Multi-element protocol on IR experiments stability: Application to the TREC-COVID test collection⋆

Gabriela Gonzalez-Saez, Philippe Mulhem, Lorraine Goeuriot and Petra Galuščáková

Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG. Grenoble, France.

Abstract
The evaluation of information retrieval systems is performed using test collections. The classical Cranfield evaluation paradigm is defined on one fixed corpus of documents and topics. Following this paradigm, several systems can only be compared over the same test collections (documents, topics, assessments). In this work, we explore in a systematic way the impact of similarity of test collections on the comparability of the experiments: characterizing the minimal changes between the collections upon which the performance of IR system evaluated can be compared. To do that, we create pair instances of sub-test collections from one reference collection with controlled overlapping elements, and we compare the Ranking of Systems (RoS) of a defined list of IR systems. We can then compute the probability that the RoS are the same across the sub-test collections. We experiment with our framework proposed on the TREC-COVID collections, and two of our findings show that: a) the ranking of systems, according to the MaP, is very stable even for overlaps smaller than 10% for documents, relevance assessments and positive relevance assessments sub-collections, and b) stability is not ensured for MaP, Rprec, Bpref and ndcg evaluation measures even when considering large overlap for the topics.

Keywords
Comparability, Rank of systems

1. Introduction

Classical evaluation of information retrieval systems follows the Cranfield paradigm, based on the use of a common test collection to evaluate all the systems in comparison. One evaluation is then a snapshot of the behaviour of systems on a fixed dataset. In a way to study in the large the quality of a system, a common approach is to test a system on several test collections. Testing a system on several test collections assesses then the system's ability to answer diverse information needs and to cope with various type of datasets. However, differences between test collections, according to the content, the structure, or the way the collection has been compiled can have a huge impact on the results of a single's system evaluation [1]. In this paper, we study the stability of a system's evaluation across varying datasets by creating multi-dimensional variations of a test collection.

The question we focus may have an impact to other fields of IR than the evaluation:

- knowing how a test collection evolution affects the stability of the systems evaluation

measures may be of great value for Web search engines, where documents, topics and relevance assessments are constantly changing ;

- deep learning approaches for IR [2] commonly use N-fold validation techniques during the training. The question to use accurate folds for such validation has to be answered. Our proposal allows to control what fold to use regarding documents, topics, assessments of a test collection.

This problem has been partly studied in the state of the art but, to our knowledge, a comprehensive study on the documents, topics and assessments does not exist yet. Our proposal is then a) to create, from one test collection, multiple controlled pairs of sub-collections according to documents, topics and assessments, and b) to study the stability of the ranking of several systems between theses sub-collections. With this, we are then able to explore the impact of these pairs. We show that the topics dimension has a greater impact than the assessments and the documents dimensions.

In the following, we present first the state of the art in section 2, before detailing our proposal in section 3. In section 4, we present the experimental setting. Section 5 details the results, before the discussion in part 6 and the conclusion.

2. Related works

We present here existing works that focus on the impact of variation of test collections on the evaluation of the quality of systems.

Classically, different test collections are used to measure and test the reproducibility of system results [3]. To do that, the same systems have to actually be applied on each collection, otherwise they cannot be compared. The problem of similarity between corpus of documents has been applied for the transfer of relevance assessments across test collections [4], but not for the comparison of systems across such collections. Such works do not tackle the problem to compare systems across test collections.

Few works are focusing specifically on the impact of topics in test collections from a ranking of systems perspective. Robertson and Kanoulas [5] find that topics are *not all equal* when evaluating document retrieval, but they do not provide answers on how make use of their findings.

Other works study the impact of corpus, assessments or topics changes on the performance of the systems. Sanderson et al. [6] and Ferro and Sanderson [7] show that evaluations conducted on several sub-collections (splits of the document corpus) lead to substantial and statistically significant differences in the relative performance of retrieval systems. In the same line, Ferro and Sanderson [8] and Voorhees et al. [9] model the performance metrics as several factors that represent the effects of the system and test collection used in the evaluation. They found significant effects in the evaluation from the topics, documents and the components of systems used. Recently, Zobel and Rashidi [10] have shown the experimental variability, using bootstrapping techniques on the corpus of documents across different performance metrics. These works consider only random corpus splits, and they do not focus specifically, as we do here, on detecting when the same ranking of systems are achieved.

