An Analysis of Variations in the Effectiveness of Query Performance Prediction

No Author Given

No Institute Given

Abstract. A query performance predictor estimates the retrieval effectiveness of a system for a given query. Query performance prediction (QPP) algorithms are themselves evaluated by measuring the correlation between the predicted effectiveness and the actual effectiveness of a system for a set of queries. This generally accepted framework for judging the usefulness of a QPP method includes a number of sources of variability. For example, "actual effectiveness" can be measured using different metrics, for different rank cut-offs. The objective of this study is to identify some of these sources, and investigate how variations in the framework can affect the outcomes of QPP experiments. We consider this issue not only in terms of the absolute values of the evaluation metrics being reported (e.g., Pearson's r, Kendall's τ), but also with respect to the changes in the ranks of different QPP systems when ordered by the QPP metric scores. Our experiments reveal that the observed QPP outcomes can vary considerably, both in terms of the absolute evaluation metric values and also in terms of the relative system ranks. Through our analysis, we report the combinations of QPP evaluation metric and experimental settings that are likely to lead to smaller variations in the observed results.

Keywords: Query Performance Prediction, Variations in QPP Results, QPP Reproducibility

1 Introduction

The problem of *query performance prediction* (QPP) [5,7,8,9,12,18,16,19,26,27] has attracted the attention of the Information Retrieval (IR) community over a number of years. QPP involves estimating the retrieval quality of an IR system. A diverse range of pre-retrieval (e.g. avgIDF [9]) and post-retrieval approaches (e.g. WIG [27], NQC [19], UEF [18]) have been proposed for the task of QPP over the years.

The primary use-case of QPP can be described as follows: "If we could determine in advance which retrieval approach would work well for a given query, then hopefully, selecting the appropriate retrieval method on a [per] query basis could improve the retrieval effectiveness significantly." [6]. In other words, the objective of QPP would be to predict how easy or difficult a given query is for an IR system. This prediction could either be a categorical label (e.g., EASY,

MODERATE, HARD), or a numerical estimate of a standard IR evaluation metric (which generally lie in $[0, 1]$).

QPP is a challenging problem, however, and this eventual objective has remained elusive thus far. Given a query and an IR system, well-known QPP methods simply compute a real-valued score that is meant to be indicative of the effectiveness of the system for the given query. While this score is typically not interpreted as a statistical estimate of a specific evaluation metric (e.g. AP or nDCG [11]), it is expected to be highly correlated with a standard evaluation measure. Indeed, the quality of a QPP method is usually determined by measuring the correlation between its predicted effectiveness scores and the values of some standard evaluation metric for a set of queries.

Consider a proposed QPP algorithm $\mathcal{P}$. Given an IR system S, and a set of queries $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_n\}$, S is used to retrieve a ranked list L_i of documents for each $Q_i \in \mathcal{Q}$. For each L_i, $\mathcal{P}$ computes a predicted effectiveness score ϕ_i. Using available relevance assessments as ground-truth, a standard IR metric g_i is also computed for L_i. The correlation between the lists $\{\phi_1, \phi_2, \ldots, \phi_n\}$ and $\{g_1, g_2, \ldots, g_n\}$ is taken to be a measure of how effective $\mathcal{P}$ is as a query performance predictor.

In this study, we analyse the above approach for evaluating and comparing different QPP methods. We identify the sources of variability within this generally accepted framework, and show that these variations can lead to differences in the computed correlations. This, in turn, can lead to differences in

- the absolute values of reported QPP evaluation measures (e.g., the ρ value for NQC [19] measured with AP@100 as the target metric and LM-Dirichlet as the retrieval model can be substantially different from that measured with AP@1000 as the target metric and BM25 as the retrieval model on the *same* set of queries); and also in
- the comparative effectiveness of a number of different QPP measures (e.g., NQC turns out to be better than WIG with AP@100, whereas WIG outperforms NQC when QPP effectiveness is measured using nDCG@10).

Thus, these variations can lead to difficulties in reproducing QPP results, both at the level of the correlation values being reported, and also in terms of the relative performance of different competing methods on standard datasets.

Contributions. We conduct a range of experiments to analyze the potential variations in QPP effectiveness results under different experimental conditions. Specifically, we consider different combinations of IR metrics and IR models (as well as rank cut-off values). The experiments described in Section 5 reveal that the results of QPP depend significantly on these settings. Thus, it may be difficult to reproduce QPP experiments without a precise description of the experimental context. While variations in other factors, such as the choice of indexing implementation and set of pre-processing steps, may also matter, we recommend that any empirical study of QPP include a precise description of at least the above experimental settings in order to reduce variations in reported results. More importantly, our findings suggest that it may even be worthwhile to systematically revisit reported comparisons between competing QPP approaches.

