QPPTK@TIREx: Simplified Query Performance Prediction for Ad-Hoc Retrieval Experiments

Oleg Zendel[1], Maik Fröbe[2] and Guglielmo Faggioli[3]

[1]*RMIT University, Melbourne, Australia*
[2]*Friedrich-Schiller-Universität Jena, Jena, Germany*
[3]*University of Padua, Padua, Italy*

Abstract

We describe our software submission to the ECIR 2024 Workshop on Open Web Search [1]. We submit the query performance prediction toolkit QPPTK that comes with 12 performance predictors to substantially simplify the re-use of those predictors in ad-hoc retrieval experiments. Therefore, we have extended QPPTK so that it can run in TIREx and that it can process arbitrary workloads using ir_datasets and PyTerrier indices as inputs. We execute QPPTK on all 23 test collections in TIREx that have PyTerrier indices available and make their predictions and the Docker image publicly available. Thereby, subsequent retrieval experiments can easily re-use the predictions by just downloading a few kilobytes, instead of having to run the system on their machine (which is still possible, e.g., by executing the Docker image on new or modified inputs). Our analysis on the 23 test collections highlights predictor performance variability, emphasizing the importance of standardized baselines and evaluation methods. Openly sharing predictions aims to enhance research accessibility, promoting broader utilization of query performance predictors and inspiring the development of novel prediction and evaluation techniques.

1. Introduction

Query Performance Prediction (QPP) is defined as assessing the quality of the query effectiveness in the absence of relevance judgements [2, 3]. This task is particularly important under three major aspects. First, it allows for some form of automatic evaluation of the Information Retrieval (IR) system, thus reducing the cost of collecting highly expensive manual relevance annotations. Secondly, it can be used as a feature for IR tasks such as reranking, model selection [3, 4], and rank fusion [5]. From the query perspective, it can be used to carry out query suggestion [3, 4] or to identify particularly challenging queries, so that the system administrators can operate failure analysis and enhance the model on those specific queries [3].

While QPP holds significant potential for enhancing IR, its progress faces a major challenge linked to the evaluation and comparison methodologies employed. Building on discussions from the recent QPP++ workshop [6], we assert that a fundamental issue in QPP research lies in the lack of result reproducibility and stability. First of all, most of the traditional QPP approaches have been designed and tested considering the Query Language Model [7] as the underlying retrieval engine. Therefore, this might induce a general instability of the models when considering different retrieval models, especially if they are based on a widely different rationale, such as neural IR models [8, 9, 10]. This also reflects on the possible instability

1st International Workshop on Open Web Search (WOWS), in: Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024), March 28, 2024, Glasgow, UK

due to the usage of different hyper-parameters for the IR model, different implementations of it, or different query normalization strategies, such as the stemmer used or stopping list considered [4, 11, 12]. Secondly, QPP models are typically characterized by one or more parameters, such as the length of the retrieved list when considering post-retrieval models. If not properly tuned, such parameters might result in particularly ineffective predictions. Finally, QPP models are known to operate differently on different collections [2]: using the wrong QPP for a given collection might lead to low prediction performance and thus to a generally weak baseline. To summarize, when experimenting with QPP, three major aspects might substantially impact the performance of (baseline) predictors: the underlying retrieval model, the hyperparameters of the predictor, and the collection considered. Furthermore, the same aspects hinder the reproducibility of the QPP: it is not uncommon in the QPP scenario to observe the same model performing differently from paper to paper. This is often due to naturally occurring differences in the implementation, but also due to different experimental settings (i.e., the retrieval model, and/or hyperparameters).

