Evaluating the Predictivity of IR Experiments

Lida Rashidi
The University of Melbourne
Melbourne, Australia

Justin Zobel
The University of Melbourne
Melbourne, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

ABSTRACT

Experimental evaluation is regarded as a critical element of any research activity in Information Retrieval, and is typically used to support assertions of the form "Technique A provides better retrieval effectiveness than does Technique B". Implicit in such claims are the characteristics of the data to which the results apply, in terms of both the queries used and the documents they were applied to. Here we explore the role of evaluation on a collection as a *prediction* of relative performance on collections that have different characteristics. In particular, by synthesizing new collections that vary from each other in a controlled way, we show that it is possible to explore the reliability of an IR evaluation pipeline, and to better understand the complex interrelationship between documents, queries, and metrics that is an important part of any experimental validation. Our results show that predictivity declines as the collection is varied, even in simple ways such as shifting in focus from one document source to another similar source.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Test collections**; **Relevance assessment**; **Retrieval effectiveness**.

KEYWORDS

Evaluation; significance testing

ACM Reference Format:
Lida Rashidi, Justin Zobel, and Alistair Moffat. 2021. Evaluating the Predictivity of IR Experiments. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463040

1 INTRODUCTION

Experimental evaluation is a critical component of Information Retrieval. In *on-line* experiments, a live system is employed in A–B mode and system effectiveness inferred from user behavior [4]. In *off-line* evaluation, a corpus of documents, a set of topics or queries, a set of relevance judgments, and one or more effectiveness metrics are employed [10]. A primary goal of evaluation is to determine whether one system can be identified as being superior to others.

Superiority is demonstrated via consistently higher effectiveness scores from the selected metrics, each corresponding to a model of how search users behave. For example, reciprocal rank is a shallow metric that reflects the search experience of users who stop once they have encountered a relevant document; whereas P@100 is a deep metric, and reflects the experience of users who always examine exactly 100 documents in the results. A range of other metrics have been developed based on such user models [6, 8].

As well as the direct goal of establishing what might be called "here and now" superiority, a primary purpose of experimental evaluation is *prediction* – an expectation that on other data "of this type" the same relative outcomes will continue to hold. It is the predictive component of an evaluation that gives the confidence to test a set of systems on a corpus, choose a "winner", and then deploy that winner against a live query stream.

The importance of statistical testing is well understood (see, for example, Sakai [9]). In the case of most IR experimentation, the statistical test is applied assuming that the employed topics are a subset of a universe of topics, and hence the "p-value" that emerges relates to variability of topics. A range of experiments have used topic subsets, comparing (for example) inferred system orderings derived from one half of the available topics with the system orderings derived from the other half [1, 11, 13, 14].

The small number of document collections available to academic researchers means that statistical testing over sets of collections is much less common. For example, while referees might be sceptical of experimental conclusions based on only one collection, they are likely to be comfortable if (say) three collections all show similar system relativities. As a consequence, there has been rather less attention given to collection variability than there has to topic variability. Instead, collection splitting has been primarily used as a device to allow investigation of the relative size of system, topic, and system-topic effects, and hence allow improved statistical confidence. For example, Voorhees et al. [15] and Ferro et al. [3] make use of the fact that random assignment of documents across a set of partitions generates replicate collections with identical statistical properties; Sanderson et al. [12] split collections based on source, domain, and content type, to generate random sub-collections; and Ferro and Sanderson [2] incorporate the effects of topic, system, and sub-collections into a model to allow multiple system comparisons, measuring the impact of sub-corpora on effectiveness scores.

