Towards the evaluation of Information Retrieval Systems on evolving datasets with Pivot systems

Gabriela Nicole González-Sáez[1⋆], Philippe Mulhem[1], and Lorraine Goeuriot[1]

Univ. Grenoble Alpes, CNRS, Grenoble INP⋆⋆⋆, LIG, 38000 Grenoble, France
gabriela-nicole.gonzalez-saez, philippe.mulhem,
lorraine.goeuriot@univ-grenoble-alpes.fr

Abstract. Evaluation of information retrieval systems follows the Cranfield paradigm, where the evaluation of several IR systems relies on a common *evaluation environment* (test collection and evaluation settings). The Cranfield paradigm requires the evaluation environment (EE) to be strictly identical to compare system's performances. For those cases where such paradigm cannot be used, e.g. when we do not have access to the code of the systems, we consider an evaluation framework that allows for slight changes in the EEs, as the evolution of the document corpus or topics. To do so, we propose to compare systems evaluated on different environments using a reference system, called *pivot*. In this paper, we present and validate a method to select a pivot, which is used to construct a correct ranking of systems evaluated in different environments. We test our framework on the TREC-COVID test collection, which is composed of five rounds of growing topics, documents and relevance judgments. The results of our experiments show that the pivot strategy can propose a correct ranking of systems evaluated in an evolving test collection.

Keywords: Information retrieval evaluation · Test collection · Result delta.

1 Introduction

Classical evaluation of Information Retrieval (IR) Systems is made using a common test collection: a set of documents, a set of queries, and a set of relevance judgments. Evaluation campaigns aim at building such test collections and help to improve search systems. At the end of an evaluation campaign, a Ranking of Systems (RoS) based on their performances is built. A search task defined in an evaluation campaign dictates the topics creation, the corpus of documents, the relevance judgments (such as pooling parameters, guidelines to measure the relevance), and the metrics used to rank systems. All these elements define an *Evaluation Environment (EE)*. Changes in the EEs may lead to changes in the results of the systems. Sanderson et al. [7] has shown that evaluating IR systems on different subsets of the document collection affects the performance of

⋆⋆⋆ Institute of Engineering Univ. Grenoble Alpes
⋆ Corresponding author

the system. Hence, results obtained on varying document collections are not comparable.

In the Web search, the topics searched and the set of documents continuously evolve. In such settings, getting a regular update on a system's performances is very challenging. Constant evolution of the test collection makes it nearly impossible to apply a classical offline evaluation following the Cranfield paradigm. We address the case where the different versions of the system are no longer available, therefore we have different systems evaluated in different test collections without the possibility to re-evaluate older versions of itself.

How can we compare a set of systems evaluated on evolving versions of the evaluation environment? We hypothesize it is possible to create a ranking of systems evaluated in different EEs by measuring the difference between the evaluated systems and a pivot system that is evaluated on all the EEs.

This paper presents a method to select a pivot system from several candidates to create a correct ranking of systems evaluated on different test collections. Our experiments use the TREC-COVID test collection, that ran in five rounds. We test the pivot strategy over one round of the TREC-COVID and select one pivot to compare all the systems taking part in the five rounds.

2 State of the art

We present now works on three topics related to our study: section 2.1 focuses on comparing systems on dynamic test collections; section 2.2 details the impact of different evaluation settings on the performance of the systems; and section 2.3 presents works that evaluate systems in changing test collections.

2.1 Dynamic test collections

One of the most important constraints of the Cranfield evaluation is the use of a common test collection for all the systems in comparison. Assessing the quality of Web search needs a repeated or continuous evaluation given incremental document collections [4]. We present two papers that describe evaluation methods tackling the problem of continuous evaluation over evolving test collections.

