Bootstrap confidence intervals for Mean Average Precision

Laurence A. F. Park

School of Computing and Mathematics, University of Western Sydney, Australia
lapark@scm.uws.edu.au

Abstract

Due to the unconstrained nature of language, search engines (such as the Google search engine) are developed and compared by obtaining a document set, a sample set of queries and the associated relevance judgements for the queries on the document set. The de facto standard function used to measure the accuracy of each search engine on the test data is called mean Average Precision (AP). It is common practice to report mean AP scores and the results of paired significance tests against baseline search engines, but the confidence in the mean AP score is never reported. In this article, we investigate the utility of bootstrap confidence intervals for mean AP. We find that our Standardised logit bootstrap confidence intervals are very accurate for all levels of confidence examined and sample sizes.

Key words: bootstrap, confidence interval, average precision

1. Introduction

Text based search engines (such as the Google search engine), also known as text retrieval systems, have been developed for the past fifty years. During that time, many systems have been constructed based on various models. Each retrieval system is a function that takes a set of key words (the query) and returns a vector of relevance judgements, where each relevance judgement is the predicted relevance of an associated document in the systems database to the query. Rather than providing the complete list of relevance judgements to the user, the search system usually returns the ten documents with greatest associated relevance judgements (in order) to the user and provides the remaining documents if requested.

To evaluate the accuracy of a retrieval system, a sample set of queries and their associated true relevance judgements (the set of correct relevance scores for each document, for each query) must be obtained. For each query, the system computed relevance judgements and true relevance judgements are compared using an evaluation function. The most widely used retrieval system evaluation function is Average Precision (AP) [1]. AP is a function of both precision (the proportion of correct documents in the retrieved set) and recall (the proportion of correct documents retrieved). Each AP value falls within the range $[0, 1]$, 0 meaning the system has not found any relevant documents, and 1 meaning all documents predicted as relevant are relevant and all predicted as irrelevant are irrelevant.

To evaluate a system, we should obtain many queries and their associated true relevance judgements to construct the system AP distribution. Unfortunately, it is costly (in terms of time) to obtain the set of true relevance judgements for a single query, since each document must be manually judged to build the list [2], and it is common for retrieval systems to have over one million documents in their database. Therefore retrieval experiments are performed using a small sample of queries and the sample mean is reported along with paired significance test results with baseline systems.

Using this experimental method, a reader of a publication is able to identify which system has performed best in the experiment, but we are unable to compare systems across publications from other experiments unless we obtain the systems and run the experiments ourselves. To compare systems across publications, the confidence interval of the mean AP should be reported. A recent study showed that accurate confidence intervals can be produced for mean AP by fitting a t distribution to the samples, as long as the queries used were standardised using five other systems and that all authors used the same standardising systems [3]. Since there are no defined set of "standard" systems, it would be unlikely that experimental results from different authors would use the same standardising systems, and hence obtain confidence intervals that are not comparable.

In this article we will investigate the accuracy of bootstrap confidence intervals on mean Average Precision. We examine the accuracy of Percentile and Accelerated bootstrap, and we introduce the Studentised logit bootstrap, based on the analysis of the system distributions. The article will proceed as follows: Sec-

tion 2 describes the experimental environment, section 3 examines set of system AP distributions, and Section 4 provides details of the experiments and results.

2. Experimental Environment

To conduct our experiments, we will use the set of 110 systems that participated in the TREC (`http://trec.nist.gov`, Text REtrieval Conference) 2004 Robust track. The Robust track consists of 249 queries and 528,155 documents. We have obtained the AP of each query on each system. We will approximate the population AP distribution with the set of 249 AP values for each system. Our experiments will involve taking 1,000 random samples of $n = 5, 10$ and 20 AP values without replacement for each system, computing the confidence interval for the mean AP and evaluating the coverage probability of the confidence interval. The bootstrap distribution is computed by taking 1,000 random samples of size n, with replacement, from the AP sample. For each experiment, we will examine the confidence intervals at $\alpha = 0.05$ to 0.50 in steps of 0.05, where α is the proposed under coverage probability.