2 Related Work

Analyzing the sensitivity of reported results on the experiment settings is important for an empirical discipline such as IR. Buckley and Voorhees while examining the stability of commonly used evaluation measures in IR [3], reported observations, such as P@30 has about twice the average error rate as compared to average precision (AP), or that a stable measurement of P@10 requires an aggregation of over 50 queries etc.

Similar to our investigation of examining the relative QPP system ranks, previous studies have investigated the sensitivity of relative ranks of IR systems to the pooling depth used for relevance assessments. It is reported that smaller samples of the relevance ground-truth obtained with smaller pool depths usually do not lead to significant changes in the relative performance of IR systems [2,22,21,4]. In another work related to analysis of pooling, Buckley et. al. demonstrated that pools created during the TREC 2005 workshop exhibit a specific bias in favor of relevant documents that specifically contain the title words.

The study in [17] analyzed the sensitivity of variations in word embedded vectors for IR models. The work in [1] stressed the importance of reproducibility in IR research by noting that most of the improvements reported over the years were not statistically significant over their predecessors. Recently, this observation has also been reinforced for neural models by arguing that most of the neural approaches have compared their results against relatively weak baselines [14,20].

3 Anatomy of a QPP Evaluation Framework

In this section, we formally define the various components in a standard QPP evaluation framework. As we demonstrate later, variations in these components can potentially lead to different experimental outcomes.

Definition 1. *The context, $\mathcal{C}(Q)$, of a QPP experiment on a query Q, is a 3-tuple of the form of $(\theta, \mathcal{S}, \kappa)$, where κ is a positive integer; the function $\mathcal{S} : Q \times D \mapsto \mathbb{R}$ is a scoring function that computes query-document similarities, and is used to retrieve $L = (D_1, \ldots, D_\kappa)$, the list of κ top-ranked documents for Q from a collection; and $\theta : L \mapsto [0, 1]$ is an evaluation metric function that, given a query Q, a list L of top-ranked documents, and $R(Q)$, the relevance assessments for Q, outputs a measure of usefulness of L.*

Definition 2. *The ground-truth or reference value of retrieval effectiveness of a query Q in relation to a QPP context, $\mathcal{C}(Q)$ of Definition 1, is a function of the form $g : \mathcal{C}(Q) \mapsto [0, 1]$.*

Definition 3. *A QPP method is a function of the form $\phi(Q, D_1, \ldots, D_k) \mapsto [0, 1]$, which, given a query Q and a list of top-k retrieved documents*[1]*, outputs*

[1] For pre-retrieval QPP approaches, $(D_1, \ldots, D_k) = \emptyset$.

a number that is indicative of how relevant the retrieved list is. In other words, the output of the predictor $\phi(Q)$ is some measure of the ground-truth retrieval effectiveness measure $g(\mathcal{C}(Q))$ from Definition 2.

For example, NQC [19] or WIG [27] compute $\phi(Q)$ by examining a set of k top-ranked documents[2] and estimating how distinct it is from the rest of the collection. The intuition behind NQC and WIG is that the higher the distinctiveness, the higher the likelihood of finding more relevant documents in the retrieved list.

The next step in QPP evaluation is to measure the correlation between the predicted retrieval effectiveness, $\phi(Q)$, and the ground-truth retrieval effectiveness, $g(\mathcal{C}(Q))$ over a set of benchmark queries $\mathcal{Q}$, using a correlation function, $\chi : (\Phi, \mathcal{G}(\mathcal{C})) \mapsto [0, 1]$, where $\Phi = \bigcup_{Q \in \mathcal{Q}} \phi(Q)$ and $\mathcal{G}(\mathcal{C}) = \bigcup_{Q \in \mathcal{Q}} g(\mathcal{C}(Q)))$. Common choices for χ are Pearson's r, which computes a correlation between the values themselves, and rank correlation measures, such as Spearman's ρ or Kendall's τ, which compute the correlation between the ordinals of the members of Φ and $\mathcal{G}(\mathcal{C})$.

It is clear from Definitions 1-3 that the QPP outcome, $\chi(\Phi, \mathcal{G})(\mathcal{C})$, depends on the context $\mathcal{C}(Q)$ used for each $Q \in \mathcal{Q}$. Our first objective is to quantify the relative changes in QPP outcomes χ with changes in the context $\mathcal{C}(Q)$. In other words, we wish to compute the relative changes of the form $|\chi(\Phi, \mathcal{G}(\mathcal{C}_i)) - \chi(\Phi, \mathcal{G}(\mathcal{C}_j))|$, for two different instances of QPP contexts $\mathcal{C}_i = (\theta_i, \mathcal{S}_i, \kappa_i)$ and $\mathcal{C}_j = (\theta_j, \mathcal{S}_j, \kappa_j)$. Thus, our first research question is the following:

> **RQ1:** Do **variations in the QPP context**, $\mathcal{C}$, in terms of the IR metric (θ), the IR model ($\mathcal{S}$) and the rank cut-off (κ) used to construct the QPP evaluation ground-truth, $g(\mathcal{C})$, lead to **significant differences in outcome of a QPP method** ϕ?

