RMIT at the NTCIR-13 We Want Web Task

Luke Gallagher
RMIT University
Melbourne, Australia
luke.gallagher@rmit.edu.au

Joel Mackenzie
RMIT University
Melbourne, Australia
joel.mackenzie@rmit.edu.au

Rodger Benham
RMIT University
Melbourne, Australia
rodger.benham@rmit.edu.au

Ruey-Cheng Chen
RMIT University
Melbourne, Australia
ruey-cheng.chen@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

ABSTRACT

Furthering the state-of-the-art in adhoc web search is one of the underlying goals for the NTCIR-13 We Want Web (WWW) task. Adhoc search can be viewed as a bridge connecting many of the specialized sub-fields that are a result of the way people connect to and use information access systems. Since this is the first year of the WWW task, and no training data was provided for the English subtask, we focused on classic techniques for improving effectiveness in lieu of modern techniques based on supervised learning. In particular, we explored the use of Markov Random Field Models (MRFs), static document features, field-based weighting, and query expansion. This round we made extensive use of the Indri search system and the flexible query language it provides to produce effective results.

Team Name

RMIT

Subtasks

WWW (English)

Keywords

Term Dependency Models; Query Expansion

1. INTRODUCTION

The RMIT team participated in the We Want Web (WWW) English subtask of NTCIR-13 [16]. The aim of this task is to revitalize interest in adhoc web search. Continued development of novel research in this area may be complementary to other information access tasks that benefit the wider information access community. With the adhoc task planned to run for three rounds, certain unique opportunities may arise, for instance, subsequent rounds will pool participating systems of the current round in conjunction with previous tasks. RMIT's interest in this track is driven by two related lines of research in our group: efficient multi-stage retrieval [2, 3, 4, 6, 7, 9, 17], and more reliable deep evaluation when using shallow judgments [8, 11, 13, 14, 15]. We plan to use the lessons we learn in the track this year to perform more reliable evaluation and search in multi-stage search systems. The remainder of this paper outlines the experiments and results conducted for the English subtask.

2. SYSTEMS

Four different system configurations were submitted by RMIT this round:

- R1 = SDM Fields $+$ RM3 Query Expansion
- R2 = Linear combination of R1 $+ 0.25 \times$ PageRank Priors
- R3 = FDM $+$ RM3 Query Expansion
- R4 = n-gram fields $+$ RM3 Query Expansion

All system configurations made use of query expansion and relevance feedback as initially described by Lavrenko and Croft [12]. The model we used is generally referred to as RM3, and is a common competitive baseline used by researchers working on query expansion. Specific feedback parameters are detailed for each system below, such as the number of feedback documents (R_d), the number of feedback terms (R_t), and the interpolation weight between the expansion terms and original query (R_w).

Another common theme among the system configurations was the use of term dependency models. Identical smoothing parameters were used for all systems, with $\mu = 2000$ and $\mu_{prox} = 2000$. Post-retrieval spam filtering was applied to systems R2, R3, and R4. Documents with a spam score less than 70 were removed from retrieved results.[1]

For system R1, unstructured queries were transformed into structured queries in the Indri query language using a field-based sequential dependency model [18]. Recent experiments have shown that this configuration works very well on the 2009-2012 TREC ClueWeb09 test collections. For example, the three-term query "big red house" was represented as:

```
#weight(
  $\alpha_1$ #combine(big.title red.title house.title)
  $\alpha_2$ #combine(big.inlink red.inlink house.inlink)
  $\alpha_3$ #weight(
    $\beta_1$ #combine(big.body red.body house.body)
    $\beta_2$ #combine(#1(big.body red.body)
               #1(red.body house.body))
    $\beta_3$ #combine(#uw8(big.body red.body)
                #uw8(red.body house.body))
  )
)
```

Here, α_1, α_2 and α_3 control the weight given to each field. The sequential dependency model is used for matching against the body representation, and β_1, β_2 and β_3 control the weight