To alleviate these limitations and foster simplified, reproducible and stable experimentation in QPP, we describe our approach to embed the Query Performance Prediction ToolKit (qpptk) [11, 12, 13] into TIRA [14] / TIREx [15]. This setup allows us to provide a solid and stable set of baselines for future experiments: when carrying out new experiments, we can ensure that the implementation and hyper-parameters of the IR models underneath remain the same as they are archived within TIREx. Secondly, it allows us to test the same QPP over the abundance of collections that are already available in TIREx. This, in turn, allows us to provide the research community with a large amount of – already computed – shared baselines for future QPP experiments. Finally, re-using cached outputs can help fast prototyping: when developing a new QPP, the practitioner can delegate secondary aspects, such as setting up the IR system or processing the corpus, to cached TIREx outputs, focusing exclusively on the development of the QPP. Our implementation is publicly available.[1]

The remainder of the paper is organized as follows: Section 2 provides the background on the QPP methods, and describes the main QPP models that have been implemented in qpptk. Section 3 details how our qpptk component submitted to TIREx can be used to obtain reproducible QPP results. Section 4 reports our experimental analysis. Finally, section 5 draws the conclusions and outlines our future work.

2. Background

We describe here the predictors implemented within qpptk. According to the classical separation, predictors are divided into pre- and post-retrieval.

Before introducing the predictors, we provide here the notation adopted in the remainder of this work. Let q be a query and d a document belonging to a corpus of documents C with $|C| = N$. Without loss of generality, we call $s(q, d)$ the score assigned by an arbitrary ranking model to the document d in response to the query q. We call $\mathcal{D}@k$ the list of the top-k documents retrieved in response to the query. Additionally, given a term t, we define $f_t = |\{d \in C : t \in d\}|$ the document frequency (i.e., the number of documents the term appears

[1] Code: https://github.com/Zendelo/QPP-EnhancedEval/tree/qpptk-dev

in), while $f_{d,t} = |\{w \in d : w = t\}|$ is the term frequency (i.e., the number of times the term appear in document d).

2.1. Pre-Retrieval Predictors

Pre-retrieval predictors are those predictors that base their prediction based only on the query tokens and the (indexed) corpus. We implemented three major categories of pre-retrieval predictors: those based on Inverse Document Frequency (IDF), those relying on Similarity between the query and the collection (SCQ), and those based on score variability (VAR).

Inverse Document Frequency (IDF)-based Predictors [16, 17] The IDF predictors rely on computing the IDF for each query term:

$$IDF(t) = \ln\left(1 + \frac{N}{f_t}\right)$$

then, this predictor can be instantiated in two ways, either by computing the average IDF over all query terms (avgIDF) or by computing the maximum IDF (maxIDF) over the query terms. The rationale is that if the terms have high inverse document frequency, they are highly characterizing (i.e., they are contained in a few documents and are very specific). Therefore, it is more likely that documents containing such terms will be relevant, indicating how the retrieval will perform.

Similarity between a Query and a Collection (SCQ)-based Predictors [16] The term-wise SCQ score is defined as follows:

$$SCQ(t) = (1 + \ln(f_{c,t})) \cdot \left(1 + \frac{N}{f_t}\right)$$

where $f_{c,t}$ is the number of times the term appears in the corpus. Once the SCQ has been computed for each term, it is possible to aggregate it either by summing the SCQ for all the query terms (SCQ), averaging (avgSCQ), or by computing the maximum (maxSCQ).

Variability (VAR)-Based Predictors [16] To compute the VAR predictors, it is first necessary to compute the weight $w_{d,t}$ of each query term t with respect to each document d as the TFIDF score $w_{d,t} = 1 + \ln(f_{d,t}) \cdot IDF(t)$. Called $\mathcal{D}_t$ the set of documents containing t, the prediction weight of each term is defined as follows:

$$VAR(t) = \sum_{t \in q} \sqrt{\frac{1}{f_t} \sum_{d \in \mathcal{D}_t} (w_{d,t} - \overline{w}_t)^2}$$

where $\overline{w}_t$ is the average weight $w_{d,t}$ over the documents in $\mathcal{D}_t$. As for the previous cases, the VAR score is aggregated over the query terms using either the sum (sumVAR), maximum (maxVAR), or the average (avgVAR).

2.2. Post-Retrieval Predictors

Post-retrieval predictors utilize both the query and the top-k retrieved documents to formulate their prediction. Among post-retrieval predictors, we recognize three major classes: those based on the coherence of the retrieved list (e.g., clarity), those based on the distribution of the scores (e.g., NQC, WIG, SMV), and those relying on the robustness of the results – i.e., how much the introduction of noise in the query or the collection changes the retrieved ranked list – such as the UEF framework.

