Combining Query Performance Predictors: A Reproducibility Study

Sourav Saha[1][0000−0001−8091−0685] Suchana Datta[2][0000−0001−9220−6652], Dwaipayan Roy[3][0000−0002−5962−5983], Mandar Mitra[1][0000−0001−9045−9971], and Derek Greene[2][0000−0001−8065−5418]

[1] Indian Statistical Institute, Kolkata, India
[2] University College Dublin, Ireland
[3] Indian Institute of Science Education and Research, Kolkata, India
souravsaha.juit@gmail.com, suchana.datta@ucd.ie, dwaipayan.roy@iiserkol.ac.in, mandar@isical.ac.in, derek.greene@ucd.ie

Abstract. A large number of approaches to Query Performance Prediction (QPP) have been proposed over the last two decades. As early as 2009, Hauff et al. [28] explored whether different QPP methods may be combined to improve prediction quality. Since then, significant research has been done both on QPP approaches, as well as their evaluation. This study revisits Hauff et al.'s work to assess the reproducibility of their findings in the light of new prediction methods, evaluation metrics, and datasets. We expand the scope of the earlier investigation by: (i) considering post-retrieval methods, including supervised neural techniques (only pre-retrieval techniques were studied in [28]); (ii) using *sMARE* for evaluation, in addition to the traditional correlation coefficients and *RMSE*; and (iii) experimenting with additional datasets (Clueweb09B and TREC DL). Our results largely support previous claims, but we also present several interesting findings. We interpret these findings by taking a more nuanced look at the correlation *between* QPP methods, examining whether they capture diverse information or rely on overlapping factors.

Keywords: Query Performance Prediction · Fusion · Evaluation · Reproducibility · Empirical Study.

1 Introduction

Query Performance Prediction (QPP) has long been an important area of study within information retrieval (IR) [9,15,17,29,34,45,50,52,60,61]. QPP focuses on estimating (without utilizing relevance judgments) how well a search engine will perform on a given query. Accurate predictions of query performance can guide decisions regarding search strategies, resource allocation, and query reformulation [25,36,44,1]. Over the years, various QPP approaches have been proposed [15,59,6,42,41,32,39,2,16,10,61,51,45,52]. These methods can be broadly categorized as pre-retrieval or post-retrieval techniques. Pre-retrieval methods [29,31,59] estimate query difficulty using only the query and collection statistics, while post-retrieval methods [15,52,61,50,46,45,51] also use retrieval results (ranked lists of documents) to make predictions.

Given the plethora of methods proposed in the literature, researchers have also investigated whether combining different QPP methods yields better predictions [54,61,28,50,34,35]. In a comprehensive study involving 19 pre-retrieval predictors (but no post-retrieval methods) and 3 test collections, Hauff et al. [28] showed that combining predictors yields significant improvements in correlation between predicted and actual performance. However, the observed gains in terms of Root Mean Squared Error (*RMSE*) were modest, or even non-existent. Based on these observations, the authors proposed two important guidelines regarding QPP evaluation:

i) When reporting the correlation[1] between predictions and actual performance, a 95% confidence interval for the observed correlation should also be reported. This helps in assessing whether observed differences between methods are statistically significant.
ii) In addition to ρ, a listwise metric, the Squared Error (SE) between the predicted and measured retrieval effectiveness should be used as a more interpretable, pointwise metric for QPP, with Root Mean Squared Error (*RMSE*) representing the aggregation of these values across a test collection.

In the 15 years since the study by Hauff et al., QPP has witnessed a great deal of research activity. Our study revisits QPP method fusion in the light of these advancements, in order to address the following research questions:

RQ1 - Does the combination of pre-retrieval QPP methods consistently enhance correlation with Average Precision (*AP*) across different datasets, including recent web collections, as suggested in [28]? Additionally, does combining predictors significantly affect *RMSE*? (Hauff et al.'s results suggest a negative answer to this question.)

RQ2 - Do the answers to **RQ1** change if we consider post-retrieval predictors instead of pre-retrieval estimators?

RQ3 - Is there a relationship between the predictions of different QPP methods, both within and across pre-retrieval and post-retrieval categories?

To answer the above questions, and to assess whether Hauff et al.'s observations generalize to newer QPP methods, evaluation metrics, and more modern datasets, we expand the scope of our experiments along the directions listed below:

- **Incorporation of post-retrieval methods:** While Hauff et al. [28] limited their analysis to pre-retrieval methods, we also consider post-retrieval methods. This includes not only traditional unsupervised methods but also more recent supervised neural approaches.

[1] In Sections 1–2, we abuse notation, and use ρ as an abbreviation for any of the correlation coefficients (Pearson's or Spearman's Rho and Kendall's Tau) that are commonly used when evaluating QPP methods. In Sections 3–5 (Experimental setup, results and analysis), ρ and τ are used in accordance with standard convention.

– **Evaluation metrics:** When measuring prediction quality, we use *scaled Mean Absolute Rank Error* (*sMARE*) [24], in addition to the conventional correlation measures and *RMSE*. *sMARE* is a pointwise metric that is seeing increased use in recent QPP research. One advantage of *sMARE* over *RMSE* is that calculating *RMSE* involves fitting a regression model, since QPP scores are generally not intended to be interpreted as predicted values of standard metrics like Average Precision (AP), and are often not even in the $[0, 1]$ range; in contrast, *sMARE* is parameter free.
– **Test collections:** Hauff et al. used one news collection (TREC 678) and two web collections (WT10G and Gov2) for their experiments. In our study, we repeat their experiments on the TREC 678 collection (also reporting figures on the Robust collection with 100 more queries), and extend the evaluation to ClueWeb09B, a much larger Web collection, as well as the more modern MS MARCO collections, thus providing a more comprehensive evaluation.

Thus, our study has elements of both reproducibility (different team, same experimental setup) and replicability (different team, different experimental setup)[2]. Given that the original experiments that we seek to reproduce and extend are 15 years old, we have not attempted to compute quantitative reproducibility metrics for our efforts using recently proposed tools [7,37]. Instead, when summarizing the answers to our research questions later in Section 6, we categorize our observations as either new insights, confirmations of, or deviations from those presented in [28]. A GitHub repository[3] provides the artifacts (retrieval results/ ranked lists, code for QPP methods, fusion, and subsequent analysis) necessary for other researchers to independently corroborate our findings.

2 Related Work

2.1 An overview of QPP methods

In this section, we provide a brief review of two categories of QPP methods, namely, *pre-retrieval* and *post-retrieval* QPPs.