In recent work Zobel and Rashidi [17] introduce the idea of *collection bootstrapping*, an approach to developing fresh collections that are "like" the seed collection but not identical to it, as a way of investigating the experimental variability that might be attributable to the collection. The large sizes of current document collections means that an N document collection is readily bootstrapped via the $Poisson(k; 1) = [(1/e)/k! \mid k \in \{0, 1, 2, 3, \ldots\}] = [0.368, 0.368, 0.184, 0.061, 0.015, 0.003, \ldots]$ distribution. That is, each document identifier should be thought of as being uniformly hashed to the real interval $[0, 1]$, and then replicated into the output collection $k \in \{0, 1, 2, 3, \ldots\}$ times, based on where that hashed value

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463040

falls in the CDF of $Poisson(k; 1)$. Statistics gathered over a large number of collection replicates (each generated using a different hash key) can then be used to infer statistics in regard to the original collection, including the variability of metric scores, and the stability of system-versus-system comparisons for any particular metric [17]. Using bootstrapping, hundreds or even thousands of "like" collections can be generated.

In the major TREC test collections, there are large numbers of contributing systems, and for each system a *run* for each query lists the sequence of documents in the order they were ranked by that system for that query. Under bootstrapping, the runs are also modified, to reflect the corpus contents; with some documents present multiple times and others absent, and with the original ordering of the runs' documents respected. These runs can then be scored using qrels that likewise have some entries duplicated, and others removed. That is, corpus bootstrapping can be used to construct simulated distributions of both system score and system comparison outcomes [17].

Our contribution. We explore a different use of bootstrapping. Instead of creating replicate collections that are identical to each other, we create complementary subcollections that differ in a controlled characteristic but share query sets and sets of runs, and then create random paired partitions of the complementary subcollections according to a *meld factor*. By adjusting the meld factor, each pair of partitions can be made either statistically identical to each other (a meld of 1.0), allowing testing of the false positive behavior of statistical tests, or controllably different from each other (melds < 1.0), allowing exploration of the sensitivity of metrics and statistical tests with differing data but identical queries.

Our findings show that some effectiveness measures are more sensitive than others to collection characteristics, and vary in the extent to which they report that the same system is significantly different to itself on different data – demonstrating that the measures do behave in different ways, but also that the collections are identifiably distinct. More importantly, they show that predictivity has strong limits. Even for the relatively homogeneous collections that we use (subcollections of TREC newswire data), predictivity notably degrades as the meld factor is increased.

2 EXPERIMENTS

This section describes the datasets used and the splitting-melding-bootstrapping approach we developed.

Collections and measures. We use Disks 4 and 5 from TREC, the queries from TREC 7 and TREC 8, and the runs from the top 50 systems in each of these TREC rounds, with "top" defined by system average for the metric RBP using a persistence value of $\phi = 0.95$. We show several measures here: average precision (AP), precision at 10 (P@10); and rank-biased precision with persistence of $\phi = 0.95$. Space limits prevent reporting of all results but those shown are representative (we also tested NDCG [5], INSQ [7], and reciprocal rank). The statistical test used is a one-sided paired t-test except where noted otherwise.

Forming subcollections. To create subcollections with different degrees of "difference", we assume a base pair of subcollections where the documents in one are labeled "L" (left) and in the other "R" (right). To create a pair with meld factor $0 \le mf \le 1.0$, each document is considered in turn, and its label flipped (from L to R or from R to L) with probability $mf/2$. Those documents now labeled "L" form one collection, and those now labeled "R" form the other.

When $mf = 1.0$, the two generated collections are statistically identical, each containing a random half of the original collection. At the other extreme, when $mf = 0.0$, the two collections are still the initial L/R division. The underlying "L or R" document labels can be based on any desired factor, with (preferably) around half of the documents labeled L and half R. We make use of the following starting-point labelings:

- *Length*: the documents in the collection are first ordered by their length in words. They are then labeled as L, C, and R, based on whether their length was in the lowest, middle, or highest tertile. Documents in the C tertile were discarded (to avoid issues caused by having many documents near the overall median length).
- *Content*: documents from financial or legislative sources (Financial Times and the Federal Register) were labeled L, while generic newswire sources (for example, the Foreign Broadcast Information Service and the Los Angeles Times), were labeled R.
- *Rank*: each document was assigned a value r, its shallowest rank position in any run from any system for any topic, considering only the documents for which $r \le 100$. Label "L" was then assigned to the documents with r less than the median of that set (that is, documents that appeared "early" in at least one run) and R to the remaining retrieved documents. Since relevant documents tend to be denser near the start of runs, this creates a partitioning in which L is rich in relevant documents compared to R, while the volume of judgments is similar.

