrow definition may improve statistical precision while a broad one may improve external validity.

External validity – the applicability of the model to the target population – is established, not by statistical inference, but by scientific inquiry in which (a) predictions about other data are made and tested by experiment, and (b) sources of possible systematic or random error are identified and tested by experiment.

4. STATISTICS IN IR EVALUATION

Tague-Sutcliffe [12] argues that of validity, reliability and efficiency should be considered in a qualitative assessment of various design issues. Validity is used to mean internal validity; reliability[2] subsumes (statistical) precision and external validity; efficiency relates to the resources that are expended in achieving validity and reliability.

Tague-Sutcliffe [13] performed a statistical analysis of the TREC-3 results, under the assumption that the set of topics was a random sample of "all possible queries that might be asked of the database." Paired testing was rejected so as to avoid the fallacy of multiple hypothesis testing (cherry-picking); analysis of variance (ANOVA) was used to compute significant differences among systems according to a number of performance measures. Very large differences in performance – spanning approximately three-quarters of the tested systems – were necessary to distinguish systems with 95% confidence (i.e. $p < .05$). Neither the choice of measure nor an arcsine transformation had substantial impact on the results.

Savoy [11], under the assumption that topics are a random sample, examines the use of classical and bootstrap methods [3] to test the relative performance of pairs of systems. The bootstrap builds a concrete model for a hypothetical population in which each element of the sample is replicated an equal and infinite number of times; this population may, in effect, be sampled any number of times by drawing elements from the original sample, with replacement. Savoy performs significance tests to support the proposition that

[2]In testing, reliability is the degree to which the same test, administered to the same subject, will yield a consistent score ([8], p. 507). Assuming one interprets "the same" literally, IR tests are 100% reliable. Figurative interpretations are captured by Fisher's hypothetical population.

the bootstrap yields higher statistical precision than parametric approaches, and that median, as opposed to mean, is a better summary statistic.

Voorhees and Buckley [17] explore the effect of topic set size, also assuming the topics to be a random sample. Rather than building a statistical model, they measure the proportion of discordant results between evaluations performed using disjoint sample subsets. Results are stratified by the difference in evaluation measure between each pair of systems. For each stratum an exponential curve on two parameters is used to estimate the proportion of discordant pairs.

Sanderson and Zobel [10] measure discordance proportion stratified by p-value of a significance test and at the same time by the magnitude of the difference between MAP scores, and observe that a large difference coupled with a small p-value predicts low discordance.

Several studies [15, 2, 20, 9, 18, 10] have considered the effect of variations in relevance judgments and judging pools on retrieval evaluation. Buckley and Voorhees [1] consider the effect of using differently formed queries to represent the same information need.

Reports on IR evaluations often[3] include standard tests such as paired t-tests, Wilcoxon signed-ranked tests, sign tests, or analysis of variance, notwithstanding questions as to their applicability [6]. Reports typically include significance judgements based on a fixed α threshold; p-values are less common and confidence intervals are rarer still. The vast majority, if not all, assume that topic variation is the only source of random error.

Statistical hypothesis testing in general, particularly that based on a fixed α threshold, has come under criticism lately ([8], pp. 183-199). H_0 – the null hypothesis that two populations are the same – is a strawman that is too easy to refute. In the real world, no two distinct things are the same[5], and a large enough sample will show this. Such a hypothesis should be replaced by an estimate of the magnitude of the difference and an argument as to whether or not that difference is important.

5. COLLECTION VARIABILITY

To measure statistical imprecision due to collection variability, we use the hypothetical population of all collections that are materially similar to the test collection. We formulated a simple characterization of this population and conducted a pilot experiment in which we constructed confidence intervals for AP using a bootstrap estimation of model parameters, and predicted the number of AP values in a second test that should fall in this interval, according to the model. The results show good precision but for some outliers which led us to examine the special cases which they represent, and to adapt the model to account for them.

Our initial model assumes that D, the set of documents in the test collection, is an independent and identically distributed (i.i.d.) sample of a population of similar documents. By similar, we mean having the same relevance value, and yielding a comparable score (or at least a comparable ranking relative to other documents) when retrieved by the IR system under test. The hypothetical population to which D belongs is the set of all such samples.

[3]Not often enough, according to Sanderson and Zobel [10], who surveyed published SIGIR papers and found that 14 of 28 claiming retrieval results reported no statistical tests.

