Quantifying Bias and Variance of System Rankings

Gordon V. Cormack
University of Waterloo
gvcormac@uwaterloo.ca

Maura R. Grossman
University of Waterloo
maura.grossman@uwaterloo.ca

ABSTRACT

When used to assess the accuracy of system rankings, Kendall's τ and other rank correlation measures conflate bias and variance as sources of error. We derive from τ a distance between rankings in Euclidean space, from which we can determine the magnitude of bias, variance, and error. Using bootstrap estimation, we show that shallow pooling has substantially higher bias and insubstantially lower variance than probability-proportional-to-size sampling, coupled with the recently released dynAP estimator.

ACM Reference Format:
Gordon V. Cormack and Maura R. Grossman. 2019. Quantifying Bias and Variance of System Rankings. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331356

1 INTRODUCTION

Since large-scale test collections were first applied in the context of TREC [10], concern has been expressed regarding the extent to which they afford "fair" or "unbiased" rankings of information-retrieval (IR) systems, and the extent to which those rankings are "stable" or "reliable" [3, 14]. We frame the problem in terms of measurement accuracy [11], where bias b and standard deviation σ are orthogonal, dimensionally meaningful meta-measurements of (lack of) fairness and (lack of) stability, and RMS error RMSE = $\sqrt{b^2 + \sigma^2}$ quantifies overall (lack of) accuracy.

This work derives an amenable definition of b and σ from Kendall's τ rank correlation (§2). The same derivation applies to any of the plethora of rank-similarity scores that have been been employed in ad-hoc strategies to evaluate the fairness and stability of test collections [4, 7, 12, 16, 17]. We show how to measure $|b|$, σ, and RMSE using bootstrap re-sampling (§3).

Using the TREC 8 Ad Hoc test collection [15] as a reference standard, we evaluate four techniques—two statistical and two non-statistical—for building test collections constrained by an assessment budget (§4). Our evaluation reveals substantial differences in bias that would be masked by RMSE or τ alone. Orthogonal measurements of b and σ allow us to predict the effect of different budget-allocation strategies, which are borne out by the bootstrap results (§4.1). Through the use of adversarial testing, we show that one method is substantially unbiased, even when ranking results that are dissimilar to the TREC submissions (§4.2).

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331356

We use bootstrap sampling to compare the official TREC 2018 Common Core test collection [1] to an alternate statistical collection created in advance by the University of Waterloo [9], along with its companion dynAP estimator[1] [6]. We find that there are two sources of bias between the collections: (i) different relevance assessments for the same documents; and (ii) additional documents in the TREC collection selected using shallow sampling methods (§5).

This work contributes a new methodology for test-collection evaluation, and uses that methodology to show that shallow pooling introduces bias far beyond what is shown by rank-correlation measures over submitted TREC results. Of the statistical estimators, the older and more well established infAP [18] shows clear bias, while dynAP shows negligible bias. The results indicate that a statistical test collection can yield comparable accuracy over 50 topics, and considerably better accuracy over more, compared to exhaustive assessment.

2 BIAS AND VARIANCE FOR RANKING

In this paper, a "test result" [11] is a ranking of systems according to their measured effectiveness. Closeness of agreement is quantified by a rank-similarity coefficient like Kendall's τ, with the maximum value 1 denoting equality.

We interpret system rankings to be points in Euclidean space, where $\delta(x, y) = 1 - \tau(x, y)$ is the the distance between x and y, whose location in space is unspecified. For test results X and Y, we define the expected squared distance Δ between them:

$$\Delta(X, Y) = \mathbb{E}\,\delta^2(X, Y)\,.$$

The variance σ^2 of a given X is one-half the squared distance between itself and an independent and identically distributed test result:

$$\sigma^2(X) = \frac{1}{2}\Delta(X, X')\ (i.i.d.\ X, X')\,.$$

When G is the gold-standard ranking, squared bias b^2 is the amount by which the squared distance between X and G exceeds that which is attributable to variance:

$$b^2(X) = \Delta(X, G) - \sigma^2(X) - \sigma^2(G)\,.$$

It follows that mean-squared error

$$\mathrm{MSE}(X) = b^2(X) + \sigma^2(X)\,.$$

Bias $b(X)$ is a vector whose direction is unspecified; however, its magnitude $|b|(X)$ is sufficient for our purpose.