Table 1 summarizes these six pairs of starting points. In each case there are more relevance judgments associated with the "R" collection, but in experiments not reported here we examined the position of the first unjudged document in each run, and in general this position is sufficiently deep in the ranking that the effectiveness measures used should be reasonably well-behaved, that is, not confounded by missing information.

Table 1: Collections used, their L–R splits, and the qrel split. The two length-based splits use one-third of the respective collections.

Collection	Subcollection	Documents	%Judg.
TREC7		358,493	
Length	$L \leftarrow$ short docs.	119,498	29.2
	$R \leftarrow$ long docs.	119,498	42.1
Content	$L \leftarrow$ financial, federal	175,957	47.8
	$R \leftarrow$ other newswire	182,536	52.2
Rank	$L \leftarrow$ high-ranking docs.	34,921	46.9
	$R \leftarrow$ low-ranking docs.	34,661	52.8
TREC8		347,598	
Length	$L \leftarrow$ short docs.	115,866	24.8
	$R \leftarrow$ long docs.	115,866	45.5
Content	$L \leftarrow$ financial, federal	169,329	44.8
	$R \leftarrow$ other newswire	178,269	55.2
Rank	$L \leftarrow$ high-ranking docs.	41,163	45.8
	$R \leftarrow$ low-ranking docs.	39,856	54.2

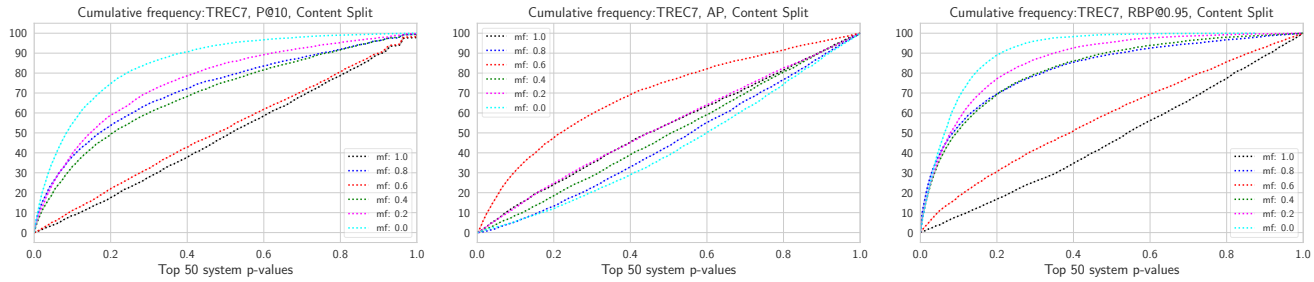

Figure 1: Cumulative frequency distribution of 50,000 p-values for system-vs-self comparison based on scores for bootstraps of L and R, using P@10 (left), AP (center), and RBP ($\phi = 0.95$, right). Subcollections were generated from the "TREC 7, Content" starting partition.