Pre-retrieval approaches. Pre-retrieval performance predictors can, in turn, be classified based on the query term features used to compute them. One category of predictors estimates query-term *specificity*: they assume that more discriminative terms (i.e., those with higher inverse document frequency (IDF) or inverse collection frequency (ICTF)) are better at identifying relevant content. Average IDF (`avgIDF`) and the maximum IDF values (`MaxIDF`) [15] fall under this category. Some studies compute the standard deviation of IDF from the average of ICTF values (`AvgICTF`) as specificity-based predictors [32]. Zhao et al. [59] proposed predictors using collection-based specificity to predict query difficulty, including methods like the sum of collection and query term similarity (`SumSCQ`), normalized query and collection similarity (`AvgSCQ`), and maximum collection and query term similarity (`MaxSCQ`).

[2] https://www.acm.org/publications/policies/artifact-review-and-badging-current
[3] https://github.com/souravsaha/qpp-comb

Another category is *rank sensitivity*, which estimates query difficulty based on the variation of term weights across documents. Common techniques here include `SumVAR` (the total of query term weight variations), `AvgVAR` (the normalized sum of these variations), and `MaxVAR` (the maximum variation among the terms) [59]. *Term relatedness* predictors leverage estimates of how related the query terms are to each other. Methods like `AvLesk` [6], `AvPath` [42], and `AvVP` [41] use external resources, typically WordNet [38] or co-occurrence statistics, to determine semantic distances between query terms. The presence of an ambiguous term in a query can lead to imprecise information needs [49], often resulting in difficulty for retrieval functions [9,39]. Predictors in this category measure *ambiguity* by computing the average inter-document similarity of all pairs of documents containing at least one query term [32] at a cost of quadratic time complexity which can be mitigated by sampling a subset of document pairs [28]. Variants like `AvQCG` include pairs containing all query terms. Averaged Polysemy (`AvP`) and Averaged Noun Polysemy (`AvNP`) utilize WordNet to assess query ambiguity by considering the number of senses associated with query terms [39].

Post-retrieval approaches. The post-retrieval predictors infer their predictions by taking the ranked list of documents in response to the query as input. These can be broadly classified into *coherence*-based, *score*-based, and *robustness*-based predictors. Coherence-based predictors measure how strongly documents retrieved are clustered together, such as `Clarity` [15]. This pioneering approach influenced a series of studies that built on the idea of query clarity [2,10,16,61]. Score-based predictors, e.g., `WIG` [61], `NQC` [52], and `SMV` [53], employ heuristics related to the retrieval scores of retrieved documents. Finally, robustness-based predictors, e.g., the `UEF` family [50], the Reference Lists framework [45,51], and `RSD` [47], compare the original ranking of documents with one produced by introducing noise in the query, the index, or documents. In parallel, several researchers explored robustness-based predictors to assess result quality, focusing on the stability of the top-ranked documents [4,8,54,55,60].

More recently, researchers show that supervised QPP approaches outperform their unsupervised counterparts [3,11,19,21,33,56]. These methods, in general, use deep learning techniques trained on large-scale datasets, resulting in improved accuracy. The first (weakly) supervised approach, `NeuralQPP` [56], uses pairs of queries to learn a binary indicator denoting which one of the pair leads to a better (or worse) retrieval effectiveness. A key drawback of this pairwise technique is that, depending on the size of the training set, the number of pairs grows quadratically, resulting in a substantial increase in training time. A solution to this increased training time problem was proposed in [3] that introduces `BERT-QPP`, an adoption of contextual embeddings to perform pointwise QPP. In contrast, the authors in [19] proposed `Deep-QPP`, a model employing a 3-dimensional convolutional neural network to train an end-to-end pairwise predictor. Inspired by the time efficient `BERT-QPP`, researchers have further proposed `qppBERT-PL` [21] which makes use of cross-encoding based query-document interactions in the form of BERT vectors, trained pointwise on individual queries while employing a listwise strategy by dividing the top documents into chunks.

2.2 Combining QPP estimators

While many QPP methods have been proposed, relatively less effort has been dedicated to effectively combining them. Empirical studies have consistently shown that combining pre-retrieval and post-retrieval predictors can improve the accuracy of query performance prediction. An early study proposed a model to enhance prediction by combining pre-retrieval features (e.g., the rounded logarithm of document frequency), with post-retrieval features, including overlap between ranked lists generated from sub-queries [55]. The first exploratory analysis of combined predictors was reported by Hauff et al. [28]. Their experiments with pre-retrieval predictors showed that fusing improves the performance of the predictors in terms of correlation coefficients (denoted ρ). Subsequent research investigated the integration of pre-retrieval and post-retrieval predictors within a probabilistic framework. The methods proposed in [34,35] achieved improved prediction accuracy by combining post-retrieval predictors. Kurland et al. also demonstrated improvements in predictive accuracy through the combination of post-retrieval predictors, such as Clarity Score [15] and Drift-based [16] predictors, highlighting the synergistic potential of these methods in QPP [34]. Other research has approached QPP enhancement through the fusion of scores from multiple rankings derived from different retrieval models. In [22], the authors combined regularized document scores across multiple rankings, further substantiating the benefit of predictor fusion. Additional studies support this trend, with researchers showing that combining post-retrieval predictors outperforms the use of individual methods, reinforcing the value of integrated predictor models in query performance prediction [43,61,26,51,48].

Our work builds upon these studies by extending the scope of predictor fusion, specifically focusing on post-retrieval methods, including modern neural approaches. Additionally, we explore the reproducibility of earlier findings using larger collections, such as ClueWeb09B [12] and TREC DL [40]. By addressing gaps in the literature, we aim to provide a better understanding of the effectiveness of combined QPP methods.

2.3 Evaluation measures

Evaluating the effectiveness of QPP methods requires robust metrics that capture various aspects of predictive accuracy and randomness. Over time, several metrics have become standard in QPP research, each offering unique insights into different facets of predictor performance. The primary metrics we apply for evaluation include Pearson's ρ and Kendall's τ, *RMSE*, together with the recently proposed *sMARE*. Correlation coefficients remain one of the most widely used metrics; they measure the association between the predicted and the actual performance values, mostly in terms of Average Precision [15,18,47,52,50,61] across a set of queries. As discussed in Section 1, Hauff et al. [28] highlighted the drawbacks of using ρ, and suggested reporting confidence intervals.

Table 1: Datasets used in our experiments. 'avg.$|Q|$' and 'avg.#rel' denote average query length and average number of relevant documents, respectively. 'Query ids' column for TREC DL is left empty as the topic identifiers do not follow a particular order.

