Addressing Different Evaluation Environments for Information Retrieval through Pivot Systems

Gabriela González Sáez[1]* — **Lorraine Goeuriot**[1]* — **Philippe Mulhem**[1]*

Univ. Grenoble Alpes, CNRS, Grenoble INP[1]*, LIG, 38000 Grenoble, France*
[1]*Institute of Engineering Univ. Grenoble Alpes*

RÉSUMÉ. -

ABSTRACT. Classical evaluations of Information Retrieval systems, under the Cranfield Paradigm, compare several systems within one evaluation environment, defined by its settings (document collection, topics, assessments and evaluation measures). In this paper, we propose a framework to handle the comparison of systems across several evaluation environments. To achieve this goal, we investigate the use of pivot systems, allowing an indirect comparison of systems across evaluation environments by computing Result Deltas, i.e. the differences between their evaluation measures values. We detail the proposed pivot-based methodology, define a pivot characteristics and present experiments to validate our proposal (and in particular the pivot characteristics). We create altered environments that differ from their topic sets using the 2018 and 2020 CLEF eHealth evaluation campaigns (Goeuriot et al.*, 2020). We explore the behaviour of the metrics and pivots measuring the correlation between the result deltas, and the ranking of systems through the pivots compared to the official ranking of the systems. Our experiment show that correlations can greatly vary according to the chosen pivot and metric. We show that some pivot/metric pairs achieve high correlation values across the altered environments, with a ranking of systems similar to the official ranking.*

MOTS-CLÉS : -

KEYWORDS: Information Retrieval, Evaluation, Test Collection, Result Delta

1. Introduction

An Information retrieval (IR) system has the aim of retrieve relevant resources (i.e. documents) given an information need (i.e. query). Then, is required the evaluation of how relevant is the list of retrieved resources. The evaluation of IR systems traditionally relies on a document collection, a set of topics, relevance judgements linking topics and documents, and an assessment protocol with the set of evaluation measures. In our paper, we define the context of an evaluation as an Evaluation Environment (EE). EEs are composed of the test collection (documents, topics, and relevance assessments), the assessment protocol and the set of evaluation measures. The Cranfield paradigm (Cleverdon, 1997) allows the comparison of several systems under the same EE. To see if one system outperforms others, it is usual to run it under several EEs, and to compare it with other systems in each EE considered (Soboroff, 2006). In many cases, e.g. when we do not have access to the code of the systems, such paradigm cannot be used, as it does not support the comparison of systems evaluated across different EEs.

We propose a framework to compare IR systems over evolving evaluation environment. We call *Result Deltas*, noted $\mathcal{R}\Delta$ the comparison between two systems evaluation using a common metric. We define a *Pivot System* as a reference system to measure $\mathcal{R}\Delta$ over systems evaluated with different EEs. We claim that using such a pivot under each considered EE can indicate the systems ranking. Our ultimate goal is to study the feasibility of the proposed framework through the comparison of different features of selected pairs of metric and pivot, and validate it on ground truth data.

This paper is organized as follows: Section 2 presents the state of the art ; in Section 3 we detail our evaluation framework ; in Section 4 we describe the experiment to evaluate the feasibility of our proposal ; in Section 5 we show the results of our experiment ; in Section 6 we discuss our results ; we conclude and present future works in Section 7.

2. State of the Art

In the Information Retrieval Evaluation task, both IR systems and Evaluation Environment may evolve, creating changing results of the metrics evaluated. It is well known that the evolution of the components of an IR system can have a great impact on the outcome of an evaluation, and is possible to measure the effect of each of the changes in the evaluation results (Ferro et Silvello, 2016 ; Ferro et Silvello, 2018). Also, any change in the EE may impact the performance measurements: the document set, the topic set, the relevance judgments. As shown in (Sanderson *et al.*, 2012), 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. Further, the set of relevance judgments may also be modified through the pooling process or the as-

sessment process. In this paper, our goal is to design a framework to compare different systems over evolving EE, assuming that some or all of the considered IR systems cannot be evaluated on new EEs. Consequently, we focus on the impact changes in the EE can have on the performances, in particular alterations of the document collection and the topic set. In the following we present papers studying such changes.

