Dynamic Sampling Meets Pooling

Gordon V. Cormack, Haotian Zhang, Nimesh Ghelani, Mustafa Abualsaud, Mark D. Smucker, Maura R. Grossman, Shahin Rahbariasl, and Amira Ghenai
University of Waterloo, Ontario, Canada

ABSTRACT

A team of six assessors used Dynamic Sampling (Cormack and Grossman 2018) and one hour of assessment effort per topic to form, without pooling, a test collection for the TREC 2018 Common Core Track. Later, official relevance assessments were rendered by NIST for documents selected by depth-10 pooling augmented by move-to-front (MTF) pooling (Cormack et al. 1998), as well as the documents selected by our Dynamic Sampling effort. MAP estimates rendered from dynamically sampled assessments using the `xinfAP` statistical evaluator are comparable to those rendered from the complete set of official assessments using the standard `trec_eval` tool. MAP estimates rendered using only documents selected by pooling, on the other hand, differ substantially. The results suggest that the use of Dynamic Sampling without pooling can, for an order of magnitude less assessment effort, yield information-retrieval effectiveness estimates that exhibit lower bias, lower error, and comparable ability to rank system effectiveness.

ACM Reference Format:
Gordon V. Cormack, Haotian Zhang, Nimesh Ghelani, Mustafa Abualsaud, Mark D. Smucker, Maura R. Grossman, Shahin Rahbariasl, and Amira Ghenai. 2019. Dynamic Sampling Meets Pooling. 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/10.1145/3331184.3331354

1 INTRODUCTION

This work evaluates the first *in situ* application of Dynamic Sampling (DS) [3] to construct a test collection for large-scale information retrieval (IR) evaluation (*see* [9]). In contrast to the de facto standard pooling method, DS estimates IR effectiveness measures like MAP by applying a statistical estimator to a sample of a very large universe of documents, drawn independently from any system to be evaluated. DS promises to yield unbiased estimates not only for the methods known at the time of construction, but also for methods yet-to-be invented. Prior to this effort, DS had been validated only through post-hoc simulation.

In preparation for the TREC 2018 Common Core Track [2], we built a test collection by employing a combination of Continuous Active Learning (CAL) [5], Interactive Search and Judging (ISJ) [7], and DS to identify a sample of 19,161 documents from a universe of 39,214, each of which we assessed to be *relevant* or *not relevant*. The 19,161 documents were shared with NIST, where they were again assessed, along with 8,902 documents identified by depth-10 pooling, and 1,010 identified by move-to-front (MTF) pooling [7] to form the 26,233 official TREC relevance assessments (qrels).

A primary concern was the speed with which we could render the assessments necessary to form a viable test collection, so as to conform to our particular resource constraints, as well as those of anyone else who might wish to employ our method. After an exploratory phase in which one author used a rudimentary CAL tool and search engine to find as many relevant documents as possible in about 13 minutes per topic, a separate team of five of the authors spent about 45 minutes per topic using HiCAL [1] [1, 12], which we adapted for this purpose, to conduct further searches and to draw a dynamic stratified sample of 300 additional documents per topic for assessment. In total, about 50 person-hours (*i.e.,* one hour per topic) were devoted to compiling our relevance assessments.

To evaluate the outcome of our DS effort, we consider the following five questions:

- How nearly *all* of the relevant documents were included in the universe of documents identified using DS, from which the sample was drawn?
- What is the accuracy of MAP estimates derived from the sample, measured by the average difference (*bias*) and RMS difference between the estimates and ground truth, per the official NIST assessments?
- What is the accuracy with which systems are ranked according to MAP estimates, measured by the rank correlation (τ or τ_{AP}), with the ranking afforded by ground truth?
- How do these outcomes compare to the NIST depth-10 and MTF pooling efforts?
- How might our or NIST's efforts have yielded better outcomes, with a different allocation of assessment effort?

Overall, our results indicate that assessment budgets would be better spent if allocated to DS rather than to pooling.

