RMITB at TREC COVID 2020

Rodger Benham
RMIT University
Melbourne, Australia

Alistair Moffat
The University of
Melbourne, Australia

J. Shane Culpepper
RMIT University
Melbourne, Australia

Abstract

Search engine users rarely express an information need using the same query, and small differences in queries can lead to very different result sets. These *user query variations* have been exploited in past TREC CORE tracks to contribute diverse, highly-effective runs in offline evaluation campaigns with the goal of producing reusable test collections. In this paper, we document the query fusion runs submitted to the first and second round of TREC COVID, using ten queries per topic created by the first author. In our analysis, we focus primarily on the effects of having our second priority run omitted from the judgment pool. This run is of particular interest, as it surfaced a number of relevant documents that were not judged until later rounds of the task. If the additional judgments were included in the first round, the performance of this run increased by 35 rank positions when using RBP $\phi = 0.5$, highlighting the importance of judgment depth and coverage in assessment tasks.

1 Introduction

Harnessing the variability of user queries for a topic to improve search effectiveness has been validated in many studies. Bailey et al. [2016] created a test collection for ClueWeb12-B using 100 topics and 10,835 collected query variations; and the recent CC-News English newswire corpus has also supplied query variations [Mackenzie et al., 2020]. Inspired by the double fusion experiments of Bailey et al. [2017], experts solicited query variations for the Robust 2004 topics [Voorhees, 2004] as well as a few new topics, to be used for the RMIT [Benham et al., 2017] runs submitted to the TREC CORE 2017 track [Allan et al., 2017]. RMIT participated again the following year [Benham et al., 2018], producing the second-best run by AP, using shallow judgments to identify and fuse the most effective query variations. In follow-up work, Benham et al. [2019] show that query fusion can be applied to support boosting effectiveness at query time using CombSUM.

In this work, we apply query fusion to efficiently and effectively retrieve answers to questions from the scientific literature collected during the COVID-19 pandemic. We dissect the decisions made between submission rounds, using additional judgments gathered for topics appearing in previous rounds. In late 2019, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged as a health and economic disaster worldwide. The severity of the pandemic has led to a spike in scientific publications about the COVID-19 disease caused by the virus, prompting Wang et al. [2020] to create the CORD-19 dataset in order to encourage IR-related explorations into COVID-related scientific outcomes.

TREC COVID is the first TREC track to use the *residual collection scoring* pooling methodology described by Salton and Buckley [1990] (see also Moffat et al. [2007]). Voorhees [2020] found that the effectiveness of our second first round run `RMITBFuseM2` increased 33 positions in the overall system ranking when using P@5 and after additional judgments were gathered for those original topics during later rounds. Here we continue this line of inquiry using a post-hoc analysis and the full judgment set from all six rounds of the challenge.

2 First Round

The COVID-19 pandemic differs from previously explored IR contexts by virtue of the rapidly changing new information published daily over an extended period of time. When combined with the wide variety of questions being asked about the disease, creating a practical IR evaluation exercise with the potential to produce valuable insights for future pandemics is a challenge. In the first round of the task RMIT submitted two double fusion runs. The first run RMITBM1 re-weighted relevance scores of documents by freshness, and RMITBFuseM2 served as a control to determine the efficacy of that time-biased approach. Since RMITBFuseM2 contained a proper subset of the techniques applied to RMITBM1, we first describe how that run was built.

Processing The Corpus. The first round uses the CORD-19 dataset as at April 10, 2020, with the commercial, non-commercial, custom license, and bioRxiv subsets. The corpus includes a metadata CSV file with various attributes about publications, with fields including but not limited to: the title, authors, an abstract, and the filename of the associated PMC and/or PDF JSON parse of each publication.

Participants were instructed to prefer the PMC parse over the PDF parse. Both of these parses were supplied in JSON, where the document text is dispersed in JSON objects with metadata about the context of each sentence in relation to its place in the document. That level of detail is superfluous for the retrieval models we employed, and we extracted the document text from these objects and transformed each parsed document into plain-text.

