Improving test collection pools with machine learning

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

IR experiments typically use test collections for evaluation. Such test collections are formed by judging a pool of documents retrieved by a combination of automatic and manual runs for each topic. The proportion of relevant documents found for each topic depends on the diversity across each of the runs submitted and the depth to which runs are assessed (pool depth). Manual runs are commonly believed to reduce bias in test collections when evaluating new IR systems.

In this work, we explore alternative approaches to improving test collection reliability. Using fully automated approaches, we are able to recognise a large portion of relevant documents that would normally only be found through manual runs. Our approach combines simple fusion methods with machine learning. The approach demonstrates the potential to find many more relevant documents than are found using traditional pooling approaches. Our initial results are promising and can be extended in future studies to help test collection curators ensure proper judgment coverage is maintained across the entire document collection.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance Evaluation*

General Terms

Experimentation, Measurement, Performance

Keywords

IR evaluation, IR test collections, relevance judgments, uncertainty, test collection bias

1. INTRODUCTION

Most IR experiments are evaluated on standard test collections composed of a corpus of documents, a sampled set of topics, and relevance judgments stating which documents from the corpus are relevant for each sampled topic. Ideally every document in the corpus is judged for each topic. However, this approach is not cost effective. So only a carefully selected subset of the corpus, known as a pool, is judged for each topic. When evaluating IR experiments, any remaining unjudged documents are considered non-relevant. Therefore, the pools should contain as many relevant documents as possible to be reliable. The origins of pooling run back to the early 1970s [27, 26].

IR systems which are explicitly designed to retrieve relevant documents provide an obvious way to initiate pooling. Therefore, the organizers of the TREC, CLEF, NTCIR, and other similar conferences invite researchers to submit ranked retrieval results from state-of-the-art systems to a depth of Z. These runs are commonly known as the *automatic runs* for a set of topics and corpus. A subset of each of these runs are then gathered for assessment. The exact cutoff z used for each run is referred to as the *pool depth*. This strategy tends to find most relevant documents for each topic, but provides no guarantees particularly when entirely new systems are evaluated [33, 8, 24, 23].

These test collections tend to favour IR systems similar to the ones used to create the original pools over new systems which may retrieve a higher proportion of unjudged documents. In order to minimise this bias, test collection curators encourage *manual runs* where queries are reformulated and results are merged before constructing a ranked list of Z documents [31, 3]. Manual runs tend to add relevant documents which were not found by automatic runs alone [25]. However, manual runs are not always obtainable when forming test collections and the number of manual runs continues to decrease in many of the TREC adhoc query tracks.

So, we investigate alternative approaches to create low bias test collections using only automatic runs in this paper. We address the following research questions:

1. How can IR test collections be future proofed when only automatic retrieval runs are available?
2. How deep should the pool depth be in order for the results to be reliable when using only automatic runs?

Our contribution: We propose an approach that combines a simple voting method with machine learning to find relevant documents that would otherwise only be found using manual runs in Section 3.1. The proposed approach can be used to construct a low bias reusable test collection without manual runs. We evaluate the approach and show that the pool coverage is similar to the pools generated with manual runs in Section 3.4. We then illustrate that the approach is effective in finding relevant documents that are not found using only automatic runs even when the pool depth in the new approaches is shallow.

ADCS '14, November 27–28 2014, Melbourne, Victoria, Australia

ACM 978-1-4503-3000-8/14/11 ...$15.00.
http://dx.doi.org/10.1145/2682862.2682864.

In this paper, we extend the prior work of Jayasinghe et al. [16]. We present and analyse alternative fusion-based solutions using two common TREC datasets. In addition, we manually judge a large pool of previously unjudged documents for the topics in order to conclusively compare the effectiveness of the new methods. Finally, we analyse the viability of the methods at various pool depths.

2. RELATED WORK

The suitability of a test collection to evaluate novel IR systems in the presence of incomplete judgments has been the subject of several other research studies. To assess the impact of bias due to incomplete judgments, Zobel [33] used *leave one run out* simulations. Unique relevant documents contributed by a run would not have been judged for relevance and assumed not relevant if the run had not been pooled. The effectiveness for every pooled run assessed using relevance judgments produced with and without the run in the pool is used to quantify the impact of incomplete judgments on evaluation. On early test collections, Zobel [33] found no conclusive evidence against reuse.

By assuming multiple submissions from the same group tend to retrieve similar documents, Voorhees [29] adapted the *leave one run out* method to use only a single run from each group, and referred to the method as *leave one group out*. In this formulation a group of runs are assessed using relevance judgments produced with and without each group of runs being in the pool. As a result of this work, the leave one group out methodology has been adopted as the de facto standard when evaluating test collections for reusability. The leave one group out approach also did not produce any evidence against reuse in early test collections.

Evidence against reusability gradually started to appear as the size of the test collections continued to increase. Buckley et al. [3] identified one such manual run using the GOV2 collection which scored lower than expected if the run had not been used in the original pooling process. Further investigation revealed that large test collections tend to contain many more potentially relevant documents (documents which contain one or more of the original query terms) than a practically assessable pool can hold. Hence, relevant documents containing a subset of the original query terms are often left out of the initial pool.

Further evidence was gathered in another experiment where runs were evaluated by holding out all of the manual runs from the pool [5]. The new system ranking was found to be different to the ranking produced with the original relevance judgments. Whether the test collection bias has any impact on ranking new systems remains unknown since a comprehensive comparison is not possible without additional judgments. As such, IR researchers now seek efficient and effective ways to locate as many potentially relevant documents as possible in new test collections.