2 SPEEDY ASSESSMENT

In the first of three distinct assessment phases, one of the authors used CAL, as implemented in the TREC 2015 Total Recall Track Baseline Implementation (BMI),[2] to identify potentially relevant paragraphs, which were rendered for assessment, along with the document containing them, on an ASCII terminal. Of 3,601 documents rendered during this phase, 1,391 were assessed relevant. For a few topics where no relevant documents were seen within the first several dozen rendered documents, the assessor reverted to a search engine,[3] finding a total of 47 more relevant documents over all of these topics. Nevertheless, the assessor found no relevant documents for several topics, and fewer than ten relevant

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

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331354

[1] *See* https://github.com/hical.
[2] *See* http://cormack.uwaterloo.ca/trecvm/.
[3] *See* http://stefan.buettcher.org/cs/wumpus/index.html.

Method:	DS	NIST	Depth-10	Depth-10 + MTF
Avg. Coverage:	0.88	0.82	0.47	0.49
Min. Coverage:	0.58	0.48	0.11	0.13
Max. Coverage:	1.00	1.00	1.00	1.00

Table 1: Coverage of document-selection methods.

documents for 20 topics. Phase 1 consumed a total of 11.1 hours (*i.e.*, 13.2 minutes per topic).

The second and third phases were conducted by a team of five different authors, who used the HiCAL system to render assessments for 10 topics each. Phase 2 involved the use of HiCAL's search and full-document-display mode to find more documents relevant to the 20 topics for which the first phase had found fewer than 10. Phase 2 was allocated 30 minutes of assessment time for each of these 20 topics (*i.e.*, 12 minutes per topic overall).

Phase 3 involved the use of HiCAL's CAL and paragraph-only-display mode to present paragraphs, excerpted from potentially relevant documents, for assessment. HiCAL was modified to present excerpts from only a sample of documents, selected using DS with strategy D25 [3] (*see* Section 7), and a budget of 300 assessments per topic. The initial training set consisted of all assessments from phases 1 and 2. Assessment time for phase 3 averaged 33 minutes per topic, bringing the total for all three phases to about one hour per topic.[4]

3 COVERAGE

We define *coverage* to be the fraction of all relevant documents in the DS universe. In other words, coverage is the recall of the DS effort, before sampling and before relevance assessment. In order to estimate coverage of our assessment strategy compared to others, it is necessary to estimate the number of relevant documents for each topic. To this end, DS provides an unbiased statistical estimate of the number of relevant documents in the universe it identifies as the sample space from which its sample is drawn. Here, the size of the universe was 39,214 documents, from which a sample of 19,161 was drawn. Pooling and MTF independently identified 1,305 documents—of which 503 were relevant—from the DS universe, which were counted directly and removed from the sample space. Pooling and MTF identified 404 additional relevant documents outside the universe. Overall, our best estimate indicates that there are 5,457 relevant documents in the corpus.

Table 1 shows the average, minimum, and maximum coverage over 50 topics achieved by the construction methods under consideration. The DS universe covers 88% of all relevant documents, on average, and 58%, in the worst case, for this collection. The NIST qrels have lower coverage because, although they include more assessments, they do not constitute a statistical sample and cannot be extrapolated to a larger population. The depth-10 pool, whether augmented or not by MTF, has substantially inferior recall.

[4]Due to a system misconfiguration, assessors spent several additional minutes in a false start for phase 3, the results of which were discarded.

4 ACCURACY

Figure 1 shows scatterplots and summary statistics comparing the MAP estimates afforded by DS using the xinfAP estimator[5] [11] to the official TREC results, for each of the runs submitted to the TREC 2018 Common Core Track. Points denoted "M" represent so-called manual runs; points denoted "R" indicate runs that relied on legacy relevance assessments for the same topics but a different corpus; points denoted "○" indicate fully automatic runs. We see that our DS assessments yield low bias[6] (b = 0.02) and RMS error (RMSE = 0.02), with negligible variance ($\sigma^2 = b^2 - \text{RMSE}^2 \approx 0$). $\tau = 0.88$ and $\tau_{AP} = 0.83$ correlation scores [10] are typical of those arising from inter-assessor disagreement.

The top right panel shows the result of substituting NIST assessments in place of our own, to eliminate inter-assessor disagreement. $\tau = 0.95$ and $\tau_{AP} = 0.92$ are substantially higher, while bias essentially vanishes, exposing a small variance as evidenced by RMSE = 0.01.

