Extending test collection pools without manual runs

Gaya K. Jayasinghe
RMIT University
Melbourne, Australia
gaya.jayasinghe@rmit.edu.au

William Webber
William Webber Consulting
Melbourne, Australia
william@williamwebber.com

Mark Sanderson
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

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

ABSTRACT

Information retrieval test collections traditionally use a combination of automatic and manual runs to create a pool of documents to be judged. The quality of the final judgments produced for a collection is a product of the variety across each of the runs submitted and the pool depth. In this work, we explore fully automated approaches to generating a pool. By combining a simple voting approach with machine learning from documents retrieved by automatic runs, we are able to identify a large portion of relevant documents that would normally only be found through manual runs. Our initial results are promising and can be extended in future studies to help test collection curators ensure proper judgment coverage is maintained across complete document collections.

Categories and Subject Descriptors

H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval—*clustering, retrieval models, search & selection process*

General Terms

Information retrieval, Evaluation, Test collection construction

1. INTRODUCTION

Successful evaluation and reproducibility of experiments in information retrieval (IR) depends on building reusable test collections composed of documents, topics, and relevance judgments. Ideally every document in a collection would be assessed against each topic, but this approach does not scale. So judgments are normally produced for a sample of the corpus, known as a *pool*, all other documents are assumed to be not relevant. This sample needs to be representative of the entire collection and robust enough to evaluate entirely new search algorithms. The genesis of pooling dates back to the 1970s [12].

To produce relevance judgments, the organizers of TREC, CLEF, NTCIR, and other such conferences invite researchers to submit the top-i documents retrieved for a set of topics from a specified corpus [14, 15] (typically $i = 1,000$). The sets of documents are known as *automatic runs*. Across the runs, the top-j ranked documents for each topic are gathered for relevance assessment (typically j is set to 50 or 100). Such a practice seems to consistently identify most of the relevant documents, but provides no guarantee on the judgment coverage for documents retrieved by new IR approaches [4, 9, 16]. Test collections tend to have a bias towards the systems contributing to the pool, and may not reliably evaluate novel IR systems that retrieve unjudged but relevant documents.

In an attempt to "future proof" test collections, the organizers of the evaluation conferences commonly encourage submissions of *manual runs*, where humans can reformulate queries and/or merge results from multiple queries [1] before a final set of top-i documents is submitted. Such runs are generally highly effective and contribute many unique relevant documents to the judgment pool. However, manual runs are not always available when building a collection, so in this short paper we ask:

Research question: Can we construct reliable IR test collections using only automatic retrieval runs?

Our contribution: We describe a methodology that can be used to construct reusable test collections in the absence of manual retrieval runs. We evaluate a simple voting approach combined with machine learning to show that we can achieve collection coverage similar to pooling generated with manual runs.

2. BACKGROUND

Efficiently building test collections for evaluation of IR systems is a well-studied problem [10]. Early research concentrated on more efficient ways for assessors to scan pools, with the objective of judging more documents with a given budget or identifying a sufficient number of relevant documents as quickly as possible. Zobel [16] showed that the number of relevant documents in a collection varies from topic to topic. He suggested that assessors should focus their effort on judging topics with more relevant documents. For each topic, the number of relevant documents found so far were used to estimate the expected ratio of relevant documents in the remaining unjudged block. Each topic was assessed until relevant documents were depleted beyond an economically viable limit to assess the block.

The idea of focusing assessor effort on the most fruitful sources of relevant documents was also applied to IR systems that contribute to a pool. Just as some topics have more relevant documents than others, some systems retrieve more relevant documents than others. Using this insight, Cormack et al. [5] described a move-to-front pooling approach which ensured that documents from the IR systems producing the most relevant documents were moved to the

SIGIR '14, July 06 – 11, 2014, Gold Coast, QLD, Australia.
http://dx.doi.org/10.1145/2600428.2609473.

front of the queue of documents to be judged. Cormack et al. [5] went on to show that combining these two insights enabled more assessing effort to be spent on the most effective systems and on the most effective topics.