Using the ANalysis Of Variance (ANOVA) model, (Ferro et Sanderson, 2017) showed that changing the test collection (splits of the documents corpus with a sub-set of relevance topics on the corpus) leads to varying system performances (inconsistently across metrics). In the same line, (Ferro et Sanderson, 2019 ; Voorhees *et al.*, 2017) models 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 improve 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.

The large difficulty variability we find in the set of topics does not allow to compare the performance of systems across different evaluation environments (Urbano *et al.*, 2013). (Sakai, 2016 ; Webber *et al.*, 2008) investigates variations within topic sets and states that systems can be compared across different test collections without worrying about topic difficulty. (Sakai, 2016) proposes a simple linear transformation of the standardized scores, given by $A \times \frac{x-\mu}{\sigma} + B$, where μ is the mean and σ is the standard deviation of the topic performance, and A and B are constant parameters, with this transformation all topics contribute equally to the final system score. (Urbano *et al.*, 2019) propose a standardization schema based on a transformation of the empirical distribution of the topic scores They show that previous standardization methods such as Sakai's are special cases of a general class of standardization, based on the assumption of a specific distribution for the per-topic score. (Soboroff, 2018) extends the score-standardization techniques with a meta-analysis of a system's performances over multiple test collections. This meta-analysis consists in comparing one baseline to a treated system, as a modified baseline, resulting in a delta measure over multiple collections, and a mean difference between the systems with a confidence interval. The Meta-analysis technique is strongly related to the measurement of the improvement across multiple test collections of a system with a specific modification that difference it from the baseline system. The difference with our proposal is that we address the problem of evaluation of different systems over different Evaluation Environments. Therefore the deltas are not computed over one, but several retrieval systems.

The works cited helps us to understand that (i) the elements of the evaluation environment affect the measures of effectiveness, (ii) there is an effort to standardize system's performances over different topics sets, (iii) Meta-analysis allows to compare one system's performances over several topics (iv) to the best of our knowledge,

there is no methodology taking into consideration variations of the EE to compare several systems evaluated with different settings.

3. Proposal

The evaluation of an Information Retrieval system uses a test collection, that is the core of each **Evaluation Environment**. An EE contains all the resources that have to be set to perform the evaluation task: the set of topics, the corpus of documents, the relevance judgments, the set of metrics to evaluate, and other process elements as the systems that create the pooling and their parameters, and the manual relevance judgement process. Our goal is to investigate the feasibility of comparing systems evaluated on varying EEs, given one or several changes from one EE to another, like different sets of topics, different corpuses of documents, or different pooling parameters. To do so, we propose an evaluation framework and a validation process.

We propose to estimate the difference between systems evaluated in different EEs with **Result Deltas**. A Result Delta, $\mathcal{R}\Delta$, estimates the difference between the performance of two systems measured with a similar metric.

$$\begin{aligned} &\mathcal{R}\Delta(Metric_k(System_i, EE_j), Metric_k(System_u, EE_v)) \\ &= Metric_k(System_i, EE_j) \ominus Metric_k(System_u, EE_v) \end{aligned} \quad [1]$$

where $\ominus$ denotes the difference between the system performance, measured by the $Metric_k$, of two configurations.

Three kinds of $\mathcal{R}\Delta$ can be measured, according the element that change in the evaluation task, as depicted on Figure 1:

– $\mathcal{R}_s\Delta$ (orange triangles in Figure 1a): When we have two different IR systems evaluated in the same EE, as a classical IR evaluation.

– $\mathcal{R}_e\Delta$ (yellow triangles in Figure 1a): If the same IR system is evaluated in two EEs, extracting mainly the environment effect that impact on the system.