For our next research question, instead of computing the relative change in the outcome values (correlations) of individual QPP methods, we seek to measure the relative change in the rankings (in terms of effectiveness) of a number of different QPP methods. Formally, given a set of m QPP functions $\{\phi_1, \ldots, \phi_m\}$, we compute the effectiveness of each with respect to a number of different QPP contexts, $\chi(\Phi_i, \mathcal{G}(\mathcal{C}_j))$ for $j = 1, \ldots, n$. The objective is to investigate whether or not the ranking of QPP systems computed with different contexts is relatively stable. For instance, if NQC is the best method for a context that used LM-Dirichlet as the retrieval model and AP@100 as the evaluation metric, we might wish to investigate whether it remains the best method for a different QPP context that uses BM25 as the retrieval model and nDCG@10 as the evaluation metric. Stated explicitly,

> **RQ2:** Do **variations in the QPP context**, $\mathcal{C}$, in terms of the IR metric (θ), the IR model ($\mathcal{S}$) and the rank cut-off (κ) used to construct the QPP evaluation ground-truth, $g(\mathcal{C})$, lead to **significant differences in the relative ranks of different QPP methods** $\phi_1, \ldots, \phi_m$?

[2] k is a parameter of a post-retrieval QPP method, and can be different from κ, the number of top documents used for QPP evaluation.

Collection	#Docs	Topic Set	#Queries	Avg.$\|Q\|$	Avg.#Rel
Disks 4,5 (w/o CR)	528,155	TREC-Robust	249	2.68	71.21

Table 1: Characteristics of the TREC-Robust dataset used in our QPP experiments. 'Avg.$|Q|$' and 'Avg.#Rel' denote the average number of terms in a query, and the average number of relevant documents for a query, respectively.

4 Experimental Setup

To investigate the research questions from the last section, we conduct QPP experiments on a widely-used dataset, the TREC Robust dataset, which consists of 249 queries. To address RQ1 and RQ2, we first define the set of possible QPP contexts that we explore in our experiments.

IR evaluation metrics investigated. As choices for the IR evaluation metric (i.e., the function θ), we consider 'AP', 'nDCG', 'P@10', and 'recall'. The evaluation functions explored represent a mixture of both precision- and recall-oriented metrics. While AP and nDCG address both the aspects of precision and recall (leaning towards favouring precision), P@10 is a solely precision-oriented metric. To investigate RQ1, we set the cut-off for AP, nDCG, and recall to 100, as is common in the literature on QPP [19,23,24].

IR models investigated. IR models represent the second component of a QPP context as per Definition 1. We explore three such models: a) language modeling with Jelinek-Mercer smoothing (LMJM) [25,10], b) language modeling with Dirichlet smoothing (LMDir) [25], and c) Okapi BM25 [15]. The values of the IR model parameters were chosen after a grid search to optimize the MAP values on the TREC-Robust queries. Unless otherwise specified, for LMJM, we used $\lambda = 0.6$, the value of k_1 and b in BM25 were set to 0.7 and 0.3, respectively, and the value of the smoothing parameter μ for LMDir was set to 1000.

QPP methods tested. To compare the relative perturbations in preferential ordering of the QPP systems in terms of the evaluated effectiveness, we employ a total of seven different QPP methods, as outlined below:

- **AvgIDF** [9] is a pre-retrieval QPP method that uses the average idfs of the constituent query terms as the predicted query performance estimate.
- **Clarity** [7] estimates a relevance model (RLM) [13] distribution of term weights from a set of top-ranked documents, and then computes its KL divergence with the collection model.
- **WIG** [27] uses the aggregated value of the information gain of each document (with respect to the collection) in the top-retrieved set as a specificity estimate.
- **NQC** [19] or normalized query commitment estimates the specificity of a query as the standard deviation of the RSVs of the top-retrieved documents.
- **UEF** [18] assumes that information from some top-retrieved sets of documents are more reliable than others. A high perturbation of a ranked list after feedback indicates a poor retrieval effectiveness of the initial list. This,

in turn, suggests that a smaller confidence should be associated with the QPP estimate of such a query. Formally,

$$\text{UEF}(Q, \phi) = \xi(R_M(Q), R_M(\theta_Q))\phi(Q) \quad (1)$$

where $\phi(Q)$ is the predicted score of a base QPP estimator (e.g. WIG or NQC), $R_M(\theta_Q)$ denotes the re-ranked set of documents post-RLM feedback, the RLM being estimated on $R_M(Q)$ - the top-M documents, and ξ is a rank correlation coefficient of two ordered sets, for which we specifically use Pearson's-ρ, as suggested in [18]. We experiment with three specific instances of the base estimators, namely Clarity, WIG and NQC for UEF, which we denote as UEF(Clarity), UEF(WIG) and UEF(NQC), respectively.