Moffat et al. [8] proposed an alternative document ordering for assessment based on the likelihood of a document being relevant. This method assumed that documents retrieved higher in the ranking by a retrieval run had a higher chance of being relevant. Each document retrieved was scored by the retrieval run using a function that exponentially decayed along with the current rank position. The documents were ordered based on the highest score they received from any retrieval run, and were fed into the retrieval pool until a fixed judging capacity of the pool was reached.

An alternative method, which used no pooling of runs, *Interactive Search and Judging (ISJ)* required assessors to manually find relevant documents by searching while judging at the same time [5]. The method identified a similar number of relevant documents to the traditional pooling approach with 75% less work than standard TREC pooling. However, the approach was not adopted by TREC as the assessors were found to be better at judging documents than as query reformulators [11]. Soboroff and Robertson [11] went on to argue that relevance feedback could be used to replace manual reformulation of queries in ISJ. After assessing the initial pool of documents, relevance feedback was applied to the relevant documents in order to obtain a new set of retrieval results which were iteratively judged and re-ranked.

The papers represent the few efforts to construct test collections with only automatic runs. Next, we investigate a different approach to building a diverse and robust pool from automatic runs.

3. METHODS AND DATA

For our investigation, we needed a test collection with both automatic and manual runs. We used TREC GOV2, with topics (801 – 850) [2]. In total, 80 runs, composed of the top-1,000 documents, were submitted. Up to three runs from a group were fully assessed down to rank 50, the collection pool depth. In total, 49% of runs were fully assessed. The ratio of automatic to manual runs was 2.5 : 1. Out of the total relevance judgments made, 27% of the documents were uniquely pooled by manual retrieval runs of which 18% were relevant. The following subsets of the GOV2 document collection are of note (Figure 1).

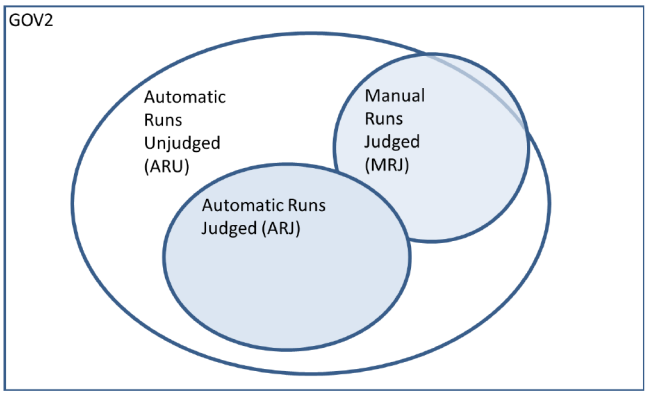

Figure 1: The document sets of the TREC GOV2 collection

- ARJ – Documents retrieved by Automatic Runs and Judged as they were ranked above rank 50. In these experiments, this is the set of existing judgments we're seeking to extend.
- MRJ – Documents retrieved by the Manual Runs and Judged minus any documents in ARJ. In these experiments, MRJ represents the additional relevance assessments we wish to build methods to find.
- ARU – Documents retrieved by Automatic Runs, but Unjudged as they were ranked below 50, the GOV2 pool depth.

Our method takes two steps: a *first pool* is created from set ARJ. Then some part of the documents in the collection but that are not in ARJ are ranked by one of the methods described below and the top-k are drawn into a *second pool*, the relevance of which is assessed using the unique judged documents in MRJ. Here, we describe three ranking approaches: a naïve baseline voting method – *Borda count*; a machine learning classifier (ML); and a combination of Borda and ML. Here, the automatic runs are being used as a form of training data and the manual runs as test data.

3.1 Borda count

Borda count is a widely used voting method for consolidating multiple preference lists into a single list. In our method, each automatic run produces a list of top-i ranked documents ($i = 1,000$). In each list, the highest ranked document receives i votes, the second highest receives $i - 1$ votes, and so forth. In the consolidated list documents are ranked in descending order by the total number of votes received across the runs. Ties are broken randomly (similar to the approach taken by Moffat et al. [8]).