Clarity [17] This represents one of the seminal efforts in the QPP domain. The approach consists of computing the language model for the first top-k documents, which we refer to as $\theta_{\mathcal{D}@k}$. Then, this language model is compared with the language model of the entire corpus θ_C. The rationale is that the language model for the first top-k is highly divergent from the language model of the entire collection, then documents are highly coherent internally and this hints at an effective retrieval. More in detail:

$$Clarity(q) = \sum_{w \in V} p(w|\theta_{\mathcal{D}@k}) \frac{p(w|\theta_{\mathcal{D}@k})}{p(w|\theta_C)},$$

where V is the vocabulary and $p(w|\theta)$ is the probability of observing the token w according to θ the language model.

Weighted Information Gain (WIG) [18] This predictor represents one of the first efforts in utilizing the distribution of the scores for the top-k retrieved documents to determine the retrieval performance. More in detail, the prediction is given as the average difference between the score of the top-k documents retrieved and the score that the entire corpus would obtain in response to the query, which acts as a regularization component.

$$WIG(q) = \frac{1}{k\sqrt{|q|}} \sum_{d \in \mathcal{D}@k} \left(s(q,d) - s(q,C)\right).$$

Normalized Query Commitment (NQC) [19] This predictor is in line with WIG, with the main difference that, in this case, the statistic of interest is the variance of the scores of the first top-k documents, normalized by the score that the entire corpus would obtain in response to the query:

$$NQC(q) = \frac{\sqrt{\frac{1}{k}\sum_{d \in \mathcal{D}@k}(s(q,d) - \hat{\mu}_{\mathcal{D}@k})^2}}{s(q,C)},$$

where $\hat{\mu}_{\mathcal{D}@k} = \frac{1}{k} \cdot \sum_{d \in \mathcal{D}@k} s(q,d)$.

Score Magnitude and Variance (SMV) [20] This QPP combines NQC and WIG, by taking into account both the magnitude of the scores, as well as their variance and it is defined as follows:

$$SMV(q) = \frac{\frac{1}{k}\sum_{d \in \mathcal{D}@k}\left(s(q,d) \cdot \left|\ln \frac{s(q,d)}{\hat{\mu}_{\mathcal{D}@k}}\right|\right)}{s(q,C)}.$$

3. Porting QPPTK to TIREx for Simplified QPP Experiments

The QPPTK toolkit currently implements eight pre-retrieval and four post-retrieval QPP methods. Leveraging a PyTerrier index, the toolkit is designed with extensibility in mind, allowing for easy integration of additional methods. You can access QPPTK on GitHub.[2] In this work, we integrate QPPTK into the TIREx framework [15].[3] We dockerize QPPTK so that it can run in the TIRA sandbox in TIREx and adopt it so that it can use arbitrary inputs from ir_datasets [21]. To reduce the effort of running performance predictions, we configure QPPTK in TIREx so that it uses the the PyTerrier [22] Indexer that is dockerized in TIRA as previous stage so that it runs against prebuilt PyTerrier indices which makes its execution faster.

The incorporation of QPPTK into TIREx serves multiple purposes. Firstly, it establishes a stable and reproducible baseline for future QPP experiments. Secondly, it allows the testing of identical QPP methods across all collections available in TIREx, contributing multiple shared baselines to the research community. Lastly, it facilitates swift and seamless prototyping. Developers can focus exclusively on QPP method development and evaluation, delegating secondary tasks like setting up the IR system or processing the corpus to TIREx.

To use QPPTK in TIREx, users must install the TIRA package and execute the following code:[4]

```python
from tira.rest_api_client import Client
from tira.third_party_integrations import ensure_pyterrier_is_loaded

ensure_pyterrier_is_loaded()

import pyterrier as pt

tira = Client()

dataset = 'antique-test-20230107-training'
# load the dataset
pt_dataset = pt.get_dataset(f"irds:ir-benchmarks/{dataset}")

# initialize QPP transformer
qpp_predictions = tira.pt.transform_queries(
    'ir-benchmarks/qpptk/all-predictors',
    dataset)
# returns a DataFrame with all the predictions
qpp_predictions(pt_dataset.get_topics('query'))
```

Listing 1: Example of using QPPTK within TIREx.