[1] https://www.mansci.uwaterloo.ca/~msmucker/cw12spam/

that the sequential dependency model allocates to each proximity feature for matching within the entire document, where bigrams (`#1`), and 8-term unordered windows (`#uw8`) are used as originally described by Metzler and Croft [18]. Parameters were determined using a parameter sweep on 200 queries from the TREC 2009-2012 Web Track topics for ClueWeb09-B, and the values were $(\alpha_1, \alpha_2, \alpha_3) = (0.20, 0.05, 0.75)$, and $(\beta_1, \beta_2, \beta_3) = (0.8, 0.1, 0.1)$. Query expansion parameters for R1 were set as $(R_d, R_t, R_w) = (10, 50, 0.6)$. System R2 was configured the same as R1, but also included PageRank priors to form a linear combination of R1 and PageRank applied with a weight of 0.25.

The configuration of the third system R3 utilized the full dependency model (FDM) as described by Metzler and Croft [18] with the suggested weights of $(\alpha_1, \alpha_2, \alpha_3) = (0.8, 0.1, 0.1)$ Query expansion was also applied with the following parameters $(R_d, R_t, R_w) = (20, 10, 0.8)$. System R4 was configured with pseudo-relevance feedback parameters $(R_d, R_t, R_w) = (10, 50, 0.6)$, in line with R1.

System R4 differs from the others in how the Indri query language was utilized to form a field-based n-gram dependency model. This can be seen as a further generalization of the query structure applied in system R1. The idea was to try and capture higher order dependencies when the queries were longer than two terms.

Using the same example query "`big red house`", the expanded query would be:

```
#weight(
  $\alpha_1$ #combine(big.title red.title house.title)
  $\alpha_2$ #combine(big.inlink red.inlink house.inlink)
  $\alpha_3$ #combine(big.body red.body house.body)
  $\alpha_4$ #combine(#1(big.body red.body)
       #1(red.body house.body))
  $\alpha_1$ #combine(#uw8(big.body red.body)
       #uw8(big.body house.body)
       #uw8(red.body house.body))
  $\alpha_1$ #combine(#1(big.title red.title)
       #1(red.title house.title))
  $\alpha_1$ #combine(#uw8(big.title red.title)
       #uw8(big.title house.title)
       #uw8(red.title house.title))
  $\alpha_4$ #combine(#1(big.body red.body house.body))
  $\alpha_1$ #combine(#uw12(big.body red.body house.body))
  $\alpha_1$ #combine(#1(big.title red.title house.title))
  $\alpha_1$ #combine(#uw12(big.title red.title house.title))
)
```

As before α_1, α_2, α_3, α_4 control the weight given to each field within the structured query, although the ordered (`#1`) and unordered (`#uw`) operators extend into n-gram combinations in addition, non-adjacent bigram pairs, unlike R1. Parameter settings were identified in a similar fashion to R1, having values $(\alpha_1, \alpha_2, \alpha_3) = (0.2, 0.05, 0.75)$ and $\alpha_4 = 0.1$. An alternative approach would have been to construct this query using the weighted sum operator (`#wsum`), however, we leave this for future investigation.

3. EVALUATION

For the evaluation of retrieval effectiveness, we use metrics targeting early precision: ERR, NDCG and RBP. When reporting ERR and NDCG the cutoffs 5, 10, and 20 are used. These are consistent with the recommendations of Lu et al. [13], based on the pooling depth, and number of systems in the pool. For RBP the user persistence parameter p is shown at values 0.8 and 0.9, corresponding to expected viewing depths of 5 and 10, respectively. The NTCIR-13 WWW task overview paper outlines the official metrics and associated evaluation tools that were used for evaluation of the task [16]. In the evaluation that follows, we report additional results in order to investigate how consistent early precision metrics are to the official reported results. ERR and NDCG are computed using `gdeval`,[2] while RBP is computed using `rbp_eval`.[3] The disparity in the size of the collection and the pooling depth presents us with an opportunity to utilize one of the strengths of RBP – calculating a residual that measures the level of uncertainty in a point estimate of system effectiveness due to the presence of unjudged documents in a system's ranked results list. This is also useful for the analysis of consistency, and inter-system comparison.