3. System AP distribution

Before we proceed, we will examine the bias and skewness of each system AP distribution. Both bias and skewness are known to affect the accuracy of confidence intervals when computed using the bootstrap [6, 7, 8]. Bias is computed as the expected difference between the sample mean and the population mean. Using 1,000 samples of size of $n = 5$ queries, we computed the bias for each system AP and provided the distribution in Figure 1. Given that AP ranges from 0 to 1, we can see that the bias is small and unlikely to affect our experiments.

Skewness is a measure of the asymmetry of the distribution, where a symmetric distribution has no skewness, and a asymmetric distribution can be positively or negatively skewed. We computed the skewness for each system AP population distribution and provided the distribution in Figure 2. The histogram shows that all systems are positively skewed, meaning that lower AP values are more likely than higher AP values.

To examine the skewness further, we have provided the histograms of the systems with the least, median, and greatest skewness in Figure 3. We can see that none of the system AP distributions are symmetric. The least skewed system is more likely to provide greater AP values than the other two. We can also see that the most skewed system has obtained AP values between 0 and 0.1 for most of the 249 queries, making it a poor system.

From this analysis, we have found that there is little sampling bias, but there is high skewness in each system distribution.

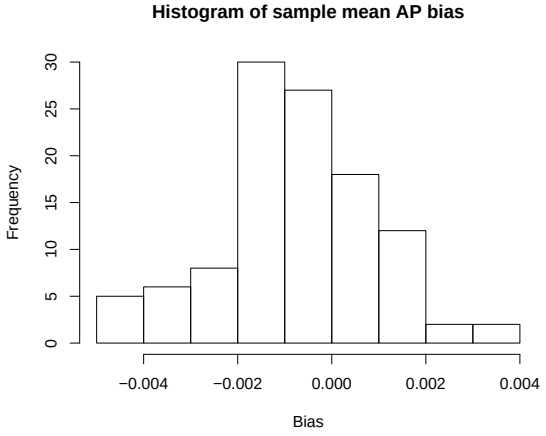

Figure 1: The distribution of sample mean AP bias. The sample mean AP bias from each system AP distribution was measured, using a sample size of 5 and the distribution of the bias across all systems is shown above.

4. Experiments

In this section we examine the accuracy of Percentile and Accelerated bootstrap confidence intervals on our experimental environment. We also derive the novel Studentised logit bootstrap from our analysis of the system distributions.

4.1. Percentile Bootstrap

To begin our experiments, we will compute the Percentile bootstrap confidence interval of the mean AP. The percentile bootstrap confidence interval is computed as follows:

1. Compute the bootstrap distribution of the sample mean AP.
2. Use the $\alpha/2$ and $1-\alpha/2$ quantiles as the $(1-\alpha)\times 100\%$ confidence interval boundary.

It is known that the Percentile bootstrap does not provide the correct coverage when the population is skewed [4, 5]. Therefore, we will measure the accuracy of the confidence intervals and use them as a baseline. The results are provided in Table 1.

We can see from Table 1 that there is a large difference between α and the under coverage probability for $n = 5$ and 10. For $n = 20$, we can see that the under coverage probability is similar to the associated value of α. This is to be expected since the distribution of the sample mean will be approximately Normal for large values of n.

4.2. Bias Corrected Accelerated Bootstrap

The bias corrected accelerated bootstrap confidence interval (BC_a) [6, 7, 8] was developed to provide good confidence intervals for a sample taking into account the bias and skewness.

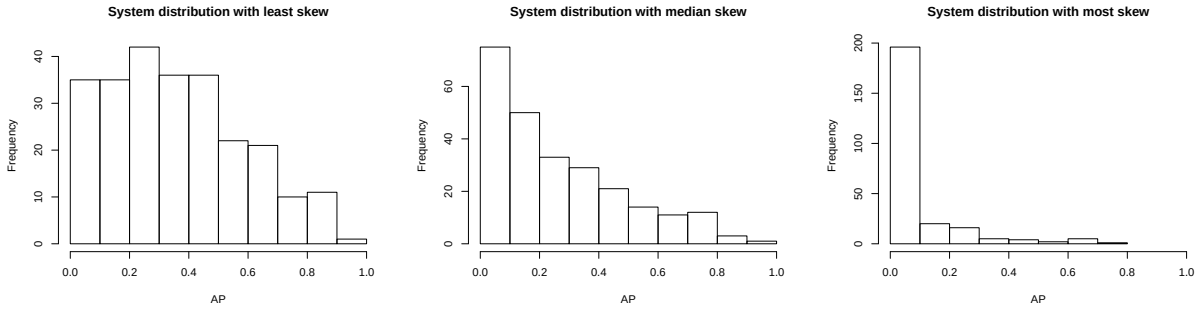

Figure 3: The AP distribution from three systems where the left-most, central and right-most distributions are associated to the systems with the least, median and most AP skewness.

