Report on the 1st Workshop on Query Performance Prediction and Its Evaluation in New Tasks (QPP++ 2023) at ECIR 2023

Guglielmo Faggioli
University of Padua
Italy
guglielmo.faggioli@unipd.it

Nicola Ferro
University of Padua
Italy
nicola.ferro@unipd.it

Josiane Mothe
INSPE, Université de Toulouse, IRIT UMR5505 CNRS
France
josiane.mothe@irit.fr

Fiana Raiber
Yahoo Research
Israel
fiana@yahooinc.com

Maik Fröbe
Friedrich-Schiller-Universität Jena
Germany
maik.froebe@uni-jena.de

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.

Date: 6 April 2023.

Website: https://qpp.dei.unipd.it/.

1 Introduction

QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks is the first edition of a workshop that aims to foster a discussion within the community on how Query Performance Prediction (QPP) can be applied to new techniques in Information Retrieval (IR) and how such techniques can be exploited to define new QPP models. This first edition was hosted by the European Conference on Information Retrieval (ECIR) 2023 in Dublin (Ireland).

QPP++ 2023 received nine scientific submissions, of which seven papers (four long and three short) were accepted. Two to three program committee members reviewed each submission, and the program chairs oversaw the reviewing. The accepted papers included authors from 8 countries and 14 institutions, as some publications resulted from international collaborations. Researchers addressed the following challenges: QPP for conversational search [Meng et al., 2023], known-item search and passage retrieval [Fröbe et al., 2023], QPP in the learning-to-rank and neural information retrieval domains [Datta et al., 2023b], issues with using correlation metrics to evaluate QPP [Mothe, 2023], QPP evaluation using pointwise approaches [Datta et al., 2023a], continuous evaluation [González-Sáez et al., 2023], and using information theory for QPP [Zendel et al., 2023].

The proceedings of the QPP++ 2023 workshop are publicly available online[1].

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 [Faggioli et al., 2023a] 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 without human relevance judgments [Carmel and Yom-Tov, 2010]. Since its introduction at the beginning of the 21st century, QPP has established itself as an essential tool in numerous tasks, including model selection [Carmel and Yom-Tov, 2010; Thomas et al., 2017], query suggestion [Carmel and Yom-Tov, 2010; Thomas et al., 2017], and rank fusion [Roitman, 2018]. 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.

2 Motivation

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 [Hauff, 2010], while new IR models often use indexes of embeddings or apply machine learning to re-rank documents [Mitra and Craswell, 2018]. Second, the vast majority of the recently developed retrieval models in IR utilize semantic information that, with a few notable exceptions [Mothe and Tanguy, 2005; Shtok et al., 2010], is rarely exploited by QPP models. This, in turn, impairs the performance of traditional QPP models applied on IR systems based on new

[1] https://ceur-ws.org/Vol-3366/

paradigms [Faggioli et al., 2023d]. Finally, QPP can be used for new processes such as selective query processing [Deveaud et al., 2018].

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 sought 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 judgments. Previous research demonstrated the unreliability of this approach when multiple experimental factors (i.e., IR models, corpora, and predictors) are considered [Hauff et al., 2009; Scholer and Garcia, 2009; Faggioli et al., 2021]. In addition, researchers demonstrated that high correlation does not necessarily translate to improved retrieval effectiveness [Raiber and Kurland, 2014; Hauff et al., 2009]. These issues are further exacerbated in new domains, such as question answering or conversational search [Faggioli et al., 2023c], where the evaluation of the retrieval models is often more challenging. The QPP++ 2023 workshop aimed to foster community discussion regarding these challenges.

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 image retrieval, 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?

2.1 QPP++ organization and execution

The QPP++ workshop was divided into two parts. During the first part, after a brief introduction of the themes and objectives given by the workshop's organizers, the authors of the accepted contributions presented their work to the audience. The goal was to set a common ground for the subsequent discussion and identify the main challenges that might arise in the future in applying QPP techniques.

The main challenge identified by the audience concerned the evaluation of QPP methods. Therefore, this was the main topic of the second part of the workshop that was organized following the world café methodology[2]. The workshop participants (approximately 20 people) were split into three small groups (5-7 people each) to discuss the theme chosen. Each group identified a

[2] https://theworldcafe.com/key-concepts-resources/world-cafe-method/

rapporteur responsible for annotating the comments and observations made in each discussion group. At the end of the first discussion turn, after approximately 20 minutes, members of each group – except the rapporteur – moved to a different group to further disseminate ideas and share opinions. At the end of the second turn, the three rapporteurs summarized the conclusions and ideas from the discussion to the whole QPP++ audience. We report in the next section the outcome of this discussion.

