Too Many Relevants: Whither Cranfield Test Collections?

Ellen M. Voorhees
National Institute of Standards and Technology
Gaithersburg, MD, USA

Nick Craswell
Microsoft
Bellevue, WA, USA

Jimmy Lin
David R. Cheriton School of Computer Science
University of Waterloo
Canada

ABSTRACT

This paper presents the lessons regarding the construction and use of large Cranfield-style test collections learned from the TREC 2021 Deep Learning track. The corpus used in the 2021 edition of the track was much bigger than the corpus used previously and it contains many more relevant documents. The process used to select documents to judge that had been used in earlier years of the track failed to produce a reliable collection because most topics have too many relevant documents. Judgment budgets were exceeded before an adequate sample of the relevant set could be found, so there are likely many unknown relevant documents in the unjudged portion of the corpus. As a result, the collection is not reusable, and furthermore, recall-based measures are unreliable even for the retrieval systems that were used to build the collection. Yet, early-precision measures cannot distinguish among system results because the maximum score is easily obtained for many topics. And since the existing tools for appraising the quality of test collections depend on systems' scores, they also fail when there are too many relevant documents. Collection builders will need new strategies and tools for building reliable test collections for continued use of the Cranfield paradigm on ever-larger corpora. Ensuring that the definition of 'relevant' truly reflects the desired systems' rankings is a provisional strategy for continued collection building.

CCS CONCEPTS

• **Information systems** → **Relevance assessment**; **Test collections**.

KEYWORDS

Cranfield, reusability, score saturation, test collections

ACM Reference Format:
Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too Many Relevants: Whither Cranfield Test Collections?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531728

1 INTRODUCTION

The Text REtrieval Conference (TREC) project has a long-standing goal of building general-purpose, reusable information retrieval test collections to support research in the field. By *general-purpose* we mean that the collection captures an abstraction of diverse user tasks (as encoded in the evaluation measures supported); by *reusable* we mean that evaluation scores induced by the collection are unbiased even for systems that did not participate in the construction of the collection. And since the collection-building process must be implemented in the real world, the building process is subject to a budget, which we measure in the number of human relevance judgments required.

The ad hoc collections built in the early years of TREC are examples of the type of Cranfield-style collections we wish to build. Those collections were built using *pooling* [16] to deep ranks over many diverse runs (retrieval results)—in other words, they are high-quality collections but they were expensive to build. Since then, document corpus size has continued to grow, and because the effectiveness of pooling depends in part on corpus size [2], building test collections using sufficiently deep pools is no longer feasible. There has been wide-ranging research on how to appropriately evaluate retrieval systems on large document corpora since pooling alone ceased to be viable. These approaches generally required forfeiting one of the other desirable attributes such as sampling in support of specific evaluation measures (as in inferred measures [21] or the TREC Legal track [17]), which sacrifices generality, or restricting comparisons to a known set of retrieval systems (such as in Minimal Test Collection processing [4]), which sacrifices reusability.

Another line of research has focused on keeping the generality and reusability attributes of the test collections while controlling the budget by forming judgment sets through dynamic sampling of runs [6, 7, 11–13]. In dynamic sampling, the selection of which document to judge next is made after the current document is judged; which document is selected depends on the ranks at which relevant documents were retrieved across the set of submitted runs. The test collections built in each year of the TREC Common Core (2017–2018) and Deep Learning (2019–2021) tracks were built using dynamic judging [8, 19]. The quality of the resulting collections was satisfactory until the TREC 2021 Deep Learning track where a similar process and budget as earlier years resulted in a clearly inferior collection. This paper provides a postmortem of why the building process failed and the implications for future test collections. Simply put, the document corpus for the TREC 2021 track was significantly larger than in previous years and this resulted in too many relevant documents. The large number of relevant documents not only prevented an unbiased sample of the relevant set from being obtained in the allotted budget, making estimates of recall-based measures such as Mean Average Precision (MAP) unreliable, but also saturated high-precision measures, rendering them unable to discriminate among systems.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531728

Increasing the judgment budget to obtain a reliable collection may work in the short term for the TREC Deep Learning track, though we estimate we would need at least several times the budget as was available in 2021 and the assumption of a constantly growing budget is as untenable now as it was for pooling. More judgment budget also does not address the problem of high-precision saturation (which is caused by too many known relevant documents); it only makes deeper measures more reliable. Defining 'relevant' to be more representative of desired system outcomes is a solution to the score saturation problem, provided the definition continues to evolve with corpus size. Unfortunately, more stringent definitions of relevance do not ameliorate the reusability concerns unless search systems can reliably retrieve the new target documents at high ranks.

2 THE 2021 DEEP LEARNING TRACK

The TREC Deep Learning track studies information retrieval in a large training data regime [8]. The track uses the MS MARCO data set[1] that contains relevance judgments for hundreds of thousands of queries (generally a single positive judgment per query). MS MARCO contains two separate corpora, one containing web documents and the other containing passages extracted from web documents, and the track has two corresponding ad hoc retrieval tasks: Document Ranking and Passage Ranking.[2] Both tasks use the same test set of several hundred queries.

The Deep Learning track has run for three years so far and is designed to build a Cranfield-style test collection over the data by generating more comprehensive relevance judgments (called *qrels*) for a smaller number of queries. The overall strategy to obtain relevance judgments was the same in each year, and was similar to the strategy used to produce judgments for the earlier Common Core TREC track [19]. Once the participant runs were submitted, track organizers down-sampled the test set of queries by using the sparse MS MARCO judgments to eliminate queries that had a median Reciprocal Rank (RR) score of either 1.0 or 0.0. Queries randomly selected from the remainder of the test set entered the judgment process. For each query, TREC assessors first judged depth-10 pools created from all submissions and then started a dynamic judgment protocol that ended when 1) the fraction of judged documents that were relevant (called the *relevant density*) was small enough, 2) when the number of relevant documents found was so large the query was abandoned, or 3) when the total judgment budget was exhausted. The final evaluation set of queries was that set of queries whose relevant density was small enough and that had at least a minimum number (most often, three) of relevant documents. Each query that started the process was judged for both Documents and Passages by the same assessor, but otherwise the Documents and Passages judgment process was separate. The collections built for the Document Ranking task and Passage Ranking task could thus have different numbers of topics in the final evaluation set and must be treated as independent test collections.

The main difference of the 2021 track from earlier years was the introduction of new versions of the corpora (MS MARCO v2), which were significantly larger than the earlier version. The 2021 Documents corpus contains just under 12M documents—3.7 times as large as the version 1 Documents corpus—and the Passages corpus contains just over 138M passages—15.6 times as large as the version 1 corpus. The track received 66 submissions to the Document Ranking task and 63 submissions to the Passage Ranking task. These run counts include a set of Baseline runs submitted to facilitate comparisons among traditional and neural retrieval methods. The remainder of this section provides the details of the judging process for the 2021 track and demonstrates the score saturation for high-precision measures that resulted.