Recent work of Rashidi et al. [11] detailed the impact of three document corpus characteristics: documents length, document source, and high/low rank of the document. They control the test collection splits by a "meld factor" of the characteristics (level of difference between the splits) and they show that each characteristic impacts differently the performance of the systems. However, this work does not define thresholds upon which we can rely to define similar collections. In conclusion, the state of the art shows that the performance of the systems is affected by changes in the test collection, but to our knowledge no focus was made on finding when collections can be assessed as comparable according to changes of several of their features.

Compared to the state of the art, we investigate here how do the changes, not limited to document corpus but also including topics and assessments, may affect the comparability of sets of systems. Moreover, we investigate to what extent it is possible to compare systems evaluated in changing test collections. Our research questions are: How to quantify the difference/similarity between test collections? And what differences in the test collection do guarantee the comparability of the systems results? We hypothesize that similar test collections produce the same Ranking of Systems (RoS), similarly to Voorhees et al. [9] and Voorhees et al. [12], as a generalization of the A-vs-B-comparison from Rashidi et al. [11] to more than 2 systems. We investigate if there exists a measurable level of similarity between the elements of test collection (documents, topics, and assessments) upon which the sub-collections are considered comparable.

3. Comparing Test Collections

Our goal in this paper is to propose a way to estimate to which extent changes in a test collection implies changes in the ranking of systems tests on it. Such problem is important to solve, as it may be used to evaluate the stability of a test collection.

Before going into detail into the framework that we build, we first define formally what are comparable test collections.

Definition 1. *Two test collections T_1 and T_2 are comparable according to an evaluation measure m, if for a given set S of information retrieval systems, the ranking of the systems in S according to m is the same in T_1 and T_2.*

The performance of systems evaluated in one test collection depends on the features of this test collection [7, 6, 9]: systems may not have the same ranking across several test collections. A test collection T is classically defined by the following *components*: a set of topics $T.Q$, a set of documents $T.D$, a set of the Relevance Assessments (RA) $T.RA$ (triplets $\in T.D \times T.Q \times \{0, 1\}$ for binary relevance assessments) and a set of evaluation measures $T.M$.

Based on these components, we will study the impact of changes using the chosen *elements*, i.e., components or subsets of them, The idea of using elements that may differ from the components allow us to study more closely specific parts of the test collections: we are then using an approach similar to Rashidi et al. [11] and Sanderson et al. [6]. We study the comparability of test collections based on changes according to these elements, assuming a single fixed evaluation measure m from $T.M$ (see Definition 1). In order to evaluate the stability of IR systems, we create artificial test sub-collection pairs, built from T. These pairs of sub-collections allow us to study controlled overlaps between the elements.

Definition 2. *For one element e under consideration from a test collection T, let T_1 and T_2 be sub-collections of T that differ only by the element e, with $|T_1.e| = |T_2.e|$, all the other elements being equal. The overlapping level o of T_1 and T_2 is defined as $o = \frac{|I|}{|T_1.e|}$ with $I = T_1.e \cap T_2.e$.*

Such overlap, in [0,1], denotes the *similarity* between the elements $T_1.e$ and $T_2.e$. If $|T_1.e| = |T_2.e|$ over several overlapping levels. We force the size of the varying elements to be constant across the different overlapping levels to avoid potential biases due to differences in size of the elements considered. When studying the impact of one element, the others are impacted in a way to ensure consistency in a test collection. In our case, a *consistent* test sub-collection T_i from a collection T, with respect to the elements $T.D$, defines $T_i.RA \subset T.RA$, so that $T_i.RA = \{(d, q, a)|(d, q, a) \in T.D, d \in T_i.D, q \in T_i.Q\}$.

According to this, we define an experimental protocol that assesses the comparability of test collections according to one element e of one test collection.

Definition 3. *The protocol that studies the threshold of comparability for one test collection T, one evaluation measure m, for a given set of overlap values O, according to one similarity measure for the ranks of systems Δ applied on i sub-collections pairs for a set of systems S and a threshold ρ, is defined as follows:*

- *for each overlapping level $o \in O$, build n controlled overlapping pairs ($T_{n,1}$,$T_{n,2}$) of subsets of T according to the element e ;*
- *compare the RoS of a given set of systems in S evaluated on one side on $T_{n,1}$ and on the other side on $T_{n,2}$, using m, is done using the ranked lists $L_{n,1}$ and $L_{n,2}$, in a way to assess the impact of the overlapping o over the element e. This is done by a function Δ which estimates the similarity between the lists ;*
- *compute the probability $p_{e,o}(\Delta(L_{.,1}, L_{.,2}) >= \rho)$ for which the $\Delta(L_{i,1}, L_{i,2})$ is larger than ρ for an overlap o on a given element e, for $i \in [1, n]$. This may be computed using classical maximum likelihood estimate on the n pairs generated, i.e.*
  $$p_{e,o}(\Delta(L_{.,1}, L_{.,2}) >= \rho) = \frac{|\{i|i \in [1,n], \Delta(L_{i,1}, L_{i,2}) >= \rho\}|}{n}.$$

Following this protocol, we are able to define the minimal overlap for which the probability of having the same RoS is large enough.