The bottom two panels show results for depth-10 pooling, and depth-10 pooling augmented by MTF. Rank correlations are slightly lower, while bias and error are substantially higher.

5 STATISTICAL GROUND TRUTH

The method we employed to estimate coverage (Section 3) yields a revised but statistically valid sample of the DS universe, augmented by 404 relevant documents outside the DS universe. Statistical estimates derived from this sample using xinfAP—which we dub *statistical ground truth*—should, in theory, be more accurate than the *official ground truth* derived from from the same assessments using `trec_eval`.

We can evaluate the extent to which the statistical and official ground truths agree. The left panel of Figure 2 shows strong but imperfect agreement. The right panel of Figure 2 and the top-right panel of Figure 1, evaluate accuracy of the same NIST-assessed DS sample, according to the statistical and official ground truth, respectively. The statistical ground truth indicates much higher accuracy. If the statistical ground truth is indeed the gold standard, this result suggests that the dynamic sample alone—without any documents discovered by pooling or MTF—may also yield ground truth more accurate than the official ground truth.

6 INCREASING COVERAGE

We used a fixed budget of 300 assessments for the third phase of our assessments, which had previously been shown to achieve good results for the D25 sampling strategy [3] (*see* Section 7). Arguably good results were achieved here; however, Table 1 indicates that coverage for at least one topic was less than 60%. We investigated whether the cause of occasionally poor coverage was assessor disagreement or inadequacy of the assessment budget.

Figure 3 shows coverage for each of the 50 topics, as well as the average, as a function of the total number of assessments for the three phases. We see that the slope for most of the curves is near-zero when the budget is exhausted. But for several of the topics—notably those with coverage of less than about 80%—the slope is noticeably positive, indicating that, had the budget been extended

[5]trec.nist.gov/data/clinical/sample_eval.pl

[6]The arithmetic mean of error.