The code in Listing 1 demonstrates how using only a few lines of code, users can access the TIRA API, load a dataset, and initialize a QPP transformer. The transformer then returns a DataFrame with all the predictions. This example illustrates the ease of use and the potential of the TIREx framework in facilitating QPP experiments.

[2] https://github.com/Zendelo/QPP-EnhancedEval.

[3] See https://www.tira.io/tirex. Accessed on 28-03-2024.

[4] Assuming that PyTerrier and `ir_datasets` are already installed.

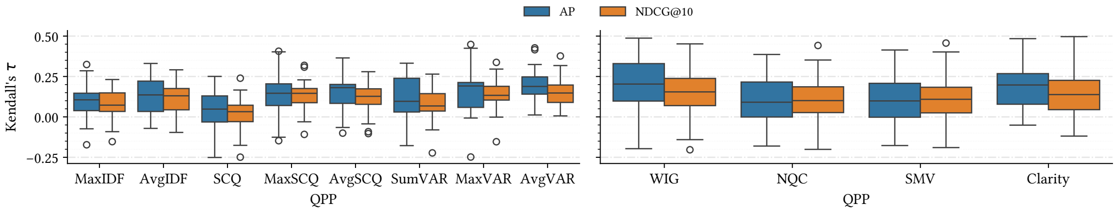

Figure 1: Boxplots detailing the Kendall's τ coefficients for various QPP methods. Each boxplot encapsulates the τ coefficient's distribution for the given QPP method across the 23 diverse datasets.

4. Experimental Analysis

In the experimental phase, we perform retrievals using BM25, followed by the application of QPPTK to the 23 test collections within TIREx with PyTerrier indices, covering diverse retrieval scenarios, such as argumentation [23, 24, 23, 24, 25], web search [26, 27, 28, 29, 30, 31], news search [32, 33, 34, 35] , question answering [36, 37, 38], and medical domains [39, 40, 41, 42, 43, 44, 45], and the historical Cranfield collection [46, 47]. The retrieval evaluation is conducted using the Average Precision (AP) metric, in line with prior work on QPP. The initial analysis involves comparing the outputs of QPP methods through the computation of the Kendall's τ correlation coefficient with the AP values per-query. Kendall's τ serves as a common metric in scrutinizing and comparing QPP method performances. The results are presented in Figure 1. Each cell encapsulates the τ coefficient's dispersion for the respective QPP method for the corresponding dataset. For post-retrieval methods, these boxplots additionally incorporate the distinct parameters utilized for each method.

The observed boxplots highlight a substantial variation in prediction quality across datasets, covering a correlation spectrum from -0.25 to 0.5. This outcome underscores the dataset-centric nature of QPP method performance, aligning with established observations in prior research.[5] Given that the correlation coefficients are strongly dependent on the sample, we also make all the results available on TIREx, enabling anyone to reproduce and extend our analysis with minimal effort, thus fostering the reproducibility of QPP research.

To investigate the performance of QPP methods in more detail, we present a heatmap in Figure 2. This heatmap offers a detailed perspective on the Kendall's τ coefficients for the pre-retrieval QPP methods, with each cell denoting the τ coefficient for a particular QPP method and dataset.

The heatmap visually illustrates the performance variations of QPP methods across the diverse datasets. Our observation from the heatmap highlights the dataset-centric nature of QPP method performance – no single method consistently outperforms others across all datasets. However, certain methods demonstrate superior performance on average, such as MaxVAR, AvgVAR and MaxSCQ. Notably, the pre-retrieval QPP methods operate independently of the retrieval model, making the prediction values agnostic to the retrieval model used. Nevertheless, it is essential to recognize that the correlation is measured with the retrieval model, introducing a dependency on the retrieval model employed. Consequently, direct comparisons across different publications are challenging unless the same index, retrieval model, and associated components

[5] See https://github.com/chauff/QPP-Overview, accessed on 28-03-2024.