The stipulated list of document identifiers for the first round excluded previously judged documents in an initial pool of three Anserini baselines judged to depth-40 [Voorhees et al., 2020]. Many of the documents to be assessed did not have an associated PMC or PDF parse, with an abstract only. Since these records had no document to process we did not index them. (We subsequently found this decision to be detrimental, as abstract-only document records were judged in the first round).

We used Terrier to index the corpus, as Kurland and Culpepper [2018] show that it can produce double fusion runs with higher effectiveness than the best submitted run to Robust04. Lin and Zhang [2020] recently showed that Terrier configurations are the most reproducible out of a series of tests applied to popular IR retrieval tools.

Query Variations. The TREC COVID task operated with short preparation windows. Each team had roughly a week to submit runs after receiving judgments for the prior round. That short time-frame prevented the design of a crowdworker study to solicit query variations. Instead, the first author of the paper took on the task of creating 10 query variations for each of the initial set of 30 topics.

For example the second topic was the query *coronavirus response to weather changes*, with the associated narrative:

> Seeking range of information about the SARS-CoV-2 virus viability in different weather/climate conditions as well as information related to transmission of the virus in different climate conditions.

After having read that narrative and interpreting the question in the released topic set, these ten queries were created:

1. *coronavirus climate change*
2. *coronavirus weather*
3. *coronavirus humidity*
4. *coronavirus cool dry climate*
5. *coronavirus cool humid climate*
6. *coronavirus viability cool temperatures*
7. *coronavirus winter temperatures viable*
8. *coronavirus summer temperatures viable*
9. *coronavirus seasonal climate*
10. *coronavirus standard laboratory conditions*

Double Fusion. Bailey et al. [2017] proposed double fusion as a technique to submit query variations to many retrieval models and fuse the results with a rank fusion algorithm. The technique was used previously to create the second-most effective TREC CORE 2018 adhoc run [Benham et al., 2018].

Initially, the aim was to use 16 different retrieval models with and without query expansion, 32 rankings in total. However limitations in the fusion script, and the short time-lines being worked to (meaning that software modifications were risky), meant that only eight retrieval models were used in the first round.

Many Terrier retrieval models are derived from the divergence from randomness (DFR) model [Amati and van Rijsbergen, 2002]. Of the available options, the following variants were used to form rankings in the first round submission:

- BB2 (DFR) [Plachouras et al., 2004]
- BM25 [Robertson et al., 1995]
- DFR_BM25 (DFR) [Amati, 2003]
- DLH (DFR) [Macdonald et al., 2005]
- DLH13 (DFR) [Macdonald et al., 2005]
- DPH (DFR) [Amati et al., 2008]
- DFRee (DFR) [Amati et al., 2011]
- Hiemstra_LM [Hiemstra, 2001]

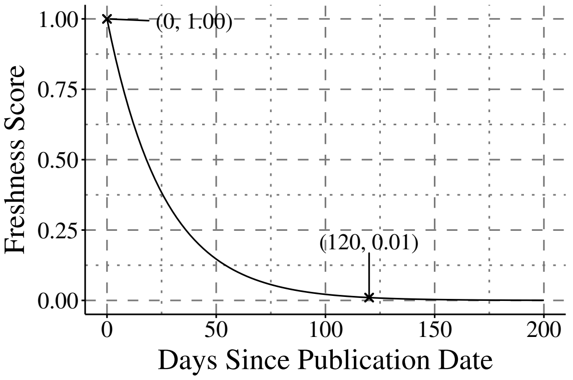


Figure 1: Experimental freshness scoring function employed in RMITBM1. The function is an exponential decay through the two points marked with crosses.

