RMIT at the 2017 TREC CORE Track

Rodger Benham
RMIT University
Melbourne, Australia

Luke Gallagher
RMIT University
Melbourne, Australia

Joel Mackenzie
RMIT University
Melbourne, Australia

Tadele T. Damessie
RMIT University
Melbourne, Australia

Ruey-Cheng Chen
RMIT University
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

TEAM NAME

RMIT

1 INTRODUCTION

The TREC 2017 CORE Track[1] is a re-run of the classic TREC ad hoc search evaluation campaign, with the vision of establishing new methodologies for creating IR test collections. The previous TREC newswire ad hoc task was the 2004 Robust Track, where the emphasis was on improving the effectiveness of poorly performing topics in previous tracks [16]. The TREC CORE 2017 track reuses the Robust 2004 topic set, for the development of relevance judgments over a new *New York Times* corpus, composed of newswire articles published between 1987 and 2007.

In this track, our interest is driven by two related lines of research: efficient multi-stage retrieval [3–7, 9, 14], where it is believed that improving recall in early stage retrieval can improve end-to-end-effectiveness; and more reliable deep evaluation when using shallow judgments [8, 10–13]. By participating in CORE, we attempted to develop a recall-oriented approach that exploits user query variations and rank fusion. We venture that, in an evaluation campaign such as the TREC CORE Track, which typically attracts runs of a high effectiveness caliber from research groups worldwide, the ability to retrieve a large number of relevant documents that other systems fail to find is indicative of a high-recall system.

A useful consequence of this approach is the ability to compare query variation phenomena across corpora. The UQV100 test collection contains one hundred single-faceted topics with over five thousand unique query variations [1], but to date has only judgments against the ClueWeb12B collection. The variation-rich collection produced for participation in the CORE track, while smaller in scope than the UQV100 collection, now enables comparisons of query variations to be made across different document representations and editorial quality. Observing the relative effectiveness across ClueWeb12B and Robust04, one consisting of websites, and the other composed of journalistic content, is of considerable interest, as is the question of the benefit of rank fusion mechanisms based on those variations.

Research Goals. We focus on these research questions:

[1]TREC 2017 CORE Track: http://trec-core.github.io/2017/

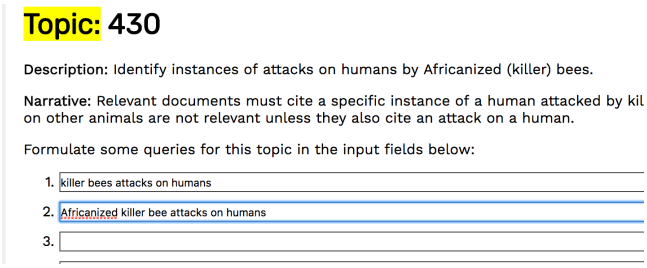

Figure 1: The query variation submission interface.

- **RQ1:** *Can tuned and parameterized query fusion be used to improve the number of unique relevant documents found relative to other participants?*
- **RQ2:** *Do query variations that are good in one collection also perform well in another collection?*
- **RQ3:** *Are the score ranges caused by user query variations consistent across different test collections?*

In the next section, we describe the process used to collect query variations. In Section 3 we discuss our submitted runs in more detail, and place our query variations into context with the existing UQV100 test collection; and then in Section 4 we provide the results of each of our submitted runs using the fifty queries that were assessed by NIST. As of writing, crowd-sourced relevance assessments for two hundred additional topics remain unreleased, reducing the scope of our exploration.

2 GENERATING QUERY VARIATIONS

In recent work Moffat et al. [15] demonstrate that the way in which queries are posed can have a substantial effect on retrieval outcomes. The UQV100 collection was created out of that same work, a set of qrels for the ClueWeb12B documents, built to provide coverage for more than 5,000 unique queries over a set of 100 topics [1]. The UQV100 collection then provided a framework in which further work was possible, including the introduction of *retrieval consistency* as an attribute of a search system, and measurement of the benefits achieved by combining the runs generated as a result of query variations into a single "smoothed" run [2].

