QPP++ 2023: Query-Performance Prediction and Its Evaluation in New Tasks

Guglielmo Faggioli[1][0000−0002−5070−2049], Nicola Ferro[1][0000−0001−9219−6239], Josiane Mothe[2][0000−0001−9273−2193], and Fiana Raiber[3]

[1] University of Padova, Italy
faggioli@dei.unipd.it,ferro@dei.unipd.it
[2] INSPE, Université de Toulouse, IRIT UMR5505 CNRS, France
Josiane.Mothe@irit.fr
[3] Yahoo Research, Israel
fiana@yahooinc.com

Abstract. Query-Performance Prediction (QPP) is currently primarily applied to ad-hoc retrieval tasks. The Information Retrieval (IR) field is reaching new heights thanks to recent advances in large language models and neural networks, as well as emerging new ways of searching, such as conversational search. Such advancements are quickly spreading to adjacent research areas, including QPP, necessitating a reconsideration of how we perform and evaluate QPP. This workshop sought to elicit discussion on three topics related to the future of QPP: exploiting advances in IR to improve QPP, instantiating QPP on new search paradigms, and evaluating QPP on new tasks.

Keywords: Query Performance Prediction · Neural IR · Conversational Search · Evaluation.

1 Introduction

The advent of large language models and the rise of new tasks, such as conversational search, semantic search and question answering, enabled by the availability of new powerful technological tools, have led to a previously unseen rapid growth in the variety and quality of Information Retrieval (IR) systems. Several ancillary research fields have also flourished due to the scientific uptake of new Natural Language Processing (NLP) methodologies, facilitating advancement in new IR tasks. The Query-Performance Prediction and Its Evaluation in New Tasks (QPP++ 2023) workshop aimed to further fuel such growth in the renowned and important area of Query-Performance Prediction (QPP).

The QPP task is defined as estimating search effectiveness in the absence of human relevance judgments [1]. Since its introduction at the beginning of the 21st century, QPP has established itself as an essential tool in numerous tasks, including model selection [1, 13], query suggestion [1, 13], and rank fusion [10]. The QPP++ 2023 workshop was a collaborative effort of researchers to master

the new tools made available by the NLP community and learn how to effectively use them for the QPP task. The workshop focused on applying QPP in traditional scenarios, such as ad-hoc retrieval, and in new domains, including conversational and semantic search, passage retrieval, and question answering. QPP++ 2023 also allowed the community to reexamine past weaknesses and challenges linked to the QPP task, such as its evaluation, while establishing a roadmap to organize and guide the community's future efforts to advance the QPP research field.

QPP and Novel Search Paradigms Given the recent developments in IR, the prediction quality of existing QPP approaches may be significantly affected in new domains and scenarios for the following three reasons. First, some of the traditional predictors exploit statistics derived from the collection [5], while new IR models often use indexes of embeddings or apply machine learning to re-rank documents [7]. Second, the vast majority of the recently developed retrieval models in IR utilize semantic information that, with a few notable exceptions [8, 12], is rarely exploited by QPP models. This, in turn, impairs the performance of traditional QPP models applied on IR systems based on new paradigms [3]. Finally, QPP can be used for new processes such as selective query processing [2].

The QPP++ 2023 workshop aimed to provide a platform for the community to jointly discuss ways to address these challenges and create a better alignment between the latest technologies, retrieval models, and QPP approaches. Along with the challenges mentioned above, the recent advances in NLP present great opportunities for enhancing the state of the art in QPP. The workshop also seeked to encourage collaboration between researchers to exploit these opportunities.

QPP and its Evaluation on New Tasks The quality of QPP methods is typically evaluated by computing the correlation between the scores assigned to queries by a QPP method and the true performance values, e.g., Average Precision (AP), attained for these queries using relevance judgements. Previous research demonstrated the unreliability of this approach when multiple experimental factors (i.e., IR models, corpora, and predictors) are considered [6, 11, 4]. In addition, researchers demonstrated that high correlation does not necessarily translate to improved retrieval effectiveness [9, 6]. These issues are further exacerbated in new domains, such as question answering or conversational search, where the evaluation of the retrieval models is often more challenging. The QPP++ 2023 workshop aimed at fostering discussion in the community regarding these challenges.

2 Workshop Topics and Goals

The workshop provided a forum for researchers and practitioners to discuss the following key research challenges emerging following the recent advances in IR:

- Can existing QPP techniques be exploited, or which new QPP theories and models need to be devised, for new tasks, such as passage-retrieval, question answering, and conversational search?
- How can new technologies, such as contextualized embeddings, large language models, and neural networks be exploited to improve QPP?
- How should QPP techniques be evaluated, including best practices, datasets, and resources?
- Should QPP be evaluated in the same manner for different IR tasks?
- What changes should we make to the QPP evaluation paradigm to accommodate new domains and IR techniques?