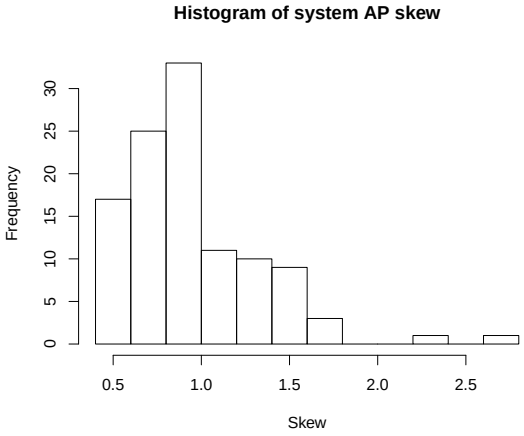

Figure 2: The distribution of system AP skewness. The skewness from each system AP distribution was measured and the distribution of the skewness across all systems is shown above.

α	Under Coverage Probability		
	$n = 5$	$n = 10$	$n = 20$
0.05	0.1814	0.1075	0.0701
0.10	0.2279	0.1576	0.1191
0.15	0.2816	0.2045	0.1672
0.20	0.3295	0.2514	0.2151
0.25	0.3658	0.2981	0.2628
0.30	0.4015	0.3435	0.3098
0.35	0.4380	0.3901	0.3587
0.40	0.4789	0.4357	0.4054
0.45	0.5166	0.4811	0.4552
0.50	0.5573	0.5269	0.5035

Table 1: Under coverage probability when computing $(1-\alpha)\times 100\%$ confidence intervals of mean AP from n AP samples using the Percentile Bootstrap method.

The BC_a bootstrap confidence interval is intended to be a general purpose method and includes many steps to compute the confidence interval bounds, therefore we refer the reader to [6] for further information.

In this experiment we use the nonparametric form of the BC_a bootstrap (where the bias correction and acceleration statistic are derived from the sample). The results are shown in Table 2.

From Table 2, we can see that the difference between the under coverage probability and α is slightly smaller when compared to the Percentile bootstrap confidence intervals, but the confidence intervals are still inaccurate for most values of α when $n = 5$, and large values of α when $n = 10$, but accurate for $n = 20$.

4.3. Studentised logit Bootstrap

It is a concern that the AP values are constrained to the domain $[0, 1]$, while this contraint is not explicitly provided when computing the confidence interval. To map the AP samples to the real domain, we can use the logit function:

$$\text{logit}(x) = \log_e\left(\frac{x}{1-x}\right)$$

The logit transform takes data from the $[0, 1]$ domain to the $(-\infty, \infty)$ domain. By observing the AP distributions in Figure 3, we can see that applying the logit transform may reduce the skewness and and provide a more Normal distribution.

Unfortunately, we can't apply the logit transform to the samples since there may be scores of 0 or 1 which are transformed to $-\infty$ and ∞ respectively. However, we are able to transform the sample mean AP. The sample mean can only be 0 or 1 if all of the samples are 0 or 1 respectively. If this is the case, then we are unable to compute a confidence interval due to the lack of variation in the sample.

To reduce the skewness, we will compute the sample mean AP bootstrap distribution and apply the logit transformation to the bootstrap distribution. Note that if a sample contains a 0 or 1, there is a chance that the a bootstrap sample mean will be 0 or 1 respectively. In this case, we remove the associated bootstrap sample.

$1-\alpha$	Under Coverage Probability		
	$n=5$	$n=10$	$n=20$
0.05	0.1704	0.0932	0.0592
0.10	0.2186	0.1442	0.1067
0.15	0.2613	0.1927	0.1552
0.20	0.3085	0.2392	0.2030
0.25	0.3515	0.2848	0.2500
0.30	0.3903	0.3310	0.2995
0.35	0.4278	0.3777	0.3477
0.40	0.4664	0.4242	0.3958
0.45	0.5050	0.4702	0.4445
0.50	0.5450	0.5172	0.4937

Table 2: Under coverage probability when computing $(1-\alpha)\times 100\%$ confidence intervals of mean AP from n AP samples using the Accelerated Bootstrap method.