– $\mathcal{R}_{se}\Delta$ (purple triangles in Figure 1a): If both EEs and the IR systems are different.

$\mathcal{R}_{se}\Delta$ can hardly be measured: the two systems are not directly comparable, because 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 with the two EEs considered[1] (Figure 1b). $\mathcal{R}_s\Delta$ 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 critical part of the evaluation of one pivot system on the EEs is the availability of a usable test collection. The pivot could be a common baseline evaluated on both

1. The pivot must be a system that has been evaluated in all EEs, or that can be reproduced on each EE.

Figure 1. *a) Two set of systems (S1, S2 and S3 on one side, and S4, S5, S6 on the other side) are evaluated in different EEs ; then, given an evaluation metric, $\mathcal{R}_s\Delta$, $\mathcal{R}_e\Delta$ and $\mathcal{R}_{se}\Delta$ are exposed. b) Pivot strategy to achieve $\mathcal{R}_{se}\Delta$.*

EEs (as BM25 is classically applied in several evaluation campaigns) or a new system applied in both EEs using a new system configuration (PM: remove because we do not talk about this later on : or a following a reproducibility strategy of one of the evaluated systems).

Given a metric M that evaluates the performance of a system S in a evaluation environment EE, we want to compare S1 and S2 as illustrated in Figure 1 b. We have $M(S_1, EE_1)$ and $M(S_2, EE_2)$, with EE1 and EE2 being comparable EEs: we assume that comparable EEs are different but the changes do not impact the ranking of systems. In this case the pivot system will help us to relate the systems across the EE 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 $\mathcal{R}_s\Delta(Pivot, S_2, EE_2)$, finally we can compare the result deltas to relate S1 and S2.

This pivot strategy is only valid if the pivot has the following two properties. Firstly, a pivot P must behave consistently across two EEs, so that the ranking obtained using P is correct. Secondly, the ranking of the systems obtained using P must be the same than the one obtained in a reference EE. These properties are detailed below:

– Consistency: The $Pivot$ system *behaves consistently* across two Evaluation Environments EE1 and EE2, for a given set of systems Set_S, if there is a high correlation between the results deltas between the systems of Set_S and P in EE1 and in EE2. This means that the deltas between the pivot and the systems are proportional to each others. In this case, the corollary it that the ranking of the systems in Set_S, relatively to the pivot P, is the same in both EEs.

– Correctness: The $Pivot$ P behaves *correctly* according to a reference environment (EE_{ref}) if using the result deltas measured with P to compare different systems evaluated across various EE_s (EE_{subs}) gives the same ranking of systems as being evaluated EE_{ref}, where EE_{ref} contains all the resources of the EE_{subs}. This property uses the capability of our proposal to build a ranking of systems based on the evaluation results of all the systems evaluated on different EEs.

The consistency property guaranties that comparison to a pivot allows to rank systems that have been evaluated on different EEs. The correctness property goes beyond consistency, as it ensures that the ranking of systems obtained with the pivot is similar to the reference ranking. In the following, the consistency will be evaluated using correlations values of the result deltas, and the correctness is evaluated through Kendall τ coefficients between the rankings.

4. Experimental Design

The use of a pivot implies that all the systems have to be compared against the performance of one pivot system. The performance of the pivot may behave differently in several EEs. We will assume that all the systems can be used as pivot, then, we want to evaluate their consistency and correctness.