2.1 TREC 2021 Qrels

The dynamic assessing protocol used for the 2021 track was a series of Continuous Active Learning® (CAL) [5] iterations. For each topic, the CAL process was given all of the judgments that had been made up to that point and it then ranked the remainder of the collection by likelihood of relevance. The top X documents from that ranking were given to the assessor to judge next, where X varied based on the number of known relevant documents and on practical concerns like giving assessors a reasonable increment each day; most frequently, X was 20 or 25. Assessors continued to judge a topic if the relevant density was more than one half; if fewer than 150 documents had been judged for it so far; if the topic had not yet had any documents suggested by CAL judged (i.e., each topic had at least one CAL iteration performed); or if more than 20% (10% in early iterations) of the documents in the most recent previous iteration had been judged relevant. Once a topic exited the process, it was not restarted.

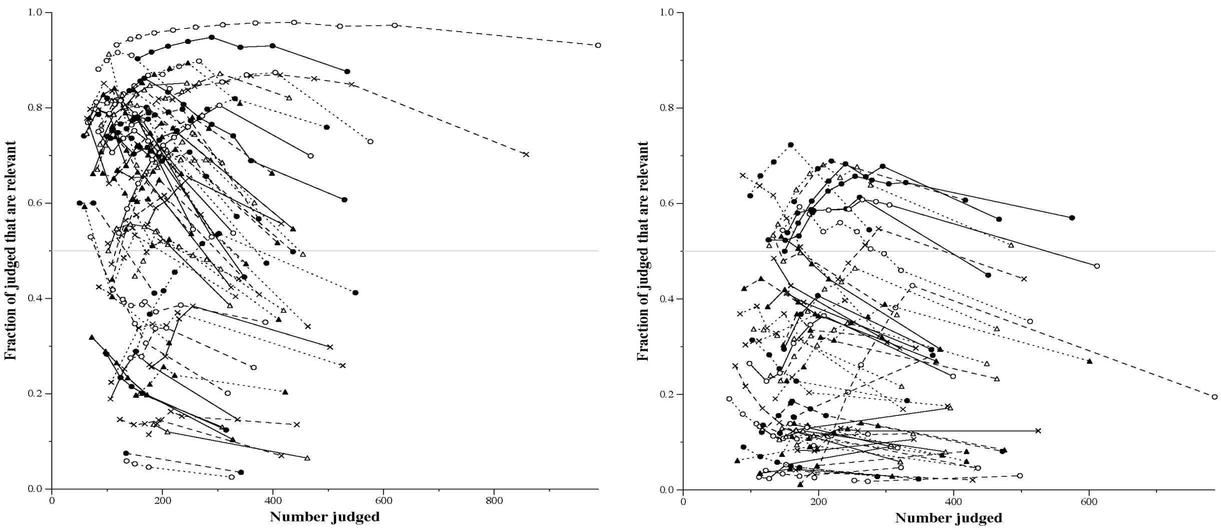
The Document Ranking task used a four-point judgment scale of Not Relevant (0), Relevant (1), Highly Relevant (2), and Perfect (3). Since the aim of the Passage Ranking task was to have systems retrieve succinct answers to the question implied by the query, the Passage Ranking task used a different four-point scale: Not Relevant (0), Related (1), Highly Relevant (2), and Perfect (3). All variants of "Relevant" were treated as relevant by CAL and the evaluation measures that use binary relevance. Since "related" means that the passage was on-topic but didn't actually answer the question, it was treated as *not* relevant for the Passage Ranking task. For variants of the NDCG measure, the gain values used are the scale values (i.e., 0–3) for both tasks.

Continuous Active Learning is most frequently used in a workflow in which the remainder of the collection is reranked after each individual judgment is made. However, TREC assessors work asynchronously, and balancing the workload across assessors and appropriately allocating the overall judgment budget to individual topics is logistically challenging when using a fully interactive version of CAL. We introduced batch sizes of 20 or so documents between CAL iterations in 2021 as a compromise between logistical necessity and theoretical best performance of CAL.

Judging on a topic stopped if all of the previously mentioned conditions were met or when the overall judgment budget was exhausted. Most topics, especially in the Document Ranking task, continued until the budget was exhausted, resulting in a large number of topics for which the relevant density remained greater than one half. The relevant density is a frequently used heuristic for

[1] https://microsoft.github.io/msmarco/

[2] For convenience, this paper will use 'document' as a generic unit of retrieval except when the distinction between the tasks is important.



Document Ranking task

Passage Ranking task

Figure 1: Relevant density for topics in the evaluation set for the Document Ranking task collection (left) and the Passage Ranking task collection (right).

deciding whether a topic is likely to have more relevant documents remaining in the as-yet-unjudged set [19]. Most of the topics in the high-quality TREC ad hoc collections have densities less than one third, and topics with densities greater than one half are almost certain to have more relevant.[3] In previous years we dropped topics with densities greater than one half from the evaluation set, but could not do that this year because it would leave too few topics in the evaluation set.

Figure 1 plots the relevant density against the number of judged documents for each CAL iteration for both the Document Ranking (left) and Passage Ranking (right) tasks. Each line in a plot connects the density at the conclusion of a round of judging for a single topic. The first (left-most) point on a line is the density for documents in the depth-10 pool. Each subsequent point, except the last point, is the density after an additional CAL iteration. The last point is the density after a post-TREC round of judging took place as discussed in Section 4. The Passage Ranking task had more stringent relevance criteria and consequently had fewer relevant passages overall and fewer topics with densities greater than one half. For the Document Ranking task, 40 of the 57 topics had relevant densities greater than one half at the end of track judging.

Since documents are judged by decreasing likelihood of being relevant (either by the runs as represented by the depth-10 pools or explicitly by CAL), the expected trajectory of the relevant density is to decrease as more documents are judged. Cormack and Grossman make use of this trajectory to define the "knee" method for determining when to stop searching in a high-recall retrieval task [6]. The knee is a region of the line where the rate of finding new relevant documents begins leveling off after decreasing sharply. As shown in Figure 1, there is little evidence of a knee for most topics, and in fact both the Document Ranking and Passage Ranking tasks have topics for which the density *increases* as more judgments are made. This is strong evidence that many more relevant documents remain in the unjudged portion of the corpus.

Table 1: Judgment counts per task.

	Document Ranking	Passage Ranking
# topics judged	57	57
# topics in eval set (qrels)	57	53
min judgments per topic	75	80
max judgments per topic	620	339
mean judgments for qrels	229.1	204.3
total judgments in qrels	13,058	10,828
total judgments made	13,058	11,556

With little alternative, we retained all topics that were judged in the evaluation set for the Document Ranking task collection. For the Passage Ranking collection, four of the 57 topics had fewer than five relevant passages, so they were dropped from the evaluation. Table 1 gives statistics for the number of documents judged for each task. We will refer to the qrels created from this set of judgments as the TRACK judgments. The test collections can be downloaded from the TREC web site at https://trec.nist.gov/data/deep2021.html.