Parameter and settings. The standard practice in QPP research is to optimize the common hyper-parameter - the number of top documents of post-retrieval QPP approaches (denoted as k in Definition 3). This hyper-parameter is tuned via a grid search on a development set of queries and the optimal setting is used to report the performance on a test set. A common approach is to employ a 50:50 split of the set of queries into development and test sets. This process is usually repeated 30 times and the average results over the test folds are reported [19,24,27].

The focus of our research is different, however, in the sense that we seek to analyze the variations caused due to different settings for constructing the QPP ground-truth, instead of demonstrating that a particular QPP method outperforms others. Moreover, an optimal tuning of the hyper-parameters for each QPP method would require averaging over 30 different experiments for a single way of defining the QPP context for constructing the ground-truth. Hence, to keep the number of experiments tractable, we set $k = 20$, as frequently prescribed in the literature [7,19,24,27]. Another hyper-parameter, specific to UEF, is the number of times a subset of size k is sampled from a set of top-K ($K > k$) documents. We use a total of 10 random samples of $k = 20$ documents from the set of $K = 100$ top documents, as prescribed in [18].

5 Results

5.1 RQ1: Variations in QPP Evaluations

Table 2 reports the standard deviations in the observed values for the QPP experiments. In Tables 2a-d, the value of $\sigma(\theta)$ in each row indicates the standard deviation of the QPP outcome values observed in that row, i.e., these values indicate the standard deviation resulting from the use of different IR metrics for QPP evaluation. Similarly, the value of $\sigma(\mathcal{S})$ in each column is the standard deviation of the r, ρ or τ values reported in that column, i.e., this value denotes the standard deviations in QPP correlations across different IR models. The lowest standard deviations for each QPP correlation type are shown bold-faced. We now discuss the observations that can be made from Table 2.

	IR Evaluation Metric (θ)				
Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r LMJM	0.3795	0.3966	0.3869	0.3311	**0.0291**
r BM25	0.5006	0.4879	0.4813	0.2525	**0.1190**
r LMDir	0.5208	0.5062	0.4989	0.2851	0.1121
$\sigma(\mathcal{S})$	**0.0764**	0.0587	0.0602	0.0395	
ρ LMJM	0.4553	0.4697	0.4663	0.3067	0.0788
ρ BM25	0.4526	0.4700	0.4736	0.2842	0.0911
ρ LMDir	0.4695	0.4848	0.4893	0.3017	0.0902
$\sigma(\mathcal{S})$	0.0091	0.0086	0.0118	0.0114	
τ LMJM	0.3175	0.3285	0.3278	0.2193	0.0529
τ BM25	0.3144	0.3162	0.3319	0.2040	0.0589
τ LMDir	0.3307	0.3407	0.3440	0.2155	0.0617
$\sigma(\mathcal{S})$	0.0087	0.0123	**0.0084**	0.0120	

(a) AvgIDF

	IR Evaluation Metric (θ)				
Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r LMJM	0.3652	0.4169	0.4503	0.2548	0.0855
r BM25	0.3563	0.4118	0.4495	0.2707	0.0777
r LMDir	0.4354	0.4583	0.4854	0.2842	0.0901
$\sigma(\mathcal{S})$	**0.0433**	0.0255	0.0205	0.0147	
ρ LMJM	0.4545	0.4843	0.5248	0.2918	**0.1022**
ρ BM25	0.4618	0.4887	0.5137	0.3308	0.0814
ρ LMDir	0.5024	0.5260	0.5453	0.3340	0.0969
$\sigma(\mathcal{S})$	0.0258	0.0229	0.0160	0.0235	
τ LMJM	0.3100	0.3319	0.3657	0.2061	0.0688
τ BM25	0.3170	0.3370	0.3551	0.2374	**0.0519**
τ LMDir	0.3539	0.3713	0.3828	0.2379	0.0668
$\sigma(\mathcal{S})$	0.0236	0.0214	**0.0140**	0.0182	

(b) NQC

	IR Evaluation Metric (θ)				
Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r LMJM	0.4056	0.4071	0.3971	0.3054	0.0491
r BM25	0.4488	0.4563	0.4386	0.3485	0.0502
r LMDir	0.4908	0.4798	0.4632	0.3423	**0.0688**
$\sigma(\mathcal{S})$	0.0426	0.0371	0.0334	0.0233	
ρ LMJM	0.3716	0.3794	0.3790	0.3120	0.0325
ρ BM25	0.4520	0.4601	0.4505	0.3586	0.0480
ρ LMDir	0.4582	0.4688	0.4667	0.3528	0.0561
$\sigma(\mathcal{S})$	0.0483	**0.0493**	0.0467	0.0254	
τ LMJM	0.2514	0.2567	0.2607	0.2209	**0.0181**
τ BM25	0.3116	0.3181	0.3125	0.2549	0.0297
τ LMDir	0.3194	0.3267	0.3259	0.2493	0.0375
$\sigma(\mathcal{S})$	0.0372	0.0382	0.0344	**0.0182**	