$1-\alpha$	Under Coverage Probability		
	$n=5$	$n=10$	$n=20$
0.05	0.0546	0.0541	0.0466
0.10	0.1097	0.1075	0.0934
0.15	0.1646	0.1592	0.1420
0.20	0.2190	0.2101	0.1910
0.25	0.2724	0.2606	0.2406
0.30	0.3244	0.3103	0.2915
0.35	0.3742	0.3601	0.3424
0.40	0.4232	0.4089	0.3924
0.45	0.4730	0.4580	0.4431
0.50	0.5235	0.5074	0.4937

Table 3: Under coverage probability when computing $(1-\alpha)\times 100\%$ confidence intervals of mean AP from n AP samples using the Studentised logit Bootstrap method.

The percentile bootstrap is invariant to monotone transformations, therefore computing the percentiles gives us no benefit over the percentile bootstrap baseline.

Assuming that the skewness has been removed, we compute the mean and standard deviation of the transformed bootstrap distribution and obtain the confidence interval boundary using the t distribution.

The Studentised logit Bootstrap is computed as follows:

1. Compute the bootstrap distribution of the sample mean.
2. Reduce the distribution skewness by applying the logit transformation.
3. Obtain the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ of the Normal distribution parameters μ and σ.
4. Compute the mean AP confidence interval boundary using $\hat{\mu} \pm t_{\alpha/2,n-1}\hat{\sigma}$
5. Apply the inverse logit function to the boundary to convert it back to the AP domain.

The accuracy of the confidence intervals is shown in Table 3.

We can see from Table 3 that the under coverage probability of the confidence intervals produced using Studentised logit Bootstrap is very close to the provided $1-\alpha$ for all values of n. We can see the difference grows as $1-\alpha$ decreases for $n=5$, but it is most accurate for small $1-\alpha$ (being the usual confidence range).

5. Conclusion

Empirical evaluation of the accuracy of document retrieval systems is performed using a sample set of queries. The sample is usually small due to the work involved in providing manual relevance judgements for all documents for each query.

It is common place for document retrieval system evaluation to report the sample mean Average Precision (AP), but the fact that we are only working with a sample set of queries is usually ignored, making results across publications incomparable.

In this article we examined the accuracy of bootstrap confidence intervals for mean AP. We found that Percentile and Accelerated bootstrap confidence intervals had poor coverage for small α and small number of samples (5 queries). We also found that our Standardised logit bootstrap confidence intervals were very accurate for all levels of confidence examined and sample sizes. We believe the accuracy of the method comes from the logit transform removing most of the skewness from the bootstrap distribution.

References

[1] C. Buckley, E. M. Voorhees, Evaluating evaluation measure stability, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '00, ACM, New York, NY, USA, 2000, pp. 33–40. doi:10.1145/345508.345543.

[2] S. Keenan, A. F. Smeaton, G. Keogh, The effect of pool depth on system evaluation in TREC, Journal of the American Society for Information Science and Technology 52 (7) (2001) 570–574. doi:10.1002/asi.1096.

[3] L. A. F. Park, Confidence intervals for information retrieval evaluation, in: A. Turpin, F. Scholer, A. Trotman (Eds.), Proceedings of the Fifteenth Australasian Document Computing Symposium (to appear), 2010.

[4] B. Efron, R. J. Tibshirani, An Introduction to the Bootstrap, CRC Monographs on Statistics and Applied Probability, Chapman and Hall, 1994.

[5] P. M. Dixon, Bootstrap resampling, in: Encyclopedia of Environmetrics, John Wiley and Sons, Ltd, 2006. doi:10.1002/9780470057339.vab028.

[6] B. Efron, Better bootstrap confidence intervals, Journal of the American Statistical Association 82 (397) (1987) 171–185. doi:10.2307/2289144.

[7] T. J. DiCiccio, B. Efron, Bootstrap confidence intervals, Statistical Science 11 (3) (1996) 189–228.

[8] F. T. Burbrink, R. A. Pyron, The Taming of the Skew: Estimating Proper Confidence Intervals for Divergence Dates, Systematic Biology 57 (2) (2008) 317–328. doi:10.1080/10635150802040605.