4.1. *Evaluation Environments Set-up*

To evaluate our proposal, we use the data for the AdHoc task of the campaign eHealth2020 from the CLEF initiative (Goeuriot *et al.*, 2020), using the official relevance judgments of the 50 topics of the eHealth 2018 and 2020 task. The set of runs considered consists of the 9 official runs (i.e. ; 9 systems) for clef-eHealth 2020[2]. We base our experiments on two simulated evaluation environment by randomly splitting the 50 topics into two groups of 25 topics, namely S_{Q1} and S_{Q2}, for each EE. So, here, two evaluation environments are defined: EE_1 with the corpus, the evaluation measures and the queries S_{Q1}, and EE_2 with the corpus, the evaluation measures and the queries S_{Q2}.The full set of queries corresponds to the reference EE, EE_{ref}. Our study averages results over 50 splits. We evaluate the performance of the systems using six metrics: Mean Average Precision (MAP), Precision at Recall level (Rprec), Binary Preference (bpref), reciprocal rank (recip_rank) and normalized discount cumulative gain at 10 (ndcg) and Precision at 10 (P_10). Where ndcg@10 and bpref are official metrics used on the task.

Our experiments evaluate the pivots in two steps. First, we check the pivots consistency, computing the correlation of the result deltas obtained with each pivot and the other systems in a couple of EEs. For the correctness property, we compute Kendall Tau comparing the ranking of systems constructed with the result deltas with the official ranking results of the task, using the full set of topics (EE_{ref}).

Pairs of EEs could be achieved through a split of the topics, or of the the document corpus or the documents pooled on the relevance assessment. In both our experiments

2. Runs submitted for clef-eHealth2020: bm25_orig_IMS, original_rm3_rrf_IMS, original_rrf_IMS, variant_rrf_IMS, FT_Straight_LIG, Noexp_Straight_LIG, UMLS_RF_LIG, UMLS_Straight_LIG, tfidf_sandiDoc.

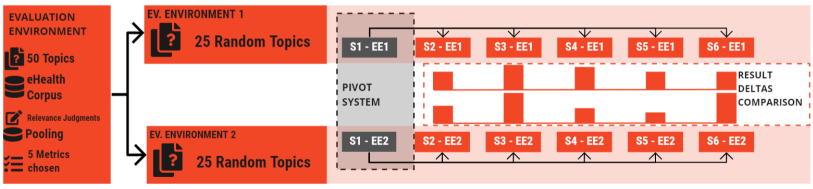

Figure 2. *Method to measure and compare the result deltas of two evaluation environments using pivot S1.*

we define the pair of EEs according to one 50%-50% split of the topics, given that different queries can change the performance results (Ferro et Sanderson, 2019).

4.2. *Consistency*

Considering the consistency property presented above, we are interested in checking if the $\mathcal{R}_e\Delta(Pivot, S, EE_1)$ has a similar behaviour to $\mathcal{R}_e\Delta(Pivot, S, E_E2)$ for all the systems evaluated on both EE_1 and EE_2. To evaluate the pivot consistency, we measure the correlation between the result deltas of the pivot and the systems evaluated in both EEs.

Figure 2 describes the process to compare the result deltas: (i) the EE is split into two groups EE_1 and EE_2, (ii) the pivot is defined and the result deltas are measured using a specific metric across all the systems in both environments (iii) the two sets of result delta values are compared using correlation to evaluate the capability of the pivot system to measure proportional result deltas. The ranking across the EEs should be the same, we expect the pivot to give with highly correlated result deltas among the EEs.

4.3. *Correctness*

To test the correctness property, we study the similarity between $Rank_{ref}$ and $Rank_1$, where: i) $Rank_1$ is the ranking of systems obtained from the computation of result delta using the pivot over two set of systems, the first set evaluated into EE_1 and the second on EE_2, where EE_1 and EE_2 are two splits of EE_{ref}, and ii) $Rank_{ref}$ is the ranking formed by all the evaluated systems over EE_{ref}. We evaluate the correctness of the pivot by the Kendall τ coefficient between $Rank_1$ and $Rank_{ref}$. This experiment uses the eHealth2020 data to simulate the real problem of comparing systems evaluated in different environments.