We sought to explore and build on those themes in our 2017 TREC CORE submission. To collect query variations that could be

Participant	Queries	Avg. terms	Avg. chars	% Unique
1	367	3.90	26.4	83.7
2	270	4.44	29.4	85.2
3	175	3.39	24.2	85.7
4	340	5.59	34.6	96.5
5	1261	6.46	41.7	97.4
6	342	4.85	31.6	95.9
7	158	8.73	57.8	100.0
8	238	3.99	25.1	94.5
Overall	3151	5.48	33.9	93.7

Table 1: Number of query variations submitted by participants, their average lengths in words and non-space characters, and the fraction of each participants' queries (and overall) that were unique across the set of variations collected.

used in a range of experimental settings, such as query fusion, and query rewriting over three distinct document collections, a tool to collect variations was developed, and over a ten-day period each of the authors (eight participants in total) contributed a set of up to ten query variations per 2017 topic (of which there are 250) using that tool. The information-need descriptions used as the basis for query solicitation were taken from the modernized Robust04 topic narratives and descriptions supplied by the track organizers, with the title of the topic not shown at all, to avoid biasing the new queries. Figure 1 shows a screenshot taken while filling out query variations for one of the 250 topics.

No query transformations were applied on the queries collected using the tool, and the onus was on each user to supply one or more queries that they thought would be accurate and useful representations of the corresponding information need. Query normalization is an additional step that might be applied in future experiments, although we note that many browsers already offer spell checking functionality which may have been used during the creation of query variants.

Although the text-boxes were presented to the user in an ordinal numbered list (Figure 1), this was not used to indicate a preference for that user's "best" query. Another opportunity for future analysis will be to determine whether there was any consistent pattern of "getting better" or "getting worse" evident in the sets of queries authored by each of the participants. About half-way through the solicitation period, it was clear that there was a heavy focus on the early topics, including the set of 50 that were to be NIST-assessed. This was useful, to allow a focus on the human judgment task; but thereafter we also sought to balance the collection process, by circulating lists of topics that had the smallest numbers of variations so far.

By the end of the collection phase, a minimum of eight query variations were captured for each of the 250 topics (not all participants covered all of the topics). The author of each query was recorded as they submitted their suggestions, and Table 1 shows the relative contributions made, with the "participant" value an arbitrary identifier. Generating a more representative sample of query variations across users remains an option to be explored in future analysis.

While the authors are familiar with many of these 2017 topics and their nominal "title queries" as a result of other TREC-based activities (including, for two of them, working on the UQV100 queries), we nevertheless engaged fully with the spirit of the new task, and made best efforts to provide queries based on the description and narratives that were presented via the solicitation interface.

3 APPROACH

We used Indri 5.11[2] to index the collection and form our runs, converting the NYT corpus from XML to SGML by parsing the fields using Nokogiri – a libxml2 wrapper. Table 2 displays an abridged description of each run; details are provided later in this section. An automatic run is formed by an IR system where there is no human involvement when retrieving ranked documents, outside of issuing the topics supplied by NIST to the system. Manual runs are all other types of runs; as two of our runs were built on multi-human involvement in the query construction stage, they are marked manual as to allow them to be treated appropriately.

Oracle-Based Manual Query Rewriting. To illustrate the power of using a better query to satisfy an information need, we observe the effectiveness of individual queries on a per-topic basis using AP on the Robust04 collection (computed using `trec_eval`[3]), and NDCG@10 (computed using `gdeval`[4]) using the UQV100 judgments. The *New York Times* collection could not be utilized, as no judgments for this collection existed at the time of run submission. Instead, we use the Robust04 collection as there are relevance assessments available for the same topics, and the documents were formed with similar editorial quality to those in the NYT corpus, allowing inferences to be made about the spread of retrieval effectiveness using different queries.