The number of relevant documents available in a corpus varies from topic to topic. Zobel [33] used this insight to show that more documents should be judged for topics having more relevant documents. However, finding the exact cutoff depth for a given topic can be problematic with no prior knowledge of how many relevant documents there are for the topic. To circumvent this problem, Zobel represented the number of relevant documents found as a function of pool depth. Before a pool with an incremented pool depth is assessed, the number of relevant documents expected in the expanded pool is estimated based on the proportion of relevant documents found in the previous pool depths. Assessment is terminated when an acceptable threshold is reached.

Just as a topic can contain more relevant documents than another, the number of relevant documents in pooled runs can also vary widely. Hence, Cormack et al. [13] place the runs yielding more relevant documents in the most recent block assessed in the front of a queue for assessment. Furthermore Cormack et al. go on to show that relevant documents can efficiently be found by spending assessor effort on the most effective systems and on the best topics.

Alternatively, ranked retrieval results from multiple IR systems can be merged to derive a fused ranking of documents. Each IR system produces a list of top-Z documents ranked by preference. Fusion schemes score each retrieved document according to a specific criteria. The documents are then ranked in descending order by the fusion score. Ties are broken randomly. Documents in the new ranking are assessed until a fixed judging capacity of the pool is reached. Popular fusion schemes are: Borda Count (BC) [1], CombSum, CombMNZ (CMNZ), CombANZ [15], and Static Judgment Orderings [19]. The criteria for scoring documents with each fusion method is presented next.

Borda Count scores documents using a simple voting method. For each list the highest ranked document receives Z votes, the second highest document receives $Z - 1$ votes, and so forth. For fused ranking each document is scored with the total number of votes received from all systems. CombSum is also a voting method. Each list votes for each document retrieved in the list with the normalised relevance score given by the IR system for the document. Again, the total votes received from all systems for each document defines the score for ranking. CombMNZ defines the fusion score as the CombSum score multiplied by the number of systems that retrieved the document. The fusion scores as defined by the CombSum are averaged over the number of systems that retrieved the document for CombANZ. Static Judgment Ordering computes the fusion score as the sum of $(1 - \rho) \cdot \rho^{rank(d)-1}$ from all systems, where $rank(d)$ is the rank of document d in a ranked list. An appropriate value for ρ is empirically determined to maximise effectiveness of the fused ranking. Depending on the dataset, different fusion approaches have shown to be effective [1]. However, CombMNZ is the most widely used.

An online learning algorithm for selecting the next document to judge was proposed by Aslam et al. [2]. Here the next document to be judged is chosen as a weighted expert opinion with each IR system viewed as an expert. Each system's *expertness* score increases with each predicted relevant document and decreases when the predicted document is non-relevant. Each system recommends documents based on the rank of the document weighted by expertness. The next document to judge is the one with the highest aggregated recommendation across all systems. The method found relevant documents at an approximate rate of 50% higher than conventional pooling.

An alternative approach required assessors to search for relevant documents with reformulated queries and do judgments at the same time [13]. A similar number of relevant documents are found using this method but with 75% less judging effort than traditional pooling approaches. This method is similar in spirit to using only manual runs. However, the assessors were found to be better at judging documents than as query reformulators. Therefore, the approach is not used by test collection curators [25].

Soboroff and Robertson [25] argued relevance feedback can be used instead of manual query reformulation. After assessing the initial pool of documents produced with an automatic run, relevance feedback is applied using seven ranking strategies including a few machine learning rankers in order to obtain a newer set of retrieval results. The retrieval results are fused using CombMNZ, and the top ranked results in the fused list are assessed. Relevance feedback is applied iteratively to find additional documents for judgment. For the machine learning rankers SVM and Naïve Bayes

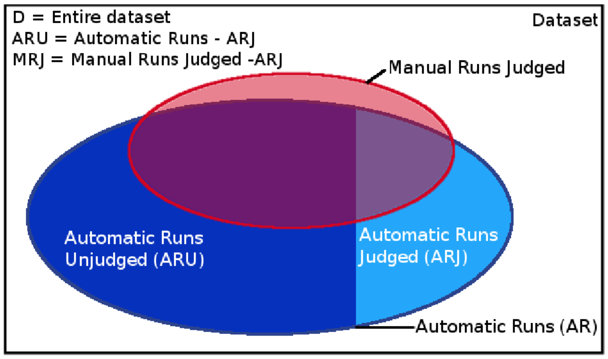
D = Entire dataset
ARU = Automatic Runs - ARJ
MRJ = Manual Runs Judged -ARJ
Dataset
Manual Runs Judged
Automatic Runs Unjudged (ARU)
Automatic Runs Judged (ARJ)
Automatic Runs (AR)

Figure 1: The subsets of the document corpus corresponding to formation of a pool for judging documents.

classifiers are used with a BOW feature space. This is the first known attempt by TREC to construct test collections (for the filtering track) using only automatic runs. Additional approaches to constructing low biased test collections without manual runs are described in the next section.