Running the 300 queries against the eight models with and without expansion resulted in the generation of 4,800 rankings to depth 1,000. To fuse each of the 160 rankings per-topic, CombSUM was then employed [Shaw and Fox, 1995], as described by Benham et al. [2019]. The run RMITBFuseM2 was the result of fusing across (for each topic) systems and queries.

Freshness. Knowledge of the COVID-19 pandemic is constantly evolving. The second run, RMITBM1, was the result of combining a freshness score with the orderings present in the RMITBFuseM2 run, based on the hypothesis that facts disseminated four months ago are likely to be obsolete compared to an article published on the same topic more recently. With that, we parse the publication date out of the supplied metadata file and calculate the number of days since the paper was published. We then define an exponential decay on the variables *days* since publication, fitting freshness scores through the two values: $(0, 1)$ and $(120, 0.01)$. Figure 1 visualizes that fitted function, plotting the derived formula:

$$t(\mathit{days}) = 0.9623506264^{\mathit{days}}. \qquad (1)$$

The publication dates in the supplied corpus metadata file required cleaning. Some articles were erroneously published in the future on the last day of 2020; those dates were adjusted to be the last day of 2019. Other future dates corresponded to the date of a conference or when a journal article was to be officially made available, but not when the work was first disseminated. In such instances, where the days since publication produced negative integers, *days* was set to 0. On rare occasions, the date format would change from Y-m-d to Y, and so, it would be parsed as the first day of that year. Empty date strings were taken to be the first day of 2020.

To combine a freshness score with the relevance score of a document assuming equal importance, the document scores in a run are adjusted to be between 0 and 1 with minimax scaling to be cast in the same units of freshness.

A simple unweighted linear combination is used, and so, the adjusted document score is

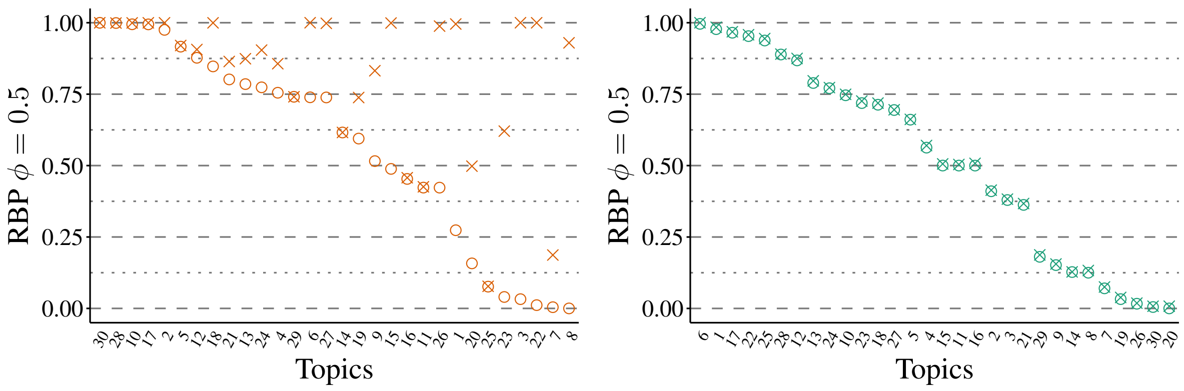
$$\mathit{score}(d, q) = \mathcal{M}(s_q, 0, 1) + t(\mathit{days}), \qquad (2)$$

where $\mathcal{M}$ represents the minimax function, and s_q refers to the relevance score of the document in response to the query q.

After applying Equation 2 to RMITBFuseM2, and sorting the document list of each topic by the respective *score*(d, q), we get RMITBM1.

Analysis. After the first round, the organizers shared the judgments and the evaluation summaries of each submitted run. Each run summary contained the measures: total relevant, total relevant retrieved, average precision, mean BPref, and mean NDCG@10. These measures are based on knowledge of how many relevant documents there are in the collection, which when judging to depth 7 in a residual collection pooling context seems unlikely to provide complete figures. A graph showing deviation of median P@5 indicated that our RMITBM1 run was less effective than anticipated.