Figure 3 describes the process of ranking creation: (i) the EE is split into two groups EE_1 and EE_2 along with the different IR systems evaluated in each EE, (ii) the pivot is defined and the result deltas are measured using a chosen metric, (iii)

the result delta values for both EEs are used to obtain the $Rank_1$ of the full set of systems ; The Kendall τ coefficient is measured between the $Rank_1$ and the reference ranking $Rank_{ref}$ to evaluate the quality of rank obtained with the pivot. We expect it to be similar to the reference ranking (ranking observed on the reference EE). In the experiment the reference ranking $Rank_{ref}$ is given by the official CLEF eHealth results (Goeuriot *et al.*, 2020).

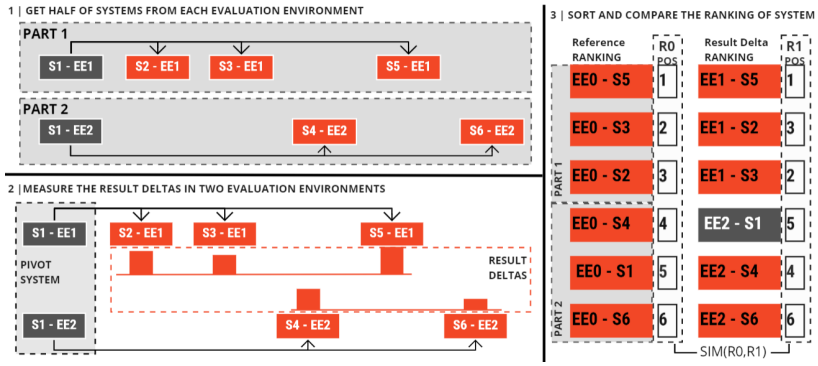

Figure 3. *Method to create a ranking of systems evaluated in different EEs using result deltas.*

5. Results

The results are organized following the pivot properties. We explore the results by selecting several systems as the pivot, and several metrics to compute the result deltas. We expect to validate the use of pivots to compare systems across EEs, we want to find a pivot system with constant and high correlation result delta values across the metrics (*consistency*) and high similarity with the official ranking of systems (*correctness*).

5.1. *Consistency*

To study if it is possible to create a fair and consistent comparison of systems across different EEs we compare the correlation of the result deltas obtained on two EEs. We expect to find a pivot system with constant and high correlation values with all the metrics.

Figure 5 shows the global results of our experiment and figure 4 shows an example that illustrates how the final matrix is computed. Figure 4 Section a) shows the result deltas of the system tfidf_sandiDoc using the metric recip_rank for EE_1 in blue and EE_2 in orange, as one split of the EE. In this case, the comparison of the result deltas is -0.20, a negative and close to zero value that shows us no relation between the result deltas in the EEs and will lead us to an unfair comparison of the systems. In

b) we see the correlation distribution of recip_rank with tfidf_sandiDoc's result deltas measured on 50 splits of the EEs, the mean is -0.26 and std is 0.27, with this result the studied pivot is not a good choice, because will behave disparately in different EEs. Finally, figure 5 summarizes the results with the metric in the rows, and the pivots in the columns, placed in decreasing order according to their map performance in the official results.

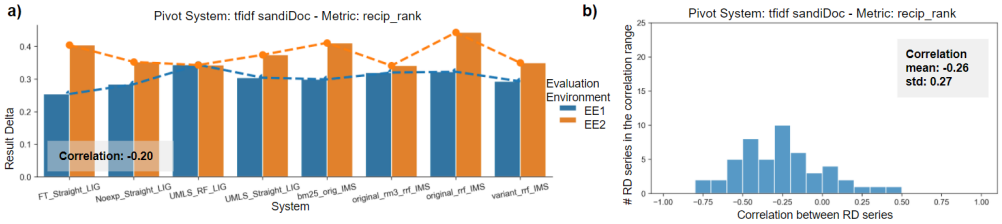

Figure 4. *a) Result Delta comparison between two EEs using the pivot tfidf_sandiDoc and the recip_rank metric ; b) distribution of the correlation of 50 pairs of EEs using a pivot and a metric.*