Soboroff [8] addresses the need to create a dynamic test collection to evaluate the web search in realistic settings. Their experiments use a changing and growing document collection, with a fixed set of topics and relevant judgments. [8] shows that it is possible to compare the performance of systems from different versions of the test collection despite the decay in relevance data due to the changing document collection. According to the Bpref evaluation measure, the rankings of the systems in different versions of the test collection are similar to the RoS of the initial version of the test collection, leading to assess that systems are comparable across these versions. The difference with our proposal is that we compare different system evaluated in different versions of a test collection.

Tonon et al. [10] proposed a method to evaluate IR systems iteratively on the same test collection, increasing the judged documents according to systems

that did not take part in the pool of documents. They focus on the bias on systems introduced by being included or not in the pooling systems. Such a bias makes it impossible to compare system accurately, because the test collection construction penalizes systems that did not take part in the pooling that might be more effective than systems that took part in the pool but retrieve different results [17]. Therefore, the pooling strategy must be considered in the EE, as being included in the pool of documents or not affects the evaluation of the system.

These papers rely on the need to create alternative methods to incorporate incremental test collections on the evaluation of IR systems, as the proper environment of the web search. Our proposal does not need to incorporate new resources into the test collection to compare systems across evolving EEs. Also, we integrate changes on any of EE elements, while guaranteeing those changes keeps the same RoS, then the EEs are comparable.

2.2 Performance on different evaluation environments

In this work we define the EE as an extended test collection, that incorporates the elements involved in the IR evaluation and may affect the performance of the IR system: the document set, the topic set, the relevance judgment, the pooling strategy and the list of metrics evaluated. The papers described in this section analyse the impact of these elements on the performance measurements.

As shown in [7], evaluations conducted on different sub-collections (splits of the document corpus with the respective relevance assessments) lead to substantial and statistically significant differences in the relative performance of retrieval systems, independently from the number of relevant documents that are available in the sub-collections. Using the ANalysis Of Variance (ANOVA) model, [2] showed that changing the test collection (splits of the documents corpus) leads to varying system performances (inconsistently across metrics). In the same line, [3, 13] model the system effect and the test collection effect on the performance metrics as separated factors, they define ANOVA models and GLMMs to analyse systems performances over several test collections with the goal of improving the measurement accuracy of retrieval system performance by better modeling the noise present in test collection scores.

Such studies are not aiming at system comparison, but rather at measuring the effect of the test collection on the system performance. They provide a better understanding of the measurement of performance, but do not allow to compare two systems that are evaluated using different EEs.

2.3 Meta-analysis of IR evaluations within evolving environments

Score standardization is an evaluation method that reduces the impact of the topic's difficulty on the IR system's performance [6, 16, 11]. It consists of normalizing the performance score for a topic by its observed mean and standard deviation over a set of runs/systems [15]. Urbano et al. [11] showed that even

when the RoS between raw and standardized scores is the same, the RoS using mean scores may differ considerably.

Meta-analysis is another approach to compare the performance of systems over multiple test collections [9]. Meta-analysis consists in measuring a delta difference between one baseline and a target system, over multiple collections. This meta analysis allows the measurement of the mean difference between the systems with a confidence interval. This technique is strongly related to the measurement of the improvement across multiple test collections of a system with a specific modification that differences it from the baseline system. Our proposal addresses the problem of evaluation of different systems over evolving EEs. Therefore, the differences are not computed over one, but several retrieval systems that need to be compared. Both techniques make use of relative measurements to compare systems evaluated in different EEs. We extend this idea in our framework of evaluation with the use of a common pivot system that defines a reference to compute the relative distance between the systems' performance to rank the systems.

3 Pivot evaluation of continuous test collections

Our proposal focuses on the comparison of systems across different EEs. We assume that running a set of IR systems on two comparable EEs should give the same RoS, as showed by Soboroff [8] when RoS is built with bpref metric.

Our main goal here is to create a single RoS with systems evaluated on different (yet comparable) EEs. To get an accurate comparison of systems evaluated on varying EEs, we detail below a framework based on the difference between systems performances across comparable EEs.