Relevance Judgments. Examining the official track relevance judgments, it is interesting to note that a high proportion of documents are labeled as relevant. The assessment process is outlined in the NTCIR-13 overview paper [16]. Of the 4 relevance levels assigned to all judgments, 6.9% (1,583/22,912) were given a relevance grade of 4. This is very high when compared with the judgments for the ClueWeb12-B collection from the TREC Web Tracks of 2013-2014 where, ignoring documents labeled as "junk" (relevance level -2), the fraction of judgments with a label of 4 is 0.001% (40/28,116). We must be clear though, that the judgment process, query set and tasks are different between the two judgment sets.

The difference in the distribution of relevance levels may also be due to the way in which queries were specified in the different evaluation campaigns; unlike TREC, the NTCIR-13 track only showed the "raw" search queries to assessors and did not include additional details about specific information needs. To examine this, we considered the case of navigational queries; even for a large collection of documents, one might expect that a navigational query such as "`fifa`" (topic 3) would have few highly relevant documents. However, this does not appear to be the case: of the 96 positive labels for this topic, 9 were given a judgment label of 4. This may indicate that the lack of a topic description and narrative presented the assessors with a challenging scenario where they were without the information required to better discriminate relevance across the topic set. On the other hand, there may be a plausible scenario where a navigational query like "`fifa`" does in fact have a high number of results that are highly relevant given the task is to diversify the results that such a query could incite in the absence of any further background information.

4. EXPERIMENTS

All experiments were conducted for the English subtask on the ClueWeb12-B corpus. Indexing was performed with Indri[4], with Krovetz stemming and no stopping of terms.

Overall results. Table 1 displays the results of all four submissions by the RMIT team for the English subtask, along with five related posthoc, non-submitted analysis runs which are detailed in later subsections. The strongest performing system R1 produced results that were significantly superior to the other submissions for metrics NDCG and RBP, apart from NDCG@5. However, the most effective system when measuring ERR was R4. Both R1 and R4 make use of field extents which may explain their effectiveness when compared against the classic full dependence model R3 with RM3 query expansion. The degraded performance of submission R2 that includes PageRank is a little surprising. Static document scoring methods such as PageRank and Spam scoring have been

[2] http://trec.nist.gov/data/web/10/gdeval.pl

[3] http://people.eng.unimelb.edu.au/ammoffat/rbp_eval-0.2.tar.gz

[4] http://lemurproject.org/indri.php

Table 1: Results for topics 1-100 for the English subtask. Holm corrected pairwise statistical tests were performed, with † indicating significance at $p = 0.05$ and ‡ indicating significance at $p = 0.01$ relative to R3.

	ERR@k			NDCG@k			RBP@p	
System	@5	@10	@20	@5	@10	@20	@0.8	@0.9
R3	0.5065	0.5207	0.5257	0.3977	0.3968	0.3970	0.8125+0.0006	0.7670+0.0242
R2	0.5285	0.5378	0.5419	0.4186	0.4069	0.3981	0.7965+0.0006	0.7533+0.0228
R4	**0.5635**	**0.5728**	**0.5775**	0.4402	0.4249	0.4175	0.7903+0.0008	0.7422+0.0270
R1	0.5548	0.5712	0.5746	**0.4670**	**0.4783**‡	**0.5006**‡	**0.8803+0.0006**†	**0.8438+0.0221**‡
Post-hoc analysis of submissions								
BM25	0.4760	0.4879	0.4922	0.3718	0.3713	0.3775	0.6966+0.1527‡	0.6509+0.1919‡
R3-NQE	0.4955	0.5096	0.5144	0.3884	0.3879	0.3881	0.8036+0.0043	0.7560+0.0348
R2-NQE	0.5279	0.5403	0.5447	0.4161	0.4125	0.4008	0.8062+0.0116	0.7537+0.0408
R4-NQE	0.5533	0.5637	0.5679	0.4276	0.4071	0.4010	0.7816+0.0069	0.7238+0.0456
RBC-14	**0.5819**	**0.5951**	**0.5984**	**0.4817**‡	0.4776‡	0.4932‡	0.8483+0.0000	**0.8263+0.0025**‡
R1-NQE	0.5743	0.5884	0.5916	0.4723†	**0.4877**‡	**0.4967**‡	**0.8677+0.0150**‡	0.8220+0.0453‡