It is worth noting that the ground truth ranking is a random variable G, rather than the particular outcome $G = g$ derived for the particular topics and assessments of the reference test collection from which G is derived. Under the assumption underlying virtually all reported statistical tests—that the set of topics in a collection is a random sample of a population of topics—the distributions of

[1] *See* cormack.uwaterloo.ca/sample/.

G and X may estimated by bootstrap re-sampling (*see* [8]). Other random factors, notably the choice of documents for assessment and the assessor's relevance determinations, may be simulated in conjunction with the bootstrap.

When presenting empirical results, we report magnitude of bias $|b|$, standard deviation σ, and root-mean-squared error RMSE.

3 THE BOOTSTRAP

To estimate bias and variance, we conduct repeated tests in which random factors associated with the selection of topics, and the selection of documents to assess, are simulated. For statistical methods like infAP and dynAP, we first repeat the sampling method 100 times for each of n_t topics in a reference collection. For each sample, we measure AP (not MAP, at this point) for each of n_s available system results, saving the results in an $n_t \times 100 \times n_s$ table. For non-statistical methods, like pooling and hedge [2], repeated per-topic measurements are unnecessary, so the table contains $n_t \times 1 \times n_s$ measurements.

We then draw $1,000$ samples of n'_t topics, with replacement, from the n_t topics of the reference collection. Typically, but not necessarily, $n'_t = n_t$, thus preserving the variance of the mean. For each topic t within a sample, and for each system s, an AP measurement is drawn from the table at position (t, r, s), where r is chosen at random. For each system, the n'_t measurements are averaged to yield a MAP score. The resulting MAP scores are used to rank the n_s systems.

Δ is estimated as the empirical distribution of $\delta(A, B)$, where A is the set of $1,000$ bootstrap rankings for one collection, and B is the set of $1,000$ rankings for another. Estimates of σ and $|b|$ follow.

4 EXPERIMENT 1: ASSESSMENT BUDGETS

For our first experiment, we assume the TREC 8 Ad Hoc test collection to yield ground truth. We are concerned with ranking not only the 129 systems whose results were submitted to TREC 8 for evaluation, but also dissimilar systems. To simulate the results of dissimilar systems, we engineered an additional 129 results, each derived from a distinct TREC result by randomly permuting the order of the relevant documents. As a consequence, corresponding submitted and engineered results have identical AP and MAP scores, and identical ground-truth ranks.

Using the ground-truth ranking for the submitted results, and additionally, the ground-truth ranking for the submitted and engineered results combined, we evaluate the bias, variance, and MSE of rankings derived from fewer assessments of the TREC collection.

The TREC collection contains 50 topics and 86,830 assessments, an average of 1,737 assessments per topic. Initially, we evaluated four methods of reducing assessment cost seventeen-fold to about 5,000 assessments:

- The well-known infAP estimator, applied to a probability-proportional-to-size (PPS) sample for each topic consisting of five documents drawn from each of 20 strata, for a total of 5,000 assessments;
- The recently released dynAP estimator, applied to the same PPS sample, for a total of 5,000 assessments;
- The trec_eval evaluator, applied to a variable number of documents per topic selected by depth-5 pooling, for a total of 5,542 assessments (avg. 111 per topic);
- trec_eval, applied to 100 documents per topic, selected by hedge, for a total of 5,000 assessments.

We then evaluated the effect of quadrupling the assessment budget in two ways: (i) by increasing the number of topics from 50 to 200; and (ii) by increasing the number of assessed documents per topic from 100 to 400. We further evaluated a budget of 80,000—about the same as at TREC—by quadrupling both the number of topics and the number of judgments per topic.

Table 1 shows ranking accuracy with respect to the submitted TREC results; Table 2 shows ranking accuracy with respect to the submitted results augmented by the engineered results. The top-left panel of each table shows accuracy for $5,000$ assessments. The result of quadrupling the number of topics is shown to the right, while the results of quadrupling the number of assessments per topic is shown below. The bottom row shows the accuracy of the reference TREC collection, under the assumption that it is unbiased.

For $5,000$ assessments and the TREC results, the RMSE results show little to choose between dynAP and hedge, with between 15% and 20% higher error than the reference collection. The $|b|$ and σ results show that dynAP has less than one-quarter the bias, but hedge has lower variance, with the net effect that their overall error is similar.