The organizers later reported rank-biased precision (RBP) [Moffat and Zobel, 2008] in subsequent rounds with an expected viewing depth of the top-2 documents ($\phi = 0.5$). RBP does not rely on knowing the count of relevant documents for each topic – a statistic that has been argued to be untrustworthy for use in pooled evaluation campaigns [Zobel et al., 2009]. Figure 3 shows the monotonically decreasing RBP $\phi = 0.5$ minimum score of each submitted run marked as a circle, with the associated score uncertainty (the *residual*) marked with a cross. Points marked in green indicate that the system was pooled, and brown points indicate otherwise, a convention used consistently throughout the paper. We were cautiously optimistic about the RMITBFuseM2 run after checking the RBP residuals, as it was more effective than RMITBM1, even though it didn't contribute to the pool of judgments. Figure 2 shows the monotonically decreasing topic



(a) `RMITBFuseM2`

(b) `RMITBM1`

Figure 2: Per-topic RBP $\phi = 0.5$ evaluation of our submitted TREC COVID round 1 runs. Circles mark the RBP score, and the corresponding crosses mark the "RBP plus residual" score, indicating the maximum possible score for that topic if all unjudged documents turned out to be relevant. On the right, run `RMITBM1` contributed to the judgment pool and has small residuals, whereas run `RMITBFuseM2` on the left did not, and hence might potentially have identified additional relevant documents and obtained an even higher mean RBP score.

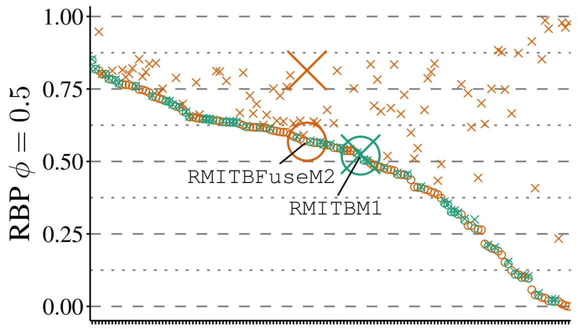


Figure 3: RBP $\phi = 0.5$ evaluation of the 143 runs submitted to TREC COVID round 1, where circles mark the RBP score, and crosses mark the corresponding "RBP plus residual" score. Run `RMITBM1` is in the middle of the pack, and the unpooled `RMITBFuseM2` run outperforms it, with the possibility of having been within the group of top runs if the judgments provided more coverage.

scores for each of our two submitted runs. In Figure 2a, although the residual and minimum score are close on some topics, for many there is a wide diverge. In particular, where the cross is near 1.0, it means that none of the document retrieved near the top of the run for that topic had been judged to be *non*-relevant either.

Figure 3 shows that the freshness reweighting function in Equation 2 was not as useful as we had anticipated. Either there were many relevant documents older than four months, or the freshness signal dominated the relevance signal in a way that harmed effectiveness. As only one run per-group had been judged 7 documents deep, there was no incentive to tune a coefficient in a linear combination of relevance and freshness scores, as we risk overfitting our model to reduce the diversity of our run, and submitting a control run might not be judged.

With the uncertainty brought about by the shallow judgments, we now move to discuss the decisions made in submitting our second round run.

3 Second Round

In assessing risk aspects of the pipeline used in the poorly performing round 1 submission of `RMITBM1`, we:

- remove freshness re-ranking;
- disable query expansion;
- add extra eight retrieval models; and
- include abstract-only documents.

Although query expansion is a powerful tool, it often requires additional parameter tuning with relevance judgments to avoid query drift, as these parameters vary across corpora [Billerbeck and Zobel, 2004]. Noting the shallow judgments issue in the analysis provided in the previous section, we abandoned query expansion, uncertain as to whether it was reducing the effectiveness of the query fusion with the default Terrier parameters. We will instead explore this option in future work.