3. CREATING TEST COLLECTIONS WITHOUT MANUAL RUNS

None of the solutions reviewed in Section 2 for forming test collections is entirely satisfying. Effective IR systems may emerge that retrieve new relevant documents [9]. To minimise the chance of unfairly evaluating such systems, proper judgment coverage must be maintained. The traditional approach for pooling relies on manual runs for diversity in coverage [25]. In the next section, new methods for future proofing test collections in the absence of manual runs are presented.

3.1 New Approaches

First, we will define a consistent document subset terminology to be used. Documents in a collection and the corresponding subsets are shown in Figure 1. The subset *Automatic runs judged* (ARJ) is the pool of top-z documents from the automatic runs. The subset *Automatic runs unjudged* (ARU) is the pool of top-Z documents minus ARJ from the automatic runs. Similarly the subset *Manual runs judged* (MRJ) is the top-z documents of the manual runs less documents in ARJ. Pooling strategies to locate relevant documents that are in ARU are of primary importance in this work.

The approaches work in two steps: a first pool is drawn with documents from set ARJ and evaluated. Then a set of documents that are not in the first pool are drawn from the collection. The documents are ranked by one of the following approaches and the top-κ are drawn into a second pool which is also evaluated.

In this section the following approaches are analysed for ranking documents to form the second pool: simple fusion methods – Borda Count and CombMNZ – naïve baselines; a machine learning classifier (ML) – similar in spirit to the approach proposed by Soboroff and Robertson [25] – another baseline; and a combination of fusion methods and machine learning.

Fusion Methods: Recall that fusion methods can be used to combine multiple ranked lists into a consolidated ranked list. Here a consolidated ranked list is derived by fusing the automatic runs. Remember only ARU documents qualify for inclusion in the second pool. Therefore, ARJ documents are removed from the fused rankings. The top-κ from the final ranking are drawn into the second pool.

Machine Learning Approach (ML)**:** An SVM classifier is trained on each topic using the documents in the first pool. Krovetz [18] stemmed documents are represented in a vector space where each unique term is a dimension using a TF×IDF weighting. A similar classification method was used by Büttcher et al. [5] to predict the relevance of unjudged documents. However, rather than classifying documents as relevant or not relevant, here documents are scored and ranked based on the likelihood of being relevant. The linear SVM classifier [14] computes a weight vector ω corresponding to the maximum-margin hyperplane that maximally separates relevant (labeled 1) and non-relevant (labeled 0) documents of the set ARJ. The decision function $h(X)$ for the classifier is given by $sign(\omega^{T} \cdot X)$, where X contains document feature vectors. The relevance score for the $i{\cdot}th$ document X_i is taken from $\omega^{T} \cdot X_i$. All documents in the corpus except for those in ARJ are ranked and the second pool is formed using the top-κ.

This method bears some similarity to the use of relevance feedback in the approach proposed by Soboroff and Robertson [25] but differs in the following aspects. In Soboroff and Robertson [25], the top ranked documents from an automatic run are judged and used as relevance feedback to rank documents using seven ranking strategies, including a few machine learning approaches. The ranked results are fused, top ranked documents are assessed, and relevance feedback is iteratively applied. Here, relevance feedback is applied to relevant documents found in the automatic runs, and uses only one machine learning ranker. Multiple machine learning rankers are not used to avoid the high computational overhead which can be substantial in large datasets. For the same reason, repeated relevance feedback is not used.

Combined Methods: The third category of approaches combine the first two methods. Here, a fusion method is used as a filter to generate a subset of the most promising documents in the collection for a given topic. The subset is then reranked using a machine learning classifier, and the top-κ form the second pool. In contrast to the method proposed by Soboroff and Robertson [25], here fusion is used to select a subset of documents to rank using the machine learning approach, rather than to form a ranking for manual assessment. The combined approaches using Borda Count and CombMNZ are respectively referred to as Combined - Borda Count (CBC) and Combined - CombMNZ (CCMNZ) henceforth.

3.2 Datasets

The TREC-8 and the TREC GOV2 datasets are used with TREC topics 401–450 [30] and 801–850 [4] respectively. Although 129 and 80 retrieval runs were submitted, not all of the runs from each group were pooled for evaluation by TREC. Pooled retrieval runs were scanned to a pool depth of 100 for TREC-8 and 50 for TREC GOV2 datasets for each run used in the pool to generate a set of documents for assessment for the corresponding topics.

3.3 Experimental Setup

For recent IR datasets pool depths lower than 100 have been used [4, 10, 11, 12]. Therefore, for both datasets a pool depth of 50 is considered in the experiments ($z = 50$). Only retrieval runs with complete relevance assessments above the pool depth are considered, and relevance assessments pooled by the selected retrieval runs are used. These filtered datasets are referred to by their original name – TREC-8 and GOV2. It was found that 53.5% and 48.8% of the runs fulfilled the above criteria with a 7.6 : 1