shown to generally improve effectiveness when using ClueWeb collections. This relationship between static document scoring and strong field-based MRF baselines warrants further study.

Bag-of-words Analysis. In order to better understand for which topics the submitted runs performed well, and which topics were more difficult, we have added an additional bag-of-words baseline run using BM25. For each of the submitted runs, we investigate per-topic performance when compared to the BM25 run. Figure 1 depicts the percentage of topics that are improved or degraded across the four submitted runs when compared to BM25 with respect to NDCG@10, with each topic binned into one of the five categories (x-axis) according to the magnitude of the change.

Firstly we note that the retrieval performed using BM25 was not a system that contributed to the pooling process, and therefore will be more likely to include unjudged documents in its ranked list of results. This is unfortunate, as it would have been relatively simple for the organizers to include several out-of-the-box baseline runs from common systems; see, for example, the system configurations available in the IR Reproducibility Challenge GitHub Repository[5]. In future iterations of the WWW Task, it might be useful to include as many of these as possible, in order to increase the diversity of the judgment pool. A secondary factor here is the shallow pooling depth for which more sophisticated models are likely to promote relevant documents from deeper ranks.

It is clear from Figure 1 that R1 consistently improves a majority of the queries over BM25, while at the same time per-query degradation resulting from R1 declines as the negative impact increases. All runs improve upon the effectiveness of BM25 by 100% or more for around $15-20\%$ of topics. Note that the topics (27,28,62,68,71) that received a BM25 score of zero were placed in the interval $[50\%, 100\%]$ so as to not inflate the perception of improvements made beyond 100%.

Per-query Breakdown. Turning our attention to individual topics, the top 5 and worst 5 performing topics relative to BM25 for NDCG@10 are shown in Table 2 for our best system configuration (R1). Even after more than two decades, the BM25 baseline is still an efficient and effective system for certain queries when compared against more sophisticated ranking models.

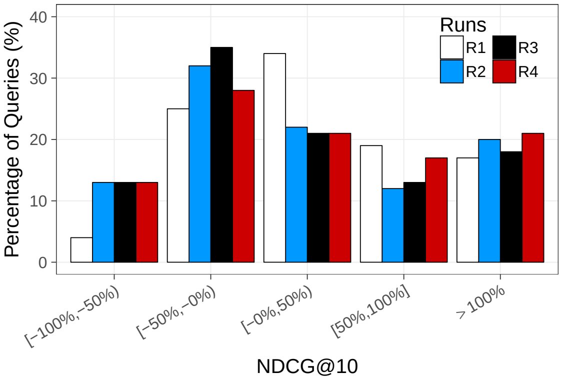

Figure 1: Per query change of the submitted runs when compared to a BM25 baseline.

Included in Table 2 are results for the non-expanded version of system R1. The 5 worst queries could have preformed better in the absence of query expansion, and in fact, topic 57 outperforms BM25. For the 5 queries that produced the most effective result relative to BM25, the trend of not applying expansion continues for three out of the five topics. However, the topics “`robot`” and “`typing practice`” are more effective with query expansion enabled. This hints at the volatile nature of query expansion and its ability to simultaneously improve and diminish effectiveness across different topics. A number of solutions have been proposed that incorporate the notion of risk [5, 10], which we may experiment with in the future. Further analysis on query expansion is presented in the next section.