As previously discussed, our implementation was able to fuse up to 16 runs for each topic, and with expansion removed, we returned to our original list of systems, adding in eight further alternatives:

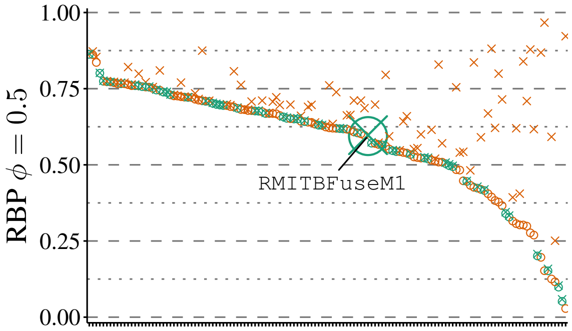


Figure 4: RBP $\phi = 0.5$ evaluation of the 136 runs submitted to TREC COVID round 2, where circles mark the RBP score, and crosses mark the RBP score plus residual. Our RMITBFuseM1 run generated only average performance.

- IFB2 (DFR) [Plachouras et al., 2004]
- In_expB2 (DFR) [Plachouras et al., 2004]
- In_expC2 (DFR) [Plachouras et al., 2004]
- InL2 (DFR) [Plachouras et al., 2004]
- LemurTF_IDF [Zhai, 2001]
- LGD [Clinchant and Gaussier, 2009]
- PL2 (DFR) [Plachouras et al., 2004]
- TF_IDF [Spärck Jones, 1972]

Combining these eight models with the original eight, and not employing query expansion, meant that there were again 16 systems being fused for each topic.

Analysis. Figure 4 shows that the resulting round 2 RMITBFuseM1 run was slightly below average in effectiveness relative to other submitted runs. Benham et al. [2018] found that a small pool of judgments used to select the top-5 most effective query variations independently per topic led to effectiveness improvements. But the tight time constraints for TREC COVID, and a lack of medical expertise, meant that undertaking manual judgments was not an option.

After inspecting the residuals shown Figure 2, where RMITBFuseM2 only has one setting turned off compared to RMITBM1, it does not appear that a failure analysis based on the run components will be fruitful at this point.

4 Post-Hoc Analysis

Voorhees [2020] provides a post-hoc analysis after the complete judgments had been shared, showing that the relative system orderings would have changed for RMITBFuseM2:, and noting:

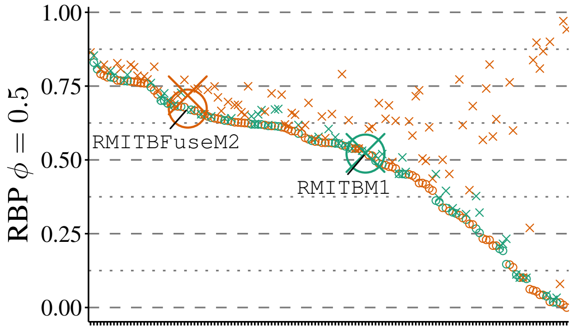


Figure 5: RBP $\phi = 0.5$ evaluation of the 143 runs submitted to TREC COVID round 1 using the complete judgments provided at the end, to be directly compared against Figure 3. Run RMITBFuseM2 was originally "average", but moves into the top-quartile.

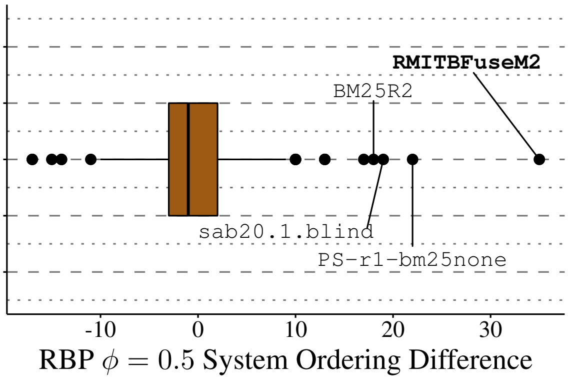


Figure 6: System rankings changes based on RBP $\phi = 0.5$ evaluation using the round 1 judgments only, and then the complete set. Run RMITBFuseM2 moves up the most places out of the submissions, where runs are labeled if they are extreme outliers (3.0 × *IQR*).

