

On the Reproducibility and Robustness of Query Performance Prediction Experiments - An Extended Abstract

Suchana Datta¹, Debasis Ganguly², Mandar Mitra³ and Derek Greene¹

¹University College Dublin, Ireland

²University of Glasgow, United Kingdom

³Indian Statistical Institute, India

Query performance prediction (QPP), i.e. the process of estimating the retrieval quality of an IR system, has attracted the attention of the IR research community for several years. A diverse range of pre-retrieval (e.g. AvgIDF) and post-retrieval approaches (e.g. WIG, NQC, UEF) have been proposed for QPP. Specifically, given a query and an IR system, QPP methods compute a score that is indicative of the effectiveness of the system for the given query. While this score is typically not interpreted as a statistical estimate of a specific evaluation metric (e.g. AP or nDCG), it is indeed expected to be correlated with a standard evaluation measure computed over a ground-truth set of assessed relevant documents. Indeed, the effectiveness of a QPP method is determined by measuring the correlation between its predicted effectiveness scores and the values of some standard evaluation metric over a set of queries.

This abstract summarizes our ECIR 2022 research article ‘An Analysis of Variations in the Effectiveness of Query Performance Prediction’ [1], where we analysed the relative stability of QPP outcomes (rank correlations) with respect to changes in the IR models used to derive the top-retrieved documents, or the IR evaluation metrics used to order a given set of queries from easy to difficult. As per our findings, we emphasize in this abstract that such variations in QPP results (both in terms of the absolute values themselves and also in terms of the relative effectiveness of different QPP systems) can lead to difficulties in reproducing QPP experiment results on standard datasets.

We now summarize the research questions and the findings of our study [1]. The *context of a QPP experiment* depends on 3 factors, which are i) a list of top- κ documents retrieved, ii) the IR model or scoring function that is used to derive this list, and iii) an IR evaluation function (e.g., average precision or AP) that is used to induce an ordering over the set of queries (e.g., low AP to high AP indicating a spectrum of easy to difficult queries).

The first research question **RQ1**, that we investigated in our paper, [1] is - ‘Do variations in the QPP contexts lead to significant differences in measured QPP outcomes?’. Next, as the second


CIRCLE'22: Joint Conference of the Information Retrieval Communities in Europe, July 04–07, 2022, Samatan, Gers, France

✉ suchana.datta@ucdconnect.ie (S. Datta); debasis.ganguly@glasgow.ac.uk (D. Ganguly); mandar@isical.ac.in (M. Mitra); derek.greene@ucd.ie (D. Greene)

🌐 <https://ucdcs-research.ucd.ie/phd-student/suchana-datta/> (S. Datta); <https://gdebasis.github.io/> (D. Ganguly); <https://www.isical.ac.in/~mandar/> (M. Mitra); <http://derekgreene.com/> (D. Greene)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

research question **RQ2**, we investigated the following: ‘Do variations in the QPP contexts lead to significant differences in the relative effectiveness of different QPP methods?’.

To investigate the above research questions in [1], we conducted QPP experiments¹ on the widely-used TREC-Robust dataset, which consists of 249 queries. To obtain diverse QPP contexts for our experiments, we tried out a number of different combinations of IR models and IR evaluation metrics. Specifically, as IR models we employed a) language modeling with Jelinek-Mercer smoothing (LMJM), b) language modeling with Dirichlet smoothing (LMDir), and c) BM25. As choices for the IR evaluation metric, we considered a) AP, b) nDCG, c) P@10, and d) recall. We compared seven different QPP methods in our experiments, namely a) AvgIDF, b) Clarity, c) WIG, d) NQC, and three variants of UEF derived from three different base QPP models - e) UEF(Clarify), f) UEF(WIG) and g) UEF(NQC). We now summarize the main findings of our experimental study (for more details see [1]).

- Observations related to RQ1 (differences in QPP outcomes):
 - With NQC as the QPP method, we observed that LMJM yielded the highest deviations in observed QPP outcomes (specifically, rank correlation computed by τ) across the 4 different IR metrics used to obtain the QPP ground-truth (a reference order of query difficulty). The difference between the highest and the lowest rank correlation values were significant (0.3657 with recall and 0.2061 with P@10). This indicates that QPP methods do not generalize well enough across different IR metrics. In other words, it is hard to consistently predict which queries are easy and which ones are difficult across the different notions of how this *query difficulty* itself is defined (e.g., via a precision or a recall oriented measure).
 - With NQC as the QPP method, we observed significant differences in the highest and lowest QPP outcomes, across different IR models. The highest difference in rank correlation was recorded between BM25 ($\tau = 0.3563$) and LMDir ($\tau = 0.4354$), the QPP ground-truth being defined with respect to AP. This shows that the effectiveness of a QPP method is also not consistent for different IR models. In other words, it is not easy to predict that which queries are easy and which ones are difficult consistently well enough for different IR systems.
- Observations related to RQ2 (differences in relative effectiveness of QPP methods):
 - We observed that the relative ranks of QPP systems (ordered by the effectiveness measure τ) changed significantly across different IR metrics. The highest disagreement in the ranks were observed between AP@10 and recall@1000. This indicates that what may be the best QPP method for predicting the query difficulty induced by AP@10 may not be so for predicting query performance with respect to recall@1000.
 - Similar trends were also observed for differences in the choice of IR models in the QPP context. The highest disagreements in the relative performance of QPP methods were observed for the P@10 metric across LMJM and LMDir. This indicates that what may be the best QPP method for predicting the retrieval effectiveness of LMJM may not be the best one when it comes to predicting the system performance of LMDir.

¹Implementation available at: <https://github.com/suchanadatta/qpp-eval.git>

The main takeaway from this extended abstract is that since our extensive investigation in [1] has shown that QPP outcomes are indeed sensitive to the experimental setup used, any future experiment on QPP should emphasize clear specification of the experimental setup to warrant better reproducibility.

Acknowledgement

The first and the fourth authors were supported by the Science Foundation Ireland (SFI) grant number SFI/12/RC/2289_P2.

References

- [1] D. Ganguly, S. Datta, M. Mitra, D. Greene, An analysis of variations in the effectiveness of query performance prediction, in: Proc. of ECIR'22, 2022, pp. 215–229.