2.2 Score Saturation

Very incomplete judgments not only cause test collections to be not reusable, but also affect the scores of participants' runs for recall-based measures. In particular, it is unlikely that Mean Average Precision (MAP) values computed for track submissions are very

[3] The TREC-COVID collection is an exception to this heuristic. Voorhees and Roberts [20] attributed the quality of the collection despite the fact that some topics have high relevant densities to the fact that nearly one percent of the entire corpus was judged per topic in that test collection.

accurate. Accordingly, we focus on Precision at ten documents retrieved (P@10) and normalized Discounted Cumulative Gain at ten document retrieved (NDCG@10) as evaluation measures for track runs since we judged top-10 pools.[4]

Unfortunately, we encounter another problem when using high-precision evaluation measures. As demonstrated by Hawking and Robertson, precision scores generally increase as the size of the document corpus grows [9]. With the size of the new corpus for the TREC 2021 Deep Learning track, many systems are now good enough to retrieve ten relevant documents in the top ten ranks, maxing out P@10 scores, for many topics. This makes the collection unable to reliably detect differences among systems.

Figure 2 shows a box-and-whisker plot of the distribution of scores for both P@10 and NDCG@10 over the 66 submissions to the Document Ranking task for each topic. Each box-and-whisker element represents one topic and plots the distribution of scores obtained by the systems for that topic. The box shows the range between the first and third quartiles of the scores, the whiskers extend to 1.5 of that range above and below the box, and any data that fall outside of that are plotted as circles. The median score is shown as the heavy black bar within the box. For P@10, 24 of the topics have a *median* score of 1.0, the maximum. The distributions of scores are more discriminative for NDCG@10, though scores are still relatively high and two topics have median scores of 1.0.

3 REUSABILITY

As described in the previous section, the relevant densities of many of the topics suggest that there are relevant documents remaining in the unjudged part of the corpora. Unknown relevant documents would not be a significant impediment to fair comparisons of system effectiveness if the known set were a random sample of the full set of relevant documents, but it is highly unlikely that the known set is a random sample since it is the relevant set that participating systems found most easily. This section therefore examines the reusability of the collections and confirms that the collections are indeed less reusable than desired.

Test collection reusability is generally gauged by the *leave-out-uniques* (LOU) test [2, 22]. The LOU test uses modified qrels for each of the individual teams that participated in the construction of the full collection to simulate the outcome of the building process had the target team not participated. The collection is deemed reusable if the evaluation scores of runs by the full and modified qrels are comparable, since the runs of the target team can be considered novel runs with respect to the modified qrels. For collections built through pooling, the qrels are modified by removing any relevant document that was retrieved within the pool depth only by the target team. These uniquely retrieved relevant documents (hence the name of the test) would not have been judged had the target team not participated. The assumption is that if some participating team had sufficiently many unique relevant documents to affect scores, than other non-participants probably also have unique relevants which are truly unknown since they did not participate. Whether scores are "comparable" is generally measured using Kendall's τ correlations between rankings of systems ordered by mean score. Since τ values computed over rankings of many items can be large even when individual items have sizable differences in rank, the maximum change in ranks in the τ computation is a complementary measure of similarity [19].

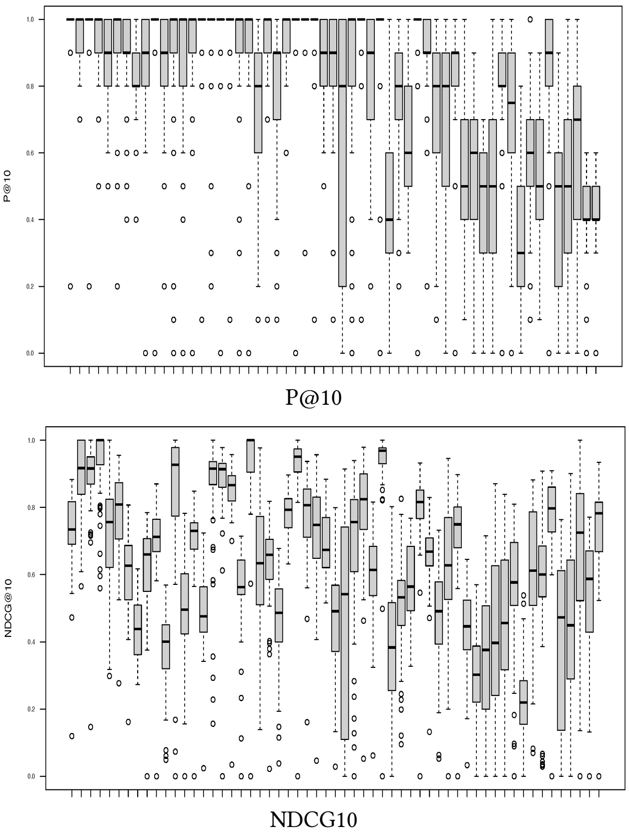


Figure 2: Distribution of scores over 66 submissions to the Document Ranking task for P@10 (top) and NDCG@10 (bottom) for the 57 topics in the evaluation set. Topics are sorted by decreasing relevant density.

The concept of uniquely retrieved relevant document is not defined for collections built through CAL since CAL can select a document to be judged independent of the ranks at which the document was retrieved by the runs. We can nevertheless use a variant of the LOU test by creating the qrels that would have resulted from a particular team not participating in the process by invoking CAL on different initial judgment sets. To keep the process manageable, we performed only a single round of CAL for each starting judgment set. Since the TRACK qrels were created from multiple rounds of CAL, we first created a baseline qrels for each topic by submitting the entire top-10 pool to CAL and selecting the next X documents from its ranked output where X is the difference between the number of documents judged in the TRACK qrels and the pool size. (That is, we used the same number of documents that were judged in the TRACK qrels for each topic.) We'll call this qrels the LOU-FULL qrels. Note that CAL returned some documents that had not been

[4] NDCG scores depend on the ideal ranking and thus the entire relevant set, but relative NDCG@k scores for runs completely judged to depth k are stable. All measures reported in the paper were computed using the `trec-eval` package.

encountered in the track judging. We obtained additional human judgments after the track qrels were released (see Section 4), so all of the LOU qrels contain exactly the same number of judgments as the TRACK qrels.