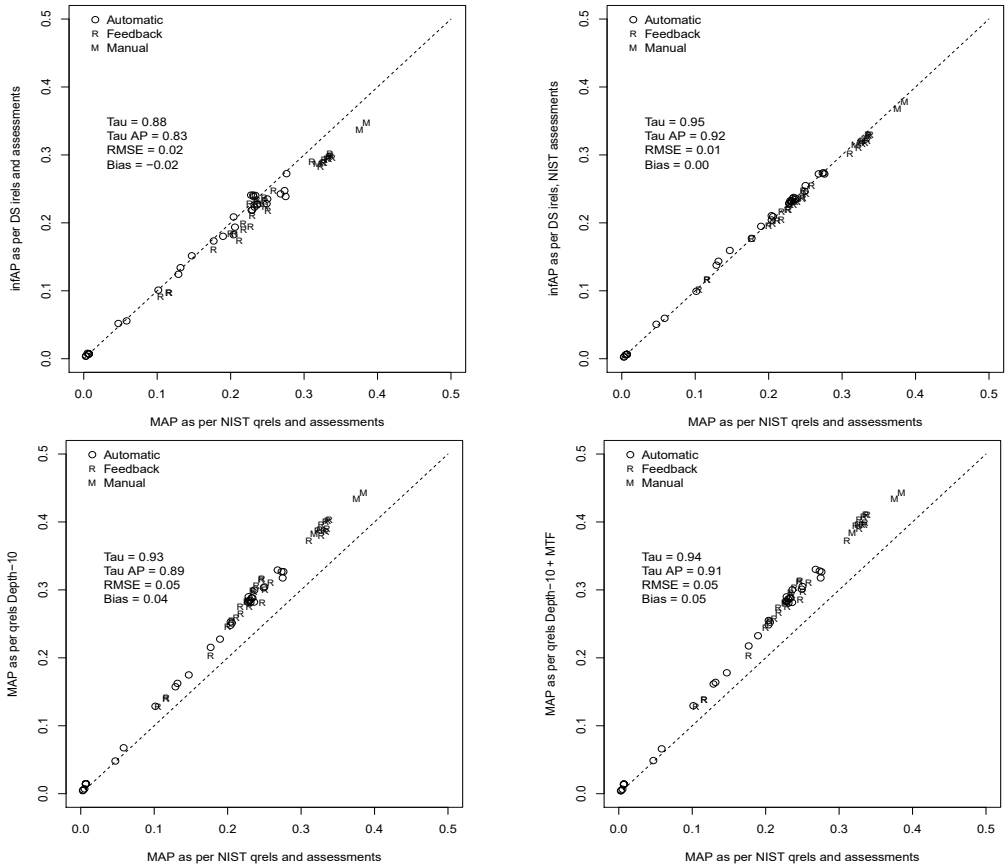

Figure 1: Accuracy of MAP estimates using DS vs. pooling methods, compared to official TREC 2018 Common Core Track evaluation. The top-left panel shows results for DS with relevance assessments by the authors; the top-right panel shows results for DS with official relevance assessments by NIST. The bottom-left panel shows results for the depth-10 pool with relevance assessments by NIST; the bottom-right shows results for the depth-10 pool augmented by MTF, with relevance assessments by NIST.

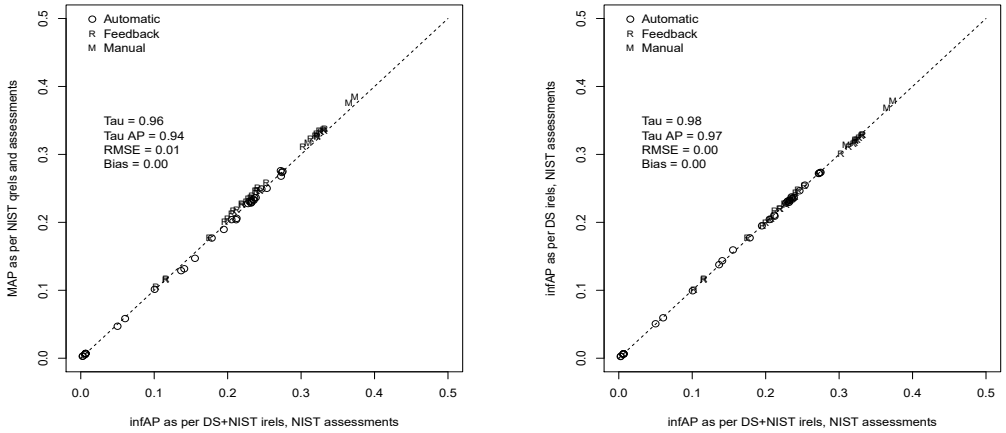

Figure 2: Accuracy of official NIST qrels (left) and DS with NIST assessments (right), according to statistical ground truth.

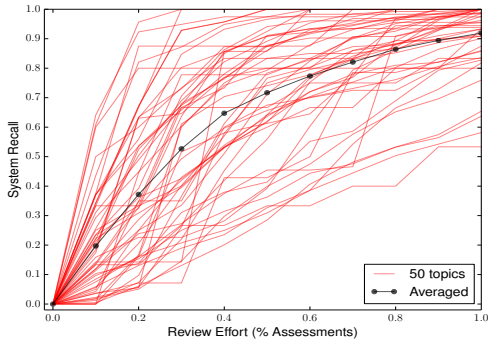

Figure 3: Coverage as a function of overall assessment effort.

for these topics, higher coverage would have been achieved. We also note that several topics achieving ostensibly high coverage have substantial slope when the budget is exhausted, suggesting a shortfall in the ground truth estimate of the number of relevant documents. These results suggest that using the "Knee Method" as a stopping criterion, which has been shown to work well for CAL [4], might be preferable to a fixed assessment budget.

A similar approach might also have yielded better results for MTF, which was constrained by a restrictive assessment budget [2].

7 DISCUSSION AND LIMITATIONS

For brevity, we present MAP results estimated by `xinfAP`, consistent with common practice. In a companion article [6], Cormack and Grossman show that better estimates of MAP, as well as precision at cutoff (P@k), rank-biased-precision (RBP), and normalized discounted cumulative gain (NDCG) may be derived from a DS test collection, using the purpose-built `dyn_eval` estimator.

Estimates derived from DS assume that the DS universe includes substantially all relevant documents; DS yields an unbiased, non-uniform statistical sample from which MAP and other evaluation measures are derived. Hence, the effectiveness of any run—not just a member of an evaluation pool—may be evaluated using a DS collection.