To analyze the volatility in query effectiveness, we used the newly created query variations and contrast with the existing UQV100 set on the separate ClueWeb12B collection. On the ClueWeb12B collection, no such editorial quality control exists, as is the nature of Web data. This gives us an additional reference point, to explore whether volatility holds constant across collections, to answer research question **RQ3**. As another point of comparison, we show how the spread of retrieval effectiveness varies with respect to the title queries published with the Robust04 set, and the most frequently submitted query variation in the UQV100 set.

Figure 2 shows the variance of the Okapi BM25 per-topic AP effectiveness, with topics sorted by decreasing 75 th percentile effectiveness, juxtaposed with the best and worst performing query variant in our new query variation collection. The spread of effectiveness for the query variations is stark. The UQV100 test collection had a similar spread of per-topic effectiveness, per query variation submitted. Figure 3 shows a contrasting view of the per-topic effectiveness (using NDCG@10 to respect the shallow pooling process used in these judgments). At the tail end of both distributions, outliers that outperform the interquartile range are surprisingly common. From this analysis, we answer **RQ3** in the

[2] https://www.lemurproject.org/indri/
[3] http://trec.nist.gov/trec_eval/
[4] http://trec.nist.gov/data/web/10/gdeval.pl

Run	Type	Description
RMITUQVBestM2	Manual	The best query variation per topic (FDM+QE), as determined by outcomes on the Robust04 collection.
RMITRBCUQVT5M1	Manual	Combines the top five runs per topic based on Robust04. Okapi and SDM+QE was executed over all variations, and the Robust04 collection used to determine the five best.
RMITFDMQEA1	Automatic	FDM plus RM3 query expansion using only the title queries supplied by the track organizers.

Table 2: A brief description of the RMIT runs.

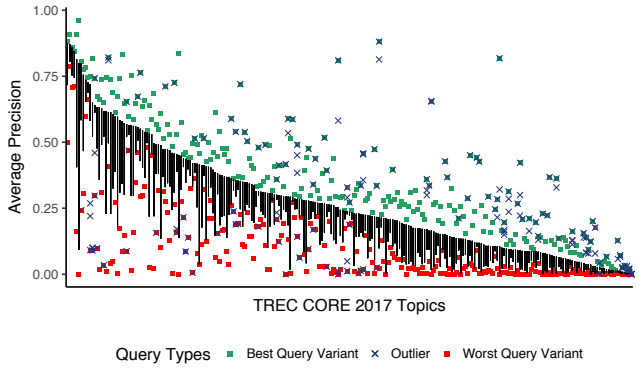

Figure 2: Average precision score distributions for all 249 topics in the author-generated CORE query variations on Robust04 corpus. Lines represent IQR, topics are sorted from greatest 75th percentile to lowest.

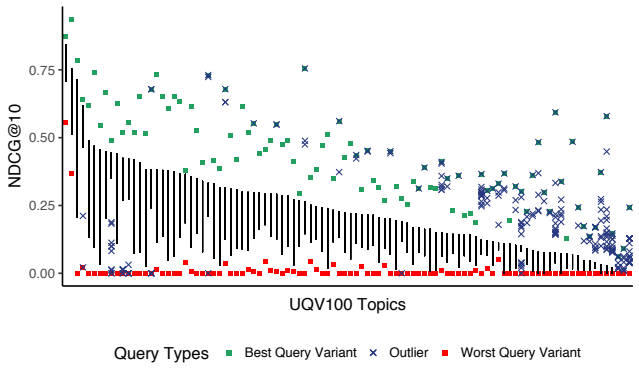

Figure 3: NDCG@10 score distributions for all 100 topics in the UQV100 judgment set and the ClueWeb12B corpus. Lines represent IQR, topics are sorted from greatest 75 th percentile to lowest.

affirmative, as we have shown that a spread of retrieval effectiveness occurs across different query variation collections and corpora.