Collection	Type	Document Set (#docs)	Topic Set	Query ids	#topics	avg.$\|Q\|$	avg.#rel
TREC 678	News	Disks 4,5 minus CR	TREC 6, 7, 8 Adhoc topics	301-450	150	2.45	89.13
TREC Robust		(528,155)	TREC 678 + Robust topics	301-450, 601-700	250	2.62	68.36
CW09B	Web	ClueWeb09B-S70 (29,038,220)	TREC Web Track 2009–2012 topics	1-200	200	2.42	16.02
TREC DL	Web	MS MARCO Passage (8,841,823)	TREC DL'19 + '20 topics	–	97 (43+54)	5.76	42.96

3 Experimental Setup

Datasets. For our investigation in this paper, we leverage three benchmark IR collections, TREC Robust (news articles with 250 topics), ClueWeb09B [14] (crawled web pages with 200 topics), and MS MARCO Passage [40] (a question answering dataset with over $100k$ Bing queries with 97 topics). Table 1 gives summary statistics for these collections. For the experiments on ClueWeb09B, spam documents are removed using the Waterloo spam filter [14] with spam confidence $> 70\%$. TREC 678 with 150 queries is used to replicate the results reported by Hauff et al. (see Table 2). Additionally, our preliminary analysis included two other web collections -WT10G [30] and Gov2 [13] (also used in [28]) - which yielded trends consistent with those found in the newer web collections in our study. Due to space limitations, we exclude those initial findings in this paper and instead, focus on the TREC Robust alongside more recent and widely used web collections, ClueWeb09B and MS MARCO passage, which have gained popularity among researchers in QPP studies.

Pre-retrieval QPP methods. We employ a range of pre-retrieval predictors targeting distinct aspects of query performance. Firstly, the specificity-based methods that include `AvgIDF`, `MaxIDF`, `SumSCQ`, `AvgSCQ`, and `MaxSCQ` which estimate query difficulty based on the informativeness of terms in the query. Complementing these methods, we employ the ambiguity-based techniques, such as, `AvP` and `AvNP`, designed to capture the potential variability or ambiguity inherent in query terms. Additionally, we incorporate rank sensitivity-based methods, like `SumVAR`, `AvgVAR` and `MaxVAR` [59] (see Section 2.1), which assess the sensitivity of document rankings to individual terms in the query. Note that, these choices are made based on the finding reported in [28] where the authors demonstrated the strong individual contribution in the fused predictors (refer to Figure 6 in [28]).

Post-retrieval QPP methods. We select a representative set of 10 post-retrieval predictors, grouped into three categories based on the level of supervision. Of these, six QPP methods, such as `NQC` [52], `Clarity`, `WIG` [61], and `UEF` [50] with the three base estimators (i.e., `NQC`, `WIG` and `Clarity`) are unsupervised. On the other hand, we employ weakly supervised neural QPP approach, namely `NeuralQPP`, that learns the relative importance of different estimators to obtain an optimal feature combination. Additionally, we choose three other fully supervised convolutional-based, `DEEP-QPP` that leverages information from the sematic interactions between the terms in the top documents and the query; and transformer-based state-of-the-art models, `BERT-QPP` which is a cross-encoder based pointwise QPP estimator; and `qppBERT-PL` which also makes use of cross-encoding based BERT vectors generated for the corresponding query and document terms (for more details see Section 2.1).

Train-test splits. In the QPP literature (e.g., [52,56,57,20]), the most common experiment setup usually involves random partitioning of the query set into two halves, *train*, *test* and repeating the process 30 times. The average outcome over 30 splits is reported. An identical setup is used for our experiments with TREC Robust and CW09B collections. However, because the MS MARCO collection has a designated train:test partition, we tuned the model hyper-parameters on a random sample (specifically 10% of train split) and reported results on the TREC DL, which is the subset of MS MARCO test set, as prescribed by [3,21,20]. All the parameters are tuned by dividing the training set into k-fold cross-validations. Additionally, to align our results with those of Hauff et al., we employ the same *leave-one-out* sampling strategy to predict the AP measure (see Table 2).

Penalized regression. Our objective is to combine m predictors $\{X_1, \ldots, X_m\}$, to predict the target variable, Y, which represents the average precision. A simple approach is to assume that the predictors are independent and apply multiple linear regression via Ordinary Least Squares (OLS). OLS aims to minimize the sum of squared residuals between the observed values and the predicted values, providing an estimate of the relationship between the predictors and the target variable. However, in practice, it is often the case that the predictors are not independent, and may be correlated with each other. When such correlations exist, the OLS method may suffer from issues such as multicollinearity, leading to unstable and overfit models. To address this issue, regularization techniques are commonly employed to improve the generalization ability of the models and enhance interpretability by enforcing sparsity. The two most common forms of regularization are the l_1 and l_2 penalties. The l_1 penalty, known as LASSO (Least Absolute Shrinkage and Selection Operator), encourages sparsity by shrinking some of the coefficients exactly to zero, thus effectively selecting a subset of the most important predictors. On the other hand, the l_2 penalty, associated with Ridge regression, encourages smaller coefficients overall but does not set any of them to zero, allowing all predictors to contribute to the model. An even more flexible approach is the ElasticNet (denoted as `E-Net`), which combines both the l_1 and l_2 penalties, offering a balance between the sparsity-inducing properties of LASSO and the stability advantages of Ridge regression. The ElasticNet is particularly useful when there are many correlated predictors, as it can handle both the selection of predictors and the regularization of those predictors in a way that neither LASSO nor Ridge can achieve individually.

Following [28], we also use `BOLASSO` [5], a computationally inexpensive and generalized version of `LASSO` that uses bootstrap samples, and `LARs` [23]. The regularization parameters of LARs may be estimated either by introducing random

Table 2: Comparison of our results with Hauff et al.'s, for pre-retrieval predictors on TREC 678. For reporting *RMSE* and *sMARE*, we use the leave-one-out based approach described in [28]. Note that, for the regression methods, ρ and *RMSE* values are not directly comparable to those in [28], as they used 19 predictors while we use 10 selected predictors. The '% imp.' columns show the relative improvement in ρ achieved by the regression methods over the best performing singleton predictor (`AvgIDF` for us, and `MaxIDF` in [28]). Values of ρ are highlighted or underlined if the value obtained by us is noticeably different from that in [28].