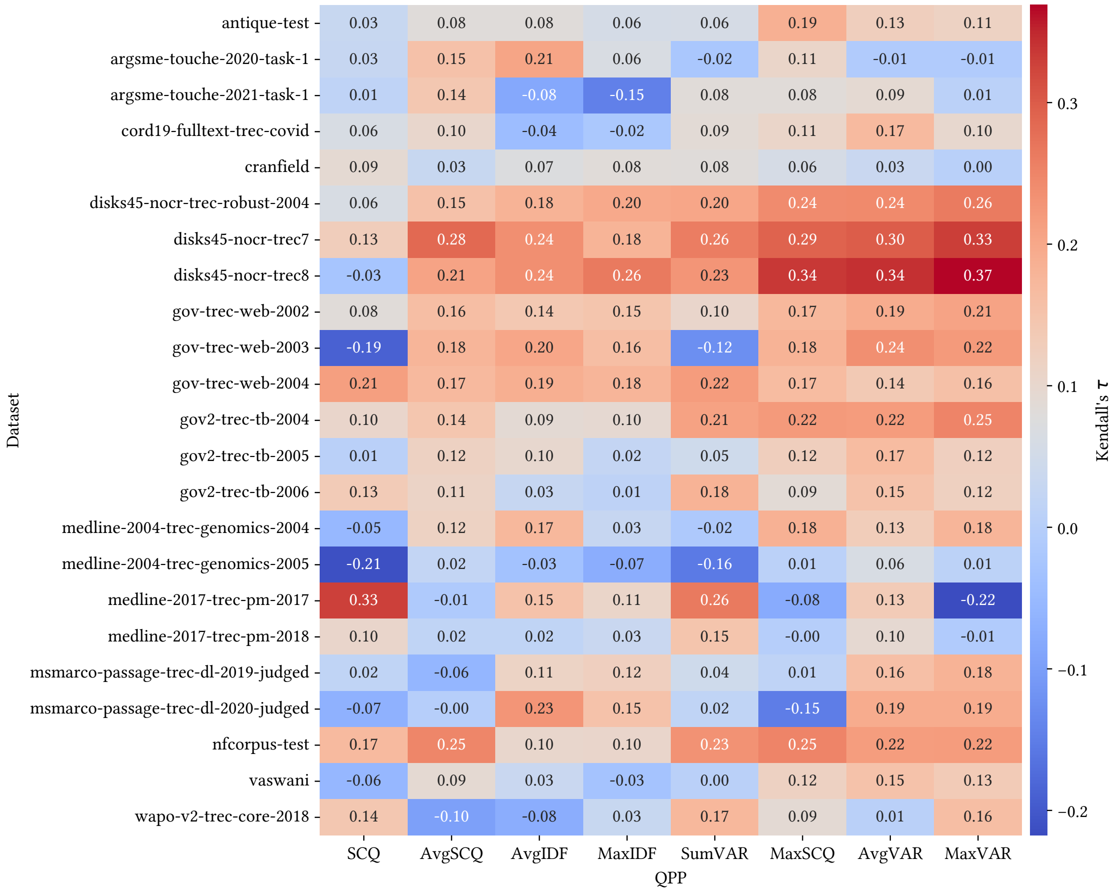

Figure 2: Heatmap illustrating the Kendall's τ coefficients for the pre-retrieval QPP methods. Each cell denotes the τ coefficient for a specific QPP method and dataset.

are employed consistently.

To further demonstrate the potential of QPPTK@TIREx for the community, we conduct a comprehensive ANalysis Of the VAriance (ANOVA) on the pre-retrieval QPP methods, employing 10 different retrieval models (rankers) and 23 diverse datasets. The rankers include BM25, LGD, PL2, TF-IDF, ANCE-Cosine, ColBERT, SBERT, MonoT5-3b, MonoT5-Base and DirichletLM. We randomly sample 30 queries from each dataset, reflecting the maximum number available across all datasets. We then compute sARE error values for each query across all datasets and rankers,[6] utilizing them for the ANOVA analysis. The results, presented in Table 1, reveal that the ranker, dataset, and QPP factors are all statistically significant, indicating that at least one of the levels of each factor has a statistically significant effect on the prediction quality. The effect sizes are small, with the dataset factor having the largest effect size on prediction quality. The interaction effects Ranker:Dataset and Dataset:QPP are also significant, with Dataset:QPP having the most substantial effect size. This underscores the dependency of pre-retrieval QPP method performance on the dataset, with the influence of the retrieval model being dataset-specific.