The only documents in ARU are used for Borda count. The top-k from the consolidated vote are drawn into the second pool.

3.2 Machine learning (ML)

A classifier is trained on the documents in the first pool, represented in a vector space using Krovetz stemming [7] and a TF×IDF weighting. A similar feature space was used by Büttcher et al. [3]. After training a linear SVM classifier [6] for each topic, documents are scored and ranked by the classifier.

The classifier computes a weighted vector ω corresponding to a hyperplane that maximally separates relevant from non-relevant. The decision function for the classifier is $sign(f(x))$, where $f(x) = \omega^T.x$. Here $x_m \in R^{\sigma^n}$, where $m = 1, ..., \ell$ represents the m·th document in ℓ documents, and y_m is the corresponding class label. The relevance score for document m is taken from $f(x_m)$. The documents not in ARJ are sorted by their score and the top-k are drawn into the second pool. The method is similar in spirit to the approach proposed by Soboroff and Robertson [11] but requires no iterations.

3.3 Combined

Finally, we combine the first two methods. The Borda count is used as a filter to generate a subset of the most promising documents in ARU for a given topic. The subset is then re-ranked using the classifier, and the top-k form the second pool.

4. EXPERIMENTAL METHODOLOGY

Recalling the motivation for our work, we wish to locate relevant documents that would normally be found only by including manual runs in the pooling process. Documents uniquely found in the manual retrieval runs (MRJ) are our surrogates of relevance assessments for the documents placed in the second pool. Because the methods rank the documents in the second pool, we can measure the quality of the pool using a retrieval effectiveness metric such as mean average precision (MAP) and precision at depth d (P@d).

We also use Kendall's τ to measure pairwise inversions between two rankings of runs, the first using full TREC relevance assessments and the second using relevance assessments generated from the union of the first and second pools formed by each of our methods. Using a convention from Voorhees [13], if the Kendall's τ correlation is ≥ 0.9, the rankings are considered equivalent.

Metric	Borda count	ML	Combined
MAP	0.0778	0.0268	0.1507*
P@10	0.1306	0.0531	0.1714
P@20	0.1122	0.0378	0.1500
P@30	0.1020	0.0361	0.1367
P@100	0.0743	0.0167	0.0916

Table 1: Effectiveness on finding relevant documents in MRJ. A *: significant improvement ($p < 0.01$) compared to Borda count.

Depth (k)	Borda count	ML	Combined
50	15.22	4.52	19.00*
100	24.19	5.45	29.83*
150	29.30	6.71	37.28*
171 ✠	31.36	7.11	39.53*
200	33.75	7.97	42.33*

Table 2: Percentage of MRJ documents found in top (k) of the proposed rankings. ✠ implies a similar assessment effort to traditional pooling method. A *: significant improvement ($p < 0.05$) compared to Borda count.

5. RESULTS

The analysis is presented in Table 1. The combined method is significantly better than the other two when evaluated with MAP. The same trend is observed when measuring using precision, but none of the differences are significant. Using only the ML method produces worse results than either Borda count or combined.

Note that the relatively low reported effectiveness numbers in Table 1 are largely a byproduct of evaluating using only the unique relevant documents in MRJ and not the entire second pool. We cannot make any claims about new documents retrieved by the ML method since a large portion of retrieved documents using this method are not judged, compared to other two approaches. In fact, $9,817$ of the top-200 documents returned across all 50 topics using only ML (98.17%) are currently unjudged. Therefore, we have to assume that these documents are not relevant until all of the documents returned are judged. In future work, we hope to investigate the full impact unjudged documents have on our classifier method in more detail.

In Table 2 we measure the proportion of documents that were found to be relevant in the second pool. Again a similar trend of differences are seen, but with significant improvements across all measurements up to $k = 200$ for the combined method.

5.1 Discussion

As indicated in Figure 1, the majority of documents uniquely judged in the manual runs (MRJ) are also retrieved by the automatic runs (ARU+ARJ). However, few appear in the first pool as they (i.e. ARJ) are not ranked highly enough to be judged. In fact, 88% of the documents judged as relevant that are uniquely pooled by manual runs could be found in the first pool, if a pool depth of $1,000$ was used.