To perform the LOU test, for each team we created the top-10 pools from all runs except those of the target team as in the traditional LOU test, and then performed a single round of CAL with the reduced-pool set of judgments as the initial judgment set. Once again we selected the top X documents from the CAL list to fill out the qrels. If the target team did not have any unique relevant documents for a topic, the reduced-pool qrels for that topic is precisely the LOU-FULL qrels. A team was defined as a TREC participant, except that the four institutions that submitted both baseline and test runs were treated as two teams each (one for baseline runs and one for test runs). With that definition of team, each task received submissions from 19 teams, though it was a slightly different set of teams per task.

Figure 3 plots the difference in evaluation scores per run where the score using the LOU-FULL qrels is plotted on the x-axis and the score using a reduced-pool qrels is plotted on the y-axis. A reduced-pool qrels is the qrels formed by omitting one team's uniquely retrieved relevant documents from the initial pool. An individual graph plots a point for each run evaluated by each reduced-pool qrels except that no point is plotted if a run's scores using the LOU-FULL qrels and reduced-pool qrels were identical. Two points with the same marker represent runs evaluated using the same reduced-pool qrels. Plots on the top show the results for the Document Ranking task and on the bottom for the Passage Ranking task; plots on the left use P@10 as the evaluation measure and plots on the right use NDCG@10.

All of the points in Figure 3 are below the equal-score line, meaning that starting with a reduced depth-10 pool always caused scores to degrade. This outcome must be true for P@10 scores since all of the documents in the top ten ranks were judged for all runs in the LOU-FULL qrels; relevant documents in those ranks could be lost by starting with reduced pools but none could be gained. The vast majority of points for P@10 for the Document Ranking task are tightly clustered close to 1.0 while the points in the remaining graphs are more dispersed and further from 1.0. Thus the figure also illustrates the compressed range of mean P@10 scores for the Document Ranking task compared to the range of mean NDCG@10 scores and compared to the Passage Ranking task scores.

A large majority of the runs rank in the same order regardless of which qrels is used (recall that only non-identical points are plotted in the figure), and this is reflected in the Kendall's τ correlations between system rankings in the LOU test where the τ is computed between the ranking of systems induced by the LOU-FULL qrels on the one hand and the reduced-pool qrels formed by omitting a team's uniquely retrieved relevant documents from the initial pool on the other. All of the τ's are greater than 0.9 for each reduced qrels, each task, and for each of P@10 and NDCG@10. But while the rankings are stable on average, some individual runs are greatly affected by the reduced qrels. Call the maximum change in rank for a run in the reduced-pool qrels' ranking compared to the LOU-FULL ranking a "drop". Nine of the nineteen reduced qrels cause drops greater than ten for the Document Ranking task when using P@10; for the Document Ranking task and NDCG@10, four reduced qrels cause drops greater than ten. The corresponding counts for the Passage Ranking task are one for P@10 and five for NDCG@10. The largest observed drop is 29 for the Document Ranking task and P@10 (the blue 'x' furthest from the equal-score line in the top left plot of Figure 3). Since there were 66 submissions to the task, this means that run compared differently to more than 40% of the other runs when its team's uniquely retrieved relevant documents were or were not considered relevant.

Given the negative impact of removing a team's uniquely retrieved relevant documents from the qrels on individual runs, we infer from the LOU test results that the collections are not generally reusable. Collections that are not reusable can still be useful for system development, but care must be taken to account for unjudged documents during system comparisons.

4 ADDITIONAL JUDGMENTS

NIST was able to procure a second round of judgments for the TREC 2021 Deep Learning track collections to support exploration of collection effects such as the LOU test described above. The original track assessing ended in September 2021 and the second round began in December 2021. Five of the six assessors who made judgments in the first round were able to return for the second round and they each judged the same set of topics they had originally judged. The sixth assessor was unavailable and was replaced by a different experienced TREC assessor for the second round. Each of the Round 2 assessors had access to the judgments made for a topic in the previous round so they could (re-)familiarize themselves with the topic. They were asked to make the Round 2 judgments as consistent with the TRACK judgments as they could. New judgments were obtained for all of the topics that were judged in the initial round of judging for both tasks, except that no additional judgments were made for the four topics dropped from the Passage Ranking task's TRACK qrels.

CAL was not used in the second round of judging. For each topic, the assessor judged the set consisting of the documents needed to obtain complete judgments for the LOU test described above plus the documents needed to complete top-10 pools over runs restricted to a subset of the collection as described in Section 6 below. A document was never judged more than once for the same topic. In total, we obtained an additional 9255 judgments for the Document Ranking task collection and an additional 10,132 judgments for the Passage Ranking task collection. We call the union of the TRACK qrels and these Round 2 judgments the UNION qrels. The rightmost point for each topic in Figure 1 is the relevant density in the final UNION qrels. The density in the UNION qrels is never greater than the density in the TRACK qrels and is considerably smaller for many topics. Nonetheless, for the Document Ranking task collection, half of the topics still have relevant densities greater than one half even after increasing the overall number of judgments by about 70%.

5 RELIABILITY

The LOU test simulates the reliability of comparisons between runs that participated in the construction of a collection and those that did not. Another way of assessing the quality of a test collection is to measure the stability of decisions regarding the relative effectiveness of arbitrary run pairs [3]. In the version of this test

