Towards Result Delta Prediction Based on Knowledge Deltas for Continuous IR Evaluation

Gabriela Gonzalez-Saez[1], Alaa El-Ebshihy[2], Tobias Fink[2], Petra Galuščáková[1], Florina Piroi[2], David Iommi[2], Lorraine Goeuriot[1] and Philippe Mulhem[1]

[1]*Univ. Grenoble Alpes, CNRS, Grenoble INP[1], LIG, Grenoble, France*
[2]*Research Studios Austria, Data Science Studio, Vienna, AT*

Abstract

The continuous evaluation of Information Retrieval Systems requires comparing IR systems both one to another, but also across collections, in other words across different evaluation environments (test collection and evaluation metrics). These evaluation environments may also be evolutionary versions of some given evaluation environment. In this work, we propose a methodology to measure and understand the impact the differences between test collection representations (i.e. knowledge delta, $\mathcal{K}\Delta$) has on system performance, and we look at the differences in their outputs (i.e. result delta, $\mathcal{R}\Delta$). We present initial experiments with various text representations on the TREC 2004 Robust Collection, and look at the relation between the $\mathcal{K}\Delta$ and the $\mathcal{R}\Delta$.

Keywords

Continuous Evaluation, Evolving Test Collections, Knowledge Delta, Result Delta

1. Introduction

Traditional offline evaluation of Information Retrieval systems uses test collections [1] which are composed of: (1) a set of documents or passages, (2) a set of queries, and (3) a set of relevance judgments indicating which document is relevant to each query. The components of a test collection together with (a set of) evaluation metrics to assess the efficiency of an IR system define the elements of an *Evaluation Environment* (EE) [2]. Changes in an EE's element then affect an IR system's performance. We analyze the differences between test collections by systematically quantifying and analyzing the differences in data representation for documents and queries test collection components. We look at document representations as these are the elements used by typical IR systems to compute relevance of documents towards a given query.

With the aim of implementing continuous evaluation for IR systems, we address the topic of measuring IR performance for evolving EEs [2]. We introduce the notion of *Results Delta* ($\mathcal{R}\Delta$) as the means to measure IR systems performance differences with respect to one metric. We introduce three types of $\mathcal{R}\Delta$: $\mathcal{R}_s\Delta$ (different IR systems evaluated in the same EE), $\mathcal{R}_e\Delta$ (the same IR system evaluated in different EEs), and $\mathcal{R}_{se}\Delta$ (both EEs and IR systems vary). We aim to understand and to quantify variations between the components of two different EEs by

[1]Institute of Engineering Univ. Grenoble Alpes.

QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks, co-located with The 45th European Conference on Information Retrieval (ECIR) April 2, 2023, Dublin, Ireland

means of defining Knowledge Delta ($\mathcal{K}\Delta$) and observing its impact on the $\mathcal{R}\Delta$. In our view, $\mathcal{K}\Delta$ for IR is a combination of a document representation delta, $\mathcal{K}_d\Delta$, and a query representations delta, $\mathcal{K}_q\Delta$, both defined as difference functions between pairs of text sequence representations.

This paper proposes a study that looks at how various simple text representations to quantify $\mathcal{K}_d\Delta$ and their impact on $\mathcal{R}_e\Delta$. Initial experiments are performed on the TREC 2004 Robust Collection [3]. As this collection stores the publishing time for each document, we consider it to be an evolving collection. That is, we can simulate the conditions of an IR system that has to provide answers to queries, answers extracted from a set of documents that changes over time.

2. Methodology

Mothe [4] analysed different approaches to understand the effectiveness of IR systems, focusing on studying the effectiveness with respect to the query and IR system parameters. In our work, we are interested in understanding the change of the IR systems performance with respect to the change of the document collection in addition to the query, in a way to predict the change in performance of the IR system for an evolving test collection. Inspired by [4], we aim to use the document representations as features for the document collection and find the correlation between the features of document collection and the change in IR system performance.