Let D' be some other collection of the same size from the same population as D, and AP' be the average precision from applying the same IR system to D'. We wish to compute a 95% confidence interval – a range of possible values such that, with 95% probability, contains the expected value $E(AP')$. We are not aware of any direct parametric method of estimating this confidence interval; therefore we use the bootstrap to sample the population to which D and D' belong. By repeated sampling, we may estimate the variance of AP' and compute parametric confidence intervals assuming a normal (Gaussian) distribution. Or we may estimate the variance of a monotonic transform $t(AP')$ which is better distributed, in effect computing confidence limits for $E(t(AP'))$ which may be more accurate. Or we may compute confidence intervals nonparametrically by selecting the $2.5th$ through $97.5th$ percentile of the bootstrap samples. A bias-corrected variant of percentile method is known as BC_a[3]. We used three methods for our pilot: variance of AP' values; variance of $logit(AP')$[4]; BC_a.

The bootstrap constructs repeated examples of D' by resampling D. That is, the elements of each example of D' are selected from D, with replacement. For this application we assume that $|D'|$ is large compared to n, the size of the ranked list of documents to be retrieved (typically $|D'| > 100n$). This assumption allows us to use the Poisson distribution to generate S', the list of documents retrieved from D', without considering the irrelevant elements of D'. Specifically, each document of S (the retrieved set of documents) is assumed to be replicated k times in S' with probability $\frac{1}{e \cdot k!}$.

So to construct S' we take each $s_i \in S$ in rank order, generate a random k according to the Poisson distribution, and replicate the element k times. We assume that the replicated elements all receive comparable scores from the IR system and thus are consecutively ranked in S'. Using this construction, $|S'| \approx |S|$. The difference in sizes is inconsequential to the AP' calculation.

It is also necessary to compute R' for the bootstrap sample. To do this we partition R:

$$R = R_{ret} + R_{not}$$

$$R_{ret} = |\{d \in S \,|\, rel(d)\}|$$

$$R_{not} = |\{d \notin S \,|\, rel(d)\}|$$

R'_{ret} is determined directly from S'; R'_{not} is computed post-hoc as:

$$R'_{not} = \sum_{i=1}^{|R - R_{ret}|} k_i$$

where each k_i is randomly generated according to the Poisson distribution.

The net effect is that we may compute as many examples of AP' as necessary to compute model parameters for our hypothetical population. For our untransformed parametric confidence interval estimate, we compute the standard deviation σ of the AP' values. The 95% confidence interval is $AP \pm 1.96\sigma$. For the logit-transformed parametric estimate, we first replace AP values of 0 and 1 by ε

[4]$logit(x) = log(\frac{x}{1-x})$

and $1-\varepsilon$ respectively, and compute the standard deviation σ_{logit} of $logit(AP')$. The 95% confidence interval is $logit^{-1}(logit(AP) \pm 1.96\sigma_{logit})$. BC_a confidence intervals were computed directly using the System R implementation of Efron's S-Plus code ([3], pp. 402-403).

6. PILOT EXPERIMENT

We used the raw results from the 74 IR system runs evaluated over 50 topics in the the TREC 6 ad hoc task[14] – 3652 non-empty ranked-result lists in total. The corpus documents were split into two subsets, A and B, of roughly equal size using an MD5 hash on the document identifier. Similarly, each result list was split into two – one representing the documents retrieved from A; the other from B. These two sets of result lists were assumed to represent the retrieval results on two independent corpora drawn from a common source population. We used 2000 bootstrap samples of the A corpus and the three bootstrap techniques to compute 95% confidence intervals.

6.1 Prediction

Recall that the confidence interval is defined to be an interval within which contains the true value $E(AP')$ with 95% probability. If we knew the value of $E(AP')$ we could simply count the proportion of times $E(AP')$ fell within the computed confidence interval, expecting this proportion to be about 95% if the intervals were accurate. Similarly, if the intervals were unbiased, we would expect an equal proportion (about 2.5%) to fall above as below the interval.