(c) WIG

	IR Evaluation Metric (θ)				
Model($\mathcal{S}$)	AP	nDCG	R	P@10	$\sigma(\theta)$
r LMJM	0.4746	0.4763	0.4646	0.3573	0.0575
r BM25	0.5386	0.5476	0.5263	0.4182	0.0603
r LMDir	0.5693	0.5566	0.5373	0.3971	**0.0797**
$\sigma(\mathcal{S})$	0.0483	0.0440	0.0392	0.0309	
ρ LMJM	0.4385	0.4477	0.4472	0.3682	0.0384
ρ BM25	0.5334	0.5429	0.5316	0.4231	0.0567
ρ LMDir	0.5407	0.5532	0.5507	0.4163	0.0662
$\sigma(\mathcal{S})$	0.0570	**0.0582**	0.0551	0.0300	
τ LMJM	0.3017	0.3080	0.3128	0.2651	**0.0217**
τ BM25	0.3677	0.3754	0.3688	0.3008	0.0351
τ LMDir	0.3833	0.3920	0.3911	0.2992	0.0450
$\sigma(\mathcal{S})$	0.0433	0.0445	0.0303	**0.0202**	

(d) UEF(WIG)

Table 2: Sensitivity of QPP results on variations in the IR evaluation metric (θ) and the IR model ($\mathcal{S}$) for the QPP methods a) AvgIDF, b) NQC, c) WIG and d) UEF(WIG). The metrics - AP, nDCG and recall (R) are measured on the top-100 retrieved documents using retrieval models LMJM($\lambda = 0.6$), BM25($k_1 = 0.7, b = 0.3$) and LMDir($\mu = 1000$) respectively. The lowest (highest) standard deviations for each group of QPP correlation measure are shown in green (red). The lowest and the highest across different correlation measures are shown bold-faced.

Variations due to IR evaluation metric. The first set of observations, listed below, is in relation to the absolute differences between two different QPP evaluations involving two different QPP contexts.

Model	Metric	AP@100	AP@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000
LMJM		0.4286	0.3333	0.9048	0.2381	**-0.1429**	1.0000	0.2381	0.3333
BM25	AP@10	1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir		1.0000	0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM			0.9048	0.5238	0.8095	0.4286	0.4286	0.8095	0.9048
BM25	AP@100		0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMDir			0.9048	1.0000	0.9048	0.4286	1.0000	1.0000	0.7143
LMJM				0.4286	0.8095	0.5238	0.3333	0.9048	1.0000
BM25	AP@1000			0.9048	0.8095	0.3333	0.9048	0.9048	0.8095
LMDir				0.9048	0.8095	0.5238	0.9048	0.9048	0.8095
LMJM					0.3333	-0.0476	0.9048	0.3333	0.4286
BM25	R@10				0.9048	0.4286	1.0000	1.0000	0.7143
LMDir					0.9048	0.4286	1.0000	1.0000	0.7143
LMJM						0.6190	0.2381	1.0000	0.9048
BM25	R@100					0.5238	0.9048	0.9048	0.6190
LMDir						0.5238	0.9048	0.9048	0.6190
LMJM							**-0.1429**	0.6190	0.5238
BM25	R@1000						0.4286	0.4286	0.5238
LMDir							0.4286	0.4286	0.5238
LMJM								0.2381	0.3333
BM25	nDCG@10							1.0000	0.7143
LMDir								1.0000	0.7143
LMJM									0.9048
BM25	nDCG@100								0.7143
LMDir									0.7143

Table 3: Each cell in the table indicates the correlation (Kendall's τ) between QPP systems ranked in order by their evaluated effectiveness (measured with the help of Pearson's r for the results of this table) for two different IR metrics corresponding to the row and the column name of the cell. A total of 7 QPP systems were used in these experiments, namely AvgIDF, Clarity, WIG, NQC, UEF(Clarity), UEF(WIG) and UEF(NQC). The lowest correlation value for each group is marked in red, and the lowest correlations, overall, are bold-faced.

