Size and Source Matter: Understanding Inconsistencies in Test Collection-Based Evaluation

Timothy Jones
RMIT University
timothy.jones@rmit.edu.au

Andrew Turpin
University of Melbourne
aturpin@unimelb.edu.au

Stefano Mizzaro
University of Udine
mizzaro@uniud.it

Falk Scholer
RMIT University
falk.scholer@rmit.edu.au

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

ABSTRACT

Past work showed that significant inconsistencies between retrieval results occurred on different test collections, even when one of the test collections contained only a subset of the documents in the other. However, the experimental methodologies in that paper made it hard to determine the cause of the inconsistencies. Using a novel methodology that eliminates the problems with uneven distribution of relevant documents, we confirm that observing a statistically significant improvement between two IR systems can be strongly influenced by the choice of documents in the test collection. We investigate two possible causes of this problem of test collections. Our results show that collection size and document source have a strong influence in the way that a test collection will rank one retrieval system relative to another. This is of particular interest when constructing test collections, as we show that using different subsets of a collection produces differing evaluation results.

Categories and Subject Descriptors

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

General Terms

Measurement; Reliability

Keywords

Information retrieval, evaluation, subcollections, TREC

1. INTRODUCTION

If a significant difference is measured between two IR systems using a test collection, is the difference real or an artifact of the collection? This question has been asked ever since test collections were first described [2]. While the reliability of relevance assessors [3], evaluation measures [4], and topics [9, 10] has been investigated, there has been little examination on whether the properties of a particular collection of documents influences search results.

While there is a general feeling in the IR community that it is preferable to demonstrate the superiority of one algorithm over another using multiple test collections, and this is often instantiated as using multiple metrics, topic sets, and relevance judgments, it is not uncommon to see published work that uses just one collection. Is this a safe experimental practice or not? There is little or no empirical evidence to show how many test collections are needed, what different types of collections are required, how one might measure the difference between collections, or indeed if it is necessary to test widely in the first place.

Sanderson et al. [6] presented evidence that the relative difference in effectiveness between retrieval systems varies across subsets (i.e. *subcollections*) of a single document collection, a result that had not been shown before. However, in that work, there were multiple differences between the subcollections being compared, making it difficult to know what the cause of this effect was. The work was also tested on relatively old retrieval systems.

In this paper, the experiments of Sanderson et al. are expanded using stricter controls on the differences between the subcollections, and a state-of-the-art retrieval system. The paper asks the following research questions:

1. What are the causes of the measured differences between subcollections described in past papers?
2. Do the measured differences still occur when relevant documents are contolled, and a state-of-the-art retrieval system is used?

2. PREVIOUS WORK

The work described here follows on from Sanderson et al. [6] who described a series of experiments examining TREC run data ranked using standard effectiveness measures based on different subcollections of a TREC test collection. They found that the ordering of the runs was substantially different between the subcollections and that these differences were much greater than would be expected by random chance.

Different ways of forming subcollections were investigated, such as splitting by document source (e.g. Financial Times, Congressional Record, etc.) or topical similarity (using k-means clustering). However, when the subcollections were formed based on any of the tested criteria, other factors also varied, which may have affected the comparisons. The most important of these factors were that the subcollections varied in size, sometimes substantially, and the number of relevant documents in each subcollection was different. In this paper, we mitigate these factors, in order to better

CIKM '14 November 03 - 07 2014, Shanghai, China

understand what properties of subcollections may play a role in leading to inconsistent retrieval results.

Additionally Sanderson et al. used IR systems that were $10-20$ years old: their experiments may have identified old problems that were resolved in newer ranking algorithms. In this paper, we use a state-of-the-art retrieval system. Finally, their method for comparing run ordering was Kendall's τ. This measure quantifies the degree of correlation between the rank ordering of two lists. However, τ may be affected by statistically insignificant swaps between systems. Here, we investigate a new agreement-based approach, which focuses only on statistically significant differences.

There is little past work on the impact of collection choice when measuring retrieval effectiveness. A recent tutorial on test collection construction [8] did not mention collection choice or the properties of the documents that should make up a collection. There has perhaps been an assumption, that an IR system will be tested on a representative sample of the documents the system will be deployed on. However, if advances in IR systems are to generalize to any document collection, rigourous evaluation is necessary.