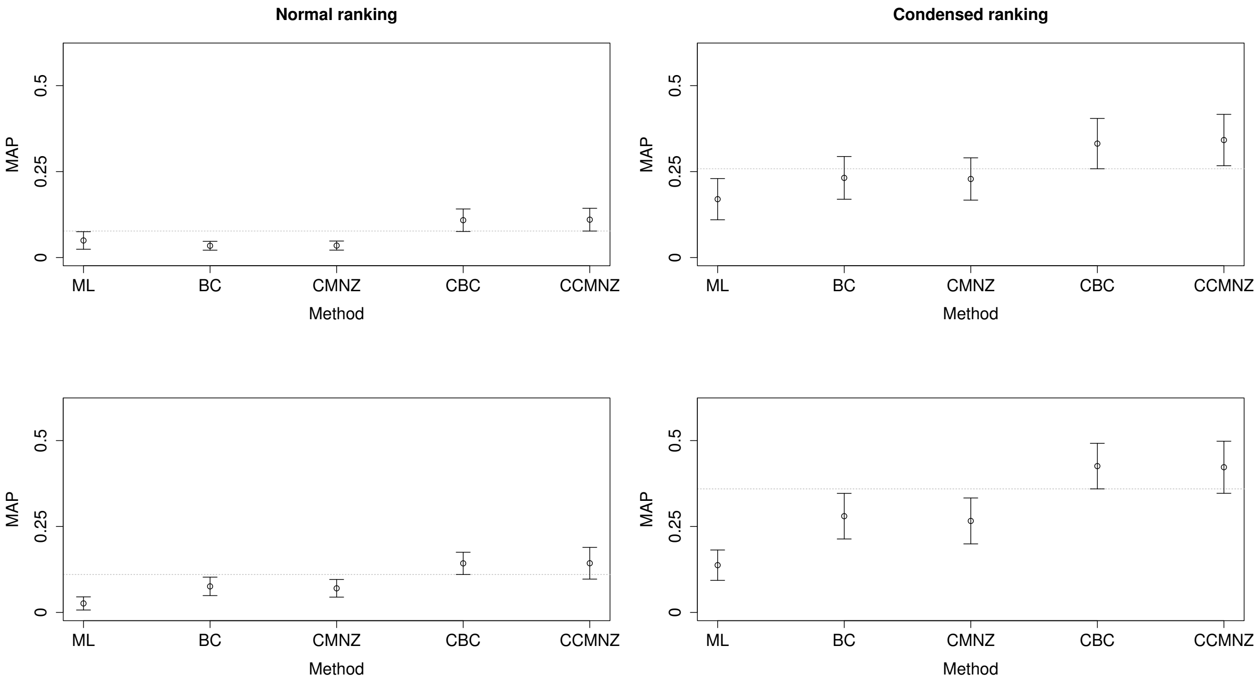

Figure 2: Retrieval effectiveness and 95% confidence interval on finding relevant documents in MRJ with traditional evaluation (left) and using a condensed ranking (right) for TREC topics 401-450 on the TREC-8 dataset (top) and TREC topics 801-850 on the TREC GOV2 dataset (bottom).

TREC-8 dataset

Depth (κ)	ML	BC	CMNZ	CBC	CCMNZ
50	10.40	13.02	12.63	$23.61^{\bullet\dagger}$	$24.90^{\bullet\ddagger}$
93 ✠	12.64	20.68	20.12	$30.26^{\bullet}$	$30.73^{\bullet\dagger}$
100	12.74	21.24	20.61	$31.11^{\bullet}$	$31.13^{\bullet\dagger}$
150	13.87	27.10	26.83	$35.61^{\bullet}$	$36.65^{\bullet}$
200	15.04	31.05	30.77	$38.86^{\bullet}$	$39.52^{\bullet}$

TREC GOV2 dataset

Depth (κ)	ML	BC	CMNZ	CBC	CCMNZ
50	5.41	15.04	13.17	$29.34^{\bullet\ddagger}$	$28.18^{\bullet\ddagger}$
100	6.45	27.49	24.14	$41.97^{\bullet\ddagger}$	$38.96^{\bullet\ddagger}$
150	7.58	32.71	31.13	$49.29^{\bullet\ddagger}$	$44.89^{\bullet\dagger}$
171 ✠	7.83	34.86	32.67	$50.14^{\bullet\ddagger}$	$46.01^{\bullet\dagger}$
200	8.48	37.28	34.66	$51.35^{\bullet\dagger}$	$47.02^{\bullet\dagger}$

Table 1: Percentage of relevant MRJ documents found per topic in the top-(κ) of the proposed rankings for TREC topics 401-450 on the TREC-8 dataset (top) and TREC topics 801-850 on the TREC GOV2 dataset (bottom). ✠ implies a similar assessment effort to a traditional pooling method. Combined approaches are tested for significance. A $^{\bullet}$ implies a significant improvement at $p < 0.01$ compared to ML. Similarly, a † and ‡ implies a significant improvement at $p < 0.05$ and $p < 0.01$ compared to the base fusion method.

and $2.5 : 1$ ratio of automatic to manual runs for the respective datasets. From the total relevance judgments 27.0% and 11.2% of the documents were only pooled by manual runs of which 19.6% and 17.6% were found to be relevant respectively.

The aim here is to locate relevant documents that would normally be found only by including manual runs in the pooling process since these runs are a reasonable surrogate for previously unseen novel systems. Documents in MRJ can be used to assess documents placed in the second pool by the automatic approaches. Because the documents in the second pool are ranked by each of the approaches, the quality of the pool can be measured using a retrieval effectiveness metric such as MAP.

The Kendall's τ correlation [17] and the AP correlation [32] are used to compute the agreement between the two mean rankings of runs evaluated with full relevance assessments and the relevance assessments generated from the union of the first and the second pools formed by the approaches. Using a convention from Voorhees [28], if the Kendall's τ correlation is greater than 0.9, the rankings are considered equivalent. The same convention is later followed by Carterette et al. [7], and Carterette and Soboroff [6].