	All Runs *(n=3652)*			Only $R > 5$ *(n=3068)*		
	below	in	above	below	in	above
linear	3.7%	77.8%	11.1%	2.8%	82.0%	15.1%
logit	11.1%	76.5%	12.5%	9.4%	81.4%	9.1%
BC_a	7.4%	77.1%	15.5%	6.7%	80.8%	12.5%
model	8.25%	83.5%	8.25%	8.25%	83.5%	8.25%

Table 1: Proportion of AP_B within interval

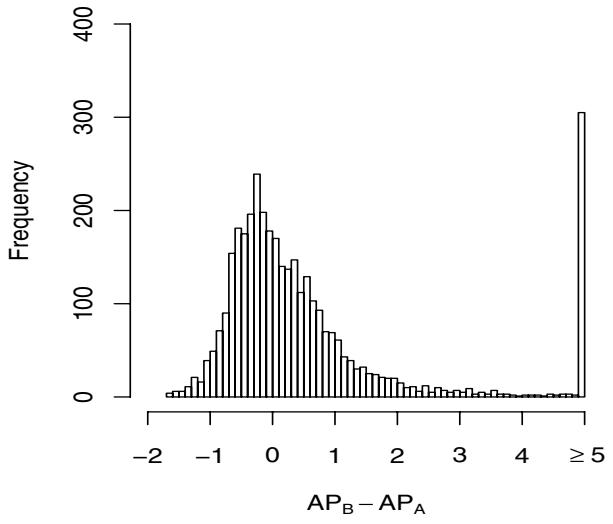


Figure 1: Distribution of Normalized $AP_B - AP_A$

But we don't know $E(AP')$; therefore, we validate the model by using it to predict how many times AP_B (AP computed on the B corpus) should fall within the confidence interval created by bootstrapping the A corpus. Note that this frequency is not 95%, as commonly assumed, but considerably lower.

For the parametric models, recall that the interval is $AP_A \pm 1.96\sigma$. We wish to predict how likely AP_B is to fall in this interval; more precisely

$$AP_A - 1.96\sigma \;\le\; AP_B \;\le\; AP_A + 1.96\sigma\,,$$

or

$$-1.96\sigma \;\le\; AP_B - AP_A \;\le\; 1.96\sigma\,.$$

Our model predicts that AP_A and AP_B both have standard deviation σ, so σ_{AB}, the standard deviation of $AP_B - AP_A$ is given by $\sigma_{AB} = \sqrt{2}\sigma$. Substituting, we have

$$-1.39\sigma_{AB} \;\le\; AP_B - AP_A \;\le\; 1.39\sigma_{AB}\,.$$

This range bounds 83.5% of the area under the normal curve for $AP_B - AP_A$ and hence we expect the inequalities to be satisfied, i.e. AP_B to fall within the confidence interval, 83.5% of the time. Furthermore, AP_B should fall about equally to the left and to the right of the interval.

The same argument holds for the logit-transformed parametric model. There is no similar mechanism for making direct prediction from the non-parametric confidence intervals. However, we may still generally compare the results to those of the parametric methods.

6.2 Pilot Results

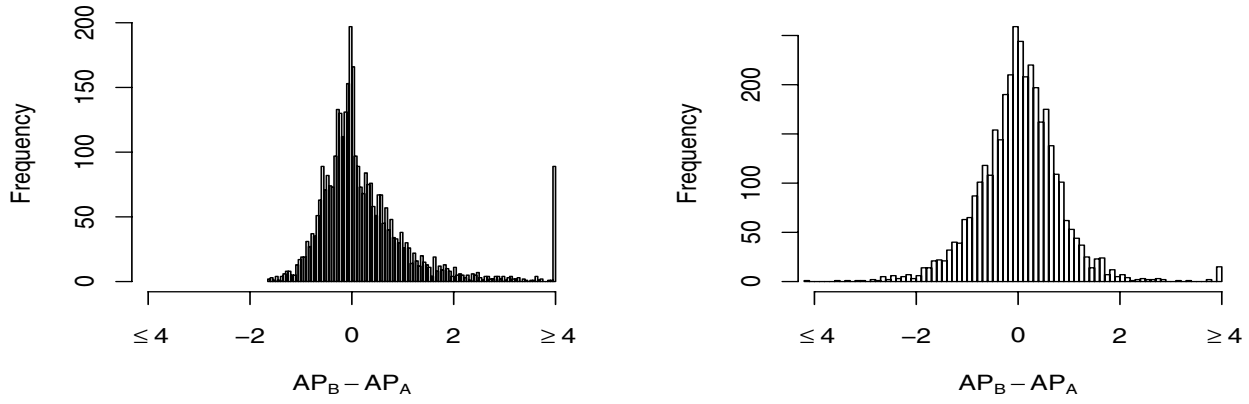
Confidence intervals were computed using the A subset of the results for each topic within each run ($n = 3652$). AP_B was computed for the corresponding B subset and compared to the confidence interval. Table 1 reports the proportion of AP_B values above, within, and below the intervals. Figure 1 shows (for the BC_a method only) the distribution of $AP_B - AP_A$, normalized so that the confidence interval occupies the range $-1 \,..\, 1$.