Apart from the work of Sanderson et al., Azzopardi and colleagues have investigated bias in ranking functions. They showed evidence that ranking algorithms appeared to be influenced by the length of the documents they retrieved (most recently in Wilkie and Azzopardi [11]). Further, they showed that different algorithms were affected by length in different ways. Although work in the past has attempted to eliminate such biases [7], it would appear, they are still present.

3. DATA AND METHODOLOGY

Here, we describe the collections, system and measures used to investigate Sanderson et al's "subcollection effect".

3.1 Collections

The TREC 4–8 ad hoc collections are used for our experiments, in-keeping with the analysis of Sanderson et al. These collections suit splitting by document source, as all five collections are composed of texts from multiple sources.

3.2 Avoiding a relevant document effect

When Sanderson et al. split subcollections by source, the number of judged relevant documents available in each subcollection varied substantially. To address this potential problem, we used the following methodology. We separated each test collection into two subsets: one containing all documents that were judged as relevant (*all-rel*); the other containing all the other documents in the collection, including documents judged as irrelevant (*not-rel*). All subcollection splits were then carried out by partitioning the *not-rel* subset of the collection based on document source. All the documents from *all-rel* were then added to each subcollection.

By forming subcollections in this way, we ask the question: is the ability of an IR system to retrieve the same relevant documents affected by distinct sets of non-relevant documents? Further, if there is an effect, are different IR systems affected in different ways?

3.3 Measuring collection agreement

Evaluating the effectiveness of a new ranking approach using a test collection typically compares two variants of an IR system—a new configuration of the system, and the previous configuration (the baseline)—over a fixed set of queries, documents, and relevance judgements. The outcome of such an experiment is typically quantified by reporting the mean effectiveness of each system accompanied by a statistical significance test. Having run such an experiment on one collection, one might ask whether the same outcome would be observed when comparing the same systems on another collection. To investigate this, we adapted an approach used by Moffat et al. [4] to investigate the agreement on IR experimental results when using different effectiveness metrics. However, here we examine the agreement on significance when using different collections. Consider the process of comparing systems, S1 and S2, on two subcollections, C1 and C2. There are a number of possible outcomes as follows.

- SS_a: Active agreement, S1 is significantly better than system S2 on both C1 and C2.
- *SN*: Passive disagreement, S1 is significantly better than S2 (or vice versa) on C1, but on C2, there is no significant difference between the systems.
- *NS*: Passive disagreement, S1 is significantly better than S2 (or vice versa) on C2, but on C1, there is no significant difference between the systems.
- SS_d: Active disagreement, S1 is significantly better than S2 on C1, but S2 is significantly better than S1 on C2.

As per Moffat et al. [4], for a given collection pair, the proportion of significant differences between systems observed for which both collections agree on which is the better system is given by

$$agree\text{-}SS_a = \frac{2 \cdot SS_a}{2 \cdot SS_a + 2 \cdot SS_d + SN + NS}.$$

3.4 State of the art IR system

In order to simulate multiple research experiments of IR systems, we used the Terrier search engine [5] configured to use sixteen different ranking models: BB2, BM25, DFR_BM25, DFRee, DLH13, DLH, DPH, Hiemstra_LM, IFB2, In_expB2, In_expC2, InL2, Lemur TF_IDF, LGD, PL2, and TF_IDF. The search engine was in default TREC mode. The models were run on each subcollection, and evaluated using MAP, producing 120 pairs of system comparisons for each subcollection.

4. RESULTS

The results of the experiments to test potential confounding factors in Sanderson et al's methodology are described here.

4.1 Investigating source-based splits

The first seven columns of Table 1 shows $agree\text{-}SS_a$ for each source-based subcollection for the TREC 4–8 ad hoc collections, with the mean (μ) and standard error (s.e.) for each subcollection reported in columns eight and nine. Recall that the original collections are split by document source, but the number of relevant documents is held constant in each subcollection. Comparisons also included comparing the whole TREC collection with each subcollection. The minimum, median, and maximum of μ are 8%, 62%, and 74% respectively. The central 50% of the data are in the range from 40% to 67%. Broadly speaking, running experiments on different subcollections may lead to inconsistent conclusions about system superiority a substantial proportion of the time.