3.4 Results

An evaluation with MAP using the original and a condensed ranking where unjudged documents are removed from the ranking for evaluation for both datasets are presented in Figure 2 left and right respectively. The combined approaches perform better than all other approaches for traditional evaluation. For instance, CBC is significantly better than ML, its counterpart fusion method BC, and CMNZ. Note that the relatively low reported effectiveness in Figure 2 for traditional evaluation is largely a byproduct of evaluating only unique relevant documents pooled by the manual runs and not the entire pool.

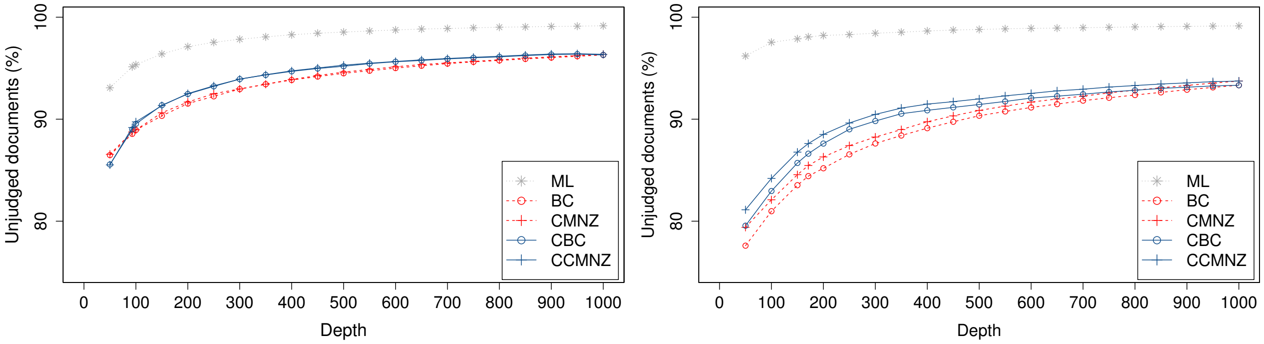

Figure 3: Percentage of unjudged documents found in the top-(κ) of the proposed rankings for TREC topics 401-450 on the TREC-8 dataset (left) and TREC topics 801-850 on the TREC GOV2 dataset (right).

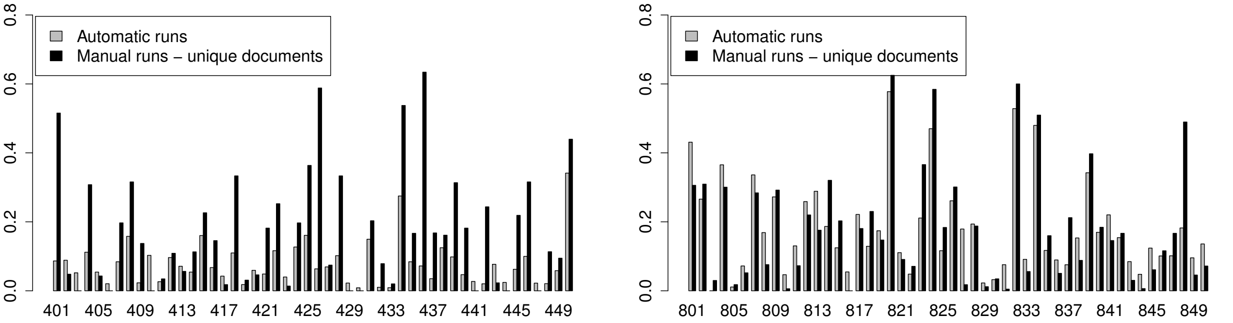

Figure 4: Proportion of documents added by automatic runs and exclusively pooled by manual runs that are relevant out of total documents pooled by the corresponding type of runs for each topic in topics 401-450 on the TREC-8 dataset (left) and TREC topics 801-850 on the TREC GOV2 dataset (right).

However, no claims can be made about the real effectiveness of these approaches since a large portion of retrieved documents using these methods remain unjudged. This is illustrated in Figure 3. The ML method retrieves a much larger proportion of unjudged documents compared to other two approaches. In fact, 97% and 98% of the top-200 documents returned across all 50 topics using only machine learning for the TREC-8 and the GOV2 datasets are currently unjudged. Therefore, each of these approaches are evaluated using a condensed ranking, for which effectiveness is shown in Figure 2 (right). Effectiveness is overestimated with condensed rankings when a large portion of documents in the ranked result lists are unjudged [20, 21, 22]. Combined approaches reported a higher effectiveness than other approaches on a condensed ranking. Surprisingly, effectiveness for the ML method on a condensed ranking is worse than the other approaches, and somewhat similar to the effectiveness of combined approaches with a normal ranking. This suggests that using only the machine learning approach for locating relevant documents from the dataset is not effective. However, a definitive assessment for ML method can only be made by judging the ranked result list.

Recall that the traditional evaluation underestimates effectiveness when retrieved documents are unjudged. Therefore, the higher effectiveness for combined approaches with a condensed ranking compared to a traditional evaluation indicate that there could be more relevant documents that are not found by the existing approaches to pooling.

In Table 1 the proportion of MRJ documents per topic that were found to be relevant in the second pool are analysed. Again a similar trend of differences are seen, but with significant improvements up to a depth of $\kappa = 200$ for combined methods.