As an example, if we consider the element $T.D$, the protocol considers the number of documents overlapping in sub-test collections. For $T.Q$ we consider the number of common topics in both test collections. For $T.RA$ we extract the proportion of common judged documents for each topic. In a way to obtain robust results, n has to be large enough (typically greater than 50 [13]).

In this part, we defined a protocol for calculating the impact of the overlap between different elements on the retrieval results. To show the feasibility of our proposal, we now show the experimental results using the TREC-COVID collection.

4. Experiments

We use the complete TREC-COVID test collection [14] to measure the comparability of the RoS. TREC-COVID[1] is composed by 191,160 different documents, 50 topics and 69,318 assessments (1,386 assessments per topic in average). This collection is modern (created in 2020), and is reasonably large. The documents, as well as the topics, are related to COVID. We chose not to use the original rounds TREC-COVID collection because it is a residual test collection: the systems evaluation measures are not comparable because the relevant documents from the previous rounds are removed from the following ones, which affects the performance of the systems.

For each overlapping level o, we create 50 test collection pairs, so $n = 50$. We evaluate 10 classical IR systems with and without Bo1 relevance feedback: S = {BM25, DLH, DirichletLM, PL2, TF_IDF, BM25_Bo1, DLH_Bo1, DirichletLM_Bo1, PL2_Bo1, TF_IDF_Bo1}, implemented using PyTerrier [15], with default parameters values. Similarly to Sanderson et al. [6] and Voorhees [16], we use the Kendall Tau similarity coefficient between different Ranking of Systems as the Δ function: it measures the minimum number of pairwise adjacent swaps required to create the same ranking. For a given set of 50 sub-collection pairs, we average the Kendall Tau coefficients. The threshold ρ of comparability (see Definition 3) between RoS is 90% [16]. The overlapping values tested are, in percentages $O = \{5, ..., 100\}$ by steps of 5%. The following classical IR evaluation measures are reported: MaP, Rprec, Bpref and ndcg.

According to the state of the art, we define the elements for a test collection T in a non-exhaustive way as follows:

- $T.D$: the set of documents (similar to Sanderson et al. [6]). For $T.D$, 5% of overlapping between to sub-collections corresponds to 4,779 documents (2.5% of the whole collection);
- $T.Q$: the set of topics (following Robertson and Kanoulas [5]). For $T.Q$, the full topic set is composed by 25 topics. We vary the number of overlapping topics from 4% to 96% (at each step we include one more topic);
- $T.RA$: the set of assessments. This is somewhat related to the idea of Yu et al. [4]. $T.RA$ contains 34,659 assessments: 5% of assessments corresponds to 1,733 assessments;
- $T.RA_R$: in a way to show that our protocol is able to cope with subsets of the components, we study the subsets of the assessments which are relevant. Namely, we study the set $T.RA_R \subset T.RA$ such that
$T.RA_R = \{(topic, document, assessment) \in T.RA, \text{ so that } assessment = 1\}$ assuming binary relevance values. Similar question was studied by Ferro and Sanderson [7]. The full set of relevant assessments $T.RA_R$ contains 13,332 assessments. We vary the number of overlapping RD from 5% to 95%, an increase of 5% of the number of overlapping relevant assessments corresponds to 667 assessments.