QPP	τ	ρ (*CI*)	% imp.	ρ (Hauff et al.)	% imp.	*sMARE*	*RMSE*	*RMSE* (Hauff et al.)
`MaxIDF`	0.4085	0.5780 [0.46,0.68]	-	0.53 [0.41,0.64]	-	0.1937	0.1766	0.186
`AvgIDF`	0.3897	0.6283 [0.52,0.72]	-	0.52 [0.39,0.62]	-	0.1997	0.1674	0.188
`SumSCQ`	0.0525	-0.0162 [-0.18,0.14]	-	0.00 [-0.16,0.16]	-	0.3123	0.2155	0.217
`MaxSCQ`	0.3825	0.4093 [0.27,0.53]	-	0.34 [0.19,0.47]	-	0.2096	0.1953	0.205
`AvgSCQ`	0.2930	0.3691 [0.22,0.50]	-	0.26 [0.10,0.40]	-	0.2252	0.1999	0.210
`SumVAR`	0.2855	0.3186 [0.17,0.46]	-	0.30 [0.14,0.44]	-	0.2383	0.2060	0.206
`AvgVAR`	0.4327	0.5933 [0.48,0.69]	-	0.51 [0.38,0.62]	-	0.1891	0.1745	0.185
`MaxVAR`	0.4505	0.5840 [0.47,0.68]	-	0.51 [0.38,0.62]	-	0.1840	0.1780	0.182
`AvP`	0.2122	0.3197 [0.17,0.46]	-	-0.12 [-0.28,0.04]	-	0.2732	0.2039	0.214
`AvNP`	0.1490	0.1965 [0.04,0.35]	-	-0.22 [-0.37,-0.06]	-	0.2909	0.2109	0.210
`OLS`	0.4619	0.6205 [0.51,0.71]	-1.24	0.69 [0.60, 0.77]	30.19	0.1905	0.1690	0.188
`LARS-Traps`	0.4297	0.6328 [0.53,0.72]	0.72	0.59 [0.47, 0.68]	11.32	0.1953	0.1702	0.179
`LARS-CV`	0.4485	0.6398 [0.53,0.73]	1.83	0.68 [0.59, 0.76]	28.30	0.1897	0.1642	0.183
`BOLASSO`	0.4504	0.6365 [0.53,0.72]	1.30	0.59 [0.47, 0.68]	11.32	0.1917	0.1641	0.181
`E-Net`	0.4444	0.6253 [0.52,0.71]	-0.48	0.69 [0.60, 0.77]	30.19	0.1966	0.1660	0.182

predictors into the model (`LARS-Traps`) or through cross-validation (`LARS-CV`). The same set of regression methods were also employed in [28].

Retrieval settings and evaluation. To ensure a fair comparison with Hauff et al., we use a language model with Dirichlet smoothing [58] (μ set to 1000) to retrieve the documents and measure AP with the top 1000 documents. Note that, this is also the most popular choice among researchers working on the QPP domain [15,18,47,52,50,61]. As the ranges of QPP scores vary across different predictors, we normalize them with min-max normalization before applying penalized regression. For evaluation, we employ the commonly used correlation measures - Pearson's ρ, Kendall's τ (ρ and τ respectively), *RMSE*, and *sMARE*.

4 Experimental Results

4.1 Reproducing results from [28]

As a validation check, we first repeat Hauff et al.'s experiments (with a selected subset of the pre-retrieval predictors) on the TREC 678 topics (see Table 1). Our results are presented in Table 2. We note that our ρ values are generally higher than those reported by Hauff et al. Following their suggestion, we include the 95% confidence interval (CI) for the observed ρ values. For two predictors (`AvP`, `AvNP`), the CIs (highlighted) are non-overlapping; since the corresponding ρs are fairly low on an absolute scale, we have not looked more carefully into this difference. For `AvgSCQ` the difference (underlined) appears numerically substantial, but is not, in fact, statistically significant. Of greater concern is the noticeable

Table 3: Results for pre-retrieval methods (top group), and their combinations (bottom group) across three datasets. We report p-values for a one-tailed t-test when comparing *RMSE*s of the combined predictors with the best-performing individual QPP method, i.e., `MaxIDF`. The best values for each metric–collection pair are in bold face.

	TREC Robust					CW09B					TREC DL				
Predictor	τ	ρ	*sMARE*	*RMSE*	*p-value*	τ	ρ	*sMARE*	*RMSE*	*p-value*	τ	ρ	*sMARE*	*RMSE*	*p-value*
`MaxIDF`	0.3269	0.4020	0.2264	0.1946	-	0.2017	0.2239	**0.2660**	0.1665	-	**0.3267**	**0.5053**	0.2285	**0.2270**	-
`AvgIDF`	0.2959	**0.4702**	0.2383	**0.1867**	-	0.1506	0.1569	0.2853	0.1684	-	0.3204	0.4884	**0.2236**	0.2276	-
`SumSCQ`	0.0977	0.0617	0.3010	0.2116	-	0.1635	0.2011	0.2839	0.1678	-	-0.0498	-0.1573	0.3535	0.2580	-
`MaxSCQ`	0.3464	0.3787	0.2194	0.1955	-	**0.2142**	**0.2914**	0.2670	**0.1625**	-	-0.0468	-0.1733	0.3552	0.2622	-
`AvgSCQ`	0.2550	0.3409	0.2539	0.1983	-	0.1472	0.1853	0.2896	0.1666	-	0.1198	0.0719	0.2921	0.2695	-
`SumVAR`	0.2852	0.2810	0.2450	0.2070	-	0.1869	0.2426	0.2768	0.1666	-	0.0825	0.0853	0.3178	0.2606	-
`AvgVAR`	**0.3726**	0.4647	**0.2146**	0.1887	-	0.1595	0.2002	0.2874	0.1666	-	0.3071	0.4093	0.2328	0.2387	-
`MaxVAR`	0.3694	0.4575	0.2148	0.1913	-	0.1923	0.2761	0.2792	0.1639	-	0.2728	0.4069	0.2572	0.2395	-
`AvP`	0.1426	0.2432	0.2885	0.2052	-	0.0787	0.0611	0.3173	0.1695	-	0.1194	0.2072	0.2972	0.2558	-
`AvNP`	0.0802	0.1203	0.3034	0.2096	-	0.0262	0.0170	0.3197	0.1700	-	0.0755	0.1160	0.3014	0.2601	-
`OLS`	**0.3389**	**0.4746**	**0.2250**	**0.1889**	0.3606	**0.1532**	0.1304	**0.2857**	0.1789	0.7048	0.2009	0.3158	0.2694	0.2607	0.8959
`LARS-Traps`	0.3002	0.3737	0.2353	0.1957	0.5297	0.1319	0.1433	0.2932	0.1666	0.5027	0.2422	0.3605	0.2546	0.2394	0.7069
`LARS-CV`	0.2997	0.3851	0.2385	0.1959	0.5049	0.1141	0.1403	0.2955	0.1665	0.5002	**0.2524**	**0.3876**	0.2529	**0.2387**	0.6779
`BOLASSO`	0.3285	0.4367	0.2291	0.1913	0.4196	0.1500	**0.1707**	0.2872	0.1682	0.5359	0.2154	0.3311	0.2628	0.2530	0.8402
`E-Net`	0.3276	0.4226	0.2289	0.1912	0.4096	0.1307	0.1517	0.2923	**0.1661**	0.4921	0.2492	0.3766	**0.2524**	0.2391	0.6832