4.1 Effect of More Topics

From these results we can predict predict the effect of quadrupling the number of topics, which is borne out by the top-right panel: $|b|$ is essentially unchanged, while σ is approximately halved. The net effect is that RMSE for dynAP is approximately halved, about 17% higher than for the reference collection. σ for hedge is similarly halved but $|b|$ is unchanged, so RMSE is reduced by only one-third, about 55% higher than for the reference collection.

From the initial results we can only bound the effect of quadrupling the number of assessments per topic: $|b|$ and σ will both generally be reduced, but $b^2 + \sigma^2$ cannot fall below σ for the reference collection. The bottom-left panel confirms this prediction, except for the fact that the bootstrap estimate of RMSE for hedge is slightly smaller than for the reference collection. This apparently preposterous result result may be explained by random error in the bootstrap estimate. Overall, this panel suggests that, with a budget of 400 assessments per topic, hedge has slightly lower overall error, but still four-times higher bias, than dynAP. Nevertheless, the results for hedge—and all other methods—are far superior when the same overall budget of $20,000$ is apportioned over 200 topics with 100 assessments per topic, instead of 50 topics with 400 assessments per topic.

The effect of quadrupling the number of topics, with a budget of 400 assessments per topic, is the same as with a budget of 100 assessments per topic: $|b|$ is invariant, while σ is halved. Overall, for a budget of $80,000$ assessments, dynAP achieves an RMSE score that is insubstantially different from that achieved by a reference collection with *331,320* assessments.

	$\|b\|$	σ	RMSE	$\|b\|$	σ	RMSE
	5,000 assessments, 50 topics			*20,000 assessments, 200 topics*		
infAP	0.0309	0.0952	0.1001	0.0293	0.0488	0.0569
dynAP	0.0117	0.0907	**0.0914**	0.0124	0.0458	**0.0475**
depth-5	0.0800	0.0839	0.1160	0.0815	0.0419	0.0916
hedge	0.0500	0.0806	0.0948	0.0498	0.0398	0.0638
	20,000 assessments, 50 topics			*80,000 assessments, 200 topics*		
infAP	0.0113	0.0826	0.0833	0.0110	0.0419	0.0433
dynAP	0.0055	0.0806	0.0808	0.0033	0.0408	**0.0409**
depth-20	0.0381	0.0783	0.0871	0.0387	0.0394	0.0552
hedge	0.0159	0.0777	**0.0794**†	0.0154	0.0394	0.0423
	86,830 assessments, 50 topics			*331,320 assessments, 200 topics*		
Reference	0	0.0797	**0.0797**	0	0.0405	**0.0405**

Table 1: Accuracy of ranking TREC system results. (†) RMSE **values less than Reference are explained by chance error in the bootstrap estimate.**

	$\|b\|$	σ	RMSE	$\|b\|$	σ	RMSE
	5,000 assessments, 50 topics			*20,000 assessments, 200 topics*		
infAP	0.1510	0.0939	0.1778	0.1517	0.0480	0.1591
dynAP	0.0319	0.1078	**0.1125**	0.0347	0.0560	**0.0658**
depth-5	0.3071	0.0695	0.3149	0.3007	0.0360	0.3029
hedge	0.1263	0.0773	0.1481	0.1267	0.0389	0.1325
	20,000 assessments, 50 topics			*80,000 assessments, 200 topics*		
infAP	0.0495	0.0847	0.0981	0.0503	0.0430	0.0662
dynAP	-0.0145†	0.0856	**0.0843**	-0.0093†	0.0442	**0.0432**
depth-20	0.2013	0.0688	0.2127	0.1961	0.0338	0.1990
hedge	0.0866	0.0780	0.1166	0.0869	0.0398	0.0956
	86,830 assessments, 50 topics			*331,320 assessments, 200 topics*		
Reference	0	0.0797	**0.0797**	0	0.0405	**0.0405**

Table 2: Accuracy of ranking TREC and dissimilar system results. (†) Cases where the $b^2(X) < 0$ are explained by chance error in the bootstrap estimate, and reported as $-\sqrt{-b^2(X)}$.