Figure 1 makes it apparent that there are a large number of outliers at the high end of the distribution. Further investigation reveals that these extreme values are almost entirely accounted for by small-sample effects. Most of these arise when $R \le 5$ ($R \le 1$ in particular). Selecting only those topics for which $R > 5$ gives the proportions listed in second half of Table 1. The number of AP_B values within the predicted interval approaches, but does not quite reach, the predicted 83.5%. Furthermore, the linear and BC_a estimates show evidence of bias.

Even when we restrict our attention to the situation in which $R > 5$, a handful of outliers remain. These consist mainly of cases with $AP_A = 0$ or $AP_A = 1$, where the Bootstrap erroneously reports $\sigma = 0$. In this situation, AP_B falls within the interval only if it is exactly equal to AP_A, which occurs in substantially less than 95% of the cases. Further investigation revealed that the high error rate among the cases with $R \le 5$ was also largely due to cases with $AP_A = 0$ or $AP_A = 1$.

6.3 Pilot Conclusions

From the pilot we conclude that the model works well for the majority of the situations, but special attention needs to be paid to situations in which R is small, or in which $AP_A = 0$ or $AP_A = 1$.



No Transform

logit Transform

Figure 2: Normalized Small-R Corrected Distribution (c.l. $= [-1, 1]$)

It is well known that many statistical methods are inapplicable to small samples. One way to deal with this problem is to design studies and experiments that yield sufficient numbers. In studying the prevalence of disease, for example, an epidemiologist would ensure that the sample population would be expected to contain a sufficient number of positive examples to be amenable to statistical inference. TREC's corpus design process pays careful attention to this consideration, rejecting topics expected to yield very low or very high values of R. The case of $R = 0$ is avoided in particular, because AP is undefined in this situation. The TREC-6 corpus that we used had 5 of its fifty topics with $R \leq 5$ (3 with $R = 5$; 1 with $R = 4$; 1 with $R = 3$; none with $R \leq 2$. Our split-sample technique created A and B corpora which effectively halved R, giving rise to substantially more (9 and 8 resp.) with $R \leq 5$ and also to several (3 and 4 resp.) with $R \leq 2$, smaller than any in the TREC-6 corpus. Therefore we would expect that the estimation techniques used in our pilot would be considerably more accurate were they applied to the full corpus.

Experimental design issues notwithstanding, the results of our pilot prompted us to investigate ways to augment our model to handle cases with small R.

7. SMALL-R CORRECTION

We modelled cases with small R (and some cases with larger R) by considering $AP = 0$ and $AP = 1$ as special cases. $AP = 0$ means that, among the R relevant documents in the collection, the retrieval system found none. But it may be that in the source population – the set of materially similar collections – there exist *silver bullets* – relevant documents that the system could retrieve but did not happen to be among the R sampled for this particular test. Using exact binomial probabilities, we establish confidence limits for the proportion of such silver bullets that may exist in the source population. The lower confidence limit is 0 and the upper confidence limit is the smallest u such that the probability $(1-u)^R$ of having no silver bullets in our sample does not exceed the significance threshold, under the assumption that u is the true proportion. For example, if $R = 4$, the 95% confidence limit for the proportion of silver bullets is $[0, 0.53]$. That is, we cannot statistically reject the hypothesis that 53% of all relevant documents are silver bullets!

We take 53% (or whatever the appropriate fraction is for a given R) to represent the extreme case that results in our upper confidence limit. But the confidence limit must be expressed as an AP' value, not a proportion of silver bullets. To compute AP', we assume that the retrieved rank of a silver bullet is uniformly distributed between 1 and n (i.e. could appear anywhere in the retrieved list) and, using dynamic programming, compute by enumeration the resulting $E(AP')$. This value is chosen as the upper confidence limit in place of 0.