difference in ρ for `AvgIDF` (highlighted): our measured ρ is at one end of the CI reported in [28], and vice versa. However, the RMSE values that we obtain are all very close to those reported in [28]. We tentatively conclude that the high-level summary of the results in Table 2 (that `MaxIDF`, `AvgIDF`, `AvgVAR` and `MaxVAR` provide the best predictions, both in terms of ρ and *RMSE*) matches the findings as in [28].

The lower part of Table 2 shows results for combined predictors. The values of ρ reported in [28, Table 5] for the combined approaches are, in some cases, substantially (though not statistically significantly) higher than those obtained via our implementation[4]. For Hauff et al., combined predictors obtained upto 30% higher ρ than the best performing singleton predictor; for us, these improvements are negligible ($< 2\%$). However, the improvements in terms of τ and *sMARE* are a little more noticeable. In terms of *RMSE*, the individual predictors perform roughly at par with the combinations; this is consistent with Hauff et al. Overall, Table 2 indicates that, across evaluation measures, combining predictors is not beneficial. This result is in contrast to Hauff et al.'s observation that "*Independent of the quality of the predictors, r [denoting Pearson's ρ] increases as more predictors are added to the model*" [28, Figure 2]. Notably, their paper does not provide any further insights into this surprising behaviour. In contrast, we offer an explanation for this major difference in Section 5.

4.2 Pre-retrieval predictors on additional collections

Table 3 presents the pre-retrieval QPP results across three datasets in terms of all the evaluation measures. The results for TREC DL are clearly in agreement with the summary for Table 2: once again, `MaxIDF`, `AvgIDF`, `MaxVAR`, and `AvgVAR`

[4] The combinations in [28] involved 19 pre-retrieval predictors, while we combine 10 of the best predictors from their set, but this does not seem to adequately explain these differences.

Table 4: Results for unsupervised (top group) and supervised (middle group) post-retrieval methods, as well as their combinations (bottom group), across three datasets. We report p-values for a one-tailed t-test when comparing *RMSE*s of the combined predictors with the best-performing individual predictor, `qppBERT-PL`. The best values are in bold face.

Predictor	TREC Robust					CW09B					TREC DL				
	τ	ρ	*sMARE*	*RMSE*	*p-value*	τ	ρ	*sMARE*	*RMSE*	*p-value*	τ	ρ	*sMARE*	*RMSE*	*p-value*
`NQC`	0.3960	0.3315	0.2025	0.2025	-	0.2423	0.1783	0.2554	0.1726	-	0.3119	0.2463	0.2330	0.2577	-
`WIG`	0.3163	0.3982	0.2367	0.1950	-	0.2624	0.3032	**0.2447**	0.1620	-	0.3295	0.4768	0.2315	0.2304	-
`Clarity`	0.2632	0.3899	0.2515	0.1948	-	0.2244	0.3134	0.2693	0.1610	-	0.2796	0.4393	0.2459	0.2358	-
`UEF-NQC`	0.3801	0.3469	0.2066	0.2015	-	0.2489	0.1853	0.2541	0.1711	-	0.3316	0.3092	0.2272	0.2505	-
`UEF-WIG`	0.2938	0.4205	0.2430	0.1924	-	**0.2697**	0.1069	0.2467	0.1789	-	0.3402	0.4887	0.2279	0.2299	-
`UEF-Clarity`	0.2313	0.2962	0.2614	0.2019	-	0.1611	0.2881	0.2887	0.1664	-	0.2822	0.4560	0.2470	0.2325	-
`NeuralQPP`	0.4204	0.3165	0.2031	0.2003	-	0.1995	0.3054	0.2691	0.1619	-	0.3857	0.5287	0.2098	0.2212	-
`qppBERT-PL`	**0.4740**	**0.6396**	**0.1759**	**0.1632**	-	0.2284	**0.3553**	0.2567	**0.1591**	-	0.3905	**0.5939**	0.2030	**0.2113**	-
`Deep-QPP`	0.4348	0.5598	0.1874	0.1747	-	0.2201	0.3332	0.2686	0.1604	-	0.3892	0.5492	0.2151	0.2176	-
`BERT-QPP`	0.4656	0.6093	0.1808	0.1680	-	0.2233	0.3424	0.2699	0.1598	-	**0.4111**	0.5459	**0.1992**	0.2205	-
`OLS`	0.5704	0.7246	0.1470	0.1469	0.1724	0.2325	0.3254	0.2610	0.1668	0.6494	0.4372	0.5946	0.1940	0.2146	0.5196
`LARS-Traps`	0.5639	0.7240	0.1507	0.1454	0.1094	0.2176	0.3527	0.2730	0.1567	0.4435	0.4169	0.5753	0.2040	0.2115	0.4981
`LARS-CV`	0.5595	0.7225	0.1512	0.1489	0.1524	0.1990	0.3073	0.2746	0.1720	0.5208	0.4471	0.6216	0.1917	0.2003	0.3032
`BOLASSO`	**0.5876**	**0.7577**	0.1490	**0.1371**	0.0328	**0.2486**	0.3581	**0.2576**	0.1615	0.5462	0.4440	0.6171	0.1930	0.2071	0.4156
`E-Net`	0.5842	0.7538	**0.1442**	0.1379	0.0347	0.2355	**0.3652**	0.2646	**0.1563**	0.4316	**0.4508**	**0.6266**	**0.1907**	**0.1997**	0.2949

appear to be the best-performing predictors in terms of all metrics. This observation mostly holds for TREC Robust as well, where `MaxSCQ` also emerges as a promising predictor, outperforming `MaxIDF` and `AvgIDF` on two of the four metrics (underlined in Table 3). CW09B presents the most confusing picture: there is little variation in the *RMSE* values across all methods; to a lesser extent, the same holds true for *sMARE*. Overall, `MaxSCQ` seems to deliver the best results, even though the `IDF` and `VAR` variants also perform reasonably well.