Kendall's Tau (τ) is shown to facilitate comparison to Sanderson et al. [6], and the values are in the same range as those obtained by Sanderson et al., and rather low. The τ values generally correlate with $agree\text{-}SS_a$, although differences exist.

For further context, the $agree\text{-}SS_a$ values obtained when measuring agreement when comparing each subcollection against itself using two different effectiveness metrics—MAP and NDCG—are shown in the final column of Table 1. Both measures were calculated over the top-1000 retrieved documents. In all but three cases,

Collection	$agree\text{-}SS_a$ 2	3	4	5	6	7	μ	s.e.	map ndcg	τ
TREC-4										
1. patents	81	63	61	54	50	56	61	2.1	92	0.39
2. fr		67	61	51	54	57	62	1.9	95	0.40
3. wsj			82	73	82	73	73	1.7	93	0.65
4. sjm				87	74	79	74	1.7	67	0.64
5. ap					69	64	66	1.9	88	0.57
6. ziff						70	66	1.8	96	0.65
7. trec4							66	2.0	80	0.62
TREC-5										
1. cr	48	64	50	48	50	61	53	2.2	85	0.62
2. ziff		48	46	37	38	48	44	2.1	94	0.36
3. wsj			82	82	84	83	74	1.6	87	0.67
4. fr				72	76	85	69	1.9	89	0.64
5. ap					87	74	67	1.8	93	0.60
6. ft						78	69	1.7	84	0.63
7. trec5							72	1.8	86	0.57
TREC-6										
1. cr	15	8	6	4	7	–	8	2.8	71	-0.00
2. fr		34	27	14	34	–	25	3.2	79	0.23
3. fbis			28	8	30	–	21	4.1	14	0.10
4. latimes				53	82	–	39	3.6	56	0.31
5. ft					41	–	24	2.9	85	0.28
6. trec6						–	39	3.9	67	0.31
TREC-7										
1. fr	48	34	39	32	–	–	38	3.0	86	0.57
2. fbis		68	77	63	–	–	64	3.2	58	0.64
3. latimes			70	68	–	–	60	3.4	72	0.67
4. ft				79	–	–	66	3.0	78	0.47
5. trec7					–	–	60	3.3	83	0.61
TREC-8										
1. fr	50	40	29	37	–	–	39	3.4	80	0.12
2. fbis		71	57	70	–	–	62	3.2	82	0.56
3. latimes			76	82	–	–	67	2.7	93	0.58
4. ft				81	–	–	61	2.7	80	0.55
5. trec8					–	–	67	2.7	75	0.43

Table 1: $agree\text{-}SS_a$ values (as percentages) for TRECs 4-8, when comparing systems across subcollections (labelled in the rows of each section and ranked by increasing size). Means (μ) are across all pairings for each subcollection, and the standard error of μ (s.e.) was derived using bootstrapping. For comparison, $agree\text{-}SS_a$ when comparing systems using NDCG and MAP is shown, and mean τ is in the last column.

the $agree\text{-}SS_a$ on different evaluation measures was higher than the mean subcollection $agree\text{-}SS_a$.

Subcollections containing identical sets of relevant documents appear to inconsistently identify many significant differences in variations of the same retrieval system.

4.2 Investigating subcollection size

While relevant documents were kept constant across subcollections in the previous experiment, each compared collection was a different size, ranging from 12,994 to 223,751 documents. It is therefore plausible that size—which represents the number of non-relevant documents in a subcollection—affected the $agree\text{-}SS_a$ rate. To further investigate the effect of size on the results, a simulation experiment was run measuring $agree\text{-}SS_a$ across subcollections of different size. We created several new subcollections of similar sizes to the TREC subcollections, but using documents selected at random from the whole TREC-8 collection rather than split by source.

Figure 1 shows the inter-size agreements. The points on the diagonal compare random subcollections of the same size, and $agree\text{-}SS_a$ between these collections is high. Towards the bottom right of the figure, where subcollection sizes differ by up to a factor of four, $agree\text{-}SS_a$ reduces.