3.1 Result delta definition

In this section, we present a method to measure the impact of EE variation on systems evaluation. Since we want to compare systems that are evaluated on different EEs, we cannot rely on absolute evaluation. Therefore, we propose to build our framework on differences between evaluation measures of performance, with **Result Deltas**. A result delta, $\mathcal{R}\Delta$, estimates the difference between the performance of two systems measured with a similar metric. Three kinds of $\mathcal{R}\Delta$ can be measured, according to the element that change in the evaluation task:

- $\mathcal{R}_s\Delta$: When we have two different IR systems evaluated in the same EE, as a classical IR evaluation.
- $\mathcal{R}_e\Delta$: If the same IR system is evaluated in two EEs, extracting mainly the environment effect on the system.
- $\mathcal{R}_{se}\Delta$: If both EEs and systems are different.

$\mathcal{R}_{se}\Delta$ can hardly be measured, as the two systems are not directly comparable: both the EEs and the systems are different. To get an estimation of this measure, we propose to use a reference system, called **Pivot system**, which

would be evaluated within the two EEs considered. $\mathcal{R}_s\Delta$ would be computed between each system and the pivot within each EE considered. Finally, both $\mathcal{R}_s\Delta$ can be used to compute $\mathcal{R}_{se}\Delta$ and compare the two systems over the two EEs. The result delta value is measured using the relative distance between the pivot system and the evaluated system S_1:

$$\mathcal{R}_s\Delta(Pivot, S_1, EE_1) = \frac{M(S_1, EE_1) - M(Pivot, EE_1)}{M(Pivot, EE_1)} \quad (1)$$

Given a metric $M(S, EE)$ that evaluates the performance of a system S in a evaluation environment EE, we want to compare S_1 evaluated in EE_1 and S_2 evaluated in EE_2 (being comparable EEs). System performances are measured with $M(S_1, EE_1)$ and $M(S_2, EE_2)$. In order to compare S_1 and S_2, using a pivot system will help relating the systems across the EEs by comparing $M(S_1, EE_1)$ with $M(Pivot, EE_1)$ as $\mathcal{R}_s\Delta(Pivot, S_1, EE_1)$ and $M(S_2, EE_2)$ with $M(Pivot, EE_2)$ as $\mathcal{R}_s\Delta(Pivot, S_2, EE_2)$. According to the EE comparability assumption, the ranking of systems should be the same in both EEs. As an illustration, if $\mathcal{R}_s\Delta(Pivot, S_1, EE_1) > \mathcal{R}_s\Delta(Pivot, S_2, EE_2)$ then, $M(S_1, EE_1) > M(S_2, EE_1) \wedge M(S_1, EE_2) > M(S_2, EE_2)$.

3.2 Pivot selection strategy

The key point in our proposal lies in the choice of the pivot. To assess the quality of a pivot, we study whether the use of a given pivot to compute the result delta measures of systems evaluated on different EEs allows to obtain a *correct* RoS. The pivot-based RoS is validated using a ground truth reference RoS.

A system P is considered to be a good pivot according to a reference EE EE_{ref} if, using the result deltas measured with P to compare different systems evaluated across various EEs (EE_{splits}, a split of the EE_{ref}) we can get the same RoS as the reference one (got on EE_{ref}). To evaluate the correctness of a pivot, we compare:

- RoS_{ref} a reference RoS according to a ground truth, namely the official RoS in an evaluation campaign based on the whole corpus and topic set, and
- RoS_{pivot}, it is artificially built from two EEs created by splitting the whole corpus and/or whole topics set, and splitting the compared systems on these two EEs. RoS_{pivot} uses the result deltas of the pivot under consideration.

If the two rankings are the same, this means that the pivot is able to correctly support the indirect comparison of systems. To evaluate the correctness of a pivot, we measure the Kendall's Tau similarity between RoS_{pivot} and RoS_{ref}.

The correctness of a pivot must be compared to a baseline. To do that, we define a $RoS_{baseline}$ that is constructed under the same EEs created for the RoS_{pivot}. The $RoS_{baseline}$ orders the absolute performance values of the two system sets evaluated on each EE split. Then, we measure the similarity between the $RoS_{baseline}$ and RoS_{ref}. We expect higher similarity values using the pivot strategy than with the absolute performance values.