Discussion: The proportion of relevant documents that are pooled by automatic and manual runs out of the total pooled by each type of runs per topic is shown in Figure 4. As shown in the plot, manual runs provide a rich source of relevant documents for judging. If documents exclusively pooled by manual runs were not judged (i.e. no MRJ), the effectiveness of IR systems producing results similar to manual runs would be judged unfairly.

However, this still provides no guarantee that manual runs alone are a sufficient surrogate for all future IR systems. In fact, increased effectiveness for combined approaches on a condensed ranking rather than using a traditional ranking suggests the possibility of finding more relevant documents not found by either automatic or

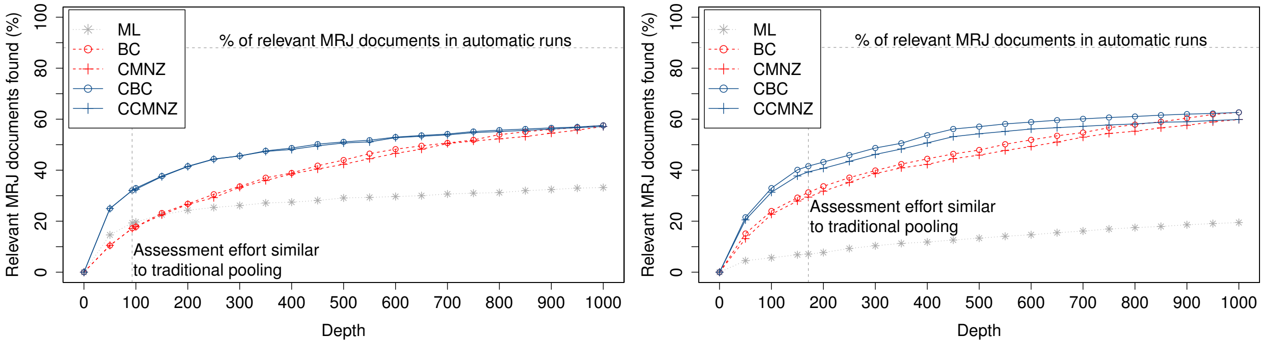

Figure 5: Percentage of relevant MRJ documents found in the top-(κ) of the proposed rankings for TREC topics 401-450 on the TREC-8 dataset (left) and TREC topics 801-850 on the TREC GOV2 dataset (right).

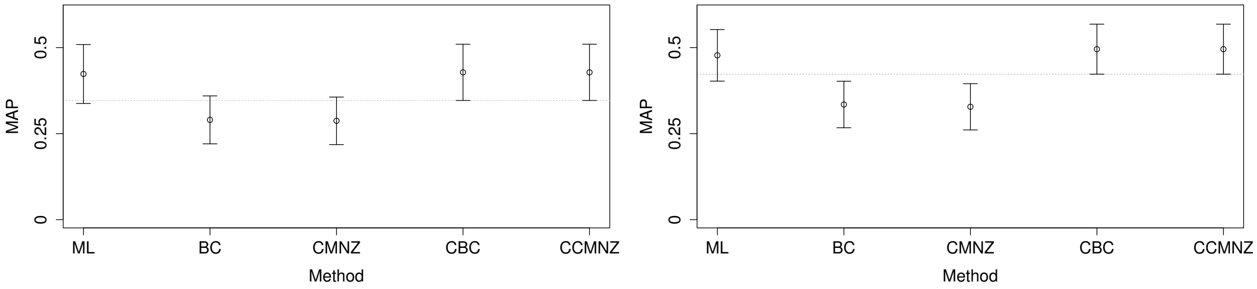

Figure 6: Just considering the documents in MRJ, how effective are ranking algorithms (MAP) on retrieving relevant documents for TREC topics 401-450 from the TREC 8 dataset (left) and for TREC topics 801-850 from the TREC GOV2 dataset (right)?

manual runs using current pooling strategies. Nonetheless, manual runs are vital to improve the reusability of test collections.

In Figure 5, the total proportion of relevant MRJ documents found by each of these approaches are analysed. The majority of documents uniquely pooled with manual runs (MRJ) also appear in automatic runs. However, they are not included in the first pool as they are ranked below the pool depth. In fact, 88.02% and 88.17% of the documents judged as relevant that are uniquely pooled by manual runs on the TREC-8 and GOV2 datasets could be found in the first pool if a pool depth of 1000 had been used. This upper threshold is the maximum proportion of relevant MRJ documents that can be found using the proposed fusion and combined methods and represented in the plots as a dashed horizontal line. The upper threshold for the maximum proportion of relevant MRJ documents that can be found by each of the combined methods is the maximum proportion of relevant MRJ documents found by the baseline fusion approach. The combined approaches effectively rerank the results produced by the fusion approaches.