To more closely examine the differences between BM25 and R1, one co-author reviewed the top 10 retrieved documents for both runs on the 5 worst queries, to specifically look for two types of errors:

[5] https://github.com/lintool/IR-Reproducibility

Table 2: R1 top 5 worst and best queries when compared to BM25.

Topic	R1	BM25	Δ	R1-NQE	Query	Error Cause	P/D
83	0.4129	0.7189	-0.3060	0.5769	`jetstar airlines hong kong`	Misaligned dependencies	2/3
88	0.3179	0.5943	-0.2764	0.5291	`mexico climate`	Biased toward short docs	0/6
57	0.3209	0.5893	-0.2684	0.6307	`axle ratio`	Query drift	0/10
54	0.4505	0.7025	-0.2519	0.4820	`anime pillow`	Query drift	0/6
41	0.2281	0.4676	-0.2395	0.2951	`autumn`	—	4/2
71	0.5948	0.0000	+0.5948	0.7863	`dog food for allergies`		
46	0.6958	0.1423	+0.5535	0.6795	`musical note`		
45	0.6399	0.0812	+0.5586	0.8193	`commendatory term`		
30	0.9458	0.3898	+0.5561	0.8045	`robot`		
28	0.5113	0.0000	+0.5113	0.3642	`typing practice`		

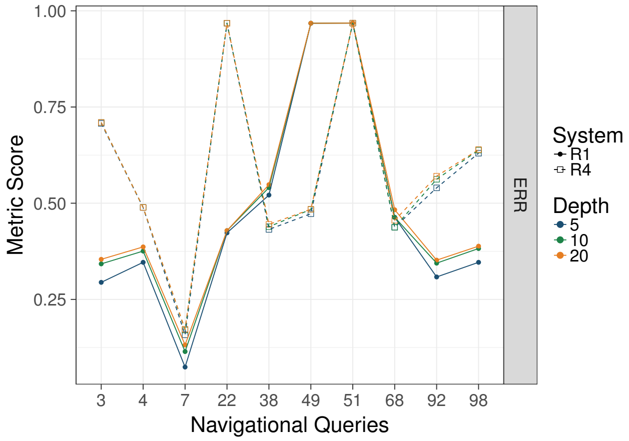

Figure 2: Results of ERR across the 10 queries labeled as navigational for the task.

- P-type error (erroneously promoting documents): non-relevant documents initially placed outside top 10 positions in the BM25 ranking, but wrong pushed into the top 10 by R1;
- D-type error (erroneously demoting documents): relevant documents originally retrieved into the top 10 positions by BM25, but wrongly placed outside the top 10 in the R1 ranking.

From the 5 worst queries, a total of 6 and 27 incidents were identified respectively for P-type and D-type errors. Three queries appeared to suffer entirely from D-type errors, indicating potential issues related to query expansion. The same co-author was then asked to comment on the qualitative aspect of the incidences (i.e. wrongly promoted/demoted documents), and finally for each query to attribute all document-level errors to one single error cause.

The identified causes, as well as the number of each error type, are given in Table 2. We found that query expansion is negatively impacting topics 57 (“`axle ratio`”) and 54 (“`anime pillow`”), suggesting some degree of query drift. In topic 83, R1 struggles with the dependencies between “`jetstar`” and “`hong kong`”, and in some cases the two query term groups can have close proximity but are entirely unrelated. Topic 88 suffers from a bias where documents of good quality are placed lower in the ranking than non-informative, short documents. The single query term topic 41 (“`autumn`”) actually has more P-type errors, but no obvious pattern is seen across all its error incidences.