- **Substantial absolute differences in the QPP outcomes**: Variations in the IR evaluation metric (i.e., the θ component of a QPP context $\mathcal{C}(Q)$ of Definition 1) keeping fixed the other two components (i.e., retrieval model and cut-off), produces considerable absolute differences in the values. As an example, compare the QPP evaluation of 0.5006 with AP@100 in Table 2a to that of 0.2525 with P@10 obtained with BM25, showing that these absolute differences can be considerably high.
- **Lower variations with τ**: In general, we observe that each QPP method (e.g. NQC, WIG etc.) exhibits considerable differences in measured outcomes specially between AP@100 and P@10. Moreover, the variations, in general, are lower when correlation is measured with the help of Kendall's τ (e.g., compare $\sigma(\theta) = 0.0181$ measured with τ vs. $\sigma(\theta) = 0.0491$ measured with r on documents retrieved with LMJM). The reason why τ exhibits a lower variance in QPP evaluation is likely due to the fact that the correlation is measured in a pairwise manner (τ being a function of the number of

Model	Metric	AP@100	AP@1000	R@10	R@100	R@1000	nDCG@10	nDCG@100	nDCG@1000
LMJM		0.5238	0.3333	0.8095	0.4286	0.2381	0.8095	0.4286	0.3333
BM25	AP@10	0.9048	0.7143	0.8095	0.8095	0.5238	1.0000	0.9048	0.5238
LMDir		0.9048	0.8095	1.0000	1.0000	0.8095	1.0000	0.9048	0.7143
LMJM			0.8095	0.5238	0.9048	0.7143	0.3333	0.9048	0.8095
BM25	AP@100		0.8095	0.9048	0.9048	0.6190	0.9048	1.0000	0.6190
LMDir			0.9048	0.9048	0.9048	0.7143	0.9048	1.0000	0.8095
LMJM				0.3333	0.9048	0.7143	**0.1429**	0.9048	1.0000
BM25	AP@1000			0.7143	0.7143	0.6190	0.7143	0.8095	0.8095
LMDir				0.8095	0.8095	0.8095	0.8095	0.9048	0.9048
LMJM					0.4286	0.2381	0.8095	0.4286	0.3333
BM25	R@10				1.0000	0.7143	0.8095	1.0000	0.5238
LMDir					1.0000	0.8095	1.0000	0.9048	0.7143
LMJM						0.8095	0.2381	1.0000	0.9048
BM25	R@100					0.7143	0.8095	0.9048	0.5238
LMDir						0.8095	1.0000	0.9048	0.7143
LMJM							0.0476	0.8095	0.7143
BM25	R@1000						0.5238	0.6190	0.8095
LMDir							0.8095	0.7143	0.9048
LMJM								0.2381	**0.1429**
BM25	nDCG@10							0.9048	0.5238
LMDir								0.9048	0.7143
LMJM									0.9048
BM25	nDCG@100								0.6190
LMDir									0.8095

Table 4: Results of relative changes in the ranks of QPP systems (similar to Table 3), the difference being that the QPP outcomes were measured with τ (instead of r).

concordant and discordant pairs). As a result of this, τ depends only on the agreements between the true and the predicted order (of query difficulty) between a query pair, and not on the absolute values of the predicted scores or the reference values of the IR evaluation metric (as in Pearson's r or Spearman's ρ).

– **Lower variances with LMJM**: Similar to the earlier observation that τ should be the preferred QPP evaluation measure (with an objective to minimize the variances in observed results due to changes in IR evaluation metric), we observe from Table 2 that LMJM, in most cases, result in low variances in QPP experiment outcomes.

Variations due to IR models. The second set of observations from Table 2 is in relation to variations in the observed QPP results with respect to variations in IR models. The standard deviations of these values correspond to column-wise calculation of standard deviations and are shown as the $\sigma(\mathcal{S})$ values. Again, similar to the $\sigma(\theta)$ values, the lowest (highest) values along each row of $\sigma(\mathcal{S})$ are colored in green (red) to reflect the situation of lower the better. The best values across different QPP correlations are bold-faced. The following are our observations.

– **Lower variations with τ**: Similar to the $\sigma(\theta)$ values it is again observed that mostly measuring QPP outcomes with τ results in the lowest variances

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (1.0, 1.0)	BM25 (0.3, 0.7)	LMDir (100)	LMDir (500)	LMDir (1000)
AP@100		1.0000	0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100	LMJM	1.0000	0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100	(0.3)	0.9048	0.8095	0.9048	1.0000	1.0000	0.9048	0.9048
P@10		1.0000	0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100			0.9048	1.0000	0.9048	0.9048	0.9048	0.9048
nDCG@100	LMJM		0.8095	0.9048	0.9048	0.9048	0.8095	0.8095
R@100	(0.6)		0.9048	1.0000	0.9048	0.9048	1.0000	1.0000
P@10			0.8095	1.0000	0.8095	**0.7143**	**0.7143**	1.0000
AP@100				0.9048	0.9048	1.0000	1.0000	1.0000
nDCG@100	BM25			0.9048	0.9048	0.9048	1.0000	1.0000
R@100	(0.7, 0.3)			0.9048	0.8095	0.8095	0.9048	0.9048
P@10				0.8095	1.0000	0.9048	0.9048	0.8095
AP@100					0.9048	0.9048	0.9048	0.9048
nDCG@100	BM25				1.0000	1.0000	0.9048	0.9048
R@100	(1.0, 1.0)				0.9048	0.9048	1.0000	1.0000
P@10					0.8095	**0.7143**	**0.7143**	1.0000
AP@100						1.0000	1.0000	1.0000
nDCG@100	BM25					1.0000	0.9048	0.9048
R@100	(0.3, 0.7)					1.0000	0.9048	0.9048
P@10						0.9048	0.9048	0.8095
AP@100							1.0000	1.0000
nDCG@100	LMDir						0.9048	0.9048
R@100	(100)						0.9048	0.9048
P@10							0.8095	**0.7143**
AP@100								1.0000
nDCG@100	LMDir							1.0000
R@100	(500)							1.0000
P@10								**0.7143**