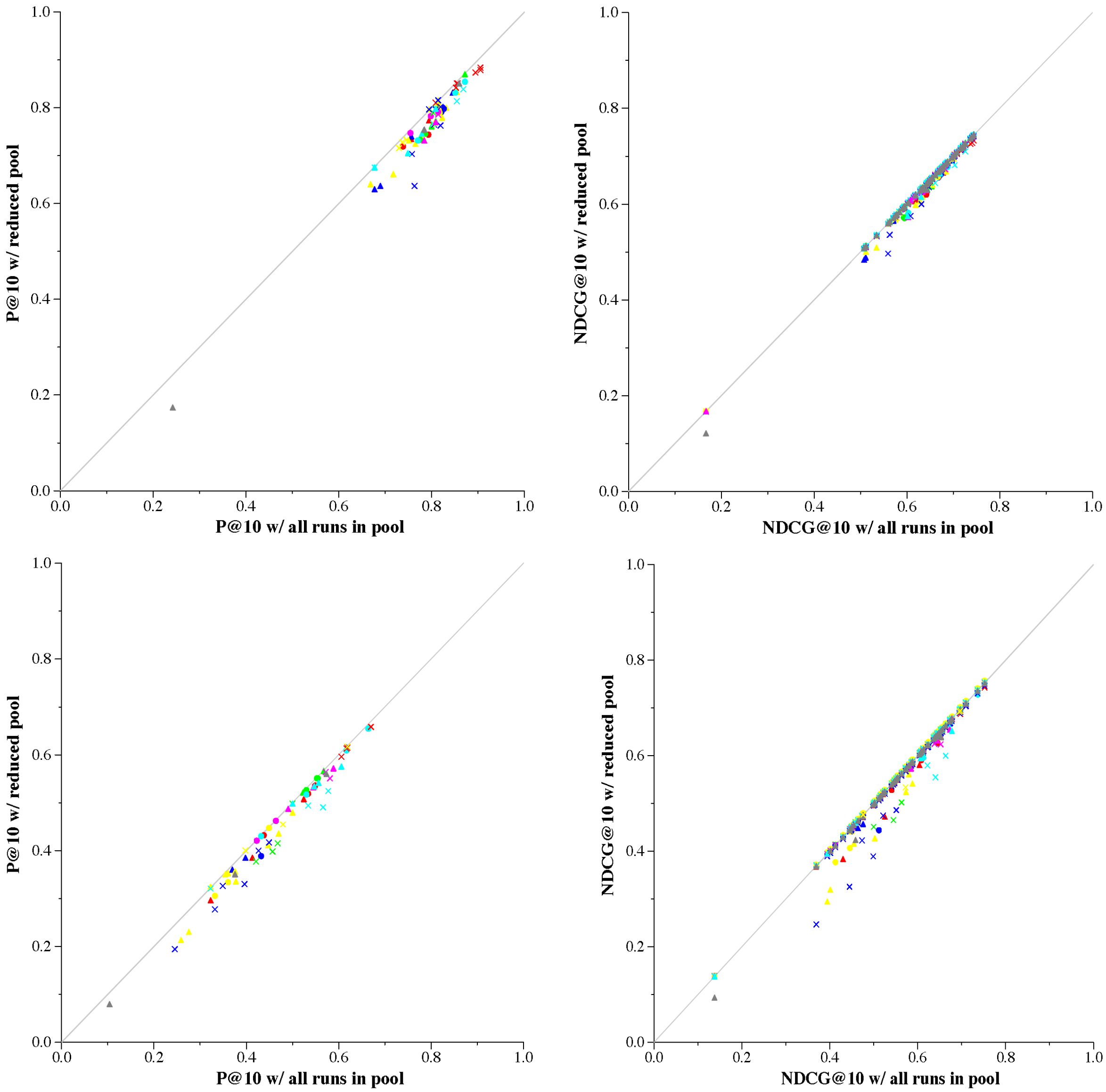


Figure 3: Change in systems' mean scores when evaluated with and without uniquely retrieved relevant documents. The Document Ranking task results are shown in the top plots and the Passage Ranking task results in the bottom plots; plots on the left side show P@10 results and plots on the right side show NDCG@10 results. Each point represents a run evaluated using the LOU-FULL qrels and a reduced-pool qrels that had a team's uniquely retrieved relevant documents removed from the initial depth-10 pools. Within a plot, points with the same marker are from the same reduced-pool qrels.

used here, we first created two independent subsets of topics of size S by drawing topics uniformly at random with replacement. We computed the mean evaluation score over the topics in each subset for all runs and compared all run pairs on both subsets. We defined a *swap* as an instance where a pair of runs evaluated in different orders on the two subsets, and computed the fraction of comparisons that were swaps binned by the size of the difference in scores. We drew 5000 pairs of subsets for each topic set size and looked at topic set sizes between five and the total number of topics in increments of five (so $S \in \{5, 10, \ldots, 55, 57\}$ for the Document Ranking task collection). There were 21 bins of score differences ranging from differences of less than 0.01 for bin 0 to differences ≥ 0.2 for bin 20 in increments of 0.01.

In previous uses of the stability test, smaller numbers of swaps implied greater stability and higher quality. Smaller topic set sizes are less stable than larger topic set sizes, and comparisons between runs with smaller differences in scores are less stable than larger score differences. When plotting swap rate against set size for

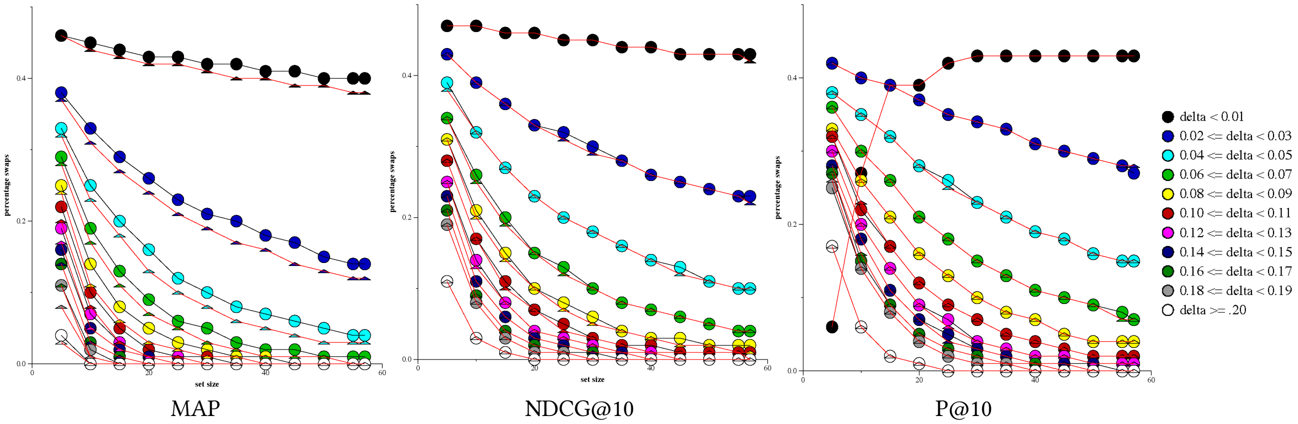


Figure 4: Swap rates of different same-sized subsets of topics as measured on the TRACK and UNION qrels for the Document Ranking task collection. Scores computed using the TRACK qrels are plotted using circles and black lines while scores computed using the UNION qrels are plotted using triangles and red lines. A line connects the swap rates for a single bin, with different bins having different marker colors.

differences within a single bin, the curve generally starts high and decreases to an asymptote, with bins representing smaller differences in scores between run pairs having larger swap rates for a given set size. Figure 4 shows plots of swap rates for the Document Ranking task collection for three different measures, MAP (left), NDCG@10 (center), P@10 (right). In the figure, scores computed using the TRACK qrels are plotted using circles and black lines while scores computed using the UNION qrels are plotted using triangles and red lines. A line connects the swap rates for a single bin, with different bins having different colors of markers; only every other bin is plotted to increase the clarity of the plots. Plots for the Passage Ranking task are very similar to the Document Ranking task plots and are not shown.