If there were no manual runs in a test collection (i.e. no MRJ), the effectiveness of IR systems producing results similar to such runs would be underestimated and any improvements would go unnoticed. It would appear that manual retrieval runs still play a critical role in improving the re-usability of test collections.

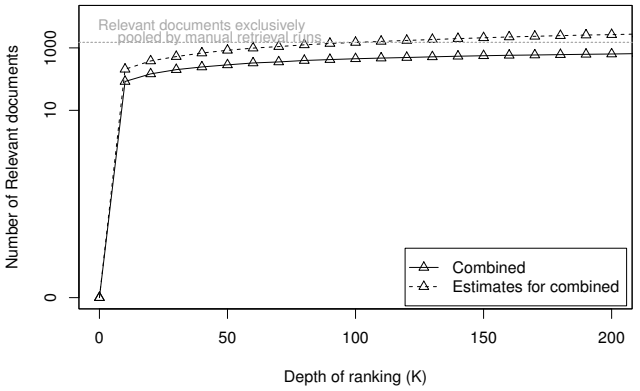

Figure 2: The number of MRJ documents, and estimated number of relevant documents in the top-k of the combined ranked list on TREC GOV2 dataset and TREC topics 801 – 850.

Metric	Borda count	ML	Combined
MAP	0.3415	0.4872*	0.5049*
P@10	0.3571	0.5082*	0.5694*
P@20	0.3551	0.4684*	0.4959*
P@30	0.3401	0.4299*	0.4497*
P@100	0.2337	0.2624*	0.2555$^\bullet$

Table 3: Just considering the documents in MRJ, how effective are ranking algorithms on retrieving relevant documents? Significant improvements ($p < 0.01$ and $p < 0.05$) compared to Borda count are denoted with a * and $^\bullet$.

Judging the ranked lists of the combined method up to a depth k identifies a subset of the relevant documents uniquely pooled by manual retrieval runs. However, we still know little about the large number of unjudged documents in the ranked lists produced by the combined method. If we assume the proportion of relevant documents among unjudged documents in these ranked lists is the same as the proportion found among judged documents in the same ranked list up to the same depth, we can estimate the total number of relevant documents that would have been found in the same depth of the ranking. Figure 2 illustrates the estimated number of relevant documents, along with the number of known relevant documents found.

Missing judgments for a large portion of the ranked lists from the proposed methods is one potential reason for the low retrieval effectiveness of those methods. Therefore, we calculate retrieval effectiveness on the intersection of the second pool with MRJ, Table 3. (Note, the first pool and the ranking functions remains the same.) The ML method now re-ranks a subset of unique documents top-j ranked by manual runs. The ranking produced by ML show significant improvements for all considered evaluation metrics compared to Borda count. The combined method achieves a better effectiveness than ML for all evaluation metrics considered, except p@100. This is due to ranking only the subset of documents top ranked by the Borda count. Re-ranking a carefully retrieved

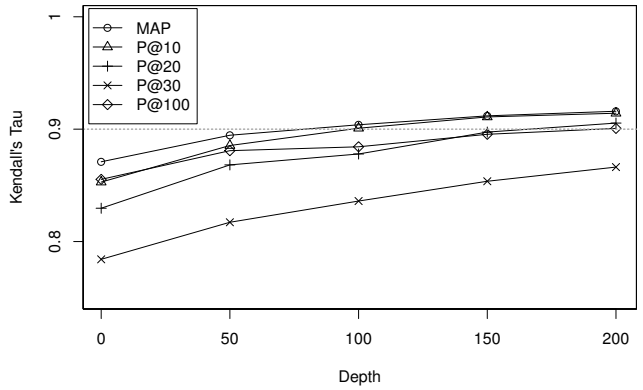

Figure 3: Kendall's τ correlation of IR system rankings for varying depths of assessing documents using combined method.