	AP_B w.r.t. A			AP_A w.r.t. B		
	below	in	above	below	in	above
linear	2.1%	82.8%	15.0%	2.4%	83.9%	13.6%
logit	8.5%	83.6%	7.9%	8.7%	82.7%	8.7%

Table 2: Small-R corrected within interval

A similar model was used for the cases of $AP = 1$. We use binomial probabilities to establish bounds for the the proportion of *lead balloons* – relevant documents that the system is unable to retrieve, but are not represented in the R relevant documents sampled for this particular test. A similar dynamic approach converts the worst-case lead-balloon proportion to a lower confidence limit.

The same corrections were applied to cases with $AP \approx 0$ and $AP \approx 1$; for such values, we use the larger of the original confidence interval and the interval under the assumption that $AP = 0$ ($AP = 1$, resp.).

Table 2 shows the result of applying small-R correction to the results of the two bootstrap methods.[5] The left

[5]We did not use the computationally intensive BC_a method as it showed poor results in the pilot and was not amenable

half shows the fraction of AP_B values that fall within the confidence intervals computed from A; the right half shows the the fraction of AP_A values that fall with intervals computed from B. In both cases, both methods yield in-interval fractions that are extremely close to those predicted by the model. However, the linear model exhibits considerable bias, as evidenced by the fact that the fraction above is roughly six times larger than the fraction below in both tests. The logit model demonstrates no apparent bias.

Figure 2 shows normalized $AP_B - AP_A$ for the linear and logit models. Logit is clearly more symmetric with fewer outliers. All of the (few) logit outliers are cases with $AP_A = 0$, for which a nonparametric model was used. Such a model predicts only the number in-interval, not the distribution of those outside. The outliers should therefore not be considered to contradict the model.

8. MEAN AVERAGE PRECISION

Mean Average Precision (MAP) – the average of AP values over several topics – is commonly reported as an overall summary measure. We investigated methods to compute the sensitivity of MAP to corpus variation. We applied the same methods to a summary measure we call *logistic* MAP (L-MAP), which averages $logit(AP)$ instead of AP. L-MAP is closely related (and nearly identical for small values) to *geometric* MAP (G-MAP) – the average of $log(AP)$ values – which has been proposed recently to increase the contribution of low AP values to the overall measure. [16]

	$(L-)MAP_B$ w.r.t. A			$(L-)MAP_A$ w.r.t. B		
	below	in	above	below	in	above
MAP	27.0%	68.9%	4.0%	21.6%	75.6%	2.7%
L-MAP	1.3%	83.7%	14.8%	16.2%	82.4%	1.3%

Table 3: Bootstrap 50 topic mean within interval

	$(L-)MAP_B$ w.r.t. A			$(L-)MAP_A$ w.r.t. B		
	below	in	above	below	in	above
MAP	1.3%	78.3%	20.2%	16.2%	81.0%	2.7%
L-MAP	2.7%	74.3%	22.9%	22.9%	74.3%	2.7%

Table 4: Parametric 50 topic mean within interval

The first row of table 3 shows the fraction of the 74 MAP_B (MAP_A resp.) fractions falling within the interval computed by bootstrap sampling A (B resp.). That is, MAP was computed for each of the 2000 bootstrap samples, and the variance of these values was used to estimate the standard error. The second row shows the same method applied to L-MAP. In the case of MAP, we see that the in-interval fraction falls considerably below that predicted by the model, suggesting that the model is inappropriate. The L-MAP in-interval fraction, on the other hand, suggests that in this case the model is appropriate. The imbalance between above and below fractions suggests random skew between A and B rather than a systematic bias in the model. Note that these fractions are determined from 74 data points, as opposed to 3652 for the AP computations. Therefore these observations should be taken as indications to be confirmed by a larger experiment.

Table 4 shows the results of using a parametric approach to combine the 74 separate AP confidence intervals into a single MAP (L-MAP) confidence interval. To compute the MAP confidence interval, we averaged the variances derived from the logistic model, but weighted them according to their relative contribution to MAP statistic. We used a multiplicative weight of $AP - AP^2$, the derivative of AP with respect to $logit(AP)$. To compute the L-MAP confidence interval, we simply averaged the unweighted variances of the 74 separate estimates. We observe that the parametric estimates for MAP are considerably better than the bootstrap estimates, further validating the logit model. However, they appear to be slightly optimistic, yielding about *80%* in-interval as opposed to the predicted *83.5%*. On the other hand, the parametric estimate for L-MAP is much worse than the bootstrap estimate. We attribute this error to the fact that the estimate weights heavily the small-R-corrected estimates, which are themselves non-parametric and therefore not suitable variance estimates. This error is much less important for MAP because most of these estimates receive extremely low weight.