Intuitively, a test collection with significantly more relevance judgments is a higher quality collection and we would expect to see the higher quality reflected in the stability test. Thus, the Document Ranking task's UNION qrels, which contains 22,313 judgments compared to the 13,058 judgments in the TRACK qrels, should be the more stable collection. The MAP stability results (Figure 4 left) exhibit the expected behavior with a consistently smaller swap rate for the UNION qrels compared to the TRACK QRELS for each bin. The NDCG@10 and P@10 results, however, are unexpected in that there is little difference between the two collections and thus the collections appear to be of equivalent quality. Both of these measures depend on the top-10 documents in each run for each topic, and these documents are all judged in both qrels. Further, the scores for individual topics are so high, especially for P@10, that drawing different topics is unlikely to affect mean scores for different subsets because the different topics all have close to the same score.[5]

It is not only the stability test that ascribes equal quality to the two collections. Kendall's τ correlations between rankings produced by the TRACK and UNION qrels for track submissions also cannot detect a difference between the collections for the high precision measures. For the Document Ranking task collection, the τ values are 1.0 for P@10, 0.9897 for NDCG@10, and 0.8959 with a maximum change in rank of 18 for MAP. The corresponding values for the Passage Ranking task are 1.0 for P@10, 0.9969 for NDCG@10, and 0.9477 with a maximum change in rank of 8 for MAP.

6 RANKS OF RELEVANT DOCUMENTS

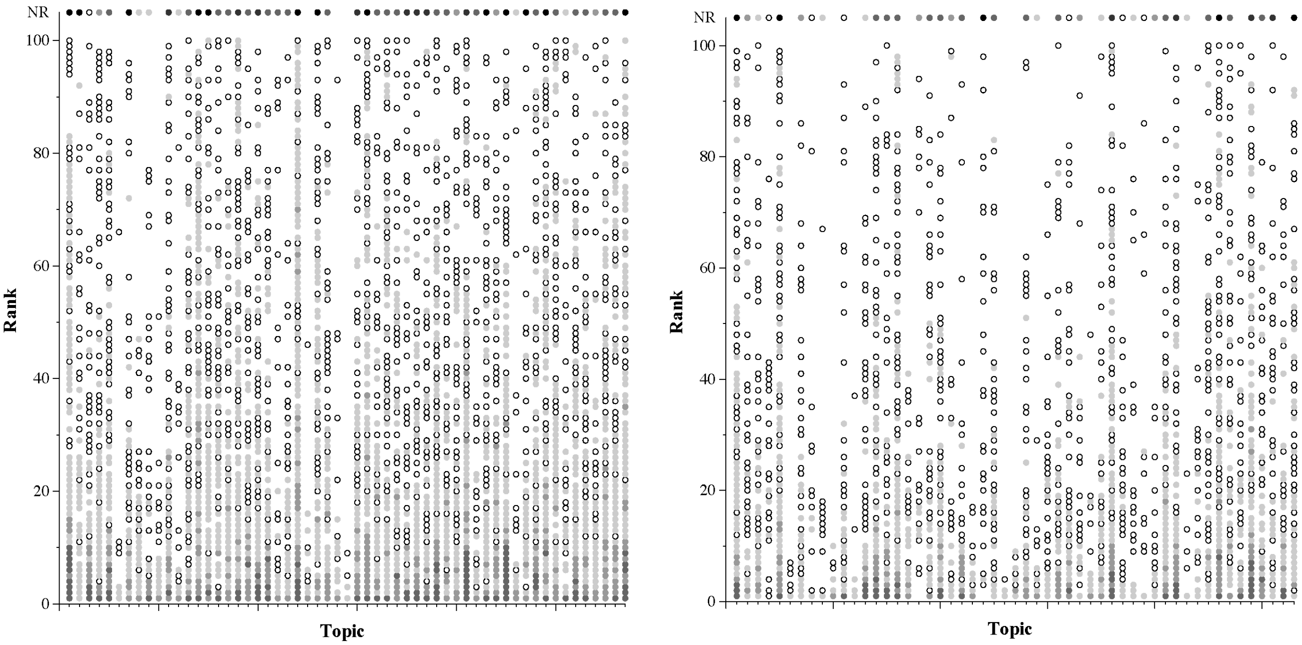
The results of the LOU and stability tests confirm that the majority of topics in the Document Ranking task collection likely have many more relevant documents than are identified in the TRACK qrels. In this section we mine deeper ranks of the original runs looking for additional relevant documents.

To explore deeper ranks of the submitted runs in a budget-friendly way, we created two random subsets of the document corpus by first creating a random shuffling of the entire corpus and then using only the first third or only the first ninth of that ordering as the corpus. We restricted each of the submitted runs to just those documents in the corresponding corpus subset and created top-10 pools from the restricted runs.[6] Any document in these pools that was not judged during the track assessment period was judged in the Round 2 assessing.

Figure 5 shows the distribution of the relevant documents in the UNION qrels. For each relevant document in the qrels, we found the minimum rank across all runs at which the document was retrieved. If no run retrieved it in the top 100 ranks (the maximum size of a submission) then it must have been contributed to the qrels by CAL and it was counted as "Not Retrieved" (NR). The figure gives a heat map of the number of relevant documents whose minimum rank is the given rank for each topic, where each item on the x-axis is a topic and the rank is plotted on the y-axis. The heat map uses a white-filled circle for one document, a light gray circle for 2–5 documents, a somewhat darker gray circle for 6–10 documents, a

[5] The inverted curve for bin 0 in the P@10 plot (Figure 4, right) is caused by small sample size issues since there are very few comparisons that have less than a 0.01 difference in mean P@10 scores for small topic set sizes.

[6] Since TREC 2021 track submissions could contain at most 100 documents per topic, some runs did not have as many as 10 documents for some topics in the restricted runs.



Document Ranking Task

Passage Ranking Task

Figure 5: Heat map of the number of relevant documents in the UNION qrels for which the given rank was the minimum rank across the track submissions where it was retrieved. The darker the circle the more documents had that rank as a minimum, with ranges of 1 document (white-filled circle), 2–5, 6–10, 11–30, 31–50, and more than 50 documents (black circle). Relevant documents encountered through CAL that were not retrieved by any run are plotted as 'NR'. The Document Ranking task collection is shown on the left and the Passage Ranking task collection on the right.

medium gray for 11–30, a dark gray for 31–50, and a black circle for more than 50 documents.

The figure shows that many topics in the Document Ranking task have minimum ranks of relevant documents throughout all 100 ranks, meaning depth-100 pools would be required to capture all of these documents if only traditional pooling were used. Depth-100 pools for the Document Ranking task would require a total of 50,590 judgments (or 2.3 times as many judgments as in the UNION qrels and 3.8 times as many as were originally made for the track) and would still not capture the NR relevant documents. But it is not only pooling that is untenable when relevant documents are throughout the rankings. Dynamic methods such as bandit algorithms that rely on system rankings process the rankings from the top down and will continue to encounter relevant documents as they delve the rankings' depths. Indeed, bandit algorithms have been shown to be unfair to runs that participate in the collection building process precisely in the case when the judgment budget is small compared to the number of relevant documents [19].