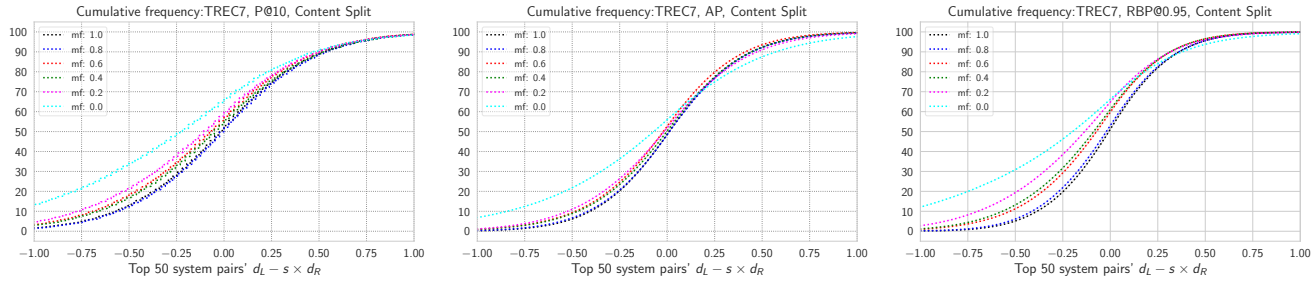

Figure 2: Cumulative frequency distribution of system score differences for P@10 (left), AP (center), and RBP ($\phi = 0.95$, right), plotting $d_L - s \times d_R$, with $s = -1$ if $d_L \times d_R < 0$, and $s = +1$ otherwise. Subcollections were generated from the "TREC 7, Content" starting partition.

Once the starting partitions are formed, a distribution of behaviors can be observed by bootstrapping the contents of each partition. In all of the experiments reported here, for each meld factor we create 10 pairs of random partitions and then bootstrap 100 times from each partition, to get 1000 pairs of collections at each measured data point. At $mf = 0.0$ the partitions are by construction identical, but the total number of bootstraps is still 1000.

As an overall observation on the experiments, behavior tended to be similar for each of Length, Content, and Rank, and on both TREC 7 and TREC 8. For reasons of space we select amongst our results, showing the complete set of cases only in the final table.

Self-comparison of systems. We first report the result of experiments in which each system is compared to itself using a range of metrics and meld factors, to determine the extent to which metric scores alter as the nature of the underlying collection changes. These experiments rest on our methodology for forming controllably "nearly the same" collections, and the verification processes used to check their consistency and applicability as the degree of meld ("nearly-ness") is changed. We report the effect of melding on predictivity in the next set of experiments, described shortly.

In these first experiments the numeric score of each system on L is compared to the numeric score of the same system on R, assessing them over the 50 topics using a t-test. If the metric behaves differently on L to R, a high fraction of small p-values should result. With 50 systems × 10 partitions × 100 bootstraps, a total of 50,000 L-vs-R system-vs-self paired comparisons and hence p-values are plotted for each metric and each starting partition (Table 1).

Figure 1 gives results for P@10, AP, and RBP, for the "TREC 7, Content" split. For P@10 and RBP, values of mf below 1.0 lead to a greater proportion of small p-values compared to $mf = 1.0$, indicating that the systems are not yielding comparable scores on "L" and "R". This behavior is a direct consequence of the way these metrics compute their scores, and the fact that the relevant documents are not evenly split between "L" and "R" (even though the judgments are). However the situation is different for the normalized metric AP (in the center pane), with system scores that tend to be more alike between the two partitions of each bootstrap, but with a less well defined pattern of behavior as mf is varied. Indeed, for AP, bootstrapping using $mf = 0.6$ is the most likely to yield statistically different scores on the "L" and "R" collections. In the case of AP, the different splits resulted in different patterns of behavior.

A-vs-B system comparisons. We now change tack, and instead of asking whether any given metric gives the same scores on different collections, we ask the extent to which each metric puts pairs of systems into the same relative ordering on the two collections, where the extent of the difference between the collections is again controlled by the meld factor mf. Starting with the top 50 systems, we examine each of the $50 \times 49/2 = 1225$ system pairs on L, and then compare the outcomes to those on R. That is, for each of the 10 random splits and 100 subsequent bootstraps, we have 1225 A-vs-B differences d, each averaged across 50 topics.

Define d_L to be the mean (over topics) score difference in L between systems A and B, and define d_R similarly. Figure 2 plots the distribution of the 1,225,000 values of the derived quantity $d_L - s \times d_R$,

Table 2: Percentage of A-vs-B system pairs "not supported in R" when restricted to $p_L \in 0.01 \pm 10\%$. There are 50 systems considered for each of TREC 7 and TREC 8; three start partitions; three metrics; and six values of $mf \in \{1.0, \ldots, 0.0\}$.