To assess the quality of a pivot, we must repeat the experiment: for instance, we may split the set of document Doc times, and the set of topics Top times, creating $Doc \times Top$ splits of the EE_{ref}. With these multiple experiments we build distributions of the correctness achieved by a pivot, and assess statistical significance of differences with the baseline. To evaluate the pivot strategy on systems already implemented, we filter the runs keeping only the documents and topics of the corresponding EE split. This process is described and validated in the work of Sanderson [7].

4 Methodology

Here, we describe how the strategy presented in section 3.2 is implemented: firstly we describe the test collection we are using in section 4.1, then we present how we validate the pivot strategy in section 4.2.

4.1 TREC-COVID evolutionary collection

The data used to validate our proposal is the TREC-COVID collection [12], created in the COVID-19 pandemic by NIST and over 60 teams and 500 runs. The created test collection is available in TREC-COVID webpage[1]. TREC-COVID is a continuous test collection, organised on five rounds, where each round is composed of a specific release of CORD-19[2] documents collection [14], a set of incremental topics and a set of relevant judgments. CORD-19 is composed of an incremental list of scientific papers related to COVID-19. Topics correspond to information needs of clinicians and biomedical researchers during the COVID-19 pandemic. Round 1 has topics is 30, and five topics are added at each round, leading to 50 topics at round 5. The relevance judgments are repeated at each round for all the topics and the non-judged documents.

While the challenge did not compare the results from different rounds, we see the opportunity to apply our framework to this incremental dataset, creating round-based splits to validate the pivot method, then we create a result delta rank that includes the systems that took part on the five rounds.

4.2 Evaluation method

The pivot selection strategy is validated on one round of the campaign, with 50%-50% splits of the topics and documents sets, over the set of participating systems. Fig.1 a) shows an example with five systems (S1,..., S5), the splits are EE1 (in orange) and EE2 (in blue). Then a ranking of the 5 systems is built using the result deltas (RD in Fig. 1) (S2, S4 and S5 in EE1, S1 and S3 in EE2), namely RoS_{pivot}. This ranking is then compared, using Kendall's tau with the reference ranking RoS_{ref} (i.e., the ranking of all the systems participating on

[1] https://ir.nist.gov/covidSubmit/archive.html
[2] https://www.semanticscholar.org/cord19

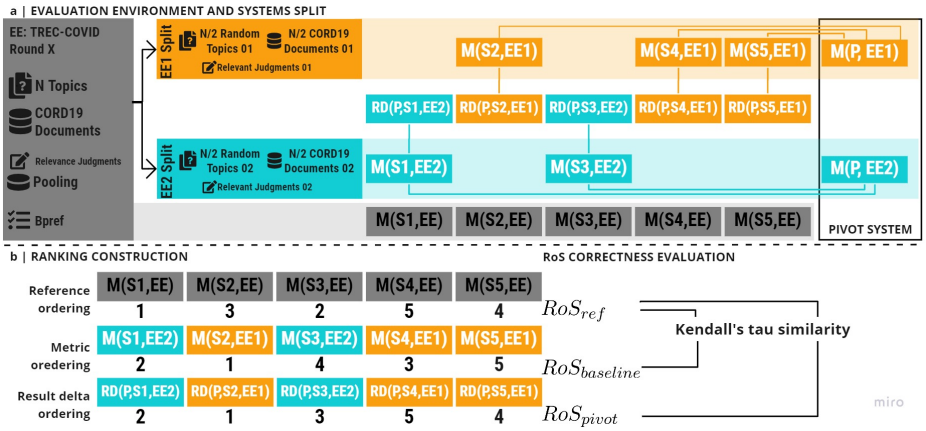

Fig. 1: a) EE split and result delta with pivot P. b) Similarity of RoS created

the corresponding round sorted by his performance metric.) and to the baseline ranking $RoS_{baseline}$. A good pivot should generate a ranking closer to the reference ranking than the baseline.