[6] Similarly to correlation, the values are computed per each combination of dataset, ranker, and QPP method.

Table 1
ANOVA analysis of the pre-retrieval QPP methods. The table presents the degrees of freedom (DF), sum of squares (SS), mean square (MS), F-statistic (F), p-value (PR(>F)), and the effect size (ω^2) for the ranker, dataset, and QPP factors. Note, that ω^2 is ill-defined for non-significant effects.

	DF	SS	MS	F	PR(>F)	ω^2
Ranker	9	6.290	0.699	14.025	<0.001	0.002
Dataset	22	23.247	1.057	21.206	<0.001	0.008
QPP	7	4.260	0.609	12.214	<0.001	0.001
Ranker:Dataset	198	26.832	0.136	2.720	<0.001	0.006
Ranker:QPP	63	1.344	0.021	0.428	1.000	-0.001
Dataset:QPP	154	33.808	0.220	4.406	<0.001	0.010
Residual	52117	2596.945	0.050			

The Ranker:QPP interaction, however, is not found to be significant by itself.

5. Conclusion and Future Work

This paper introduces the QPPTK toolkit, encompassing a diverse array of pre-retrieval and post-retrieval QPP methods. We present the integration of QPPTK into TIREx, to simplify rapid and efficient experimentation with QPP methods.

The integration helps to establish stable and reproducible baselines for forthcoming QPP experiments, testing identical QPP methods across all available TIREx collections, and contributing multiple shared baselines to the research community. Currently, available QPP methods in QPPTK include eight pre-retrieval methods and four post-retrieval methods, with the potential for further expansion. Moving forward, our future work involves expanding the QPPTK toolkit to incorporate additional QPP methods and enhancing its integration into TIREx. All code and data used in this work are accessible on GitHub and Zenodo,[7] encouraging the research community to utilize and build upon our contributions.

References

[1] S. Farzana, M. Fröbe, M. Granitzer, G. Hendriksen, D. Hiemstra, M. Potthast, S. Zerhoudi, 1st International Workshop on Open Web Search (WOWS), in: Advances in Information Retrieval. 46th European Conference on IR Research (ECIR 2024), Springer, 2024.

[2] C. Hauff, Predicting the Effectiveness of Queries and Retrieval Systems, Ph.D. thesis, University of Twente, Enschede, Netherlands, 2010. URL: http://eprints.eemcs.utwente.nl/17338/.

[3] D. Carmel, E. Yom-Tov, Estimating the Query Difficulty for Information Retrieval, Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers, 2010. URL: https://doi.org/10.2200/S00235ED1V01Y201004ICR015. doi:10.2200/S00235ED1V01Y201004ICR015.

[7] https://zenodo.org/records/10852738

[4] P. Thomas, F. Scholer, P. Bailey, A. Moffat, Tasks, queries, and rankers in pre-retrieval performance prediction, in: Proceedings of the 22nd Australasian Document Computing Symposium, ADCS 2017, 2017, pp. 1–4.

[5] H. Roitman, Enhanced performance prediction of fusion-based retrieval, in: D. Song, T. Liu, L. Sun, P. Bruza, M. Melucci, F. Sebastiani, G. H. Yang (Eds.), Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2018, Tianjin, China, September 14-17, 2018, ACM, 2018, pp. 195–198. URL: https://doi.org/10.1145/3234944.3234950. doi:10.1145/3234944.3234950.

[6] G. Faggioli, N. Ferro, J. Mothe, F. Raiber, M. Fröbe, Report on the 1st workshop on query performance prediction and its evaluation in new tasks (qpp++ 2023) at ecir 2023, SIGIR Forum 57 (2023). URL: https://doi.org/10.1145/3636341.3636356. doi:10.1145/3636341.3636356.