Figure 3 shows the MAP parametric confidence intervals computed from A, along with the MAP_A and MAP_B values, marked x and o respectively. This graphic shows that the confidence intervals generally do a good job of predicting the range of possible MAP_B values, and the out-of-interval values are near-misses rather than outliers. Figure 3 also shows the corresponding L-MAP bootstrap confidence intervals, and reflects the same general observation. Figure 4 shows the MAP and L-MAP values based on the full corpus (as opposed to the A subset). As expected, the confidence intervals are smaller, typically with a width of about 0.05.

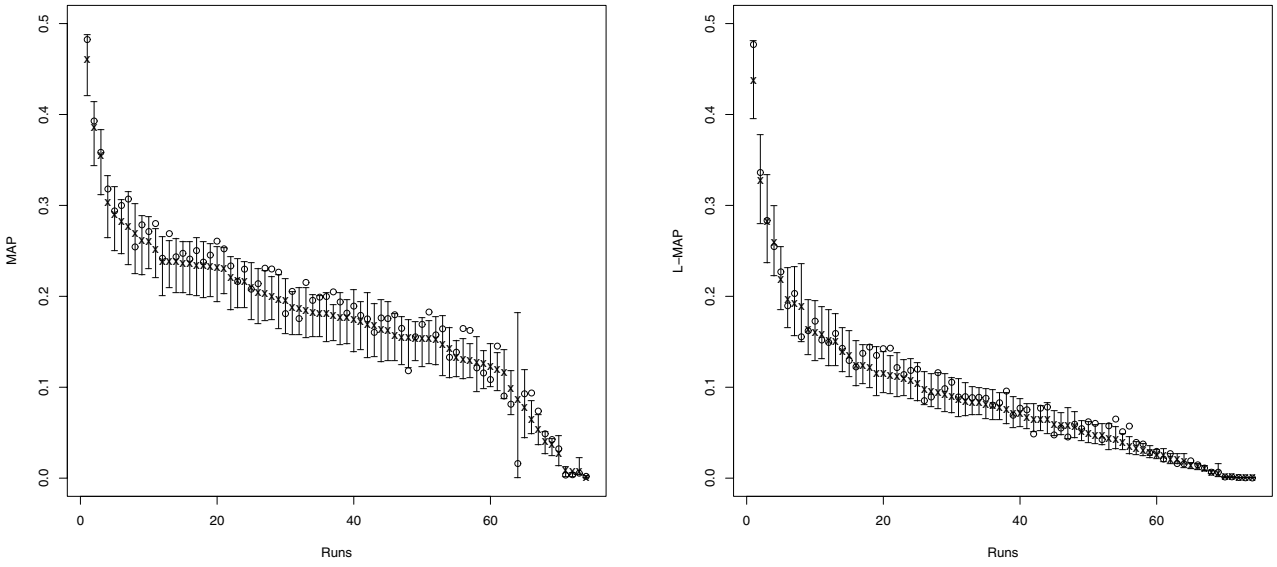
9. DISCUSSION

Experimental evidence suggests that our model for corpus variability aptly predicts confidence intervals for individual AP values. $logit(AP)$ has better algebraic properties than AP and therefore yields a better model from which AP confidence intervals can be derived. AP values close to 0 and 1 are problematic, and arise often when R – the number of relevant documents – is small. These anomalies may be addressed by using a non-parametric binomial model to predict *silver bullets* and *lead balloons* – relevant documents whose properties are not represented at all in the corpus.

MAP exhibits the same algebraic anomalies as AP; in this situation, a weighted average of $logit(AP)$ variances is used to predict indirectly the effect of averaging non-logit-transformed AP values. L-MAP, on the other hand, may be estimated directly using the bootstrap. We expect that G-MAP would exhibit similar properties to L-MAP, as they differ substantially only for values close to 1 – values which occur rarely in IR evaluation.