We use six pivot implementations, that are commonly used as baseline systems, using Terrier [5] system, BM25, DirichletLM, TF_IDF with default parameters, without and with pseudo relevance feedback (RF) using default parameters (DFR Bo1 model [1] on three documents, selecting 10 terms). We evaluate the correctness of these six candidate pivots. We run the experiment in 10 splits of documents and 10 splits of topics (leading to an overall of 10×10=100 pairs of EEs). Finally, per each EE pair we have six pivot-based RoS and one baseline RoS that are compared to the reference RoS with Kendall's tau similarity (Fig.1 b). The metric of performance used is BPref, one of the official metrics used on the campaign. Bpref is robust to incomplete relevance judgments, then it is appropriate to our experiment due to the split of documents.

Once validated, we can rank the result deltas of all the participant systems measured by the selected pivot, to create a final RoS that includes the 500 runs submitted on the five TREC-COVID rounds.

5 Experiments and results

In this section, we present the results of two experiments. In the first one, we aim at *validating the pivot strategy*, by measuring the correctness of the ranking obtained with the different pivots presented in Section 4.2. In the second one, we apply the validated pivot strategy on a dynamic test collection (presented in Section 4.1) to observe the RoS obtained across several rounds.

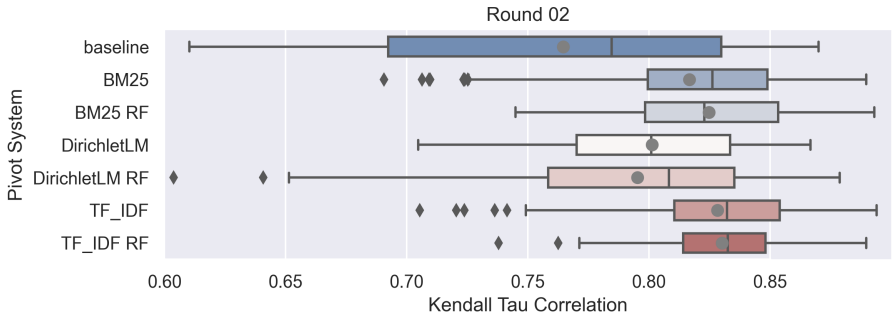

Fig. 2: Boxplot of similarity between RoS_{pivot} and RoS_{ref}, and between $RoS_{baseline}$ and RoS_{ref} on the first line. Gray circle represents mean value.

5.1 Pivot selection

We describe our pivot selection on the round 2: EE_{split} is a half of round 2 test collection. We compare the correctness of the created pivot-based RoS with the reference RoS that considers the round's full set of documents and topics.

Figure 2 shows the correctness distribution for each created RoS over 100 EE splits. Each boxplot summarizes the Kendall's tau similarity distribution measured between the pivot-based RoS (obtained with the pivot's result delta of two sets of systems, where each set is evaluated on a EE split) and the reference RoS. We compare the boxplots of the pivots versus a baseline RoS (first boxplot). We see that the correctness mean (resp. standard deviation) of the baseline RoS is lower (resp. larger) than any pivot-based RoS. This reflects a higher uncertainty of the ranking created with bpref absolute values in comparison to the rankings created using result deltas. The result deltas of TF_IDF RF formed the RoS with the highest correctness (bottom boxplot).

Table 1 summarizes the results in the five rounds of TREC-COVID, a high correctness of pivot-based RoS is repeated in the five rounds (columns). Table 1 presents the mean and standard deviation values of Kendall's Tau similarity be-

Table 1: Mean± std. dev. of the similarity values between RoS and RoS_{ref}. In **bold** the higher similarity value. '*' if distribution difference is statistical significant from $RoS_{baseline}$ (Kolmogorov-Smirnoff test, p-value<0.05).