Test collection difference, $\mathcal{K}_d\Delta$: We define the $\mathcal{K}_d\Delta$ as a quantifiable value of the differences between document representations, which may be more or less complex: bag of words, TF-IDF [5], topic detection methods (e.g. Latent Dirichlet Allocation [6] and conceptual embeddings [7]) and neural networks language models (e.g. Word2Vec [8] and BERT [9]). Any of these representations, or a combination of them, may contribute to generate the document collection representation which can then be used to quantify $\mathcal{K}_d\Delta$ and predict the $\mathcal{R}\Delta$.

Performance impact, $\mathcal{R}_e\Delta$: We define $\mathcal{R}_e\Delta$ as the absolute difference in the IR system performance in two EEs: consider $M(S_i, EE_j)$ as the performance of systems S_i evaluated in evaluation environment EE_j with metric M, we compute $\mathcal{R}_e\Delta$ as $M(S_1, EE_1) - M(S_1, EE_2)$.

Prediction model, ($\mathcal{K}_d\Delta \sim \mathcal{R}_e\Delta$): We propose to understand the impact of $\mathcal{K}\Delta$ on $\mathcal{R}\Delta$ by building a model that predicts $\mathcal{R}_e\Delta$ from $\mathcal{K}_d\Delta$. We will, first, observe the correlation between $\mathcal{K}_d\Delta$ and $\mathcal{R}_e\Delta$ using different text representation methods as $\mathcal{K}_d\Delta$. Then, we will build a prediction model based on these observations. Finally, we will analyse the impact of the $\mathcal{K}_d\Delta$ elements on the prediction of the $\mathcal{R}_e\Delta$ by feature selection techniques [10].

Dataset: We measure $\mathcal{K}_d\Delta$ and $\mathcal{R}_e\Delta$ from an evolving test collection as an example of documents changing in a real corpus. The evolving test collection is built by creating shards of a classical test collection [11] that contains timestamped documents. We use these timestamps to assign documents, according to their temporal order, to shards and to define fixed percentages of corpus overlap to control the evolution.

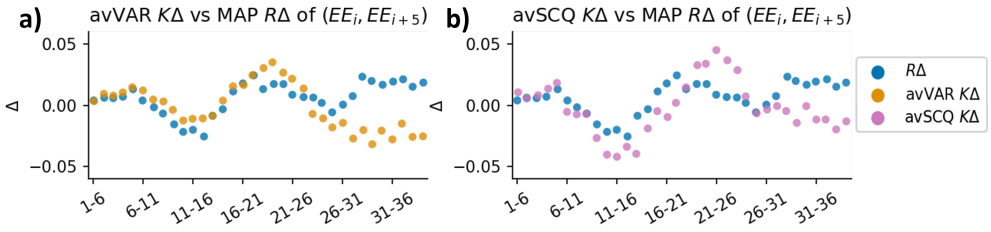

Figure 1: MAP $\mathcal{R}\Delta$ of BM25 in blue; and $\mathcal{K}\Delta$ in orange (avVAR) and purple (avSQC). The x-axis shows the compared EEs.

Initial Experiment: We evaluate pyterrier BM25 system [12] in an evolving test collection created from Robust [3] using the MAP metric. We create 41 EEs using 90% document overlaps between successive shards, with full set of topics. As text representations, we test two features used in query performance prediction [13]: Averaged Term Weight Variability (avVAR) [14] and Averaged Collection Query Similarity (avSCQ) [14]. We compare EEs with 50% of overlap (e.g. EE_1 vs. EE_6, EE_2 vs. EE_7, etc.). Figure 1 presents changes in the MAP score ($\mathcal{R}_e\Delta$) compared with the $\mathcal{K}_d\Delta$ calculated as the changes in the selected feature values: avVAR in (a) and avSCQ in (b). The pearson correlation between the Δ MAP and the features is 0.5 and 0.12 for the avVAR and avSCQ, respectively. These results confirm that the changes in $\mathcal{K}_d\Delta$ have a considerable effect $\mathcal{R}_e\Delta$ values. Moreover, they show that the effect might substantially differ for different features and over time.