Title Only. To compare the effectiveness of using the best query variations per topic on the Robust04 collection using FDM+QE, as a baseline we submitted an automatic run that only used the supplied

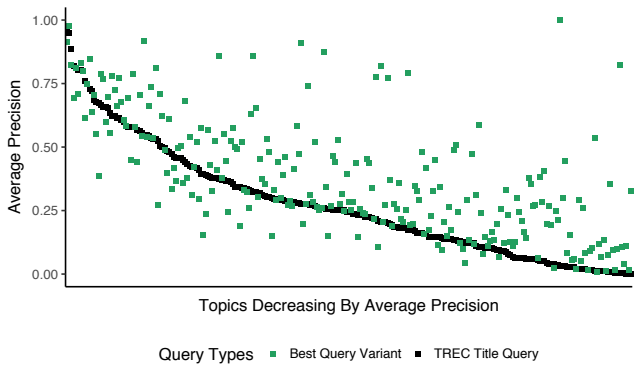

Figure 4: Comparison of FDM+QE runs using AP on the Robust04 title-only queries from most effective to least, and the best per-topic AP query of the TREC CORE query variants, according to the Robust04 collection.

TREC titles. Figure 4 compares the per-topic AP scores of the title-only automatic run, sorted by decreasing score, and against the best score for that topic achieved by any of the query variations that were generated. In most cases, the best query variation is more effective than the TREC title query counterpart, an effect also noted by Moffat et al. [15]. However there is a fraction of cases where the TREC title query is more effective than any of the variations. As we already planned to submit as a baseline an automatic run using the TREC title queries (and probably other participants did likewise), we took the best Robust04-identified query variation, despite knowledge that some are performing better than the best query variation. We did this in an attempt to find more unique-relevant documents. It is of interest how the performance profile exhibited in Figure 4 on the Robust04 collection compares with the editorially similar NYT corpus, when the profiles of the corresponding NYT runs RMITUQVBestM2 and RMITFDMQEA1 are able to be inspected with the new judgment set.

Query Fusion. Bailey et al. [2] show that fusing query variations improves retrieval effectiveness. By introducing more diversity into the final coalesced run, our hypothesis is that it also increases the chances of finding more uniquely relevant documents. Experimentation on Robust04 found that fusing the top five AP-scored query variation runs for each topic, using the RBC fusion method of Bailey et al., yielded a very high mean (across topics) AP score of 0.430.

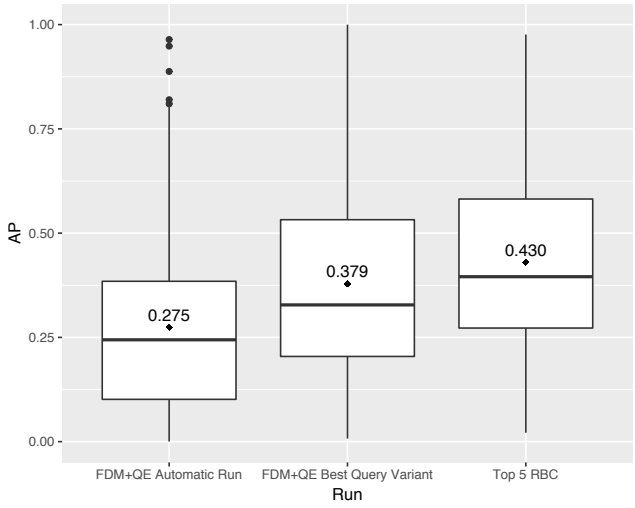

Figure 5: A comparison of effectiveness by AP, for all runs to be submitted on the Robust04 collection.

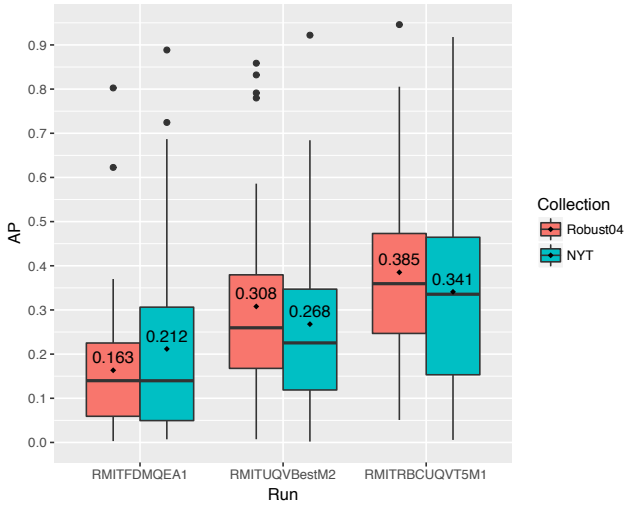

Figure 6: A comparison of effectiveness by AP, for all runs on the 50 NIST assessed topics. Our Robust04 oracle run is listed on the left – restricted to the 50 topics assessed by NIST assessors, for a like-for-like comparison with the NIST assessed plot on the right.