The workshop is expected to have two main outcomes:

- We intend to compile the workshop proceedings from the submitted papers. The proceedings will be published in the CEUR-WS.org proceedings series.
- We intend to draft a position paper describing the roadmap identified during the discussions and submit it to the SIGIR forum.

3 Workshop Organizers

Guglielmo Faggioli is a PhD student in Information Engineering at the University of Padua, Italy. His main research concerns information retrieval evaluation, focusing on performance modeling, query performance prediction models, and conversational search systems. His thesis concerns the modeling and prediction of information retrieval systems' performance. He contributed as co-editor to the Proceedings of the Twelfth and Thirteenth International Conference of the CLEF Association (CLEF 2021 and CLEF 2022).

Nicola Ferro is a full professor in computer science at the University of Padua, Italy. He works in information retrieval and its evaluation. He is the coordinator of CLEF (Conference and Labs of the Evaluation Forum) and has organized several evaluation tasks over the years. He has co-authored many papers on IR evaluation, and his current interests are reproducibility of IR experiments, IR system performance modeling and prediction, formal models, and properties of IR evaluation measures. He co-organized several workshops at major conferences, among which GLARE at CIKM 2018, acted as general chair of ECIR 2016, short paper co-chair of ECIR 2020, and resource paper co-chair at CIKM 2021.

Josiane Mothe is full professor in Computer Science at INSPE-Université Toulouse Jean-Jaurès and researcher at Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS. Her research focuses on information systems and information retrieval, including selective query processing and query performance prediction. She serves as a senior PC member in major conferences in IR, is co-editor of SIGIR Forum, associate editor at TOIS, board member of IR journal, SIGIR 2023 co-chair, ECIR 2021 short papers co-chair, and CIRCLE

2022 general chair.

Fiana Raiber is a senior manager at Yahoo Research. She earned her Ph.D. from the Technion - Israel Institute of Technology, where she currently holds a research associate position, collaborating with faculty members and graduate students. Fiana is a co-author of multiple conference and journal papers in information retrieval, including several publications on query-performance prediction. She served as the SIGIR 2018 workshops co-chair and SIGIR 2022 short papers co-chair. She is a (senior) program committee member of numerous conferences, including SIGIR, ICTIR, WSDM, and CIKM.

References

1. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services **2**(1), 1–89 (2010)
2. Deveaud, R., Mothe, J., Ullah, M.Z., Nie, J.Y.: Learning to adaptively rank document retrieval system configurations. ACM Transactions on Information Systems (TOIS) **37**(1), 1–41 (2018)
3. Faggioli, G., Formal, T., Marchesin, S., Clinchant, S., Ferro, N., Benjamin, P.: Query Performance Prediction for Neural IR: Are We There Yet? In: Proc. ECIR (2023)
4. Faggioli, G., Zendel, O., Culpepper, J.S., Ferro, N., Scholer, F.: An Enhanced Evaluation Framework for Query Performance Prediction. In: Proc. ECIR. pp. 115–129 (2021)
5. Hauff, C.: Predicting the effectiveness of queries and retrieval systems. In: Ph.D. Dissertation. University of Twente. pp. 1–179 (2010)
6. Hauff, C., Azzopardi, L., Hiemstra, D.: The Combination and Evaluation of Query Performance Prediction Methods. In: Proc. ECIR. pp. 301–312 (2009)
7. Mitra, B., Craswell, N.: An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval **13**(1), 1–126 (2018)
8. Mothe, J., Tanguy, L.: Linguistic Features to Predict Query Difficulty. In: Proc. SIGIR. pp. 7–10 (2005)
9. Raiber, F., Kurland, O.: Query-Performance Prediction: Setting the Expectations Straight. In: Proc. SIGIR. pp. 13–22 (2014)
10. Roitman, H.: Enhanced performance prediction of fusion-based retrieval. pp. 195––198. ICTIR '18 (2018)
11. Scholer, F., Garcia, S.: A Case for Improved Evaluation of Query Difficulty Prediction. In: Proc. SIGIR. pp. 640–641 (2009)
12. Shtok, A., Kurland, O., Carmel, D.: Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction. In: Proc. SIGIR. pp. 259–266 (2010)
13. Thomas, P., Scholer, F., Bailey, P., Moffat, A.: Tasks, queries, and rankers in pre-retrieval performance prediction. pp. 1–4. ADCS 2017 (2017)