Next, we will show the impact of each of these elements considered independently on the TREC-COVID collection.
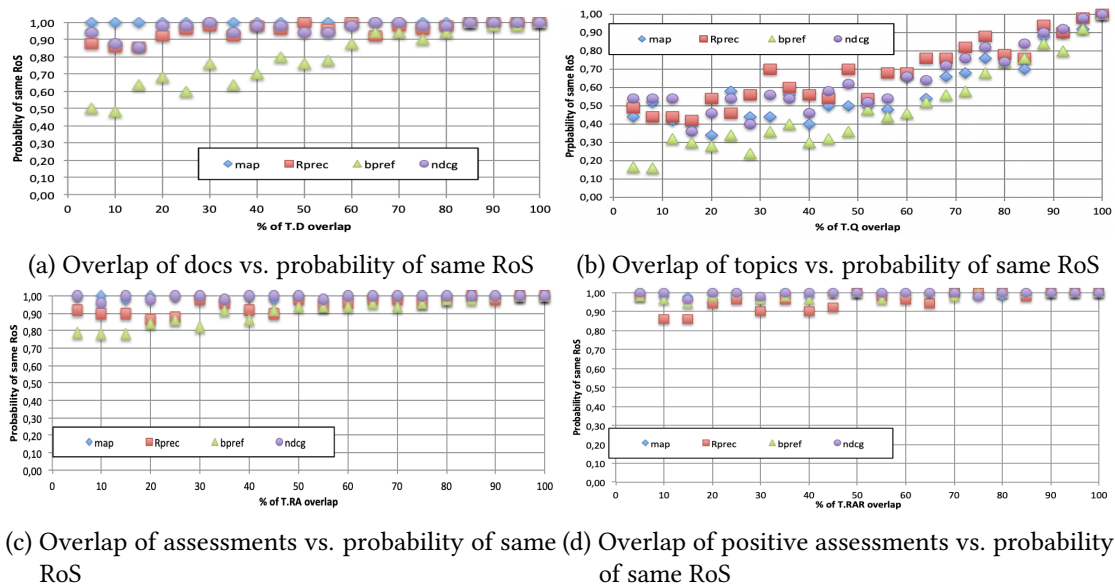
(a) Overlap of docs vs. probability of same RoS

(b) Overlap of topics vs. probability of same RoS

(c) Overlap of assessments vs. probability of same RoS

(d) Overlap of positive assessments vs. probability of same RoS

Figure 1: Similarity of sub-test collections on one element in $\{D, Q, RA, RA_R\}$ versus probability of same RoS.

5. Results

The Figure 1 presents the probabilities of having the same ranking of the considered systems, respectively on $T.D$, $T.Q$, $T.RA$ and $T.RA_R$.

For the evaluation measures presented in Figure 1a regarding the overlap of $T.D$, we see, as expected, that the probability of similar RoS increases as the overlaps increase. In this figure, the MaP is very stable, as $p_{T.D,o}(\Delta(L_{.,1}, L_{.,2}) >= 90\%) = 1$ for each document overlap o greater than 5%. This underlines the fact that the corpus is very focused on one topic area (COVID-related documents). The Bpref evaluation measure (green triangle) has the lower probability for all overlaps tested: the probability of having the same RoS in larger than 90% only for overlaps greater than 65%. A detailed look on the 50 runs for the overlaps of 5% shows that: a) on average, the average for MAP values is 0.998 and average for the Rprec values is 0.886: both values are very large, and b) overall 50 of the 50 MAP values are larger than 0.9 where only 24 of the 50 Bpref values are above 0.9. Mainly the large difference comes from the threshold ρ: if $\rho = 0.8$, then the Rprec and MAP behave similarly.

The Figure 1b, focusing on topics, exhibits expected behaviors: the more the overlap of topics, the higher the probability of the same RoS. However, we see that the MaP is not as stable as Rprec: the Rprec has almost most of the time the higher probability of RoS similarity. Here, the Bpref is still the least stable measure.

Figures 1c and 1d, corresponding to the overlaps of the assessments and the overlaps of the positive assessments respectively, are the most flat ones for all the evaluation measures. MaP and ndcg always reach the probability of 1 for each overlap values considered. For $T.RA$ the

---

less stable measure is Bpref and for $T.RA_R$ it is Rprec.

The low slopes in the graphs mean a stable comparability across overlaps, and the intercept is interpreted as the projected minimum comparability value when there is no overlapping elements. The metric with the lowest slope and highest intercept, for three of the four features, is MaP. The comparability of test collections for the RPrec is higher than for MaP only for the Topics experiments. The metric with highest slope and lowest intercept is Bpref in all the analyzed elements, leading to a larger sensitivity of this measure.

Figure 2 presents under a radar view the lower overlap values for which the probability of having a similarity of RoS larger than 0.9 is equal to 1. The overlap value of 100 means that we did not get any full stability for a partial overlap considered. We see that the only element for which we are not able to get any stability on the TREC-COVID collection is $T.Q$ (i.e., the topics splits).