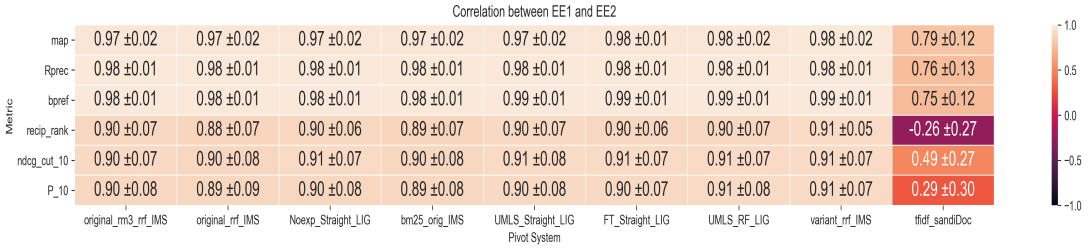

Figure 5. *Correlation (mean and std. dev.) between result deltas in two evaluation environments, using several pivots and metrics. Systems are listed in decreasing MAP value (left to right) according to the official results.*

We see from figure 5 that the result deltas have different behaviours for each metric: while MAP, Rprec and bpref have correlations close to 0.98 and constant performance (std rounding 0.02), recip_rank, ndcg and P_10 mean correlations are rounding to 0.90. In relation to the values from the pivot perspective, only one system has poor correlation values across all the metrics, this system presents the worst performance in the official results: tfidf_sandiDoc.

In a way to find out if the behaviour detected above also holds when considering a subset of the 9 systems considered, the figure 6 presents the same methodology applied only to the LIG's runs: we compare the correlation of a set of three result deltas (LIG systems except for pivot). In this case, the worse correlation values are presented using the best performance system Noexp_Straight_LIG, with negative values in almost all the metrics. From the metrics perspective, recip_rank correlation is always negative

and, as ndcg and P_10, the std values are rounding 0.60, meaning that the EEs are not comparable using these metrics due to the high probability of inconsistent result delta across EEs.

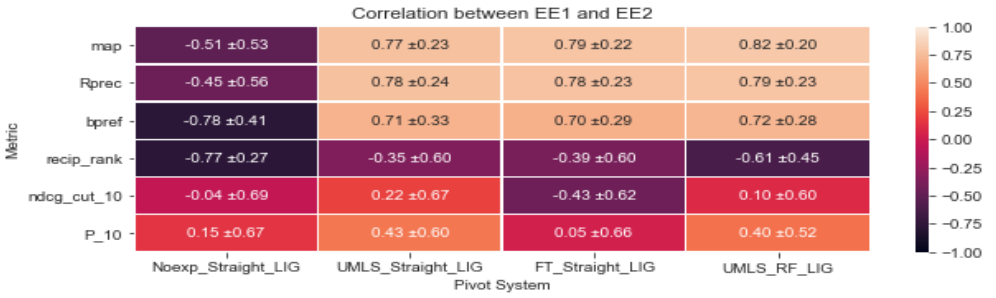

Figure 6. *Correlation (mean and std. dev.) between result deltas in two evaluation environments considering only LIG's systems. Systems are listed in decreasing MAP value (left to right) according to the official results.*

5.2. *Correctness*

As an example, figure 7 a) shows the official ranking $Rank_{ref}$ of systems in the first row, and a result delta ranking $Rank_1$ formed using the pivot systems tfidf_sandiDoc and with recip_rank metric, both rankings have a similarity of 0.67 then, the pivot is not constructing a similar to $Rank_{ref}$ ranking. Figure 7 b) shows the distribution of Kendall tau similarity measured between the official ranking results and the rankings of the 50 splits of EE, in this case, the mean similarity between the $Rank_{ref}$ and the 50 ranking generated is 0.42 and the standard deviation (std.) equals 0.22, therefore, as $Rank_1$ the delta result rankings are not similar to $Rank_{ref}$. The global results of the experiment are summarized in Figure 8. It shows the similarity between the official ranking and the Result Deltas ranking of systems, obtained with a pivot system (each column, ordered by the best MAP performance on the official results from left to right) using one specific metric (each row).