	Metric	Starting from Length partition						Starting from Content partition						Starting from Rank partition					
		1.0	0.8	0.6	0.4	0.2	0.0	1.0	0.8	0.6	0.4	0.2	0.0	1.0	0.8	0.6	0.4	0.2	0.0
TREC7	AP	0.7	0.6	0.9	1.0	2.0	11.1	0.4	0.4	0.3	0.5	0.8	2.4	0.3	0.3	0.3	0.4	0.6	0.7
	P@10	1.8	2.3	2.7	2.4	3.7	13.3	1.9	1.7	1.6	1.7	1.5	1.9	1.2	1.8	2.3	2.9	6.3	14.5
	RBP@0.95	0.2	0.3	0.5	1.0	3.1	6.4	0.2	0.1	0.2	0.3	0.4	1.8	0.1	0.4	0.7	2.2	4.7	8.8
TREC8	AP	1.6	2.2	1.4	4.4	4.3	12.5	0.9	1.2	0.9	1.3	1.4	3.3	0.7	0.8	0.6	0.9	1.6	5.1
	P@10	4.5	3.2	3.4	3.7	5.6	10.6	2.9	2.8	3.3	3.6	4.1	7.7	3.0	3.1	3.2	3.5	5.4	11.5
	RBP@0.95	0.6	0.6	0.6	1.0	3.0	8.8	0.3	0.3	0.6	0.5	0.8	4.2	0.4	0.6	0.4	0.9	1.3	3.4

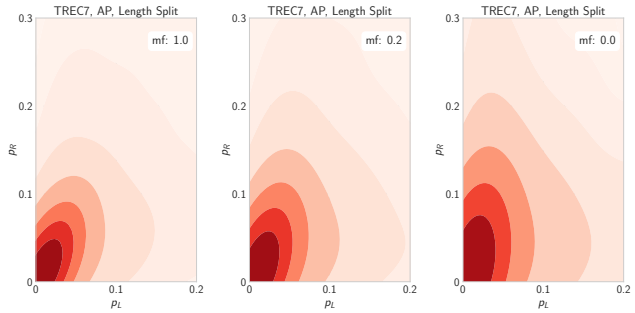

Figure 3: The relationship between p_L and p_R for A-vs-B system comparisons using AP for the "TREC 7, Length" starting partition.

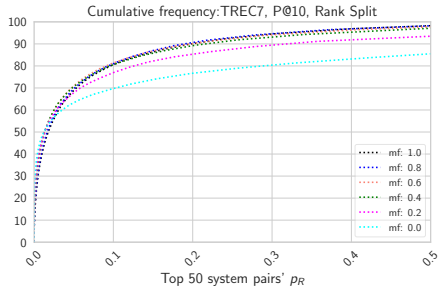

Figure 4: Cumulative frequency distribution of p_R for system pairs for which $p_L \in 0.01 \pm 10\%$, using P@10 and "TREC 7, Rank".

where s is -1 if d_L and d_R have opposite signs. If relative metric scores on L and R were always identical, the result would be a vertical line above the 0.0 point on the horizontal axis. Unsurprisingly there is a distribution across bootstraps, and even for a meld factor of 1.0 there is a spread, becoming more pronounced as the document collections become less similar. Nor is the normalized AP completely immune to this effect.

Predictivity. Each of the 1,225,000 system-collection pairs associated with any given starting partition can be t-tested to yield two p-values, p_L for A-vs-B in L, and p_R for A-vs-B in R. Figure 3 plots density over the (p_L, p_R) space, showing the interrelationship. When $mf = 1.0$, p_L and p_R are well correlated; they also tend to lie in the same regions when $mf < 1$, but with a bias due to the lower number of relevant documents in L. At $mf = 0.0$ the upward elongation is pronounced, showing a weakening correlation as the collections become more distinct. Other combinations of measure, split, and collection, show the same trend, to varying degrees.