Missing judgments for a large portion of the ranked lists from the proposed methods is one potential reason for the low retrieval effectiveness of the proposed approaches. The effectiveness for all approaches are higher when using a condensed ranking rather than a traditional evaluation, but rankings are of varying length. Therefore, retrieval effectiveness on retrieving documents from MRJ is computed in Figure 6. Note that the first pool and the ranking functions remains the same. The ML method now reranks the top-z unique documents ranked by manual runs. The ranking produced by ML shows a considerable improvement. The combined

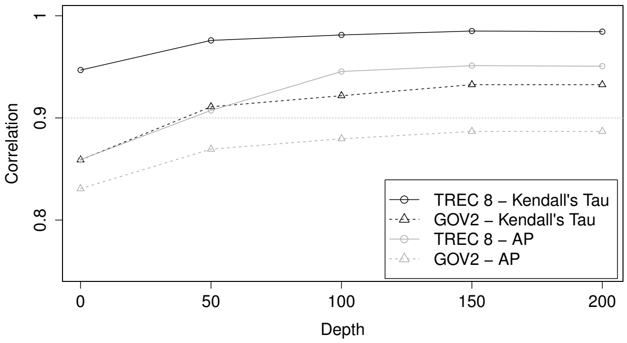

Figure 7: Kendall's τ and AP correlation of IR system rankings for varying depths of assessing documents with combined method (CBC) on the TREC-8 dataset with TREC topics 401–450 and the TREC GOV2 dataset with TREC topics 801–850.

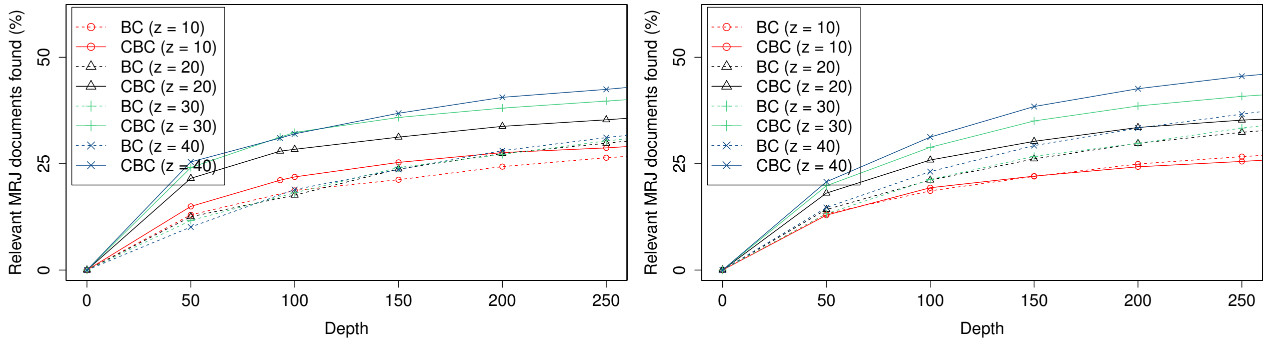

Figure 8: Percentage of relevant MRJ documents found out of total relevant MRJ documents found with a pool depth of 50 with varying pool depths for TREC topics 401-450 on the TREC 8 dataset (left) and TREC topics 801-850 on the TREC GOV2 dataset (right).

Metric	ML	BC	CMNZ	CBC	CCMNZ
P@10	0.0500	0.3375$^\bullet$	0.3250$^\bullet$	0.4187$^\bullet$	0.4000$^\bullet$
P@20	0.0406	0.3094$^\bullet$	0.3156$^\bullet$	0.4000$^\bullet$	0.3875$^\bullet$

Table 2: Effectiveness (P@10 and P@20) for each ranking approach when complete judgments are manually assessed up to a depth of 20 for the first 16 topics in the TREC GOV2 dataset. A $^\bullet$ implies a significant difference at $p < 0.01$ compared to ML.

method is more effective than fusion methods for MAP on both datasets, and the improvement is significant on the TREC GOV2 dataset. Recall that more unjudged relevant documents exist in larger test collections. Hence there is more room for improvement in the TREC GOV2 collection compared to the TREC-8 collection. Reranking a carefully retrieved subset of documents for topics with ML is an effective approach to locate new documents to be pooled and judged.

The true effectiveness of the above methods cannot be estimated without judging all of the unjudged documents for each method. Therefore, the top 20 documents produced by each method for the first 16 topics in the TREC GOV2 dataset were manually judged by an assessor, and results are shown in Table 2. The ML method is significantly worse with a p less than 0.01 than other methods even with 16 topics. The ML method is trained using ARJ documents. However, using only the ML method also ranks documents that contain similar terms to relevant documents in ARJ but not the terms related to the topic highly, and therefore the ML method alone is not effective. ML method may be improved by including query-dependent features, but not pursued further in this work. The above weakness is overcome when ranking is limited to documents that are already ranked by IR systems. This supports prior observed results. That is, effectiveness with ML method alone is low as in Figure 2 when entire document corpus is ranked. However, as observed in Figure 6 just ML method is highly effective when documents retrieved by manual runs (MRJ) are ranked. Therefore, the combined method is more effective than any of the other methods analysed.

For the rest of the discussion, the most effective method, CBC, is used. Whenever a new approach for pool composition is proposed, it is vital to quantify how well the approach ranks IR systems compared to the original method. A Kendall's τ and AP ranking correlation for varying depths of assessing documents with the CBC approach are shown in Figure 7. Manual runs are viewed as novel approaches for retrieval. The Kendall's τ correlation for MAP is above 0.9 beyond a depth of 50 on both datasets. A budget similar to original assessment permits processing up to a depth of 93 and 171 documents for TREC-8 and TREC GOV2 datasets respectively. However, the AP correlation is lower than the Kendall's τ correlation. The lower AP correlation compared to Kendall's τ correlation indicates that the top results are affected more. Yet the combined approach provides a valid method for improving reusability of test collections in the absence of manual runs.