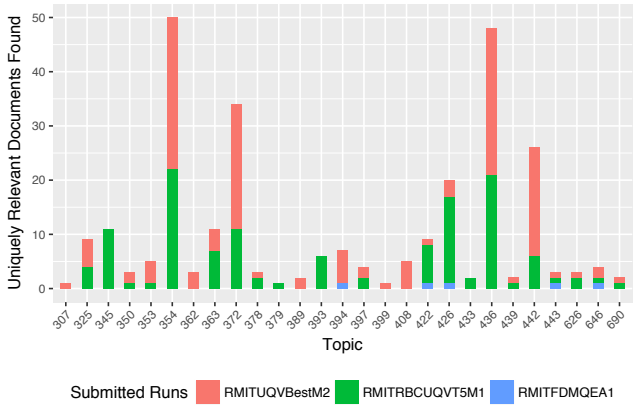

Figure 7: The distribution of uniquely relevant documents found across all NIST assessed topics.

For each query fusion performed, the RBC persistence parameter ϕ was swept to find the optimal value using the Robust04 relevance assessments. This process was conducted over two distinct retrieval systems, BM25 and SDM+QE. Finally, the best per-topic fused run was selected with only BM25 runs or SDM+QE runs. In other words, a top-five fused run at the topic level consisted of a single retrieval system only. The parameters and retrieval system used were logged, and then applied to the *New York Times* document collection to form the run RMITRBCUQVT5M1.

Figure 5 shows the retrieval effectiveness of all three experimental configurations on the previous Robust04 collection. The query variation approaches are statistically signficantly more effective than the title query run (paired two-tailed t-test, $p < 0.001$), and we anticipate that the same also holds when the same query and weighting arrangements are applied on the *New York Times* corpus. Of course, these scores must be interpreted with caution, as they are based on tuning derived from knowledge of the corresponding relevance assessments. Even so, the approach shows strong gains on the Robust04 collection; and hence our conjecture was that at least some of those gains would be preserved in the new CORE collection, for which the judgment-based feedback was not so directly applicable.

4 RESULTS

The CORE track organizers published the relevance assessments formed by NIST for their fifty topics. We restrict our analysis of our results to the NIST sample as the crowd-sourcing relevance judgments are not yet released.

All of our runs met the effectiveness requirements of the track organizers to contribute to the pool of relevance assessments. Figure 6 shows a box-plot to observe how our knowledge transfer approach worked – note that the distribution of scores on Robust04 is different to Figure 5 as only the NIST topics are compared here. This plot helps answer **RQ2**, where we find that the "best" UQV found for each topic over AP on the Robust04 collection outperforms the TREC title run. However, when the same approach of selecting the best query variation was employed over the new judgment set, only 12 of the 50 topics had the same "best" query as Robust04. Where the best query variations were selected over the judgment set for failure analysis, an AP score of 0.346 is achieved, outperforming our fusion run. This seems to suggest that there is no "best" query, and that the best query is coupled to the collection.