Table 3: Wins, ties and losses for each submitted system's non-query expansion run, compared against each submitted query expansion run counterpart, using NDCG@10. Statistical tests were performed between each pair (Sys. A, Sys. B) with † indicating significance at $p = 0.05$ and ‡ indicating significance at $p = 0.01$. Scores are tied for NDCG@10 $\Delta \pm 0.025$, and ignored in the $\sum$ *Win* and $\sum$ *Loss* columns.

NDCG@10								
System A	System B	Win	Tie	Loss	$\frac{Win}{Loss}$	$\sum Win$	$\sum Loss$	$\frac{\sum Win}{\sum Loss}$
R1-NQE	R1	32	39	29	1.103	16.033	15.391	1.042
R2-NQE	R2	29	45	26	1.115	12.957	11.518	1.125
R3-NQE	R3†	11	70	19	0.579	4.943	6.909	0.715
R4-NQE	R4‡	12	58	30	0.400	4.883	13.655	0.358

Navigational Queries. Figure 2 shows the comparison between system R1 and R4 across topics that were classed as navigational. There were 10 such queries identified manually, most of which were one or two terms in length. For example, Topic 7 “`samsung official site`” seems to clearly be navigational, and the number of highly relevant results is expected to be small. The term “`site`” is a very common term that is likely to pollute the results returned by any two systems, leading to poor effectiveness for named-page finding on this topic. Upon inspection of topics 3, 4, 92 and 98, the R1 system also exhibits performance that is less than desirable. On the other hand, system R4 results in improved effectiveness for these topics. One plausible hypothesis is that the n-gram fields may provide a surrogate for diversifying the results that is not captured by the sequential dependence model. We plan to explore this effect more carefully in future work.

Query Expansion. We now turn our attention to determining the magnitude of difference between the query expanded runs (which all submissions utilized), and their non-expanded counterparts. In order to do this, we performed a posthoc analysis of using the same system configurations without query expansion to see how much query expansion affected our overall results. Turning off query expansion for R3 (the full dependence model) consistently degraded performance as can be seen in Table 1. A similar trend can be observed when expansion is disabled in the n-gram fields system R4. This indicates that both of these systems consistently benefit from query expansion.

The performance of system R1 saw the most benefit when expansion was not used, while for system R2 effectiveness oscillates between the expanded and non-expanded configurations, favoring the original submission for ERR@5 and NDCG@5 and the non-expanded configuration for the same metrics (NDCG, ERR) at cutoffs 10 and 20.

It is interesting to observe in Table 3 that the comparisons showing a high number of ties is congruent with statistical significance. It could be argued that for systems R3-NQE and R4-NQE without expansion were already performing poorly, and enabling query expansion helped to counteract these deficient systems. However, R4-NQE is significantly improved by the use of query expansion in comparison, where there is a significantly larger total of NDCG@10 wins than R3-NQE. We deduce that the n-gram fields model is better than FDM at leveraging query expansion to additively improve retrieval effectiveness in this situation.

One could intuit that a higher number of ties between query expansion and the absence of it for a system is moot in that it improves upon a degenerate form of the retrieval model and suggests perhaps that more attention should be paid to the underlying model that query expansion is used with. On the other hand, the other two comparisons with fewer ties lead towards the assumption that there is some element of additivity that causes the query expanded systems to lose out in effectiveness more often than not.

Rank Fusion. Our final exploration involves the use of a recently developed parameterized rank-based fusion technique, Rank-Biased Centroid (RBC) by Bailey et al. [1], that discounts the re-ranking of documents in consensus with the use of an exponential distribution. Unlike the Borda model that assigns a linearly weighted value based on the rank position of documents, RBC is able to mute the aggregation of scores from documents occurring deep in the list, which is unlikely to yield any utility to the user issuing the query. Therefore, the RBC method is able to behave in a more risk-sensitive manner than the Borda model, such that there is less opportunity for non-relevant documents to rank highly in the fused result list. It is not clear how the shallow judgment pool and the size of the collection might affect the results of RBC, and therefore we performed a sweep of the persistence parameter ϕ across all combinations of our submitted runs. The result of fusing R1 and R4 was the most effective out of all combinations. Note that different values of ϕ appear to have little impact on the overall performance of the fusion. This is likely due to the shallow pool depth and possibly, additivity – it tends to be harder to improve the performance of high performing systems. The result shown in Table 1 is configured with $\phi = 0.8$. This combination yielded a marginal improvement in very early precision metrics – ERR and NDCG@5.