The data available to train a machine learning ranker is less with lower pool depths. As a consequence, the classifiers can also be less effective. In Figure 8 we investigate how shallow the pool depth can be before the combined method is equally or less effective than the simple fusion methods. As shown in the graphs, the method is more effective than the fusion only method (BC) when the pool depth is above 10 or 20 for the TREC-8 and the TREC GOV2 datasets respectively.

4. CONCLUSION

In this paper, methodologies for building reusable test collections in the absence of manual runs are investigated. Simple fusion methods, a machine learning approach, and approaches that combine fusion and machine learning are studied. Combined methods are consistently more effective than any of the methods in isolation.

A large portion of relevant documents that are uniquely added by manual runs are also retrieved but not pooled by automatic runs. Taking advantage of the above fact, the combined approaches discover a considerable proportion of relevant documents that were previously only found by manual runs. The approach demonstrates the potential of finding relevant documents that are not possible using the current pooling approaches. However, the true efficacy of the approach cannot be properly assessed until all of the newly retrieved documents are judged. By judging top ranked documents for few topics, we demonstrate using ML method (without query-dependent features) is not effective and combined methods are the most effective. Topics containing more relevant documents could also be judged to different depths to maximise the effectiveness of the proposed approaches, but was not explored here. The initial results are promising as the method is already able to achieve a system ranking close to previous approaches which depended heavily on manual runs to add the necessary diversity to the assessment pool. The combined method is more effective than simple fusion methods on finding relevant MRJ documents even when the pool depth is shallow as 20.

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] J. A. Aslam and M. Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, pages 276–284. ACM, 2001.

[2] J. A. Aslam, V. Pavlu, and R. Savell. A unified model for metasearch, pooling, and system evaluation. In *Proceedings of the 12th Annual International Conference on Information and Knowledge Management (CIKM '03)*, pages 484–491. ACM, 2003.

[3] 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.

[4] S. Büttcher, C. L. A. Clarke, and I. Soboroff. The TREC 2006 terabyte track. In *Proceedings of the 15th Text REtrieval Conference (TREC '06)*, volume 6, page 39, 2006.

[5] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pages 63–70. ACM, 2007.

[6] B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*, pages 539–546, Geneva, Switzerland, 2010. ACM.

[7] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*, pages 651–658. ACM, 2008.

[8] B. Carterette, E. Gabrilovich, V. Josifovski, and D. Metzler. Measuring the reusability of test collections. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pages 231–240. ACM, 2010.

[9] B. Carterette, E. Kanoulas, V. Pavlu, and H. Fang. Reusable test collections through experimental design. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*, pages 547–554. ACM, 2010.

[10] C. L. A. Clarke, N. Craswell, I. Soboroff, and E. M. Voorhees. Overview of the TREC 2011 web track. In *Proceedings of the 20th Text REtrieval Conference (TREC '11)*, 2011.

[11] C. L. A. Clarke, N. Craswell, and E. M. Voorhees. Overview of the TREC 2012 track. In *Proceedings of the 21st Text REtrieval Conference (TREC '12)*, 2012.

[12] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. A. Clarke, and E. M. Voorhees. Overview of the TREC 2013 web track. In *Proceedings of the 22nd Text REtrieval Conference (TREC '13)*, 2013.

[13] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 282–289. ACM, 1998.

[14] 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.

[15] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC '93)*, pages 243–243. National Institute of Standards & Technology, 1993.

[16] G. K. Jayasinghe, W. Webber, M. Sanderson, and J. S. Culpepper. Extending test collection pools without manual runs. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, pages 915–918, 2014.

[17] M. G. Kendall. *Rank correlation methods.* Griffin, 1948.

[18] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93)*, pages 191–202, Pittsburgh, Pennsylvania, USA, 1993. ACM.

[19] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pages 375–382. ACM, 2007.

[20] T. Sakai. Comparing metrics across TREC and NTCIR: The robustness to system bias. In *Proceedings of the 17th Annual International Conference on Information and Knowledge Management (CIKM '08)*, pages 581–590. ACM, 2008.

[21] T. Sakai. Comparing metrics across TREC and NTCIR: The robustness to pool depth bias. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*, pages 691–692, Singapore, 2008. ACM.

[22] T. Sakai. On the robustness of information retrieval metrics to biased relevance assessments. *Journal of Information Processing*, 17:156–166, 2009.

[23] T. Sakai. The unreusability of diversified search test collections. *Proceedings of the 5th International Workshop on Evaluating Information Access (EVIA '13)*, 2013.

[24] T. Sakai, Z. Dou, R. Song, and N. Kando. The reusability of a diversified search test collection. In *Information Retrieval Technology*, pages 26–38. Springer, 2012.

[25] I. Soboroff and S. Robertson. Building a filtering test collection for TREC 2002. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*, pages 243–250. ACM, 2003.

[26] K. Spärck Jones and R. G. Bates. Report on the design study for the 'ideal' information retrieval test collection. *British Library Research and Development Report*, 5428, 1977.

[27] 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, 1975.

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

[29] E. M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370. Springer, 2002.

[30] E. M. Voorhees and D. Harman. Overview of the 8th text retrieval conference (TREC-8). In *TREC*, 2000.

[31] E. M. Voorhees and D. K. Harman. Overview of the 7th text retrieval conference (TREC-7). In *Proceedings of the 7th Text REtrieval Conference (TREC-7)*. NIST, 1998.

[32] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*, pages 587–594. ACM, 2008.

[33] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 307–314. ACM, 1998.