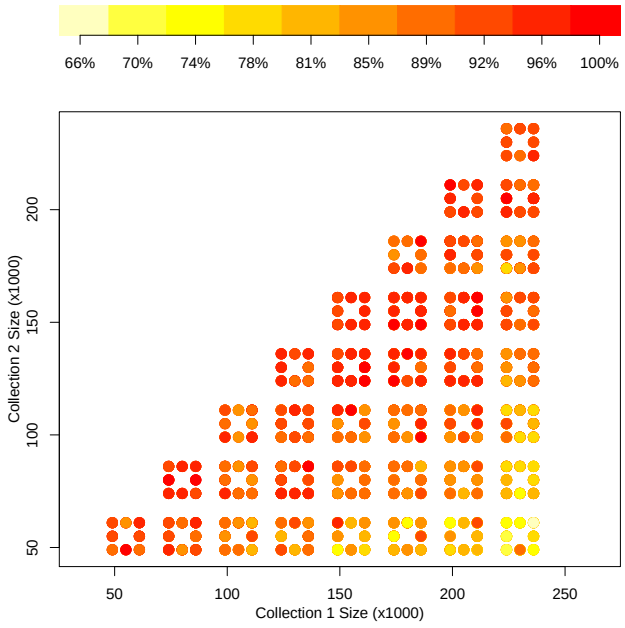

Figure 1: Agreement between collections as a function of collection size. Note the gradient of lighter points (less agreement) towards the bottom right of the figure. Eight distinct pairs of random subcollections were tested for each comparison, plotted as a square centered on the actual sizes.

However, the low $agree\text{-}SS_a$ values in Figure 1—even when subcollections were substantially different in size—are generally higher than those recorded in Table 1. While this result indicates that the relative size of subcollections could impact on $agree\text{-}SS_a$, it appears that this is not the main cause of the results shown in Table 1. Additionally, it was noted that two subcollections, fbis and latimes, were of similar size, but across TRECs-6,7,8 the $agree\text{-}SS_a$ was low (28%, 68%, 71%), which we further examine.

4.3 Comparing fbis and latimes

To investigate why fbis and latimes had such a low $agree\text{-}SS_a$, but similar sizes, a series of comparisons were made between the subcollections by varying the amount of random documents that were added to each subcollection. To facilitate exposition, we label the new subcollections fbis-20, fbis-40, fbis-60 and fbis-80, where fbis-N is the same size as fbis, but contains N% documents pulled at random from the fbis subcollection, and 100-N% of documents pulled from the rest of the TREC-8 collection. As in the other collections used in this paper, these collections all follow the construction methodology described in Section 3.2. For example, an instance of the fbis-20 collection includes a portion the same size as the non-relevant documents from fbis, but that portion is instead comprised of 20% documents selected randomly from the non-relevant documents in fbis, and 80% documents from the non-relevant documents that are not in fbis. Then, the *all-rel* collection is added, for a total size equal to the non-relevant portion of fbis, plus the size of the *all-rel* collection. Using this methodology, we also constructed latimes-20, latimes-40, latimes-60 and latimes-80.

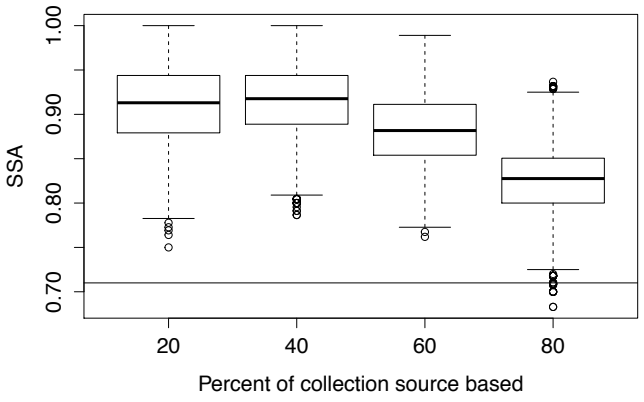

Figure 2: Agreement between collections for various percentages of source restricted splits using fbis and latimes from TREC8. The horizontal line indicates the agreement for a fully source based split.

For each value of $N \in \{20, 40, 60, 80\}$, 50 fbis and 50 latimes subcollections were formed and $agree\text{-}SS_a$ was measured (Figure 2). As collections become more source-based, there is a corresponding decrease in agreement, indicating that the source-based split is indeed contributing to the disagreement observed above.