As before, we present the penalized regression results in the lower part of Table 3. Once again, combining pre-retrieval predictors does not improve results on any of the collections; in fact, combined predictors often perform worse. For example, for CW09B, Pearson's ρ decreases from 0.2914 (`MaxSCQ`, CI=[0.16,0.41]) to 0.1707 (`BOLASSO`, CI=[0.03,0.30]), and `sMARE`, `RMSE` scores increase marginally from 0.2660 (`MaxIDF`) and 0.1625 (`MaxSCQ`) to 0.2857 (`OLS`) and 0.1661 (`E-Net`), respectively. Thus, in response to **RQ1**, we find that combining predictors using penalized regression techniques *does not* improve predictor performance (as measured by correlations with AP). For more recent and larger collections, this could even lead to degraded performance. For these newer collections, both *RMSE* and *sMARE* metrics also show a decline when compared to individual predictors.

4.3 Results for post-retrieval predictors

Table 4 shows the performance of individual post-retrieval methods, as well as their combinations. We observe a clear separation in performance between supervised and unsupervised methods, with `qppBERT-PL` performing best overall. The bottom group in Table 4 presents results obtained by combining 10 post-retrieval predictors (these include both unsupervised and supervised methods; please see Section 3 for details). We observe that both `BOLASSO` and `E-Net` approaches perform comparably well across all evaluation metrics. Compared to

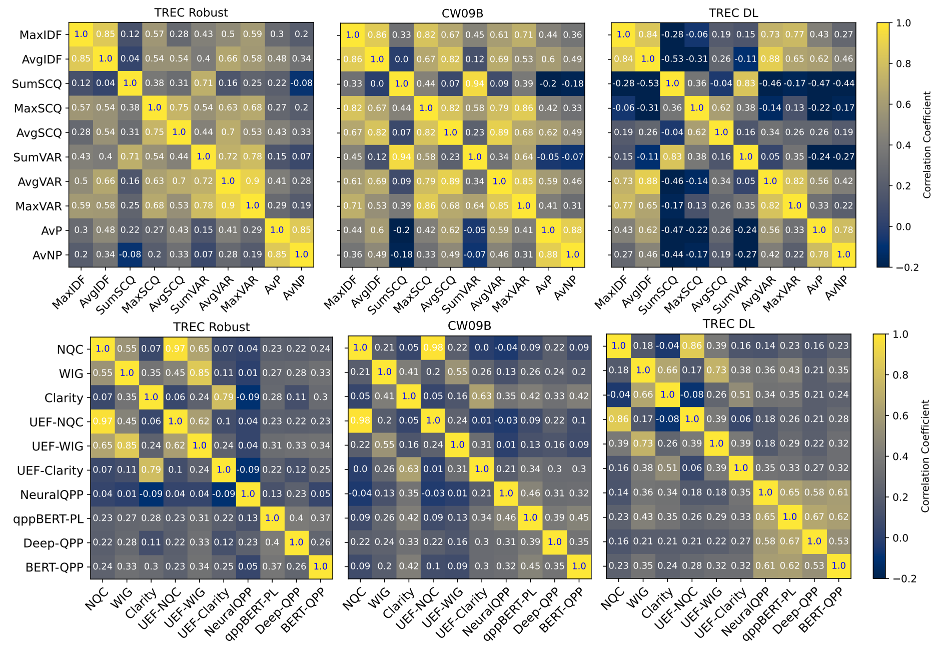

Fig. 1: Heatmap visualizing the rank correlation among QPP methods based on their individual QPP scores. Colour intensity represents the correlation values between the corresponding techniques measured with ρ; lighter intensity represents higher correlation. The upper row depicts the correlation for pre-retrieval methods and the bottom row shows the post-retrieval QPP correlations for all the collections.

the best standalone predictor (`qppBERT-PL`), these methods yield notable improvements (23% and 18%, resp.) in τ and ρ for the TREC Robust collection, but only the decline in *RMSE* (from 0.1632 (`qppBERT-PL`) to 0.1371 (`BOLASSO`)) is statistically significant (one-sided t-test, p-value $= 0.0328$). Further, these improvements fade as collection size increases, dropping to less than 5% for larger collections. None of the improvements observed on CW09B and TREC DL are significant. Indeed, in one case (rank correlation (τ) on CW09B), a performance degradation is observed. These findings address **RQ2**, and suggest that pre- and post-retrieval predictors behave differently. Specifically, combining post-retrieval predictors improves prediction accuracy, although this advantage gradually declines as collection size grows.

5 Discussion

We now turn to a deeper analysis of the behavior of combined predictors, by looking more closely at the relationships *between* different predictors. To the best of our knowledge, this issue has received limited attention, and seems to have been carefully explored only in [27], where the author reported Kendall's τ among a set of pre-retrieval predictors (see Tables 2.3 and 2.6 in [27]).

We compute pairwise ρ values between the lists of QPP scores calculated using different QPP methods[5]. Figure 1 displays these values as a heatmap for each collection for the two families of QPP approaches. Notably, we observe a high correlation ($\rho > 0.7$) among the `IDF` variants, `VAR` variants, collection-based specificity methods, and the polysemy-based approaches (`AvP` and `AvNP`) across all datasets when using pre-retrieval methods. Overall, the correlation among the pre-retrieval predictors are seen to be moderate to high with the exception of `SumSCQ` and `SumVAR`, which do not correlate well with the other predictors. However, they are mutually strongly correlated, with ρ reaching a peak of 0.94 in the case of CW09B.

In contrast, post-retrieval predictors generally show *lower* correlation between methods. Among these, only the `UEF` variants display moderate to high correlation with each other. Neural methods, particularly `NeuralQPP`, `qppBERT-PL`, `Deep-QPP`, and `BERT-QPP`, demonstrate varied correlation behavior; for collections with fewer training queries (e.g., TREC Robust and CW09B), these methods show negligible correlation, while their correlations increase with the availability of more training data, as observed with the TREC DL collection ($\rho > 0.5$).

Further, we observe nearly zero to negative intra-predictor correlations, particularly among the pre-retrieval predictors in the TREC DL collection. For example, in TREC DL collection, correlations between the pairs (`SumSCQ`, `AvgIDF`) or (`SumSCQ`, `AvgVAR`) are observed to be negative ($\rho < -0.3$). In addition, the post-retrieval predictors exhibit a significant reliance on the amount of training data, which leads to low correlations in collections with fewer training samples.

By integrating the findings from the heatmap in Figure 1 with the fused predictors presented in Tables 3 and 4, we propose the following three hypotheses.