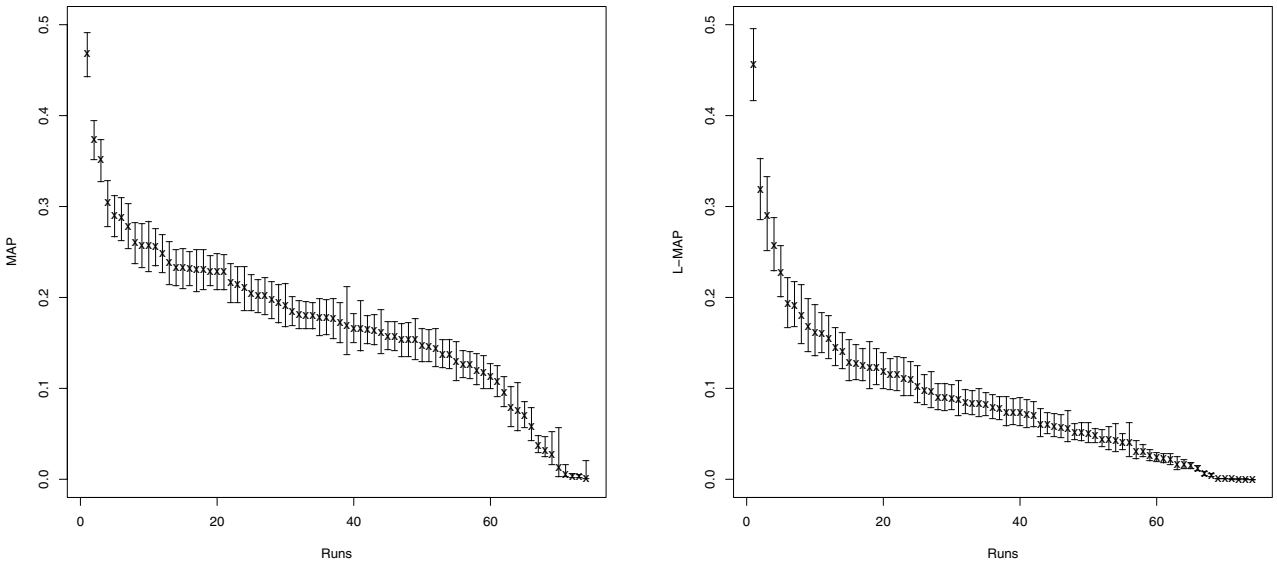
Our framework and validation technique applies equally to models for evaluating the relative effectiveness of a pair of IR systems. One simply has to model the difference d between the two systems according to some measure of interest; for example $d = AP_x - AP_y$. If we construct separate models for AP_x and AP_y with standard deviations $\sigma_x$and σ_y we may estimate $\sigma_d = \sqrt{\sigma_x^2 + \sigma_y^2}$, and a 95% confidence interval of $\pm 1.96\sigma_d$. Our results indicate that defining $d = logit(AP_x) - logit(AP_y)$ would yield a better estimate.

to the further experiments detailed below.



MAP Parametric Intervals L-MAP Bootstrap Intervals

Figure 3: MAP_B and $\boldsymbol{L}$-MAP_B with respect to A confidence intervals



MAP Parametric Intervals L-MAP Bootstrap Intervals

Figure 4: Full corpus Confidence Intervals

Confidence intervals for the difference contain strictly more information than fixed-threshold hypothesis tests; the difference between AP_x and AP_y is significant (two-tailed test, $\alpha = 0.05$) exactly when 0 does not fall within the confidence interval for d. The same approach may be applied to the difference between MAP, L-MAP or G-MAP scores.

A more powerful estimate for σ_d – which takes into account correlations between the two systems' results – may be effected by bootstrapping d. For each bootstrap sample, directly compute AP'_x, AP'_y $[logit(AP_x), logit(AP_y)]$ and hence d'; then estimate σ_d from the various d' values.

Our analysis indicates that current methods[6] poorly model random error due to topic variability. By the argument in section 6.1, if a test concluded from one sample that $AP_x > AP_y$ (p = 0.05) we would expect that for some other sample $AP'_x > AP'_y$ would occur with probability 0.835, (*not* 0.95). Furthermore, this prediction should be insensitive to the sample size and the magnitude of the difference between AP_x and AP_y. Experimental results [10] are inconsistent with these predictions, contradicting the validity of the tests. We conjecture that using the *logit* transform would mitigate but not overcome the shortcomings of the underlying model.

If the *logit* transform were to yield a reasonable model for topic variability – a proposition the experimental investigation of which we leave to future work – it would be a simple matter to use the bootstrap method developed here to model it, or to model both topic and collection variability at once. One must simply compute L-MAP (or the difference between L-MAPs) using bootstrap resampling to select both the topic and the corpus for each sample. The underlying foundation is the same.