Table 5: Each cell in the table indicates the correlation (Kendall's τ) between QPP systems ranked in order by their evaluated effectiveness (measured with the help of Pearson's r for the results presented in this table) for each pair of IR models for 7 different QPP systems. The lowest correlation value for each group is marked in red. The lowest correlation in the table is bold-faced.

in QPP results. Consequently, for better reproducibility it is better to report results with Kendall's τ.

- **Lower variations in the QPP outcomes**: Compared to variations across IR evaluation metrics, we observe that the variations occurring across IR models is lower (compare the bold-faced green $\sigma(\mathcal{S})$ values with those of $\sigma(\theta)$ ones). This entails that experiments need to put more emphasis on a precise description of the IR metrics used for QPP evaluation.
- **Lack of a consistency on which combination of QPP method with IR evaluation context yields least the variance**: While WIG and UEF(WIG) exhibit lowest variances for a precision oriented evaluation of

Metric	Model	LMJM (0.6)	BM25 (0.7, 0.3)	BM25 (1.0, 1.0)	BM25 (0.3, 0.7)	LMDir (100)	LMDir (500)	LMDir (1000)
AP@100		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
nDCG@100	LMJM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
R@100	(0.3)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
P@10		0.9048	1.0000	0.9048	0.8095	0.9095	1.0000	1.0000
AP@100			1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
nDCG@100	LMJM		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
R@100	(0.6)		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
P@10			0.9048	1.0000	0.7143	0.7143	0.9048	0.9048
AP@100				1.0000	1.0000	1.0000	1.0000	1.0000
nDCG@100	BM25			1.0000	1.0000	1.0000	1.0000	1.0000
R@100	(0.7, 0.3)			1.0000	1.0000	1.0000	1.0000	1.0000
P@10				0.9048	0.8095	0.8095	1.0000	1.0000
AP@100					1.0000	1.0000	1.0000	1.0000
nDCG@100	BM25				1.0000	1.0000	1.0000	1.0000
R@100	(1.0, 1.0)				1.0000	1.0000	1.0000	1.0000
P@10					0.7143	0.7143	0.9048	0.9048
AP@100						1.0000	1.0000	1.0000
nDCG@100	BM25					1.0000	1.0000	1.0000
R@100	(0.3, 0.7)					1.0000	1.0000	1.0000
P@10						**0.6190**	0.8095	0.8095
AP@100							1.0000	1.0000
nDCG@100	LMDir						1.0000	1.0000
R@100	(100)						1.0000	1.0000
P@10							0.8095	0.8095
AP@100								1.0000
nDCG@100	LMDir							1.0000
R@100	(500)							1.0000
P@10								1.0000

Table 6: The difference of this table with Table 5 is that the QPP effectiveness is measured with Kendall's τ (instead of Pearson's r as in Table 5).

ground-truth retrieval effectiveness, for AvgIDF and NQC methods, the least variations are noted for recall.

5.2 RQ2: Variations in the Relative Ranks of QPP Methods

We now report results in relation to the second research question, where the intention is to measure how stable are QPP system ranks (ordered by their evaluated effectiveness measures) for variations in the QPP context.

Variation due to IR metrics. Tables 3 and 4 present the pairwise contingency table for different combinations of IR metrics for three different IR models. The following observations can be made from the results.

– **LMJM leads to the most instability in the relative QPP system ranks**: This behaviour, most likely, can be attributed to the fact that this

model has a tendency to favour shorter documents in the top-retrieved in contrast to LMDir or BM25.

- **Some evaluation metrics are more sensitive to the rank cut-off values**: For instance, the QPP ground-truth measured with Recall@10 yields considerably different results when the ground-truth corresponds to Recall@1000.
- **Relative ranks of QPP systems more stable with τ**: A comparison between the values of Tables 3 and 4 reveals that a rank correlation measure such as τ leads to better stability of QPP experiments than when r is used to measure the relative effectiveness of QPP models.