3. Discussion and Future Work

We propose the definition of *Knowledge Delta* ($\mathcal{K}\Delta$) for the elements of the EEs. As a first attempt to quantify the $\mathcal{K}_d\Delta$ and its impact on the *Result Delta* ($\mathcal{R}_e\Delta$), we use two simple text representation metrics, avVAR and avSCQ. We experiment on an evolving test collection which is built by using the timestamps from the Robust test collection. The initial results show a correlation between $\mathcal{K}_d\Delta$ and the $\mathcal{R}_e\Delta$ and thus provide justification for our approach. These results motivate us to build a prediction model ($\mathcal{K}_d\Delta \sim \mathcal{R}_e\Delta$) that can predict the change of the performance of an IR systems using the $\mathcal{K}_d\Delta$ and also to quantify $\mathcal{K}_d\Delta$using different text representations (see Section 2). We either plan to construct a machine learning model that assumes $\mathcal{K}_d\Delta$ as input feature to predict $\mathcal{R}_e\Delta$ or to use time series [15] techniques to predict significant changes in $\mathcal{K}_d\Delta$, which lead to changes in the performance of the IR system. Moreover, we plan to define other types of $\mathcal{K}\Delta$ and $\mathcal{R}\Delta$, such as quantifying the differences in query representations ($\mathcal{K}_q\Delta$) and apply them in the LongEval collection [16]. This will contribute to understand the impact of the $\mathcal{K}\Delta$ on other $\mathcal{R}\Delta$, including $\mathcal{R}_s\Delta$ and $\mathcal{R}_{se}\Delta$.

Acknowledgments

This work is supported by 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 grant I4471-N.

References

[1] M. Sanderson, Test collection based evaluation of information retrieval systems, Now Publishers Inc, 2010.

[2] G. N. González-Sáez, P. Mulhem, L. Goeuriot, Towards the evaluation of information retrieval systems on evolving datasets with pivot systems, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2021, pp. 91–102.

[3] E. M. Voorhees, The trec 2005 robust track, in: ACM SIGIR Forum, volume 40, ACM New York, NY, USA, 2006, pp. 41–48.

[4] J. Mothe, Analytics methods to understand information retrieval effectiveness—a survey, Mathematics 10 (2022).

[5] G. Salton, A. Wong, C. S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (1975) 613–620.

[6] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, L. Zhao, Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multim. Tools Appl. 78 (2019) 15169–15211.

[7] K. Abdulahhad, Concept embedding for information retrieval, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, volume 10772 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 563–569.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 3111–3119.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[10] S. Déjean, R. T. Ionescu, J. Mothe, M. Z. Ullah, Forward and backward feature selection for query performance prediction, in: Proceedings of the 35th annual ACM symposium on applied computing, 2020, pp. 690–697.

[11] N. Ferro, Y. Kim, M. Sanderson, Using collection shards to study retrieval performance effect sizes, ACM Transactions on Information Systems (TOIS) 37 (2019) 1–40.

[12] 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.

[13] C. Hauff, Predicting the effectiveness of queries and retrieval systems, in: SIGIR Forum, volume 44, 2010, p. 88.

[14] Y. Zhao, F. Scholer, Y. Tsegay, Effective pre-retrieval query performance prediction using

similarity and variability evidence, in: Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30, Springer, 2008, pp. 52–64.
[15] C. Chatfield, The analysis of time series: an introduction, Chapman and hall/CRC, 2003.
[16] P. Galuščáková, R. Deveaud, G. Gonzalez-Saez, P. Mulhem, L. Goeuriot, F. Piroi, M. Popel, Longeval-retrieval: French-english dynamic test collection for continuous web search evaluation, arXiv preprint arXiv:2303.03229 (2023).