In Figure 8, we see that the ranking obtained varies according to the metric used: while MAP, Rprec and bpref have a similarity mean bigger than 0.71 with the official ranking, recip_rank and ndcg are rounding the 0.6, with the lower similarity and higher std values on recip_rank metric. This leads to different rankings of systems across EEs and far from the reference ranking results. From the pivot Perspective, the lowest similarity is achieved by the systems with the worst performance: tfidf_sandiDoc.

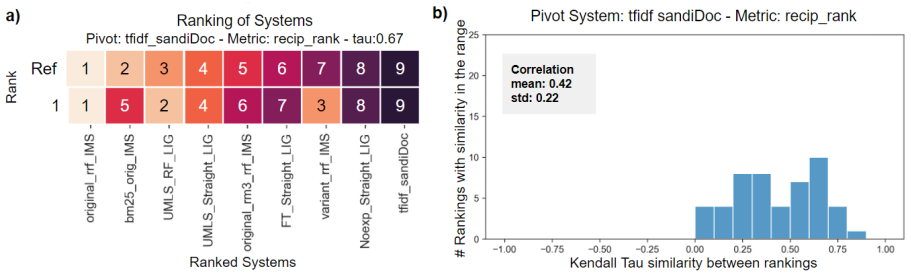

Figure 7. *a) Ranking comparison between two the official ranking of systems and the proposed using result deltas with a pivot (tfidf_sandiDoc) and a metric (Rprec) b) distribution of the Kendall tau similarity of 50 pairs of ranking formed from pairs of EEs using the same pivot and metric.*

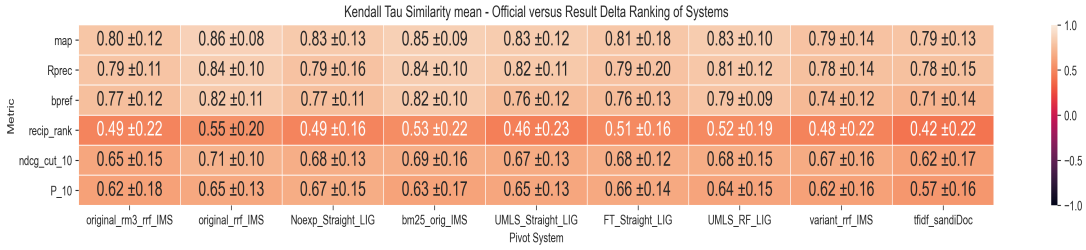

Figure 8. *Similarity (mean and std. dev.) between the official and proposed Ranking of systems using different metrics. Systems are listed in decreasing MAP value (left to right) according to the official results.*

6. Discussion

In the consistency experiment, we try to evaluate the behaviour of the Result Deltas when measured using different pivot systems through the correlation of two series of Result Deltas extracted from two EEs. When the 9 systems of the eHealth adHoc task were considered (figure 5), we found several pivots that allow to compare fairly a group of systems evaluated in two different EEs, demonstrated by a strong correlation (values far from zero) in all the metrics used. This first result was suggesting that a pivot may be defined according to the underlying models of the systems considered (as the run tfidf uses a very different model than the others, based all on BM25 according to (Goeuriot *et al.*, 2020)), then, the Figure 6 clearly contradicts this conclusion (as all these systems are based on BM25): the choice of a consistent pivot is then open for further researches. Both scenarios differ in the number of result deltas used to compute the correlation. Then, to find a robust pivot is important to test it in different splits of systems.