4.2 Adversarial Testing

Table 2 affirms the same predictions regarding assessment-budget allocation, but tells a different story regarding the accuracy of the test collections. When the engineered results are ranked with the TREC results, bias is roughly tripled for all methods subject to a 5,000-document assessment budget, while variance is increased moderately. For infAP, depth-5, and hedge, bias dominates variance, calling into question whether these methods provide reasonable accuracy, regardless of their RMSE. The dynAP results represent a closer call. $|b|$ is one-third of σ, rendering it a small but noticeable component of RMSE. Quadrupling the number of topics exacerbates the influence of $|b|$, but still yields RMSE scores better than the reference collection for 50 topics.

Quadrupling the per-topic assessment budget dramatically reduces $|b|$ for dynAP, to the point that the bootstrap estimate of b^2 is negative (shown as $|b| = -\sqrt{-b^2}$), indicating that we are unable to distinguish $|b|$ from zero. When the number of topics is also quadrupled, we are still unable to distinguish $|b|$ from zero, while σ and RMSE are about 10% greater than RMSE for the reference collection with 200 topics; and half the RMSE for the reference collection with 50 topics.

5 EXPERIMENT 2: ALTERNATE ASSESSMENTS

Before TREC 2018, the University of Waterloo used Dynamic Sampling [5], and 19,161 of their own assessments to create relevance assessments (irels) as input the dyn estimator, thereby forming a test collection for the 2018 Common Core Track, which was released, along with the dyn estimator, as UWcore18[1] [9]. To form the official TREC test collection, NIST (re)assessed the 19,161 documents, and 5,767 additional documents selected by a combination of depth-10 and move-to-front pooling, to create the official TREC relevance assessments (qrels) as input to trec_eval. To examine the impact on system ranking of using Waterloo versus TREC assessments, we compare UWcore18 to UWcore18N, in which the Waterloo assessments are replaced by NIST assessments. To examine the impact of using an additional 5,767 TREC assessments, while eschewing dyn in favour of trec_eval, we compare UWcore18N to the NIST test collection.

Table 3 shows the accuracy with which the three test collections rank 69 of the 73 system results submitted to TREC, excluding the four submitted by Waterloo. The top panel shows inter-collection

G	UWcore18	UWcore18N	NIST
		$-\|b\|-$	
UWcore18	0.0067†	0.0887	0.0752
UWcore18N	0.0364	0.0064†	0.0364
NIST	0.0710	0.0369	0.0084†
		−RMSE−	
UWcore18	0.0828	0.1210	0.1090
UWcore18N	0.1186	0.0826	0.0869
NIST	0.1089	0.0903	0.0794

Table 3: Accuracy results for alternative TREC 2018 Common Core Collections. The top panel shows pairwise bias; the bottom panel shows RMSE**. (†) In the top panel, asymmetry and differences from** 0 **on the diagonal are explained by chance error in the bootstrap estimate.**

bias $|b|$ using each collection as ground truth to measure the bias of each other collection, including itself. The top panel should be symmetric, and its diagonal should be 0. Deviations from this prediction may be attributed to random error in the bootstrap estimates. The diagonal of the bottom panel shows σ for each collection, while the non-diagonal elements show inter-system RMSE. Our theory predicts that the bottom panel will be symmetric only if the σ values are equal, which they nearly are.

There is substantial bias—about equal to σ—between UWcore18 and UWcore18N, indicating that some systems score better according to the Waterloo assessments, while others score better according to NIST's. We are not in a position to render an opinion on which set of assessments is better and suggest that evaluation using the two sets of assessments should be considered to be separate experiments, and both results considered when comparing system results.

Bias between UWcore18N and NIST, while smaller in magnitude, may be more of a concern, because it reflects different measurements of ostensibly the same value. The results from our first experiment show that shallow pooling methods exhibit strong bias, while dynAP does not. Together, these results suggest that the inter-collection bias $|b| = 0.0364$ may be attributed in large part to the NIST collection. If this were the case, it would offset an apparent advantage in σ for NIST.

We posit that better results would have been achieved, had the budget of 5,767 NIST assessments allocated to shallow pooing been used to assess an additional 15 topics using Dynamic Sampling. Using our bootstrap simulation, we project that UWcore18N, if extended to 65 topics, would achieve $\sigma = 0.0724$, lower than the official 50-topic collection.