RoS	round1	round2	round3	round4	round5
Baseline	*0.819±0.06*	*0.765±0.08*	*0.857±0.05*	*0.837±0.08*	***0.892±0.04***
BM25	0.837±0.04	0.817±0.04*	0.87±0.03*	0.886±0.03*	0.883±0.05
BM25 RF	0.844±0.03*	0.825±0.04*	0.880±0.03	0.884±0.03*	0.882±0.04
DirichletLM	0.841±0.03*	0.801±0.04*	0.865±0.04	0.870±0.05	0.880±0.05*
DirichletLM RF	0.827±0.05	0.795±0.06*	0.840±0.06	0.873±0.05	0.841±0.06*
TF_IDF	**0.852±0.02***	0.828±0.04*	0.887±0.03*	**0.895±0.03***	0.891±0.04*
TF_IDF RF	0.846±0.03*	**0.830±0.03***	**0.890±0.03***	0.888±0.03*	0.883±0.05*

Table 2: Bpref mean performance with complete test collection for each round

run	round1	round2	round3	round4	round5
Participants mean	0.31 ±0.12	0.36 ±0.10	0.43 ±0.15	0.48 ±0.13	0.44 ±0.15
BM25	0.3965	0.3691	0.4234	0.4367	0.3399
BM25 RF	0.4173	0.3757	**0.4375**	0.4365	0.3440
DirichletLM	0.3530	0.3341	0.3193	0.3316	0.2558
DirichletLM RF	0.3555	0.3116	0.3105	0.2941	0.2245
TF_IDF	0.4115	0.3733	0.4221	0.4319	0.3483
TF_IDF RF	**0.4407**	**0.3919**	0.4348	**0.4395**	**0.3548**

tween the seven created RoS (lines) and the reference RoS. The first row describes the correctness of the baseline RoS. Even when the ranking is based on absolute values, the similarity between the baseline RoS and the reference RoS is close 0.8, this high value could be related to our assumption of comparable EE (no modifications of the ranking across EEs). The standard deviation of the baseline RoS is the highest one on the first four rounds. Considering the five rounds, the pivot with the best correctness results are TF_IDF RF and TF_IDF. The similarity of these RoS and the reference RoS has the lowest standard deviation values and their distributions are significantly different from the baseline RoS similarity distribution in all rounds. In the five rounds the distribution of the Kendall's tau similarity got by sorting the systems with the result delta value measured by TF_IDF is significantly different to the similarity distribution of the rankings created with the Bpref values (denotes as * on Table 1). The distributions are significant different with 95% of confidence according to Kolmogorov-Smirnoff test, a non-parametric test useful to our experiment because the similarity data have non-normal distributions in most of the cases.

Table 2 shows the mean Bpref performance of participant runs (Participants mean row) and pivots considering the EE of reference (full set of documents and topics). TF_IDF RF is the pivot system with the highest performance in four rounds. DirichletLM RF is the pivot system with the lowest performance in all the rounds, and as is showed in table 1 the RoS created with the result deltas of this pivot achieve the worst similarity values and the biggest standard deviation on all the rounds. The pivots presented the worst bpref performance on the final round, with lower values than the participants mean performance. Only in this round the baseline RoS is more similar to the RoS of reference than any pivot-based RoS. Finally, the selected pivot is TF_IDF RF, due to its result deltas values constructed a similar to the RoS of reference. Therefore, the correctness property is achieved in more than 83% of the RoS considering the five rounds by TF_IDF RF pivot.

5.2 Exploratory experiment: Testing the pivot in real settings

The pivot strategy has proved its ability to compare systems in comparable EEs, that were created splitting in half each test collection round. Now, we are interested in apply our method in a realistic setting, with a non-artificially evolving

Table 3: Best runs of the five rounds of TREC-COVID ranked with pivot strategy

round	Best run per round	Official Bpref	Pivot result delta	Pivot rank
1	BBGhelani1	0.5294	0.2012	159
2	mpiid5_run3	0.5679	0.4491	47
3	mpiid5_run1	0.6084	0.3993	66
4	UPrrf38rrf3-r4	0.6801	0.5474	29
5	UPrrf102-wt-r5	0.6378	0.7976	1

test collection. Our purpose is to observe within a realistic setting what ranking would our method give. Therefore, We apply our method in TREC-COVID to understand if the system's performance are improving across the rounds, even when the EEs are not completely comparable (Kendall's tau similarity of the pivot's rankings across the rounds ranges between 0.6 and 0.86).