5. DISCUSSION AND CONCLUSION

This paper addresses the following two research questions. What are the causes of the measured differences between subcollections described in past papers? Do the measured differences still occur when assessing a state-of-the-art retrieval system?

Using a novel methodology, this work has confirmed what Sanderson et al. reported in an earlier paper [6]. There is good evidence that the choice of collection impacts on whether one retrieval system will be measured to be significantly better than another retrieval system. Note also, this effect is visible even when a subset of a larger collection is used. The focus of work here has been to eliminate confounding factors introduced in Sanderson et al's methodology, and to begin to understand the causes of the effect.

To that end, the work here has established that the measured inconsistency in evaluation across subcollections is not an effect of the number of relevant documents in that collection or the relative difference in collection size. Additionally, the measured differences between subcollections also appear to still be occurring when using a state-of-the-art retrieval system and comparing between variants of the same collection.

While it has long been understood that one should test on multiple collections in order to cover for unknown variations in test collections, there has been little or no work understanding what those variations are. Such an understanding could reduce the number of collections one needs to test on.

We contend, as Sanderson et al did, that the results here are quite striking. Many of the subcollections in the early years of TREC were on the surface, quite similar to each other. In TREC-7 for example, three of the four subcollections, fbis, latimes, and ft, are collections of news, just from different sources. We view this work as a starting point: the most important issue to address now is what is causing the inconsistency in measurement of effectiveness?

In addition, this work has implications when creating collections, since test collections are typically subcollections of the available documents appropriate for the collection. It is important to create test collections that agree with the real world of all documents, and therefore it is important to understand the factors that cause evaluations on subcollections to disagree with the original collection.

Towards understanding the inconsistency in evaluation, this work has demonstrated that both collection size and source of documents contribute to the reliability of experiments. Although document source is a meta-feature that does not describe specifically how the document is different, and is not typically used as a feature by ranking algorithms, document source has been used at collection creation time, such as in the UK web spam collections—where only URLs ending in `.uk` were included [1].

Although the collection size and document source both strongly affect the consistency of evaluation, these features do not explain the whole effect. Some further underlying cause(s) must be affecting the results measured. Finding these will be our next focus.

Once we understand the effect well enough to predict whether particular subcollections will agree with each other (and with the whole collection that they are sampled from), many interesting applications are possible. For example, these results will allow the creations of test collections that are confidently generalizable, i.e., subcollections that agree with the larger set of available documents. Additionally, by splitting existing collections into subcollections where different retrieval approaches perform best, increases in retrieval performance may be possible.

Acknowledgements

This work was supported in part by the Australian Research Council (DP130104007) and also by a Google Faculty Research Award.

6. REFERENCES

[1] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, Dec. 2006.

[2] C. W. Cleverdon. *Report on the Testing and Analysis of an Investigation Into the Comparative Efficiency of Indexing Systems*. ASLIB Cranfield Research Project. 1962.

[3] C. W. Cleverdon. *The effect of variations in relevance assessments in comparative experimental tests of index languages*. Number 3 in Cranfield Library Report. 1970.

[4] A. Moffat, F. Scholer, and P. Thomas. Models and metrics: IR evaluation as a user process. In *Proc. ADCS*, pages 47–54. ACM, 2012.

[5] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proc. OSIR Workshop*, pages 18–25. Citeseer, 2006.

[6] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. Differences in effectiveness across sub-collections. In *CIKM'12*, pages 1965–1969. ACM, 2012.

[7] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. SIGIR*, pages 21–29. ACM, 1996.

[8] I. Soboroff. Test collection diagnosis and treatment. *Proc. EVIA 2010*, pages 34–41, 2010.

[9] J. Urbano, M. Marrero, and D. Martín. On the measurement of test collection reliability. In *Proc. SIGIR*, SIGIR '13, pages 393–402, New York, NY, USA, 2013. ACM.

[10] E. M. Voorhees. Topic set size redux. In *Proc. SIGIR*, SIGIR '09, pages 806–807, New York, NY, USA, 2009. ACM.

[11] C. Wilkie and L. Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. In *Advances in Information Retrieval*, pages 13–25. Springer, Apr 2014.