7 DISCUSSION

The TREC 2021 Deep Learning track used significantly larger corpora than had been used in previous years, resulting in large numbers of relevant documents. The size of the relevant set caused two subsequent problems: the inability to construct an adequate sample of the relevant set for fair comparisons using recall-based measures, and the saturation of high precision measures that prevent them from being able to discriminate among runs.

7.1 Sampling the Relevant Set

The inability to construct an adequate sample of the relevant set has been faced before when pooling was found to no longer be viable because of a size dependency [2]. Among the suggestions for how to proceed then were to form pools differently (which led to the adoption of dynamic methods), to engineer the topic set such that topics would not have large relevant sets (which has been done to varying degrees in all TREC collections including the early ad hoc collections), and to engineer the judgment sets by which the authors meant to down-sample an existing judgment set to a fair set (though they left how to do that as an open problem).

These suggestions are still options, but each requires more research to be actionable. Current dynamic judging techniques are already finding relevant documents efficiently, but there are too many relevant documents. Constructing a fair sample of relevant documents when the true set is unknown remains an open problem. Selecting only those topics whose relevant density is low enough for the evaluation set and iterating until there are sufficient topics (the intended process in 2021) is expensive. Once a topic has been abandoned, all of its judgments are wasted in that they will not contribute to a test collection but still count against the budget. It is also hard to know when to make the decision to abandon a topic.

As the relevant density trajectories in Figure 1 show, some densities *increase* as more judgments are made, so a topic that appears acceptable early in the judgment process may not be.

For the Document Ranking task, 17/57 topics had relevant densities less than one half at the end of the track judging. If we assume the relevant distribution is the same for the rest of the topics in the test set and use the end of track judging as the decision point, then we could get a traditional evaluation set size of 51 topics for 39,174 judgments—three times the TRACK qrels budget. Since the 17 topics with densities less than a half account for only 26% of the judgments in the TRACK qrels, 74% of the 39,174 budget (or 29,139 judgments) would be wasted. Further, the resulting test collection would be highly skewed toward small topics. Retrieval effectiveness as measured on the collection would not be representative of effectiveness of the entire test set.

7.2 Score Saturation

The saturation of high-precision measures is also not completely new. For example, P@5 was used as an official evaluation measure in TREC-COVID until it saturated in the later rounds.[7] The solution for saturation previously was just to use different, deeper measures, but that is only acceptable if the deeper measures are reliable and are an appropriate measure for the task. P@10 provides a good view of the searcher's experience. Thus a possible response to the "problem" that search systems are good enough to retrieve ten relevant documents in the top ten ranks on large collections is to declare success as search has been solved!

However, few researchers really believe that search engines are as good as they should be or can be, which argues that the effectiveness of retrieval systems with respect to large collections requires a more nuanced definition of 'relevance'. This, too, is not a new idea. NDCG was introduced largely in reaction to the number of relevant documents in the large-for-the-time web test collections [10, 18] and Sormunen [15] warned of the 'liberal' definition of relevance used in TREC. Much more recently, Arabzadeh et al. [1] advocated for a stricter definition of relevant within the context of the sparse MS MARCO judgments. For the 2021 Deep Learning track, the Passage Ranking task used a more stringent definition of 'relevant', and thus had fewer relevant documents and correspondingly fewer problems with score saturation than the Document Ranking task.

If the problem is too many relevant documents, then schemes that reduce the number of relevant documents should provide a solution. But an arbitrary definition of relevance is not a panacea for collection building. First, for test collections to be good tools the definition of relevant needs to reflect what users in the real use case actually want to have returned. The Passage Ranking task's more stringent definition of relevance arose organically in that 'related' passages that don't answer the question really aren't what is desired. Second, all of the collection-building processes rely on the ability of systems to rank the target documents highly.

We still would have had incomplete qrels for the TREC 2021 Document Ranking task even if we had restricted 'relevant' to include only the most highly relevant documents. Figure 6 shows the minimum ranks at which relevant documents were retrieved for the Document Ranking task, similar to Figure 5, but this time the color coding indicates the relevance grade of documents retrieved at that rank. White-filled circles represent RELEVANT (grade 1) documents, gray circles represent HIGHLY RELEVANT (grade 2) documents and black circles represent PERFECT (grade 3) documents. Higher grades were plotted later and are thus the visible grade when multiple documents of different grades have the same minimum rank. Relevant documents encountered through CAL that were not retrieved by any run are plotted as 'NR'. HIGHLY RELEVANT and, to a lesser degree PERFECT documents are distributed throughout the ranks, including documents not retrieved in the top 100 ranks by any submission.

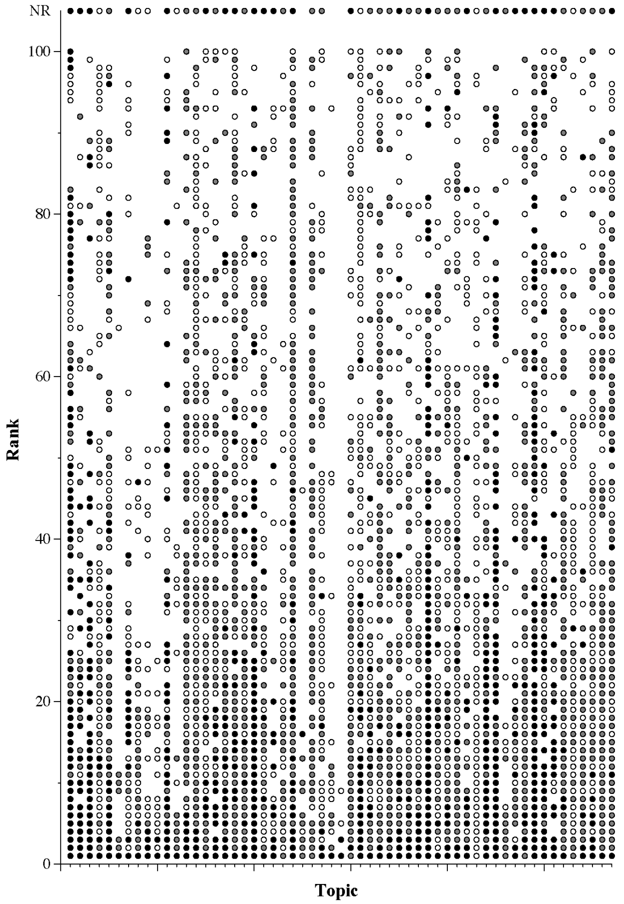


Figure 6: Minimum rank at which a relevant document was retrieved across Document Ranking task submissions color-coded by the relevance grade. White-filled circles represent RELEVANT **documents, gray circles represent** HIGHLY RELEVANT **documents and black circles represent** PERFECT **documents. Relevant documents encountered through CAL that were not retrieved by any run are plotted as 'NR'.**