Although we do not get the same performance as fusion, both manual runs exhibit scores with acceptable effectiveness and contributed 229 uniquely relevant documents that no other system found. Out of all fifteen submissions, our approach ranked fourth, over the 50 topic NIST sample. The most uniquely relevant document count came from the `Sabir` submission, with

a count of 694 from automatic runs. Our submission was the only contribution to the pool that retrieved 3 uniquely relevant documents across both our automatic run and manual runs. The number of uniquely contributed documents for each submitted run is RMITFDMQEA1: 5, RMITRBCUQVT5M1: 126 and RMITUQVBestM2: 144. Surprisingly, the RMITUQVBestM2 run extracts more unique relevant documents than our fusion run. The number of uniques for RMITUQVBestM2 and RMITRBCUQVT5M1 sum to 270, indicating that 41 uniquely relevant documents overlapped between both retrieval approaches. This is unsurprising, as the top 5 submission contained the top queries used in RMITUQVBestM2. Figure 7 shows the per-topic breakdown of uniquely relevant documents returned for each of our runs. We observe a relatively even distribution of uniquely contributed documents between the RMITUQVBestM2 and RMITRBCUQVT5M1 approaches, where both runs appear to be complementary. We therefore positively answer **RQ1** as query fusion can yield competitive results compared to other participants, however they were not the best over the top 100 – even in the case of our single query run.

Alongside knowledge of the number of uniquely relevant documents retrieved for each research group, a per-topic breakdown of the best, median and worst scores were supplied to each group for the evaluation metrics AP, NDCG and P@10. These figures were provided for relative comparison between automatic submissions and manual submissions. Table 3 shows how our automatic run performed relative to others in the same category at a per-topic level. We find that no topics in our automatic run achieved the best or worst AP or NDCG scores relative to others. We achieve the best P@10 value on four topics (336, 394, 416 and 614), but also achieve the worst scores on four topics (325, 356, 367 and 445). Note that other groups may have tied effectiveness scores. Less than half of our topics surpassed the median scores on AP and NDCG, indicating that our automatic run was relatively weak compared to other submissions.

Table 3 also shows the per-topic breakdown of our two manual runs relative to other manual runs. Our query fusion oracle run achieved the best results out of all three runs, where 6 topics performed the best out of all manual submissions for AP: 372, 379, 404, 419, 422 and 443. We also achieved the best NDCG score for ten topics: 341, 353, 379, 404, 416, 419, 443, 614, 620 and 677. Conversely, we achieve the worst performance on topic 690 for AP, and for 626, 646 and 690 for NDCG. As an example, topic 404 occurs in both of these lists, with the title query "Ireland, peace talks" – where the goal is to extract documents that discuss "How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?". The most effective query fusion configuration found on the Robust04 set, were the query variations (in order): "ireland peace talk delay violence bombing", "ireland peace talks disruption roken off violence attack threat fighter IRA republican army british ulster" [sic], "peace talks delayed Ireland violence", "Ireland peace talk delay disrupt violence" and "North Ireland peace process delayed violence". All of these query variations were fused using RBC $\phi = 0.99$, where it was found that BM25 gave more effective results than SDM+QE for this query on the Robust04 collection. Our best query variation run RMITUQVBestM2 achieved the best AP score for topic 435 "curbing population growth", with the query "population growth control".

Finally, in Table 3, we merge the best and worst scores across automatic and manual submissions to see how we compared globally against all participants. RMITRBCUQVT5M1 achieves the best AP scores on a per-topic level for topics 372; "Native American casino" with AP score 0.691 and 379; "mainstreaming" with score 0.374. The median scores for both of these queries were 0.423 and 0.199 respectively. For NDCG, 379 reappears with an NDCG score of 0.781 where the median is 0.549, and 416; "Three Gorges Project" appears with the NDCG score 0.912 where the median is 0.855.

5 CONCLUSION

All three of our runs met the track organizers' quality criteria for inclusion into the judgment pool. RMITRBCUQVT5M1 was our most effective run in terms of AP, achieving a score of 0.341, however RMITUQVBestM2 identified more unique and relevant documents than the former to cutoff 100. We were met with some fierce competition in the track, where we were outperformed in uniquely contributed documents by the Sabir run produced by Chris Buckley, and two submissions from the University of Waterloo, placing us fourth out of fifteen participants in this respect. Our automatic run did not appear to perform well compared to other automatic submissions, however we did not anticipate it to do so as it was used as a baseline for comparison with our manual query fusion approaches. Query fusion was able to produce a highly effective result list with a sub-par retrieval model in comparison to other participants. This confirms previous observations the query fusion is an effective technique for maximizing recall, and in future work we plan to explore this approach further using stronger systems. We look forward to reading about the approaches other participants utilized, and to participating in future ad hoc retrieval tracks.