One can increase coverage by increasing the size of the universe, at the expense of higher variance. The evidence presented here suggests the the DS universe does indeed contain a substantial majority of the relevant documents. Future work may explore the influence of three parameters that balance the tension between coverage, sampling budget, and skew of the sampling rate in favour of likely relevant documents. The sampling budget of 300 was occasioned by our target of one hour of assessment time per topic. Given the fact that we had no pool of runs, we were further constrained to use content-only features for the learner. Within these constraints, strategies D12, D25, and D50, reflecting different tradeoffs between coverage and sampling rate, appeared equally good [3], and we chose the median. Further investigation might yield a better tradeoff. If it were logistically feasible, a flexible assessment budget with an average of 300 documents per topic, and an amenable stopping criterion, might have yielded better results, for the same overall assessment budget.

We have measured the effect of assessor disagreement only on the MAP estimates derived from identical sets of assessments. Our results show that DS conducted by one set of assessors (the authors) can achieve high coverage in the eyes of another (the NIST assessors). If NIST assessors had conducted the DS assessments, would coverage have been higher, and, if so, how much higher?

Finally, we note that the relevance determinations of assessors are influenced by context; in particular, the order and richness of the documents they are shown [8]. DS assessors are shown the most-likely relevant documents first, which suggests they are likely be more stringent in their judgment of relevance, at least at the outset. As the density of relevant documents inevitably decreases, some assessors may have a tendency to "reach" and thus, be more likely to judge a document relevant than at the outset. Our design controls for this possible effect with respect to the NIST assessments for the DS sample, because the NIST assessors were unaware of the order in which the DS documents were discovered, or, indeed, whether a document was identified by pooling, by DS, or both. Our assessments with respect to the DS sample, and NIST's assessments with respect to MTF, might have been so influenced.

8 CONCLUSION

Independent of NIST assessments or any pool of submitted runs, a small team of researchers spent 50 hours to create a set of sampled relevance assessments that effectively scores and ranks systems according to MAP over 50 topics. This level of effort represents an order-of-magnitude reduction in human effort, compared to typical efforts like TREC, and does not rely on pooling, which is both logistically challenging and a potential source of bias against systems not contributing to the pool. DS avoids this source of bias and, as both theory and empirical evidence show, does not introduce bias against the TREC runs that had no influence on the DS selection strategy, or our relevance assessments.

ACKNOWLEDGMENTS

Special thanks to Ellen M. Voorhees for integrating the results of our DS process into the official NIST assessment effort.

REFERENCES

[1] Abualsaud, M., Ghelani, N., Zhang, H., Smucker, M. D., Cormack, G. V., and Grossman, M. R. A system for efficient high-recall retrieval. In *SIGIR 2018*.
[2] Allan, J., Harman, D., Kanoulas, E., and Voorhees, E. TREC 2018 Common Core Track overview. In *TREC 2018*.
[3] Cormack, G. V., and Grossman, M. R. Beyond pooling. In *SIGIR 2018*.
[4] Cormack, G. V., and Grossman, M. R. Engineering quality and reliability in technology-assisted review. In *SIGIR 2016*.
[5] Cormack, G. V., and Grossman, M. R. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *SIGIR 2014*.
[6] Cormack, G. V., and Grossman, M. R. Unbiased low-variance estimators for precision and related information retrieval effectiveness measures. In *SIGIR 2019* (2019).
[7] Cormack, G. V., Palmer, C. R., and Clarke, C. L. Efficient construction of large test collections. In *SIGIR 1998*.
[8] Roegiest, A., and Cormack, G. V. Impact of review-set selection on human assessment for text classification. In *SIGIR 2016*.
[9] Sanderson, M., et al. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval 4*, 4 (2010).
[10] Yilmaz, E., Aslam, J. A., and Robertson, S. A new rank correlation coefficient for information retrieval. In *SIGIR 2008*.
[11] Yilmaz, E., Kanoulas, E., and Aslam, J. A. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR 2008*.
[12] Zhang, H., Abualsaud, M., Ghelani, N., Smucker, M. D., Cormack, G. V., and Grossman, M. R. Effective user interaction for high-recall retrieval: Less is more. In *CIKM 2018*.