As TF_IDF RF was the pivot with the best results in the rounds, we rank the result delta of all the system that participated on the five rounds of TREC-COVID challenge with TF_IDF RF using Bpref metric.

Table 3 presents the best runs of each round of TREC-COVID campaign and their rank using the pivot-based RoS. The best run of the fifth round was twice better than the pivot system, this is the largest difference between the best run and the pivot, and it explains why the fifth round's best run is at the first place of the pivot-based RoS. The best system of round4 is UPrrf38rrf3-r4, this system is submitted by the team that also presented the best bpref system in round5 UPrrf102-wt-r5. These runs are produced by Reciprocal Rank Fusion of three systems for UPrrf38rrf3-r4[3] and four systems for UPrrf102-wt-r5[4]. As the pivot-based RoS takes in consideration the pivot performance to compare the systems across the rounds, the relative improvement of the best system in round five is biggest than the improvement of the round four's best run, then we conclude that it should expected that UPrrf102-wt-r5 have better bpef performance than UPrrf38rrf3-r4 if they were evaluated in the same round.

6 Discussion

The ranking created with Bpref absolute values shows high similarity with the reference RoS, this could be due to the high similarity in the CORD-19 documents, all the documents are scientific papers from PubMed Central (PMC), bioRxiv, and medRxiv [14]. This similarity on the document collection, lead to similar ranking even when we consider only the half of the documents of the test collection to be retrieved by the IR systems.

Ranking the systems using result delta measured with a common pivot across the EEs is better than rank the bpref absolute values. The pivot-based RoS are more certain, because the standard deviation is lower than using bpref values

[3] https://ir.nist.gov/covidSubmit/archive/round4/UPrrf38rrf3-r4.pdf
[4] https://ir.nist.gov/covidSubmit/archive/round5/UPrrf102-wt-r5.pdf

to rank. Nevertheless, not all pivots work the same. We found one pivot system which ranking have lower similarity values than the baseline RoS. DirichletLM RF is the system with the lowest Bpref performance. Likewise, the RoS with the higher similarity values is constructed using the result deltas of the pivot with higher Bpref performance on the rounds. This lead to interpret that the performance of the pivot is related to the correctness achieved by the ranking created using the pivot's result deltas. To confirm this relation we will continue our work using more systems as pivot to create the ranking, attempting to explore pivots with higher and lower performances.

In the fifth round the baseline RoS is more similar to the reference RoS than any RoS created with the pivot strategy. In this final round the performance of the pivot systems decreased and it is far from the Bpref values achieved by the participant runs. We will continue exploring the impact of the distance between the pivot performance and the mean performance of the rounds to improve the correctness of the pivot-based RoS.

After the validation of the pivot strategy on each round of the TREC-COVID test collection we can propose a pivot to measure the result deltas with all the participating systems and create a final ranking of systems. Using this RoS we can evaluate the evolution of the results in the growing test collection. The best Bpref performance systems were evaluated on the final round followed by the fourth round. Table 2 shows that the highest Bpref mean performance is achieved in the fourth and fifth rounds. Because the pivot's Bpref performance (TF_IDF RF) is lower in round5 than in round4, the system with the overall highest Bpref value (achieved in round4) is ranked in position 29 with our framework. The difference on the pivot's bpref value across the rounds might be a measure of the EE difficulty.