In the correctness experiment, we try to measure the similarity between the ranking of systems evaluated on a reference EE and a ranking constructed by the result delta values of a pivot across systems evaluated on two different EEs. Over the best three metrics: MAP, Rprec and Bpref the result deltas values constructed rankings similar to the $Rank_{re}$, nevertheless, none of the values is one, consequently, we were not able to find a pivot that behaves the same as the official ranking, meaning that any pivot achieves complete correctness.

As presented in the experiments, the pivot selection impacts the comparison of the systems in the EEs, so there is a need to define what is a good pivot and how to select it given changing systems and EEs. Also, in both experiments is possible to define a list of metrics that achieve, in general, high correlation and similarity values across all the pivots. MAP, Rprec and Bpref have in all the experiments better results than recip_rank, ndcg and P_10. Further research has to be developed to prove the applicability of the method in the second list of metrics, to define if it is possible to adapt these metrics allowing them to compare systems in different EEs, or if the result deltas are only measurable using the first list of metrics.

On the presented framework we establish the assumption that the Evaluation Environments are comparable if the ranking of systems, ordered by the same metric, is the same across the EEs. To improve the definition of comparable Evaluation Environments is needed to define the characteristics of them, and in respect to these characteristics and their differences define the constraints of comparables EEs.

7. Conclusion

We have presented a framework proposal to manage the evaluation of systems across evaluation environments using result deltas and pivot systems. To evaluate our proposal, we propose two pivot's properties: consistency and correctness, they lead us to two experiments where we found different behaviors in the result deltas of the metrics and pivots and also on the resulting ranking of systems constructed with them. We conclude that the pivot system and the metric set have to be defined with great care.

In the future, we will continue this work by: expanding the experiment to test other differences between the EE, as different document corpus, assessments constructions, among others; studying what make a good pivot system for evolving evaluation of systems, and its relation with the consistency and correctness properties; creating a method to select and use metrics to measure the Result Delta and compare systems.

Acknowledgements

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

8. Bibliographie

Cleverdon C., *The Cranfield Tests on Index Language Devices*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 47–59, 1997.

Ferro N., Sanderson M., « Sub-corpora impact on system effectiveness », *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 901-904, 2017.

Ferro N., Sanderson M., « Improving the Accuracy of System Performance Estimation by Using Shards », *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 805-814, 2019.

Ferro N., Silvello G., « A general linear mixed models approach to study system component effects », *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, p. 25-34, 2016.

Ferro N., Silvello G., « Toward an anatomy of IR system component performances », *Journal of the Association for Information Science and Technology*, vol. 69, n° 2, p. 187-200, 2018.

Goeuriot L., Suominen H., Kelly L., Miranda-Escalada A., Krallinger M., Liu Z., Pasi G., Saez G. G., Viviani M., Xu C., « Overview of the CLEF eHealth evaluation lab 2020 », *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, p. 255-271, 2020.

Sakai T., « A simple and effective approach to score standardisation », *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, p. 95-104, 2016.

Sanderson M., Turpin A., Zhang Y., Scholer F., « Differences in effectiveness across sub-collections », *ACM International Conference Proceeding Series*, vol. 2006, p. 1965-1969, 2012.

Soboroff I., « Dynamic test collections: measuring search effectiveness on the live web », *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 276-283, 2006.

Soboroff I., « Meta-Analysis for Retrieval Experiments Involving Multiple Test Collections », *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, p. 713-722, 2018.

Urbano J., Lima H., Hanjalic A., « A New Perspective on Score Standardization », *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 1061-1064, 2019.

Urbano J., Marrero M., Martín D., « On the measurement of test collection reliability », *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, p. 393-402, 2013.

Voorhees E. M., Samarov D., Soboroff I., « Using replicates in information retrieval evaluation », *ACM Transactions on Information Systems (TOIS)*, vol. 36, n° 2, p. 1-21, 2017.

Webber W., Moffat A., Zobel J., « Score standardization for inter-collection comparison of retrieval systems », *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, p. 51-58, 2008.