Acknowledgments. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP170102231), by an Australian Government Research Training Program Scholarship, and by a grant from the Mozilla Foundation.

REFERENCES

[1] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A test collection with query variability. In *Proc. SIGIR*. 725–728.
[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2017. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*. 395–404.
[3] R.-C. Chen, J. S. Culpepper, T. T. Damessie, T. Jones, A. Mourad, K. Ong, F. Scholer, and E. Yulanti. 2015. RMIT at the TREC 2015 LiveQA Track. In *Proc. TREC*.
[4] R.-C. Chen, L. Gallagher, R. Blanco, and J. S. Culpepper. 2017. Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proc. SIGIR*. 445–454.
[5] C. L. A. Clarke, J. S. Culpepper, and A. Moffat. 2016. Efficiency-effectiveness tradeoffs in two-stage retrieval mechanisms. *Inf. Retr.* 19, 4 (2016), 351–377.
[6] M. Crane, J. S. Culpepper, J. Lin, J. Mackenzie, and A. Trotman. 2017. A comparison of document-at-a-Time and score-at-a-Time evaluation. In *Proc. WSDM*. 201–210.
[7] J. S. Culpepper, C. L. A. Clarke, and J. Lin. 2016. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proc. Aust. Doc. Comp. Symp.* 17–24.
[8] J. S. Culpepper, S. Mizzaro, M. Sanderson, and F. Scholer. 2014. TREC: Topic engineeRing ExerCise. In *Proc. SIGIR*. 1147–1150.
[9] J. S. Culpepper, M. Yasukawa, and F. Scholer. 2011. Language independent ranked retrieval with NeWT. In *Proc. Aust. Doc. Comp. Symp.* 18–25.
[10] J. K. Jayasinghe, W. Webber, M. Sanderson, and J. S. Culpepper. 2014. Improving test collection pools with machine learning. In *Proc. Aust. Doc. Comp. Symp.* 2–9.
[11] X. Lu, A. Moffat, and J. S. Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr.* 19, 4 (2016), 416–445.
[12] X. Lu, A. Moffat, and J. S. Culpepper. 2017. Can deep effectiveness metrics be evaluated using shallow judgment pools?. In *Proc. SIGIR*. 35–44.
[13] X. Lu, A. Moffat, and J. S. Culpepper. 2016. Modeling relevance as a function of retrieval rank. In *Proc. AIRS*. 3–15.

Table 3: The per-topic effectiveness of our runs when placed into context with other submissions over different categories. The Median column represents how many topics we achieve better or equal scores than the median value across all participants. As we merge the best scores across manual and automatic runs for Auto ⋃ Manual, the median value across this set is unknown as other participant's runs are not publicly available.

	AP			NDCG			P@10		
Run	Best	⇑ Median	Worst	Best	⇑ Median	Worst	Best	⇑ Median	Worst
Automatic Runs									
RMITFDMQEA1	0	21	0	0	17	0	4	32	4
Manual Runs									
RMITRBCUQVT5M1	6	20	1	10	21	3	14	36	2
RMITUQVBestM2	1	12	14	0	12	12	9	32	8
All Runs, Best of Auto ⋃ Manual									
RMITFDMQEA1	0	—	0	0	—	0	4	—	4
RMITRBCUQVT5M1	2	—	0	2	—	0	12	—	1
RMITUQVBestM2	0	—	0	0	—	0	9	—	2

[14] J. Mackenzie, R.-C. Chen, and J. S. Culpepper. 2016. RMIT at the TREC 2016 LiveQA Track. In *Proc. TREC*.

[15] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Information Systems* 35, 3 (2017), 24:1–24:38.

[16] E. M. Voorhees. 2005. Overview of the TREC 2004 Robust Retrieval Track. In *Proc. TREC*.