6. Discussion

From the Figures 1a, 1c and 1d and 2, we see that the MaP is able to cope with the differences between sub-collections. When considering the documents $T.D$, relevance assessments $T.RA$ and relevance assessments $T.RA_R$ that are relevant, the MaP is able to cope even with very low overlaps (5%). So, MAP gives us a good comparison over two completely different collections, for the same set of topics. The only element for which none of the evaluation measure is reaching a probability of 1 is $T.Q$, reflecting the fact that there is a low stability of the ranking of systems across very similar collections according to the topics. This finding is in contradiction with Carterette et al. [17], in which the authors found that totally separated sets of topics led to the same ranking of systems: it is possible that our smaller set of considered topics is the reason. Further studies have to be conducted to validate this hypothesis.

In Figure 1a, we see that the MaP, Rprec and ndcg are very high for each overlap considered. This may be explained by the fact that the corpus focuses on one area of the topics related to Covid, and that there is a large redundancy between the documents. It might be just the case that MaP is just stable even across different collections, but the table X of Fang et al. [1] shows that this hypothesis does not hold. We guess that this behavior comes from the fact that the corpus (and topics) are related to one quite specific domain. The probabilities for the Bpref measures are much lower that the probabilities for the other measures, especially for the small overlaps (below 60%), Our guess is that the splits of documents impact the assessments (if a document is removed from a split, it is also removed from the assessment file): as there are less relevant documents, there are more chances to have non-relevant documents retrieved before relevant ones, leading to lower the Bpref values.

The Figure 1b exhibits the large impact of the overlapping on topics: as the topics do not behave similarly, the non-overlapping topics lead to very different ranking of systems. We see on the left part of Figure 1b that the probability of similar ranking for bref is very low (for instance 0.16 for an overlap of 8%). This findings is mainly caused by the fact that the topics in a test collection are classically built manually, and are supposed to be very different (as shown in Figure 2 of Banks et al. [18] for instance). Such constrain does not hold for documents, for which there is no redundancy-check achieved. Going further with our protocol by fixing
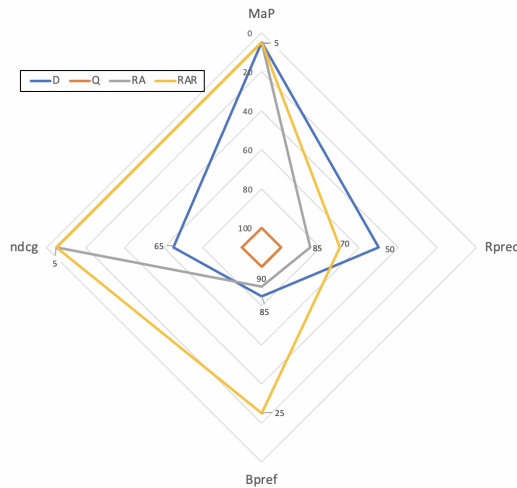
Figure 2: Radar view of $\arg\min_{o\in[5\%,100\%]}[p_{T.e,o}(\Delta(L_{.,1}, L_{.,2}) >= 90\%) = 1]$ for $e \in \{D, Q, RA, RA_R\}$.

$\rho = 0.8$ the MaP measure is able to get similar rankings for an overlap of 72% of topics. This shows that the the Kendall Tau values for the MaP obtained are still high.

The two elements that are considering the relevant assessments $T.RA$ and $T.RA_R$ have similar behaviors in Figures 1c and 1d: for almost all overlaps the probabilities are greater than 0.8. This shows that using only partial overlaps of assessments smaller than 50% leads to similar rankings of systems for MaP and ndcg evaluation measures. The Rprec (precision computed at the number of relevant documents for a topic) is more sensitive to the overlaps of positive assessments, as this measures is based on the number of positive relevant assessments. Rprec is especially sensitive to small overlaps of assessments because the positive assessments form roughly 1/3 of all available assessments.

As presented above, Figure 2 describes graphically the minimal overlapping, for each element and each evaluation measure considered, that leads to a probability of 1 to get a 90% similarity between ranking of systems: the larger the area, the smaller the overlap. From this Figure, we see on one side that the Rpec and Bpref evaluation measures are more sensitive to the overlaps, whatever element we consider, and on the other side that the topics are very sensitive to any overlap ratios (orange line, surface on 100% for all evaluation measures). For this Figure, we conclude that for MaP and ndcg evaluation measures, having a test collection composed of only 5% of the assessments and 5% of the positive assessments lead to the same ranking of systems: such result may relax the need for N-fold validation in the case of evaluation of learning-based IR systems, or may constrain an N-fold validation experiment by using splits that lead to same ranking of systems.