> *The largest change in the relative ranking of runs is the* RMITBFuseM2 *run which rises 33 ranks when using P@5 as the measure (21 ranks by NDCG@10, 7 ranks by MAP and none for BPref).*

The RBP analysis shown in Figure 3 hinted that RMITBFuseM2 could have been in the top-third of first round submissions, but all that could be concluded based on the round 1 judgments was that it was at least average in effectiveness. Figure 5 plots the same first round systems, but using the larger set of qrels, and shows that when measured by RBP, RMITBFuseM2 moves up 35 places. Enjoying the wisdom that comes with hindsight, if we had known that outcome, we might not have disabled the query expansion features mentioned in Section 3 in our second round submission.

Figure 6 shows a boxplot of the relative ranking changes of all systems on the RBP $\phi = 0.5$

System	Rank	RBP	Residual
Round 1 Judgments			
RMITBFuseM2	65	0.586	0.246
PS-r1-bm25none	60	0.592	0.205
sab20.1.blind	74	0.547	0.267
BM25R2	91	0.468	0.366
Complete Judgments			
RMITBFuseM2	30	0.674	0.045
PS-r1-bm25none	38	0.641	0.108
sab20.1.blind	55	0.615	0.110
BM25R2	73	0.557	0.098

Table 1: Comparing the effectiveness of the extreme outlier runs for RBP $\phi = 0.5$ rank changes shown in Figure 6, where rank refers to the RBP system ranking measured against the first round runs.

measure from the smaller first round judgment set to the full set. Most systems have modest changes in rank, however, there are outliers, with the extreme outliers labeled on the graph. Of these outliers: `PS-r1-bm25none` also generated a query manually from the topic descriptions; `sab20.1.blind` is a pseudo relevance feedback run without abstracts; and `BM25R2` is a BM25 run where the index contains the title, abstract, and paragraph fields combined with Anserini's *Covid-Query Generator* to generate queries.

Table 1 documents these RBP-based relative system orderings, along with the effectiveness scores and residuals for the first round judgment set compared with the complete judgment set. Although the residuals are smaller on the complete judgment set, it is possible that `PS-r1-bm25none` or `sab20.1.blind` could outrank our `RMITBFuseM2` run with complete judgments, and could, potentially, result in further jumps in system ordering of 8 and 25 places respectively.

5 Conclusion

We have documented our participation in the TREC COVID track. While early evaluation outcomes in the per-round analysis indicated that our runs were at best average, deeper judgments on the first round run `RMITBFuseM2` lifted its system ranking by 35 positions (RBP $\phi = 0.5$). We found the residual pooling approach to be a refreshing take on judgment solicitation with feedback, and welcome a similar approach being applied to future evaluation campaigns. Our recommendation would be to have fewer rounds and allow for more judgments per round, where each feedback round is evaluated with fixed pooling conditions. Not only would this reduce the volatility in judgment coverage observed in this year's track, it would also provide a higher quality test collection for detailed failure analyses of system submissions in all rounds. We also encourage the use of residuals as a way of gauging the extent to which measurements derived from pooled judgments can be considered to be reliable.

Acknowledgments

The first author was supported by an RMIT Vice-Chancellor's PhD scholarship. The second and third authors were supported by the Australian Research Council (grant DP190101113).

Code

Code and query variations to reproduce experiments are available at https://github.com/rmit-ir/rmitb-trec-covid.

