New Perspectives to Query Performance Prediction Evaluation

Oleg Zendel
RMIT University
Melbourne, Australia
oleg.zendel@rmit.edu.au

EXTENDED ABSTRACT

The research on Query Performance Prediction (QPP) focuses on estimating the effectiveness of retrieval results in the absence of human relevance judgments. Accurately estimating the result of a search performed in response to a query has been extensively studied over the past two decades [1]. With the rising popularity of virtual assistants along with evolving research on complex information needs, e.g. [4, 5], the need for reliable QPP methods as well as the number of potential applications significantly increases. In this work, we focus on improving the evaluation framework of QPP. As we see the existing evaluation as a considerable limitation in the improvement of QPP methods, a reliable and improved evaluation framework would constitute a stepping-stone for a breakthrough in QPP.

The existing evaluation framework in QPP mainly relies on the measurement of the correlation coefficient between the per-query prediction scores and the actual per-query system effectiveness measure, usually Average Precision (AP). The QPP method that achieves higher correlation is considered to be superior. However, Hauff et al. [3] demonstrate that higher correlation does not vouch for more accurate prediction. The authors additionally advocate the usage of Fisher's z transformation and Confidence Intervals (CIs) to determine statistically significant differences between multiple correlation coefficients. Furthermore, the existing evaluation methodology is true only per a specific combination of a corpus, retrieval method, and set of queries; and does not necessarily hold if any of these is changed [6, 8]. That is, the existing evaluation is not agnostic to the different components, thus any conclusions about the relative prediction quality of the QPP methods should be taken with a grain of salt.

In the proposed research we aim to develop a better evaluation technique to reliably compare the performance of QPP methods. We intend to develop a new evaluation framework and standards that will simultaneously enable the utilization of query variants and take into consideration other confounding factors in QPP evaluation. Specifically, we raise the following research questions: (i) What limitations exist in the current evaluation practices of QPP? (ii) What are the best approaches to perform detailed failure analysis of query performance predictor results? (iii) How do existing QPP methods differ in performance on a set of topics (distinct information needs) represented by a single query versus a set of multiple queries which represent the same information need? (iv) How do the existing and new evaluation methodologies align with user satisfaction?

To answer the first two research questions Faggioli et al. [2] proposed a new evaluation framework for QPP. In the proposed framework an error is calculated for each query, resulting in a distribution of per-query errors for a set of queries. The new distribution of errors enables the authors to apply an N-way ANalysis Of VAriance (ANOVA) followed by a post-hoc analysis, Tukey's Honestly Significant Difference (HSD) test, to determine statistically significant differences between the multiple factors involved in the QPP evaluation. Separating the different components in the evaluation process allows reaching more reliable conclusions regarding the effects of each component in the prediction process.

As a preliminary study, Zendel et al. [7] compared multiple existing QPP methods in the aforementioned tasks; predicting the effectiveness for different queries representing different topics, and different query variants, that represent the same topic. They found that the difference in AP between the queries is an important confounding factor, that affects the prediction quality.

Future work will focus on developing a reliable evaluation framework for QPP both for queries from different topics and query variants from the same topic. A suitable framework should enable rigorous statistical analysis with decomposition and quantification of the different factors that affect QPP. In addition, a subsequent user study will explore how the new evaluation framework aligns with user satisfaction of QPP results.

Acknowledgements. The work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP190101113).

REFERENCES

[1] David Carmel and Elad Yom-Tov. 2010. Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 2, 1 (2010), 1–89.
[2] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2021. An Enhanced Evaluation Framework for Query Performance Prediction. In *Proc. ECIR*. 115–129.
[3] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. 2009. The Combination and Evaluation of Query Performance Prediction Methods. In *Proc. ECIR*. 301–312.
[4] Sharon Hirsch, Ido Guy, Alexander Nus, Arnon Dagan, and Oren Kurland. 2020. Query Reformulation in E-Commerce Search. In *Proc. SIGIR*. 1319–1328.
[5] Haggai Roitman, Shai Erera, and Guy Feigenblat. 2019. A Study of Query Performance Prediction for Answer Quality Determination. In *Proc. ICTIR*. 43–46.
[6] Falk Scholer and Steven Garcia. 2009. A Case for Improved Evaluation of Query Difficulty Prediction. In *Proc. SIGIR*. 640–641.
[7] Oleg Zendel, J. Shane Culpepper, and Scholer Falk. 2021. Is Query Performance Prediction With Multiple Query Variations Harder Than Topic Performance Prediction?. In *Proc. SIGIR*. https://doi.org/10.1145/3404835.3463039
[8] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. In *Proc. SIGIR*. 395–404.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07.
https://doi.org/10.1145/3404835.3463270