Hypothesis 1 (H1): *If the correlation between predictors is moderate to high, combining them will not significantly improve performance.*

Based on this hypothesis, we can analyze the results in the tables, particularly for pre-retrieval in TREC Robust and post-retrieval in CW09B, to understand the lack of improvement in the combined predictors. The correlation between the predictors in these two cases (the first and fifth heatmaps in Figure 1) are moderate to strong; thus, combining predictors for TREC Robust (in Table 3) and CW09B (in Table 4) does not yield any improvements. This lack of significant improvement can be attributed to the fact that no new information is being added, as the predictors being combined are quite similar to each other (as indicated by the high correlation).

Hypothesis 2 (H2): *Low correlation among predictors may lead to an improvement in performance when they are combined.*

In the fourth and sixth heatmaps in Figure 1, we observe lower correlations among the pairs of post-retrieval predictors for the TREC Robust and TREC

[5] We also validated our results for the TREC678 collection against those reported in [27].

DL collections. Our second hypothesis **H2** suggests that in such cases, combining predictors could enhance performance. The performance improvements arise because the combined predictors are relatively unrelated, creating opportunities for complementary strengths to enhance overall accuracy. This is supported by the results in Table 4, where we note improvements in these two collections.

Hypothesis 3 (H3): *If the correlation between predictors is negative, combining them may degrade performance.*

When the pre-retrieval predictors are applied to the TREC DL collection, we observe significant negative correlations between pairs of predictors as shown in Table 3. Our third hypothesis, **H3**, posits that combining predictors that conflict with one another may degrade overall performance. This decline occurs because negatively correlated predictors provide contradictory signals, which may cancel each other out or introduce noise into the combined prediction. This is empirically verified in Table 3 for the TREC DL dataset, where all metrics against the combined predictors indicate a decline in performance.

Thus, in response to **RQ3**, we conclude that a discernible relationship, whether positive or otherwise, among the predictors is evident due to which they exhibit similar performance producing the observed correlations. This is supported by the heatmap and result tables, and is formalized by the proposed hypotheses.

6 Conclusion

In this study, we have revisited and extended earlier work on combining different QPP methods to enhance prediction accuracy. Our results demonstrate that, while combining predictors can improve correlation metrics in some cases, these benefits diminish with larger collections and specific predictor relationships. Notably, combining predictors that have low positive correlation among themselves can yield improved performance, but if the predictors are strongly positively correlated, or negatively correlated, combining them provides no benefits, and may lead to degraded results. This is a more fine-grained version of Hauff et al.'s observations.

Our experiments highlight that, while penalized regression techniques like `BOLASSO` and `E-Net` may constitute robust combination strategies, the impact of combining predictors varies significantly between dataset types and predictor categories. Thus, careful consideration of predictor relationships is essential to maximize QPP accuracy, with potential implications for future model selection and hybridization strategies in QPP research. As part of future work, we plan to explore the individual combination process more carefully considering the predictor-predictor relationship.

Acknowledgement. This publication is partially funded by (i) the European Research Council (ERC) under the Horizon 2020 research and innovation program (Grant No. 884951); (ii) Science Foundation Ireland (SFI) to the Insight Centre for Data Analytics under grant No 12/RC/2289_P2. Sourav Saha acknowledges the support from the TCS Research Scholar program.

References

1. Ahmad, W.U., Chang, K., Wang, H.: Context attentive document ranking and query suggestion. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 385–394. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019)
2. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: Advances in Information Retrieval. pp. 127–137. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
3. Arabzadeh, N., Khodabakhsh, M., Bagheri, E.: Bert-qpp: Contextualized pretrained transformers for query performance prediction. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. p. 2857–2861. CIKM '21, Association for Computing Machinery, New York, NY, USA (2021)
4. Aslam, J.A., Pavlu, V.: Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In: Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings. pp. 198–209 (2007)
5. Bach, F.R.: Bolasso: model consistent lasso estimation through the bootstrap. In: Proceedings of the 25th International Conference on Machine Learning. p. 33–40. ICML '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1390156.1390161, https://doi.org/10.1145/1390156.1390161
6. Banerjee, S., Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the 18th International Joint Conference on Artificial Intelligence. pp. 805–810. IJCAI'03, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2003)
7. Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P., Soboroff, I.: How to measure the reproducibility of system-oriented ir experiments. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 349–358 (2020)
8. Butman, O., Shtok, A., Kurland, O., Carmel, D.: Query-performance prediction using minimal relevance feedback. In: Proceedings of the 2013 Conference on the Theory of Information Retrieval. p. 14–21. ICTIR '13, Association for Computing Machinery, New York, NY, USA (2013)
9. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 911. SIGIR '10, Association for Computing Machinery, New York, NY, USA (2010)
10. Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 390–397. SIGIR '06, Association for Computing Machinery, New York, NY, USA (2006)
11. Chen, X., He, B., Sun, L.: Groupwise query performance prediction with bert. In: Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part II. p. 64–74. Springer-Verlag, Berlin, Heidelberg (2022)
12. Clarke, C., Craswell, N., Soboroff, I., Cormack, G.: Overview of the trec 2010 web track. In: Proceedings of TREC (2010)

13. Clarke, C.L., Craswell, N., Soboroff, I.: Overview of the trec 2004 terabyte track. In: Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 13-16, 2000 (2004)
14. Cormack, G.V., Smucker, M.D., Clarke, C.L.A.: Efficient and effective spam filtering and re-ranking for large web datasets. Inf. Retr. **14**(5), 441–465 (Oct 2011)
15. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 299–306. SIGIR '02, Association for Computing Machinery, New York, NY, USA (2002)
16. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A framework for selective query expansion. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. p. 236–237. CIKM '04, Association for Computing Machinery, New York, NY, USA (2004)
17. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Precision prediction based on ranked list coherence. Information Retrieval **9**(6), 723–755 (Dec 2006)
18. Cummins, R., Jose, J., O'Riordan, C.: Improved query performance prediction using standard deviation. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1089–1090. SIGIR '11 (2011)
19. Datta, S., Ganguly, D., Greene, D., Mitra, M.: Deep-qpp: A pairwise interaction-based deep learning model for supervised query performance prediction. In: WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Tempe, AZ, USA, February 21 - 25, 2022. pp. 201–209. ACM (2022)
20. Datta, S., Ganguly, D., Mitra, M., Greene, D.: A relative information gain-based query performance prediction framework with generated query variants. ACM Trans. Inf. Syst. **41**(2), 38:1–38:31 (2023)
21. Datta, S., MacAvaney, S., Ganguly, D., Greene, D.: A 'pointwise-query, listwise-document' based query performance prediction approach. In: SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022. pp. 2148–2153. ACM (2022)
22. Diaz, F.: Performance prediction using spatial autocorrelation. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 583–590. ACM, Amsterdam The Netherlands (jul 2007)
23. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. The Annals of Statistics **32**(2), 407–451 (2004)
24. Faggioli, G., Zendel, O., Culpepper, J.S., Ferro, N., Scholer, F.: An enhanced evaluation framework for query performance prediction. In: Advances in Information Retrieval. pp. 115–129. Springer International Publishing, Cham (2021)
25. Feild, H., Allan, J.: Task-aware query recommendation. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 83–92. SIGIR '13, Association for Computing Machinery, New York, NY, USA (2013)
26. Hauff, C.: Predicting the effectiveness of queries and retrieval systems. SIGIR Forum **44**(1), 88 (Aug 2010)
27. Hauff, C.: Predicting the Effectiveness of Queries and Retrieval Systems. Ph.D. thesis, University of Twente, Enschede, Netherlands (2010). `https://doi.org/10.3990/1.9789036529532`, `http://eprints.eemcs.utwente.nl/17338/`
28. Hauff, C., Azzopardi, L., Hiemstra, D.: The combination and evaluation of query performance prediction methods. In: Proc. ECIR. pp. 301–312. Springer (2009)

29. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. p. 1419–1420. CIKM '08, Association for Computing Machinery (2008)
30. Hawking, D.: Overview of the TREC-9 web track. In: Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000 (2000)
31. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: String Processing and Information Retrieval. pp. 43–54 (2004)
32. He, J., Larson, M., de Rijke, M.: Using coherence-based measures to predict query difficulty. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) Advances in Information Retrieval. pp. 689–694. Springer Berlin Heidelberg (2008)
33. Khodabakhsh, M., Zarrinkalam, F., Arabzadeh, N.: Bertpe: A bert-based pre-retrieval estimator for query performance prediction. In: Advances in Information Retrieval. pp. 354–363. Springer Nature Switzerland, Cham (2024)
34. Kurland, O., Shtok, A., Carmel, D., Hummel, S.: A unified framework for post-retrieval query-performance prediction. In: Advances in Information Retrieval Theory: Proceedings of the Third International Conference on Advances in Information Retrieval Theory (ICTIR'11). LNCS, vol. 6931, pp. 15–26. Springer (2011)
35. Kurland, O., Shtok, A., Hummel, S., Raiber, F., Carmel, D., Rom, O.: Back to the Roots: A Probabilistic Framework for Query-performance Prediction. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 823–832. CIKM '12, ACM, New York, NY, USA (2012)
36. Li, R., Kao, B., Bi, B., Cheng, R., Lo, E.: Dqr: A probabilistic approach to diversified query recommendation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. p. 16–25. CIKM '12, Association for Computing Machinery, New York, NY, USA (2012)
37. Maistro, M., Breuer, T., Schaer, P., Ferro, N.: An in-depth investigation on the behavior of measures to quantify reproducibility. Information Processing & Management **60**(3), 103332 (2023)
38. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. International journal of lexicography **3**(4), 235–244 (1990)
39. Mothe, J., Tanguy, L.: Linguistic features to predict query difficulty. In: ACM Conference on research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop. pp. 7–10 (2005)
40. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016)
41. Patwardhan, S., Pedersen, T.: Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense-Bringing Computational Linguistics and Psycholinguistics Together. vol. 1501, pp. 1–8 (2006)
42. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. In: IEEE Transactions on Systems, Man and Cybernetics. pp. 17–30 (1989)
43. Raiber, F., Kurland, O.: Query-performance prediction: setting the expectations straight. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 13–22. SIGIR '14, Association for Computing Machinery, New York, NY, USA (2014)

44. Rha, E.Y., Shi, W., Belkin, N.J.: An exploration of reasons for query reformulations. In: Diversity of Engagement: Connecting People and Information in the Physical and Virtual Worlds - Proceedings of the 80th ASIS&T Annual Meeting, ASIST 2017. vol. 54, pp. 337–346. Wiley (2017)
45. Roitman, H.: An enhanced approach to query performance prediction using reference lists. In: Proceedings of SIGIR '17. p. 869–872. Association for Computing Machinery (2017)
46. Roitman, H.: Normalized query commitment revisited. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1085–1088. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019)
47. Roitman, H., Erera, S., Weiner, B.: Robust standard deviation estimation for query performance prediction. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval. p. 245–248. ICTIR '17, Association for Computing Machinery, New York, NY, USA (2017)
48. Roy, D., Ganguly, D., Mitra, M., Jones, G.J.: Estimating gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. Information Processing & Management **56**(3), 1026–1045 (2019)
49. Sanderson, M.: Word sense disambiguation and information retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 142–151. Springer-Verlag New York, Inc. (1994)
50. Shtok, A., Kurland, O., Carmel, D.: Using statistical decision theory and relevance models for query-performance prediction. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 259–266. SIGIR '10 (2010)
51. Shtok, A., Kurland, O., Carmel, D.: Query performance prediction using reference lists. ACM Trans. Inf. Syst. **34**(4) (Jun 2016)
52. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting query performance by query-drift estimation. ACM Transactions on Information Systems **30**(2) (2012)
53. Tao, Y., Wu, S.: Query performance prediction by considering score magnitude and variance together. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. p. 1891–1894. CIKM '14, Association for Computing Machinery, New York, NY, USA (2014)
54. Vinay, V., Cox, I.J., Milic-Frayling, N., Wood, K.R.: On ranking the effectiveness of searches. In: SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006. pp. 398–404 (2006)
55. Yom-Tov, E., Fine, S., Carmel, D., Darlow, A.: Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 512–519. ACM (aug 2005)
56. Zamani, H., Croft, W.B., Culpepper, J.S.: Neural query performance prediction using weak supervision from multiple signals. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. p. 105–114. SIGIR '18, Association for Computing Machinery, New York, NY, USA (2018)
57. Zendel, O., Shtok, A., Raiber, F., Kurland, O., Culpepper, J.S.: Information needs, queries, and query performance prediction. In: Proceedings of the 42nd Interna-

tional ACM SIGIR Conference on Research and Development in Information Retrieval. p. 395–404. SIGIR'19, New York, NY, USA (2019)
58. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 334–342. SIGIR '01, Association for Computing Machinery, New York, NY, USA (2001)
59. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) Advances in Information Retrieval. pp. 52–64. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
60. Zhou, Y., Croft, W.B.: Ranking robustness: a novel framework to predict query performance. In: Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006. pp. 567–574 (2006)
61. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 543–550. SIGIR '07, Association for Computing Machinery, New York, NY, USA (2007)