Variations due to IR models. Tables 5 and 6 present the pairwise contingency between retrieval similarity scores measured with different IR evaluation metrics. For this set of experiments, the intention is also to investigate how stable are QPP system ranks with changes not only to the retrieval model itself but also with different parameter settings on the same model, e.g. BM25(0.7,0.3)[3] vs. BM25(1,1). We observe the following:

- **Relative ranks of QPP systems are somewhat stable across IR models**: The correlation values of Tables 5 and 6 are higher than those of Tables 3 and 4, which shows that the QPP experiments are less sensitive to variations in the set of top documents retrieved by different similarity scores.
- **LMJM leads to more instability in the QPP outcomes**: LMJM shows the lowest correlation with other retrieval models. Parameter variations of an IR model usually lead to relatively stable QPP outcomes. For instance, see the correlations between LMDir(500) and LMDir(1000).
- **The relative ranks of QPP systems are more stable with τ**: This observation (a comparison between the values of Tables 5 and 6) is similar to the comparison between Tables 3 and 4. However, the differences between the correlation values are smaller in comparison to the differences observed between Tables 3 and 4.

6 Concluding Remarks

We have shown via extensive experiments that QPP outcomes are indeed sensitive to the experimental configuration used. As part of our analysis, we have found that certain factors, such as variations in the IR effectiveness measures, has a greater impact in terms of QPP outcomes than other factors, such as variations in the choice of IR models. An important outcome arising from this study is that future research on QPP should place greater emphasis on a clear specification of the experimental setup to enable better reproducibility.

For the experiments reported in this paper, we have used only the TREC Robust dataset. A natural question that we would like to explore in future concerns the impact of varying $\mathcal{Q}$ (the set of benchmark queries) on relative QPP outcomes.

[3] Values of k_1 and b, respectively, in BM25 [15].

References

1. Armstrong, T.G., Moffat, A., Webber, W., Zobel, J.: Improvements that don't add up: Ad-hoc retrieval results since 1998. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. p. 601–610. CIKM '09 (2009)
2. Aslam, J.A., Yilmaz, E.: Inferring document relevance from incomplete information. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM. pp. 633–642. ACM (2007)
3. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 33–40. SIGIR '00, Association for Computing Machinery (2000)
4. Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Sanderson, M., Järvelin, K., Allan, J., Bruza, P. (eds.) SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004. pp. 25–32. ACM (2004)
5. 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, ACM, New York, NY, USA (2010)
6. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services **2**(1), 1–89 (2010)
7. 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, ACM, New York, NY, USA (2002)
8. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Precision prediction based on ranked list coherence. Inf. Retr. **9**(6), 723–755 (Dec 2006)
9. 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, ACM (2008)
10. Hiemstra, D.: Using language models for information retrieval. Univ. Twente (2001)
11. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (Oct 2002)
12. Kurland, O., Shtok, A., Carmel, D., Hummel, S.: A unified framework for post-retrieval query-performance prediction. In: Proceedings of the Third International Conference on Advances in Information Retrieval Theory. p. 15–26. ICTIR'11 (2011)
13. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proc. of SIGIR '01. pp. 120–127. ACM, New York, NY, USA (2001)
14. Lin, J.: The neural hype, justified! a recantation. SIGIR Forum **53**(2), 88–93 (2021)
15. Robertson, S., Walker, S., Beaulieu, M., Gatford, M., Payne, A.: Okapi at trec-4 (1996)
16. Roitman, H.: An enhanced approach to query performance prediction using reference lists. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 869–872. Proc. SIGIR '17, ACM, New York, NY, USA (2017)

17. Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., Mitra, M.: Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. p. 1835–1838. CIKM '18 (2018)
18. 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)
19. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting query performance by query-drift estimation. ACM Trans. Inf. Syst. **30**(2) (2012)
20. Yang, W., Lu, K., Yang, P., Lin, J.: Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1129–1132. SIGIR'19, Association for Computing Machinery (2019)
21. Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. p. 102–111. CIKM '06, Association for Computing Machinery (2006)
22. Yilmaz, E., Kanoulas, E., Aslam, J.A.: A simple and efficient sampling method for estimating AP and NDCG. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR. pp. 603–610. ACM (2008)
23. 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 and Development in Information Retrieval. p. 105–114. SIGIR '18, ACM (2018)
24. Zendel, O., Shtok, A., Raiber, F., Kurland, O., Culpepper, J.S.: Information needs, queries, and query performance prediction. In: Proc. of SIGIR '19. pp. 395–404 (2019)
25. 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 (2001)
26. Zhou, Y., Croft, W.B.: Ranking robustness: A novel framework to predict query performance. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. p. 567–574. CIKM '06, ACM, New York, NY, USA (2006)
27. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 543–550. SIGIR '07 (2007)