7 Conclusion

We have presented a framework proposal to rank systems evaluated in different evaluation environments using result deltas and pivot systems. The proposed framework is evaluated on the TREC COVID test collection by assessing the correctness of the pivot-based RoS. The results show that, using the pivot strategy we can improve the correctness of ranking of systems that were evaluated in different EEs, compared to the RoS created with bpref absolute values.

In this paper we proposed only baseline systems as pivot, because of their easy implementation that guarantee the reproducibility of our framework. We shall explore other strategies, as pivots based on the participant systems, to achieve closer performances between the pivot and the evaluated systems. With these new pivots, we will explore the effect of the pivot performance on the proposed RoS. Also, we will analyse merging the result deltas of several pivots to create a meta-pivot. Additionally, we will study further EE comparability and investigate the impact of EE changes on the evaluation framework; Finally, we will define the guidelines to create a test collection for continuous evaluation based on the characteristics of comparable EEs.

Acknowledgements. This work was supported by the ANR Kodicare bilateral project, grant ANR-19-CE23-0029 of the French Agence Nationale de la Recherche, and by the Austrian Science Fund (FWF).

References

1. Amati, G.: Probabilistic models forinformation retrieval based on divergence fromrandomness. Ph.D. thesis, Glasgow University, Glasgow (6 2003)
2. Ferro, N., Sanderson, M.: Sub-corpora impact on system effectiveness. In: Proceedings of SIGIR'2017. pp. 901–904 (2017)
3. Ferro, N., Sanderson, M.: Improving the accuracy of system performance estimation by using shards. In: Proceedings of SIGIR'2019. pp. 805–814 (2019)
4. Jensen, E.C., Beitzel, S.M., Chowdhury, A., Frieder, O.: Repeatable evaluation of search services in dynamic environments. ACM Transactions on Information Systems (TOIS) **26**(1), 1–es (2007)
5. Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research directions in terrier: a search engine for advanced retrieval on the web. CEPIS Upgrade Journal **8**(1) (2007)
6. Sakai, T.: A simple and effective approach to score standardisation. In: Proceedings of ICTIR'2016. pp. 95–104 (2016)
7. Sanderson, M., Turpin, A., Zhang, Y., Scholer, F.: Differences in effectiveness across sub-collections. In: Proceedings of CIKM'2012. pp. 1965–1969 (2012)
8. Soboroff, I.: Dynamic test collections: measuring search effectiveness on the live web. In: Proceedings of SIGIR'2006. pp. 276–283 (2006)
9. Soboroff, I.: Meta-analysis for retrieval experiments involving multiple test collections. In: Proceedings of CIKM'2018. pp. 713–722 (2018)
10. Tonon, A., Demartini, G., Cudré-Mauroux, P.: Pooling-based continuous evaluation of information retrieval systems. Information Retrieval Journal **18**(5), 445–472 (2015)
11. Urbano, J., Lima, H., Hanjalic, A.: A new perspective on score standardization. In: Proceedings of SIGIR'2019. pp. 1061–1064 (2019)
12. Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W.R., Lo, K., Roberts, K., Soboroff, I., Wang, L.L.: Trec-covid: constructing a pandemic information retrieval test collection. In: ACM SIGIR Forum. vol. 54, pp. 1–12. ACM New York, NY, USA (2021)
13. Voorhees, E.M., Samarov, D., Soboroff, I.: Using replicates in information retrieval evaluation. ACM Transactions on Information Systems (TOIS) **36**(2), 1–21 (2017)
14. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al.: Cord-19: The covid-19 open research dataset. ArXiv (2020)
15. Webber, W., Moffat, A., Zobel, J.: Score standardization for robust comparison of retrieval systems. In: Proc. 12th Australasian Document Computing Symposium. pp. 1–8 (2007)
16. Webber, W., Moffat, A., Zobel, J.: Score standardization for inter-collection comparison of retrieval systems. In: Proceedings of SIGIR'2008. pp. 51–58 (2008)
17. Webber, W., Park, L.A.: Score adjustment for correction of pooling bias. In: Proceedings of SIGIR'2009. pp. 444–451 (2009)