A more promising approach, we argue, is to regard the results from each topic as separate tests and to combine them using meta-analysis ([8], pp. 643-673; [5]). For a simple paired hypothesis test, one may simply combine the values of d_i and σ_i arising from k separate tests to compute an overall single-tailed p-value $p = 1 - \Phi \sum_{i=1}^{k} \frac{d_i}{\sqrt{k}\sigma_i}$ where Φ is the cumulative normal distribution. More sophisticated meta-analysis involves identifying a quantitative "effect" and measuring it with confidence intervals. While it is difficult to argue that $d = AP_x - AP_y$ is a meaningful quantity, the value $d = logit(AP_x) - logit(AP_y)$ represents the logarithm of the ratio of the effectiveness of the two systems, which we advance as a worthwhile measure.

Meta-analysis may also be used to estimate the performance of a single system over several tests. The most obvious measure is simply AP but our results indicate that $logit(AP)$ would be more appropriate. Even more appropriate would be a measure that compensated for topic difficulty; we suggest $d = logit(AP) - logit(X)$ where X is the performance of some baseline system or some other estimate of "normal" system performance.

Fixed-effect model meta-analysis computes the effect d and standard error σ as follows:

$$d = \frac{\sum_{i=1}^{k} d_i \sigma_i^{-2}}{\sum_{i=1}^{k} \sigma_i^{-2}} \qquad \sigma = \left(\sum_{i=1}^{k} \sigma_i^{-2} \right)^{-\frac{1}{2}}$$

The overall effect estimate is the average of the individual estimates, weighted by their statistical precision. Random-effect model meta-analysis further compensates for heterogeneity of tests such as might occur when using diverse topics or corpora.

10. REFERENCES

[1] Buckley, C., and Voorhees, E. M. Evaluating evaluation measure stability. In *SIGIR Conference 2000* (Athens, Greece, 2000).

[2] Cormack, G. V., Palmer, C. R., and Clarke, C. L. A. Efficient construction of large test collections. In *SIGIR Conference 1998* (Melbourne, Australia, 1998).

[3] Efron, B., and Tsibirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1994.

[4] Fisher, R. A. Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society 22* (1925), 700–725.

[5] Glass, G. V. Meta-analysis at 25. http://glass.ed.asu.edu/gene/papers/meta25.html, 2000.

[6] Hull, D. A. Using statistical testing in the evaluation of retrieval experiments. In *Research and Development in Information Retrieval* (1993), pp. 329–338.

[7] Lenhard, J. Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science* (2006).

[8] Rothman, K. J., and Greenland, S. *Modern Epidemiology*. Lippincott Williams & Wilkins, 1998.

[9] Sanderson, M., and Johno, H. Test collections with no system pooling. In *SIGIR Conference 2004* (Sheffield, UK, 2004).

[10] Sanderson, M., and Zobel, J. Information retrieval evaluation: Effort, sensitivity, and reliability. In *SIGIR Conference 2005* (Salvador, Brazil, 2005).

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

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

[13] Tague-Sutcliffe, J., and Blustein, J. A statistical analysis of the TREC-3 data. In *Proceedings of TREC-3, The Third Information Retrieval Conference* (1994), pp. 385–398.

[14] Voorhees, E., and Harman, D. Overview of the Sixth Text REtrieval Conference (TREC-6). In *6th Text REtrieval Conference* (Gaithersburg, MD, 1997).

[15] Voorhees, E. M. Variations in relevance judgements and the measurement of retrieval effectiveness. In *SIGIR Conference 1998* (Melbourne, Australia, 1998).

[16] Voorhees, E. M. Overview of the TREC-2004 robust track. In *13th Text REtrieval Conference* (Gaithersburg, MD, 2004).

[17] Voorhees, E. M., and Buckley, C. The effect of topic set size on retrieval experiment error. In *SIGIR Conference 2002* (Tampere, Finland, 2002).

[18] Voorhees, E. M., and Buckley, C. Retrieval evaluation with incomplete information. In *SIGIR Conference 2004* (Sheffield, UK, 2004).

[19] Voorhees, E. M., and Harman, D. K., Eds. *TREC - Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[20] Zobel, J. How reliable are the results of large-scale information retrieval experiments? In *SIGIR Conference 1998* (Melbourne, Australia, 1998).

[6] In particular, a t-test which tacitly models $d = AP_x - AP_y$ over the population of all topics with a fixed corpus.