3 Focus Groups Discussion Outcome

During the QPP++ workshop, focus groups discussed the current methodology for evaluating QPP systems. The approach involves measuring the correlation between the values assigned to queries by a predictor and the true performance attained using relevance judgments, most often measured using Average Precision (AP). One major drawback of this methodology is that it is difficult to determine the connection between the observed correlation and downstream tasks' performance. Therefore, workshop participants suggested that the evaluation of QPP systems should focus on both intrinsic and extrinsic evaluation, including downstream tasks. Possible suitable tasks include choosing retrieval models, suggesting query reformulations, identifying queries that the system may not be able to answer, deciding when not to provide an answer, and detecting ambiguous queries.

The participants pointed out that the current method of evaluating QPP lacks qualitative analysis. Usually, the evaluation only focuses on the system's overall performance and does not scrutinize individual queries. To enhance the evaluation process, the community should also pay attention to single queries to find possible weak spots. This might include not just reporting the global correlation but also using scatterplots to show the connection between predictions and actual performance. Such an approach would facilitate a more accurate analysis of failures and the identification of any abnormal or weak points.

The low reproducibility of QPP methods is a major challenge. Even if the same ranking function is used, different hyperparameters can cause the predictors to behave differently. To address this issue, the focus groups recommended that practitioners and researchers release a set of artifacts, including:

- Code: the code used to preprocess the data, compute the predictors, and evaluate the results, should be released alongside the paper.
- Ranked lists of documents used as ground truth: since different IR frameworks and libraries may handle ties differently, providing the original ranked lists used as ground truth can help practitioners achieve comparable results when attempting to reproduce the results.
- Evaluation measures: different frameworks and libraries may implement slightly different versions of the same evaluation measure, or may have different default hyperparameters. To ensure reproducibility of how the ground truth is computed in the QPP domain, evaluation measures computed on the ranked lists should be released.
- Models: the use of trained models to compute predictions is becoming more common in learned and supervised QPPs and IR at large. Practitioners should release the models used whenever possible, along with the trained parameters, or clearly detail the training process, including the negative selection procedure, loss function, and hyperparameters used.

- Domain-specific information: specific scenarios may require additional information used by the QPP model. To ensure reproducibility and fair comparisons of approaches, practitioners should release as much domain-specific information as possible. For example, in the conversational search scenario, both original and reformulated queries should be released.
- Clear description of the experimental setup: a clear description of the experimental setup should be provided to enable a fair and comparable procedure. Practitioners should clearly state the procedure followed when describing the evaluation process, including the number of folds considered, standard deviation, hyperparameters evaluated, values of random seeds, and details on the statistical tests.

During the workshop, attendees suggested creating "Living Labs" dedicated to QPP. The lab's proposed lifecycle involves participants submitting their runs, which will be shared with all track participants during the annotation period. The QPP track participants will then compute the performance predictions and submit them. The QPP approaches would then be evaluated, allowing for "leaderboard approaches" where QPP strategies are evaluated and ranked based on their effectiveness.

Another aspect considered by the focus groups is defining a user model for QPP techniques. While the traditional IR literature has a well-established definition of what it means to model user satisfaction [Carterette, 2011], this is not the case for QPP. Future research should focus on modeling what it means to predict performance from theoretical and formal perspectives. As a first step, future publications should focus more explicitly on identifying and stating the prediction target, such as whether the QPP is trying to sort queries based on expected difficulty or actually predict performance. It is also necessary to further clarify what we mean by "difficulty" and "performance" and how they may change depending on the context considered. To improve the development of QPPs, the community should invest time and effort in framing the objectives of newly designed QPP models.

Acknowledgments

We would like to thank ECIR for hosting this workshop. We would also like to thank the program committee, consisting of Negar Arabzadeh, Fabio Giachelle, Claudia Hauff, Ornella Irrera, Stefano Marchesin, Jian-Yun Nie, Haggai Roitman, Laure Soulier, and Ellen Voorhees. Final thanks are due to the paper authors, and the participants for a great and lively workshop. CEUR-WS is hosting the proceedings[3] while the University of Padua is hosting the presentations[4].

References

David Carmel and Elad Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

[3] https://ceur-ws.org/Vol-3366/
[4] https://qpp.dei.unipd.it/

Ben Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, pages 903–912, 2011.

Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. On the feasibility and robustness of pointwise evaluation of query performance prediction. In Faggioli et al. [2023b], pages 1–6. URL http://ceur-ws.org/Vol-3366/#paper-01.

Suchana Datta, Debasis Ganguly, Josiane Mothe, and Md Zia Ullah. Combining word embedding interactions and LETOR feature evidences for supervised QPP. In Faggioli et al. [2023b], pages 7–14. URL http://ceur-ws.org/Vol-3366/#paper-02.

Romain Deveaud, Josiane Mothe, Md Zia Ullah, and Jian-Yun Nie. Learning to adaptively rank document retrieval system configurations. *ACM Transactions on Information Systems (TOIS)*, 37(1):1–41, 2018.

Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. An Enhanced Evaluation Framework for Query Performance Prediction. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021*, volume 12656 of *Lecture Notes in Computer Science*, pages 115–129, 2021.

Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, and Fiana Raiber. QPP++ 2023: Query-performance prediction and its evaluation in new tasks. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023*, volume 13982, pages 388–391, 2023a.

Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, and Fiana Raiber, editors. *Proceedings of the The QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks Workshop (QPP++)*, number 3366 in CEUR Workshop Proceedings, Aachen, 2023b. URL http://ceur-ws.org/Vol-3366/.

Guglielmo Faggioli, Nicola Ferro, Cristina Muntean, Raffaele Perego, and Nicola Tonellotto. A Geometric Framework for Query Performance Prediction in Conversational Search. In *Proceedings of the 46th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023*, 2023c.

Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, and Benjamin Piwowarski. Query Performance Prediction for Neural IR: Are We There Yet? In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023*, volume 13980 of *Lecture Notes in Computer Science*, pages 232–248, 2023d.

Maik Fröbe, Eric Oliver Schmidt, and Matthias Hagen. A large-scale dataset for known-item question performance prediction. In Faggioli et al. [2023b], pages 15–19. URL http://ceur-ws.org/Vol-3366/#paper-03.

Gabriela González-Sáez, Alaa El-Ebshihy, Petra Galuščáková, David Iommi, Florina Piroi, Lorraine Goeuriot, and Philippe Mulhem. Towards result delta prediction based on knowledge deltas for continuous IR evaluation. In Faggioli et al. [2023b], pages 20–24. URL http://ceur-ws.org/Vol-3366/#paper-04.

Claudia Hauff. Predicting the effectiveness of queries and retrieval systems. In *Ph.D. Dissertation. University of Twente.*, pages 1–179, 2010.

Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. The combination and evaluation of query performance prediction methods. In *Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009*, pages 301–312, 2009.

Chuan Meng, Mohammad Aliannejadi, and Maarten de Rijke. Performance prediction for conversational search using perplexities of query rewrites. In Faggioli et al. [2023b], pages 25–28. URL http://ceur-ws.org/Vol-3366/#paper-05.

Bhaskar Mitra and Nick Craswell. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval*, 13(1):1–126, 2018.

Josiane Mothe. On correlation to evaluate QPP. In Faggioli et al. [2023b], pages 29–36. URL http://ceur-ws.org/Vol-3366/#paper-06.

Josiane Mothe and Ludovic Tanguy. Linguistic Features to Predict Query Difficulty. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005*, pages 7–10, 2005.

Fiana Raiber and Oren Kurland. Query-Performance Prediction: Setting the Expectations Straight. In *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014*, pages 13–22, 2014.

Haggai Roitman. Enhanced performance prediction of fusion-based retrieval. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China, September 14-17, 2018*, pages 195–198, 2018. doi: 10.1145/3234944.3234950. URL https://doi.org/10.1145/3234944.3234950.

Falk Scholer and Steven Garcia. A Case for Improved Evaluation of Query Difficulty Prediction. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pages 640–641, 2009.

Anna Shtok, Oren Kurland, and David Carmel. Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010*, pages 259–266, 2010.

Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. Tasks, queries, and rankers in pre-retrieval performance prediction. Proceedings of the 22nd Australasian Document Computing Symposium, ADCS 2017, pages 1–4, 2017.

Oleg Zendel, Binsheng Liu, J. Shane Culpepper, and Falk Scholer. Entropy-based query performance prediction for neural information retrieval systems. In Faggioli et al. [2023b], pages 37–44. URL http://ceur-ws.org/Vol-3366/#paper-07.