To explore these results further, Figure 4 considers a narrow band in which $p_L \approx 0.01$, often used to indicate high significance, and plots the corresponding p_R values. Reducing mf leads to notably reduced correlation. As an extreme (and concerning) case, when $mf = 0.0$ only 85.5% of comparisons put system A and system B in the same order in R as they were in L, and nearly 15% of the cases that were significant ($p \approx 0.01$) in L had no support at all in R.

The combination of settings used in Figure 4 was chosen because the effect is clearly demonstrated; in other cases it was not as strong. Table 2 shows the corresponding "non-support" rates for three metrics, all splits, and both collections. In principle, with $p = 0.01$ around 1% of observations might violate the prediction, that is if the test showed $A > B$ with $p = 0.01$ then it would not be surprising to see $\approx$ 1% of cases with $B > A$ on the new data. Indeed, given that these are tests with the same queries, an even lower rate of violation might be expected in practice. However, we observe a very different result. Precision at 10 is never adequately predictive (confirming the work of Webber et al. [16]); and, when $mf = 0.0$, the order violations are higher than might be anticipated with other metrics too. As L and R become different, predictivity declines.

3 CONCLUSION

We have described a methodology for measuring predictivity, and used it to demonstrate that experimental results may not have the universality that is often implicit or assumed. This is obviously the case for collections that dramatically differ from the standard test corpora, such as collections of tweets or of biomedical abstracts – contexts in which TREC-tuned methods can behave poorly. Nonetheless, the usual assumption is that systems will exhibit the same relative performance across a range of kinds of data; indeed, the purpose of an experiment is exactly that, to predict future performance. It is now clear that such predictions may require caveats.

Further, our results assume the same set of queries, arguably the easiest possible test case. That predictivity is not preserved in this scenario is of deep concern. Much IR research yields small but consistent improvements that accrue over time, and we need to be able to reliably recognize such outcomes. Moreover, many methods are tightly bound to collection characteristics; our results show that it is incumbent on researchers to explain and proscribe the scope of data on which any claims they make are valid.

REFERENCES

[1] B. A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. Inf. Sys.*, 30(1):4, 2012.

[2] N. Ferro and M. Sanderson. Sub-corpora impact on system effectiveness. In *Proc. SIGIR*, pages 901–904, 2017.

[3] N. Ferro, Y. Kim, and M. Sanderson. Using collection shards to study retrieval performance effect sizes. *ACM Trans. Inf. Sys.*, 37(3):30:1–30:40, 2019.

[4] K. Hofmann, L. Li, and F. Radlinski. Online evaluation for information retrieval. *Founds. Trnds. Inf. Retr.*, 10(1):1–117, 2016.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.*, 20(4):422–446, 2002.

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

[7] A. Moffat, F. Scholer, and P. Thomas. Models and metrics: IR evaluation as a user process. In *Proc. Aust. Doc. Comp. Symp.*, pages 47–54, 2012.

[8] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. Inf. Sys.*, 35(3):24:1–24:38, 2017.

[9] T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proc. SIGIR*, pages 5–14, 2016.

[10] M. Sanderson. Test collection based evaluation of information retrieval systems. *Founds. Trnds. Inf. Retr.*, 4(4):247–375, 2010.

[11] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proc. SIGIR*, pages 162–169, 2005.

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

[13] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. CIKM*, pages 623–632, 2007.

[14] J. Urbano. Test collection reliability: A study of bias and robustness to statistical assumptions via stochastic simulation. *Inf. Retr.*, 19(3):313–350, 2016.

[15] E. M. Voorhees, D. Samarov, and I. Soboroff. Using replicates in information retrieval evaluation. *ACM Trans. Inf. Sys.*, 36(2):12:1–12:21, 2017.

[16] W. Webber, A. Moffat, J. Zobel, and T. Sakai. Precision-at-ten considered redundant. In *Proc. SIGIR*, pages 695–696, 2008.

[17] J. Zobel and L. Rashidi. Corpus bootstrapping for assessment of the properties of effectiveness measures. In *Proc. CIKM*, pages 1933–1952, 2020.