References

J. Allan, D. Harman, E. Kanoulas, D. Li, C. V. Gysel, and E. M. Voorhees. TREC 2017 common core track overview. In *Proc. TREC*, 2017.

G. Amati. *Probabilistic Models for Information Retrieval Based on Divergence From Randomness*. PhD thesis, School of Computing Science, University of Glasgow, 2003.

G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Sys.*, 20(4):357–389, 2002.

G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2008 blog track. In *Proc. TREC*, 2008.

G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR, UNIVAQ at TREC 2011 microblog track. In *Proc. TREC*, 2011.

P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV100: A test collection with query variability. In *Proc. SIGIR*, pages 725–728, 2016.

P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. SIGIR*, pages 395–404, 2017.

R. Benham, L. Gallagher, J. Mackenzie, T. T. Damessie, R.-C. Chen, F. Scholer, A. Moffat, and J. S. Culpepper. RMIT at the 2017 TREC CORE track. In *Proc. TREC*, 2017.

R. Benham, L. Gallagher, J. Mackenzie, B. Liu, X. Lu, F. Scholer, A. Moffat, and J. S. Culpepper. RMIT at the 2018 TREC CORE track. In *Proc. TREC*, 2018.

R. Benham, J. Mackenzie, A. Moffat, and J. S. Culpepper. Boosting search performance using query variations. *ACM Trans. Inf. Sys.*, 37(4):41.1–41.25, 2019.

B. Billerbeck and J. Zobel. Questioning query expansion: An examination of behaviour and parameters. In *Proc. ADC*, pages 69–76, 2004.

S. Clinchant and E. Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for IR. In *Proc. ICTIR*, pages 54–65, 2009.

D. Hiemstra. *Using Language Models for Information Retrieval*. Univ. Twente, 2001. ISBN 978-90-75296-05-1.

O. Kurland and J. S. Culpepper. Fusion in information retrieval: SIGIR 2018 half-day tutorial. In *Proc. SIGIR*, pages 1383–1386, 2018.

J. Lin and Q. Zhang. Reproducibility is a process, not an achievement: The replicability of IR reproducibility experiments. In *Proc. ECIR*, pages 43–49, 2020.

C. Macdonald, B. He, V. Plachouras, and I. Ounis. University of Glasgow at TREC 2005: Experiments in terabyte and enterprise tracks with Terrier. In *Proc. TREC*, 2005.

J. Mackenzie, R. Benham, M. Petri, J. R. Trippas, J. S. Culpepper, and A. Moffat. CC-News-En: A large English news corpus. In *Proc. CIKM*, pages 3077–3084, 2020.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.*, 27(1):2.1–2.27, 2008.

A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. SIGIR*, pages 375–382, 2007.

V. Plachouras, B. He, and I. Ounis. University of Glasgow at TREC 2004: Experiments in web, robust, and terabyte tracks with Terrier. In *Proc. TREC*, 2004.

S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proc. TREC*, 1995.

G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *J. Am. Soc. Inf. Sci.*, 41(4):288–297, 1990.

J. A. Shaw and E. A. Fox. Combination of multiple searches. In *Proc. TREC*, 1995.

K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60(5):493–502, 1972.

E. Voorhees. Effect on system rankings of further extending pools for TREC-COVID round 1 submissions. https://ir.nist.gov/covidSubmit/papers/rnd1runs_j0.5-2.0.pdf, 2020.

E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, and L. L. Wang. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1–12, 2020.

E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proc. TREC*, pages 69–77, 2004.

L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. CORD-19: The COVID-19 Open Research Dataset. *arXiv*, abs/2004/10706, 2020.

C. Zhai. Notes on the Lemur TFIDF model, 2001. Unpublished report, available at http://lemurproject.org/lemur/tfidf.pdf.

J. Zobel, A. Moffat, and L. A. F. Park. Against recall: Is it persistence, cardinality, density, coverage, or totality? In *SIGIR Forum*, volume 43, pages 3–8, 2009.