[7] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to information retrieval, ACM Trans. Inf. Syst. 22 (2004) 179–214. URL: https://doi.org/10.1145/984321.984322. doi:10.1145/984321.984322.

[8] H. Hashemi, H. Zamani, W. B. Croft, Performance prediction for non-factoid question answering, in: Y. Fang, Y. Zhang, J. Allan, K. Balog, B. Carterette, J. Guo (Eds.), Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2019, Santa Clara, CA, USA, October 2-5, 2019, ACM, 2019, pp. 55–58. URL: https://doi.org/10.1145/3341981.3344249. doi:10.1145/3341981.3344249.

[9] G. Faggioli, T. Formal, S. Marchesin, S. Clinchant, N. Ferro, B. Piwowarski, Query performance prediction for neural IR: are we there yet?, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I, volume 13980 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 232–248. URL: https://doi.org/10.1007/978-3-031-28244-7_15. doi:10.1007/978-3-031-28244-7_15.

[10] S. Datta, D. Ganguly, M. Mitra, D. Greene, A relative information gain-based query performance prediction framework with generated query variants, ACM Trans. Inf. Syst. 41 (2023) 38:1–38:31. URL: https://doi.org/10.1145/3545112. doi:10.1145/3545112.

[11] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I, volume 12656 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 115–129. URL: https://doi.org/10.1007/978-3-030-72113-8_8. doi:10.1007/978-3-030-72113-8_8.

[12] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, smare: a new paradigm to evaluate and understand query performance prediction methods, Inf. Retr. J. 25 (2022) 94–122. URL: https://doi.org/10.1007/s10791-022-09407-w. doi:10.1007/S10791-022-09407-W.

[13] O. Zendel, J. S. Culpepper, F. Scholer, Is query performance prediction with multiple query variations harder than topic performance prediction?, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3404835.3463039. doi:10.1145/3404835.3463039.

[14] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[15] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The information retrieval experiment platform, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 2826–2836. URL: https://doi.org/10.1145/3539618.3591888. doi:10.1145/3539618.3591888.

[16] Y. Zhao, F. Scholer, Y. Tsegay, Effective pre-retrieval query performance prediction using similarity and variability evidence, in: C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, R. W. White (Eds.), Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings, volume 4956 of *Lecture Notes in Computer Science*, Springer, 2008, pp. 52–64. URL: https://doi.org/10.1007/978-3-540-78646-7_8. doi:10.1007/978-3-540-78646-7_8.

[17] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: K. Järvelin, M. Beaulieu, R. A. Baeza-Yates, S. Myaeng (Eds.), SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, ACM, 2002, pp. 299–306. URL: https://doi.org/10.1145/564376.564429. doi:10.1145/564376.564429.

[18] Y. Zhou, W. B. Croft, Query performance prediction in web search environments, in: W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, ACM, 2007, pp. 543–550. URL: https://doi.org/10.1145/1277741.1277835. doi:10.1145/1277741.1277835.

[19] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits, Predicting query performance by query-drift estimation, ACM Trans. Inf. Syst. 30 (2012) 11:1–11:35. URL: https://doi.org/10.1145/2180868.2180873. doi:10.1145/2180868.2180873.

[20] Y. Tao, S. Wu, Query performance prediction by considering score magnitude and variance together, in: J. Li, X. S. Wang, M. N. Garofalakis, I. Soboroff, T. Suel, M. Wang (Eds.), Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014, ACM, 2014, pp. 1891–1894. URL: https://doi.org/10.1145/2661829.2661906. doi:10.1145/2661829.2661906.

[21] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: SIGIR, 2021.

[22] C. Macdonald, N. Tonellotto, Declarative experimentation ininformation retrieval using pyterrier, in: Proceedings of ICTIR 2020, 2020.

[23] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2021: Argument Retrieval, in: K. Candan, B. Ionescu, L. Goeuriot, H. Müller, A. Joly, M. Maistro, F. Piroi,

G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 12th International Conference of the CLEF Association (CLEF 2021), volume 12880 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2021, pp. 450–467.

[24] A. Bondarenko, M. Fröbe, J. Kiesel, S. Syed, T. Gurcke, M. Beloucif, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of Touché 2022: Argument Retrieval, in: A. Barrón-Cedeño, G. D. S. Martino, M. D. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2022.