Our findings are quite consistent results with table 1 of Ferro and Sanderson [7] on the TREC Adhoc T07 and T08 test collections: MaP and ndcg measures are more sensitive to the topics splits than to the document splits.

7. Conclusion

We have presented a protocol that supports the study the impact of *elements* of a test collection on the ranking of systems. Our proposal formalizes this crucial part which needs be defined to perform such study. We then applied the protocol on the TREC-COVID collection. The outcomes of this study show that the documents and topics are the considered elements that have the most impact on the stability of the ranking of systems. We also showed that each evaluation measure behaves very differently in our experiments. A future work could extend our proposal to study sub-collections overlaps across several measures.

As a future work, we also would like to extend our proposal to be able to consider jointly several elements, so that we may detect dependencies between elements, as in [7]. Refining some parts of the protocol will be also considered in the future, as limiting the overlaps on sets does not cover semantic aspects of documents and topics. The study achieved is limited to one test collection, and we plan to asses the stability of the results on other test collections, especially collections with a wider and more general range of topics.

Acknowledgments

References

[1] H. Fang, T. Tao, C. Zhai, Diagnostic evaluation of information retrieval models, ACM Trans. Inf. Syst. 29 (2011). URL: https://doi.org/10.1145/1961209.1961210. doi:10.1145/1961209.1961210.

[2] B. Mitra, N. Craswell, An introduction to neural information retrieval, Foundations and Trends® in Information Retrieval 13 (2018) 1–126. URL: http://dx.doi.org/10.1561/1500000061.

[3] T. Breuer, N. Ferro, N. Fuhr, M. Maistro, T. Sakai, P. Schaer, I. Soboroff, How to Measure the Reproducibility of System-oriented IR Experiments, SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (2020) 349–358.

[4] R. Yu, Y. Xie, J. Lin, Simple techniques for cross-collection relevance feedback, in: European Conference on Information Retrieval, Springer, 2019, pp. 397–409.

[5] S. E. Robertson, E. Kanoulas, On per-topic variance in ir evaluation, in: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 891–900. URL: https://doi.org/10.1145/2348283.2348402. doi:10.1145/2348283.2348402.

[6] M. Sanderson, A. Turpin, Y. Zhang, F. Scholer, Differences in effectiveness across sub-collections, ACM International Conference Proceeding Series 2006 (2012) 1965–1969. doi:10.1145/2396761.2398553.

[7] N. Ferro, M. Sanderson, Sub-corpora impact on system effectiveness, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 901–904.

[8] N. Ferro, M. Sanderson, Improving the accuracy of system performance estimation by using shards, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 805–814.

[9] E. M. Voorhees, D. Samarov, I. Soboroff, Using replicates in information retrieval evaluation, ACM Transactions on Information Systems (TOIS) 36 (2017) 1–21.

[10] J. Zobel, L. Rashidi, Corpus Bootstrapping for Assessment of the Properties of Effectiveness Measures, International Conference on Information and Knowledge Management, Proceedings (2020) 1933–1952.

[11] L. Rashidi, J. Zobel, A. Moffat, Evaluating the Predictivity of IR Experiments, SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (2021) 1667–1671.

[12] E. M. Voorhees, I. Soboroff, J. Lin, Can old trec collections reliably evaluate modern neural retrieval models?, 2022. `arXiv:2201.11086`.

[13] E. M. Voorhees, C. Buckley, The effect of topic set size on retrieval experiment error, in: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '02, Association for Computing Machinery, New York, NY, USA, 2002, p. 316–323. URL: https://doi.org/10.1145/564376.564432. doi:`10.1145/564376.564432`.

[14] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, Trec-covid: constructing a pandemic information retrieval test collection, in: ACM SIGIR Forum, volume 54, ACM New York, NY, USA, 2021, pp. 1–12.

[15] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using pyterrier, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, 2020, pp. 161–168.

[16] E. M. Voorhees, Evaluation by highly relevant documents, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001, pp. 74–82.

[17] B. Carterette, J. Allan, R. Sitaraman, Minimal test collections for retrieval evaluation, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 268–275.

[18] D. Banks, P. Over, N.-F. Zhang, Blind men and elephants: Six approaches to trec data, Inf. Retr. 1 (1999) 7–34.