We also evaluated the 2021 Document Ranking task submissions using the UNION qrels and only grades two and three as relevant. One topic has no documents with those grades, so averages are over 56 topics. The box-and-whiskers plot for P@10 is shown in Figure 7. The distributions of scores is much better, though seven

[7] https://ir.nist.gov/trec-covid/round5.html

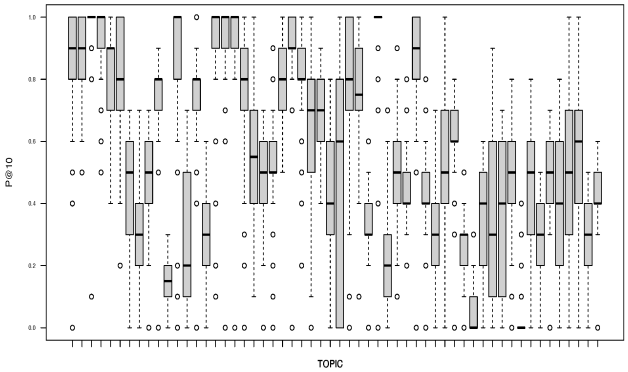


Figure 7: Distribution of scores over 66 submissions to the Document Ranking task for P@10 using only relevance grades two and three and the UNION qrels for the 56 topics with relevant documents. Topics are sorted as in Figure 2.

topics do still have median scores of 1.0. Consistent with the earlier work on evaluation by highly relevant documents [18], the ranking of systems when evaluated using P@10 for all relevance grades or top grades only are quite different. Using the UNION qrels, the Kendall's τ correlation between the two rankings is just 0.7796 with a maximum change in ranks of 26. This confirms that changing the definition of 'relevant' *does* change the user task represented by the collection (while still not necessarily reducing the number of relevant documents to manageable levels).

These results suggest that score saturation is currently better addressed by using deeper measures than by using arbitrarily narrow definitions of relevance. Varying the persistence parameter in the Rank-Biased Precision (RBP) family of measures [14] controls the effective depth used to compute the score while also signaling any effects from incomplete judgments. This makes RBP preferable to P@10 when the latter is saturated.

8 CONCLUSION

Finding a small number of relevant documents is easier for search systems as corpus size increases [9]. The Deep Learning track results suggest a wide range of search systems are now capable of consistently retrieving ten relevant documents as the first ten documents for a wide range of queries in corpora as large as the MS MARCO version 2 corpora. Distinguishing among retrieval system behavior thus requires different metrics or a more focused definition of relevance (or some combination).

Using arbitrarily narrow definitions of relevance does not solve the problem of building affordable, reusable test collections for massive corpora. Relevance grades in test collections need to reflect the desirability of documents in the actual search task for the collections to be useful tools, and all of the known collection-building techniques rely on systems being able to rank relevant documents higher than non-relevant documents. Thus relevance must be defined by the use case and not the collection building scheme. Happily, if the actual search task really is just to find a few relevant-as-currently-defined documents, then systems with saturated scores are both capable and equivalent, and no additional test collections are needed.

Using measures that evaluate to deeper ranks while accommodating (many) unjudged documents provides another way forward. While some such measures already exist, new metrics will be needed because the number of unjudged documents likely to be encountered in the system rankings will cause current metrics' error bounds to be too large to be discriminative. Future collection builders will thus need to find a balance between an affordable judgment budget and acceptable measure sensitivity.

ACKNOWLEDGMENTS

We thank Gordon Cormack for making available a batch interface to CAL that was used both to select documents to be judged in the track and to enable the post-track simulations. Thanks also to Chris Buckley for multiple discussions on test set construction.

REFERENCES

[1] Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. 2021. Shallow pooling for sparse labels. arXiv:2109.00062 [cs.IR]

[2] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the Limits of Pooling for Large Collections. *Information Retrieval* 10 (2007), 491–508.

[3] Chris Buckley and Ellen M. Voorhees. 2000. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*. 33–40. https://doi.org/10.1145/345508.345543

[4] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal Test Collections for Retrieval Evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. 268–275. https://doi.org/10.1145/1148170.1148219

[5] Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. arXiv:1504.06868 [cs.IR]

[6] Gordon V. Cormack and Maura R. Grossman. 2016. Engineering Quality and Reliability in Technology-Assisted Review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. 75–84. https://doi.org/10.1145/2911451.2911510

[7] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. 282–289. https://doi.org/10.1145/290941.291009

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, and Ian Soboroff. 2021. TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2369–2375. https://doi.org/10.1145/3404835.3463249

[9] David Hawking and Stephen Robertson. 2003. On Collection Size and Retrieval Effectiveness. *Information Retrieval* 6 (2003), 99–105. https://doi.org/10.1023/A:1022904715765

[10] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (Oct 2002), 422–446. https://doi.org/10.1145/582415.582418

[11] Dan Li and Evangelos Kanoulas. 2017. Active Sampling for Large-Scale Information Retrieval Evaluation *(CIKM '17)*. 49–58. https://doi.org/10.1145/3132847.3133015

[12] David E. Losada, Javier Parapar, and Álvaro Barreiro. 2016. Feeling Lucky? Multi-Armed Bandits for Ordering Judgements in Pooling-Based Evaluation. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC '16)*. 1027–1034. https://doi.org/10.1145/2851613.2851692

[13] Alistair Moffat, William Webber, and Justin Zobel. 2007. Strategic System Comparisons via Targeted Relevance Judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. 375–382. https://doi.org/10.1145/1277741.1277806

[14] Alistair Moffat and Justin Zobel. 2008. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems* 27, 1, Article 2 (Dec 2008), 27 pages. https://doi.org/10.1145/1416950.1416952

[15] Eero Sormunen. 2002. Liberal Relevance Criteria of TREC—Counting on Negligible Documents?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. 324–330. https://doi.org/10.1145/564376.564433

[16] Karen Spärck Jones and Cornelis J. van Rijsbergen. 1975. Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge.

[17] Stephen Tomlinson and Bruce Hedin. 2011. Measuring Effectiveness in the TREC Legal Track. In *Current Challenges in Patent Information Retrieval*, M. Lupu, K. Mayer, J. Tait, and A.J. Trippe (Eds.). The Information Retrieval Series, Vol. 29. Springer, 167–180.

[18] Ellen M. Voorhees. 2001. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. 74–82. https://doi.org/10.1145/383952.383963

[19] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections Using Bandit Techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 407–416. https://doi.org/10.1145/3269206.3271766

[20] Ellen M. Voorhees and Kirk Roberts. 2021. On the Quality of the TREC-COVID IR Test Collections. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2422–2428. https://doi.org/10.1145/3404835.3463244

[21] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. 603–610. https://doi.org/10.1145/1390334.1390437

[22] Justin Zobel. 1998. How Reliable Are the Results of Large-Scale Information Retrieval Experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*. 307–314. https://doi.org/10.1145/290941.291014