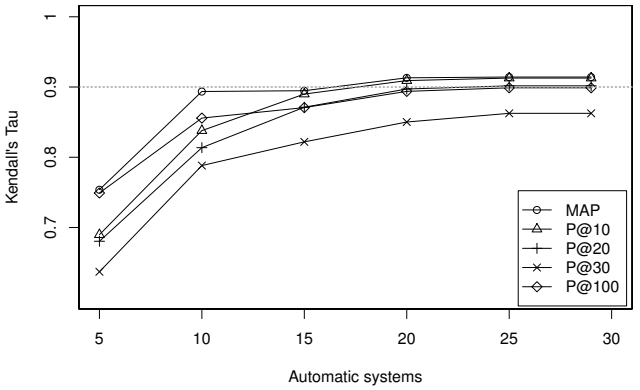

Figure 4: Kendall's τ correlation of IR system rankings with varying number of automatic systems in the pool. Automatic systems are added in the order least contributing system to most, and the ranking produced by the combined method is processed to a depth of 200.

subset of documents for topics with ML is an effective approach to locate new documents to be pooled and judged.

Whenever a new approach for pool composition is proposed, we would like to be able to quantify how well the approach ranks IR systems compared to the original method. A Kendall's τ ranking correlation for varying depths of assessing documents with the proposed approach for various evaluation metric are shown in Figure 3. Here, we consider all 80 submitted runs rather than only the subset originally used for pooling. Manual retrieval runs are viewed as novel approaches to retrieval. The Kendall's τ correlation for MAP is above 0.9 beyond a depth of 100. A budget similar to original assessment permits processing up to a depth of 171 documents, which demonstrates the validity of the proposed approach in the absence of manual retrieval runs.

Another question of interest is how small the automatic runs pool can be when there are no manual runs. In Figure 4 we introduce runs incrementally in order starting with the run contributing the fewest relevant documents. When 20 or more automatic retrieval runs are pooled the Kendall τ correlation for MAP exceeds 0.9.

6. CONCLUSION

In this paper, we present a methodology for building reusable evaluation pools in the absence of manual retrieval runs. Our approach can discover many relevant documents that were previously only found by manual retrieval runs. The approach demonstrates the potential of finding relevant documents that are not currently possible using current pooling approaches. However, the true efficacy of our approach cannot be properly assessed until all of the newly retrieved documents are judged. We plan to investigate this in future work. Nonetheless, our initial results are promising as we are already able to achieve a similar IR system ranking to previous approaches which depended heavily on manual runs to add the necessary diversity to the assessment pool.

Acknowledgments

This work was supported in part by the Australian Research Council (DP130104007). Dr. Culpepper is the recipient of an ARC DECRA Research Fellowship (DE140100275).

References

[1] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling for large collections. *Information Retrieval*, 10(6): 491–508, 2007.

[2] S. Büttcher, C. L. A. Clarke, and I. Soboroff. The TREC 2006 terabyte track. In *TREC-2006*, volume 6, page 39, 2006.

[3] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *SIGIR*, pages 63–70, 2007.

[4] B. Carterette, E. Gabrilovich, V. Josifovski, and D. Metzler. Measuring the reusability of test collections. In *WSDM*, pages 231–240, 2010.

[5] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *SIGIR*, pages 282–289, 1998.

[6] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, June 2008.

[7] R. Krovetz. Viewing morphology as an inference process. In *SIGIR*, pages 191–202, Pittsburgh, Pennsylvania, USA, 1993.

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

[9] T. Sakai. The unreusability of diversified search test collections. In *EVIA*, June 2013.

[10] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4 (4):247–375, 2010.

[11] I. Soboroff and S. Robertson. Building a filtering test collection for TREC 2002. In *SIGIR*, pages 243–250, 2003.

[12] K. Spärck Jones and C. J. Van Rijsbergen. Report on the need for and provision of an "ideal" information retrieval test collection. Technical report, British Library Research and Development Report 5266, 1975.

[13] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management*, 36(5):697–716, 2000.

[14] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 355–370. Springer, 2002.

[15] E. M. Voorhees and D. K. Harman. *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge, 2005.

[16] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *SIGIR*, pages 307–314, 1998.