The evidence suggests the current UWcore18N test collection, notwithstanding its slightly higher variance, is preferable to the official TREC test collection, because it is less likely to be biased, particularly with respect to novel IR methods.

6 DISCUSSION AND LIMITATIONS

Kutlu et al. [12] provide an excellent survey of rank-similarity measures and their shortcomings, in support of a "significance aware" approach that takes into account the results of pairwise significance tests. Our approach exposes the variances of G and X directly—separate from bias—rather than obliquely through an amalgam of test results, distilled into a dimensionless overall score. Previous work has evaluated fairness and stability separately, by perturbing the topics or the results to be ranked so as to calculate a separate summary measure for each [3, 14].

Our bootstrap sample was inadequate to quantify b when $\sigma \gg |b| \approx 0$. Further analytic and empirical work is needed to compute confidence intervals for the bootstrap estimates. In theory, the variance of the variance estimates can be determined from the distributions of G and X, but those estimates should be verified by meta-experiments.

Our experiments rely on on an assumption—contradicted by the evidence [13]—that an assessor will give the same relevance determination for a given document, regardless of the sampling strategy. Whether real assessor behaviour would have any substantive positive or negative effect on rankings remains to be determined.

Orthogonal estimates of bias and variance have predictive ability lacking in a single score conflating the two, or separate uncalibrated scores. Harnessing this predictive ability, we offer evidence that shallow pooling methods introduce unreasonable amounts of bias, while offering hardly lower variance than dynAP, which represents a substantial improvement over infAP. Based on this evidence, we have reason to suggest that the UWcore18 or UWcore18N statistical test collections are more accurate than the official TREC 2018 Common Core test collection.

REFERENCES

[1] Allan, J., Harman, D., Voorhees, E., and Kanoulas, E. TREC 2018 Common Core Track overview. In *TREC 2018*.

[2] Aslam, J. A., Pavlu, V., and Savell, R. A unified model for metasearch, pooling, and system evaluation. In *CIKM 2003*.

[3] Buckley, C., and Voorhees, E. M. Evaluating evaluation measure stability. In *SIGIR 2000*.

[4] Carterette, B. On rank correlation and the distance between rankings. In *SIGIR 2009*.

[5] Cormack, G. V., and Grossman, M. R. Beyond pooling. In *SIGIR 2018*.

[6] Cormack, G. V., and Grossman, M. R. Unbiased low-variance estimators for precision and related information retrieval effectiveness measures. In *SIGIR 2019*.

[7] Cormack, G. V., and Lynam, T. R. Power and bias of subset pooling strategies. In *SIGIR 2007*.

[8] Cormack, G. V., and Lynam, T. R. Statistical precision of information retrieval evaluation. In *SIGIR 2006*.

[9] Cormack, G. V., Zhang, H., Ghelani, N., Abualsaud, M., Smucker, M. D., Grossman, M. R., Rahbariasl, S., and Grossman, M. R. Dynamic sampling meets pooling. In *SIGIR 2019*.

[10] Harman, D. K. The TREC Test Collections. In *TREC: Experiment and Evaluation in Information Retrieval* (2005), E. M. Voorhees and D. K. Harman, Eds., pp. 21–52.

[11] International Organization for Standardization. *ISO 5725-1: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions*. International Organization for Standardization Geneva, Switzerland, 1994.

[12] Kutlu, M., Elsayed, T., Hasanain, M., and Lease, M. When rank order isn't enough: New statistical-significance-aware correlation measures. In *CIKM 2018*.

[13] Roegiest, A., and Cormack, G. V. Impact of review-set selection on human assessment for text classification. In *SIGIR 2016*.

[14] Sakai, T. On the reliability of information retrieval metrics based on graded relevance. *Inf. Process. Manag. 43*, 2 (2007), 531–548.

[15] Voorhees, E., and Harman, D. Overview of the Eighth Text REtrieval Conference. In *TREC 8* (1999).

[16] Voorhees, E. M. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manag. 36*, 5 (2000).

[17] Yilmaz, E., Aslam, J. A., and Robertson, S. A new rank correlation coefficient for information retrieval. In *SIGIR 2008*.

[18] Yilmaz, E., Kanoulas, E., and Aslam, J. A. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR 2008*.