5. CONCLUSIONS

We are pleased with the overall results we were able to achieve with relatively simple system configurations in the first year of the WWW track at NTCIR. Term dependency models and RM3 query expansion continue to be very difficult baselines to beat in adhoc retrieval tasks. In the next iteration of the track, we hope to more thoroughly explore other state-of-the-art Learning-to-Rank techniques to see how they compare to the competitive unsupervised techniques we developed for this year's task.

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

References

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

[2] R.-C. Chen, J. S. Culpepper, T. T. Damessie, T. Jones, A. Mourad, K. Ong, F. Scholer, and E. Yulanti. RMIT at the TREC 2015 LiveQA Track. In *Proc. TREC*, 2015.

[3] R.-C. Chen, L. Gallagher, R. Blanco, and J. S. Culpepper. Efficient cost-aware cascade ranking in multi-stage retrieval. In *Proc. SIGIR*, pages 445–454, Aug. 2017.

[4] C. L. A. Clarke, J. S. Culpepper, and A. Moffat. Efficiency-effectiveness tradeoffs in two-stage retrieval mechanisms. *Inf. Retr.*, 19(4):351–377, 2016.

[5] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proc. CIKM*, pages 837–846, 2009.

[6] M. Crane, J. S. Culpepper, J. Lin, J. Mackenzie, and A. Trotman. A comparison of document-at-a-time and score-at-a-time evaluation. In *Proc. WSDM*, pages 201–210, Feb. 2017.

[7] J. S. Culpepper, M. Yasukawa, and F. Scholer. Language independent ranked retrieval with NeWT. In *Proc. Aust. Doc. Comp. Symp.*, pages 18–25, 2011.

[8] J. S. Culpepper, S. Mizzaro, M. Sanderson, and F. Scholer. TREC: Topic engineeRing ExerCise. In *Proc. SIGIR*, pages 1147–1150, 2014.

[9] J. S. Culpepper, C. L. A. Clarke, and J. Lin. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proc. Aust. Doc. Comp. Symp.*, pages 17–24, Dec. 2016.

[10] J. V. Dillon and K. Collins-Thompson. A unified optimization framework for robust pseudo-relevance feedback algorithms. In *Proc. CIKM*, pages 1069–1078, 2010.

[11] J. K. Jayasinghe, W. Webber, M. Sanderson, and J. S. Culpepper. Improving test collection pools with machine learning. In *Proc. Aust. Doc. Comp. Symp.*, pages 2–9, 2014.

[12] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. SIGIR*, pages 120–127, 2001.

[13] X. Lu, A. Moffat, and J. S. Culpepper. The effect of pooling and evaluation depth on IR metrics. *Inf. Retr.*, 19(4):416–445, 2016.

[14] X. Lu, A. Moffat, and J. S. Culpepper. Modeling relevance as a function of retrieval rank. In *Proc. AIRS*, pages 3–15, 2016.

[15] X. Lu, A. Moffat, and J. S. Culpepper. Can deep effectiveness metrics be evaluated using shallow judgment pools? In *Proc. SIGIR*, pages 35–44, Aug. 2017.

[16] C. Luo, T. Sakai, Y. Liu, Z. Dou, C. Xiong, and J. Xu. Overview of the NTCIR-13 We Want Web Task. In *Proc. NTCIR-13*, 2017.

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

[18] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. SIGIR*, pages 472–479, 2005.