[25] A. Bondarenko, M. Fröbe, J. Kiesel, F. Schlatt, V. Barriere, B. Ravenet, L. Hemamou, S. Luck, J. H. Reimer, B. Stein, M. Potthast, M. Hagen, Overview of Touché 2023: Argument and Causal Retrieval, in: Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023.

[26] N. Craswell, D. Hawking, Overview of the TREC-2002 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Eleventh Text REtrieval Conference, TREC 2002, Gaithersburg, Maryland, USA, November 19-22, 2002, volume 500-251 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2002.

[27] N. Craswell, D. Hawking, R. Wilkinson, M. Wu, Overview of the TREC 2003 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Twelfth Text REtrieval Conference, TREC 2003, Gaithersburg, Maryland, USA, November 18-21, 2003, volume 500-255 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2003, pp. 78–92.

[28] N. Craswell, D. Hawking, Overview of the TREC 2004 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004.

[29] C. L. A. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2004 terabyte track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004.

[30] C. L. A. Clarke, F. Scholer, I. Soboroff, The TREC 2005 terabyte track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005, volume 500-266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2005.

[31] S. Büttcher, C. L. A. Clarke, I. Soboroff, The TREC 2006 terabyte track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14-17, 2006, volume 500-272 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2006.

[32] E. M. Voorhees, D. Harman, Overview of the seventh text retrieval conference (trec-7), in: TREC, 1998.

[33] E. M. Voorhees, D. Harman, Overview of the eight text retrieval conference (trec-8), in:

TREC, 1999.

[34] E. M. Voorhees, Nist trec disks 4 and 5: Retrieval test collections document set, 1996.

[35] E. Voorhees, Overview of the trec 2004 robust retrieval track, in: TREC, 2004.

[36] H. Hashemi, M. Aliannejadi, H. Zamani, W. B. Croft, ANTIQUE: A non-factoid question answering benchmark, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, volume 12036 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 166–173.

[37] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019 Deep Learning Track, in: E. Voorhees, A. Ellis (Eds.), 28th International Text Retrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, NIST Special Publication, National Institute of Standards and Technology (NIST), 2019.

[38] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 Deep Learning Track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the 29th Text REtrieval Conference, TREC 2020, Virtual Event, Gaithersburg, MD, USA, November 16-20, 2020, volume 1266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2020.

[39] E. M. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, TREC-COVID: constructing a pandemic information retrieval test collection, SIGIR Forum 54 (2020) 1:1–1:12.

[40] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: the covid-19 open research dataset, CoRR abs/2004.10706 (2020). arXiv:2004.10706.

[41] W. R. Hersh, R. T. Bhupatiraju, L. Ross, A. M. Cohen, D. Kraemer, P. Johnson, TREC 2004 genomics track overview, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2004.

[42] W. R. Hersh, A. M. Cohen, J. Yang, R. T. Bhupatiraju, P. M. Roberts, M. A. Hearst, TREC 2005 genomics track overview, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005, volume 500-266 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2005.

[43] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, S. Pant, Overview of the TREC 2017 precision medicine track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, volume 500-324 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2017.

[44] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, Overview of the TREC 2018 precision medicine track, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018, volume 500-331 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2018.

[45] V. Boteva, D. G. Ghalandari, A. Sokolov, S. Riezler, A full-text learning to rank dataset for

medical information retrieval, in: N. Ferro, F. Crestani, M. Moens, J. Mothe, F. Silvestri, G. M. D. Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings, volume 9626 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 716–722.

[46] C. Cleverdon, The Cranfield tests on index language devices, in: ASLIB Proceedings, MCB UP Ltd. (Reprinted in Readings in Information Retrieval, Karen Sparck-Jones and Peter Willett, editors, Morgan Kaufmann, 1997), 1967, pp. 173–192.

[47] C. W. Cleverdon, The significance of the Cranfield tests on index languages, in: A. Bookstein, Y. Chiaramella, G. Salton, V. V. Raghavan (Eds.), Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum), ACM, 1991, pp. 3–12.