On the Feasibility and Robustness of Pointwise Evaluation of Query Performance Prediction

Suchana Datta[1], Debasis Ganguly[2], Derek Greene[3] and Mandar Mitra[4]

[1] *University College Dublin, Ireland*
[2] *University of Glasgow, UK*
[3] *University College Dublin, Ireland*
[4] *Indian Statistical Institute, Kolkata, India*

Abstract

Despite the retrieval effectiveness of queries being mutually independent of one another, the evaluation of query performance prediction (QPP) systems has been carried out by measuring rank correlation over an entire set of queries. Such a listwise approach has a number of disadvantages, notably that it does not support the common requirement of assessing QPP for individual queries. In this paper, we propose a pointwise QPP framework that allows us to evaluate the quality of a QPP system for individual queries by measuring the deviations between each prediction versus the corresponding true value, and then aggregating the results over a set of queries. Our experiments demonstrate that this new approach leads to smaller variances in QPP evaluations across a range of different target metrics and retrieval models.

1. Introduction

Query performance prediction (QPP) methods have been proposed to automatically estimate the retrieval effectiveness for queries without making use of any true relevance information (e.g. [1, 2]). In practice, a QPP method allows us to dynamically adjust the processing steps for a query, depending on its initial performance estimate. Although estimating the performance of individual queries independently is a common requirement in many downstream tasks (e.g., adaptive query processing [3]), the standard QPP evaluation methodology adopted by the IR research community has previously involved a **listwise** approach, rather than a **pointwise** one. This is despite the fact that the latter represents a more appropriate strategy for use in downstream applications. To elaborate, a listwise approach operates on a *set of queries* $\mathcal{Q}$ by first converting it into an ordered set as induced by the QPP estimated scores $\phi(Q)\ \forall Q \in \mathcal{Q}$. It then computes a rank correlation measure, such as Kendall's τ, between the ground-truth ordering of the queries as induced by their average precision (AP) values [4] or by any other IR metric, such as nDCG [5].

QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks, co-located with 45th European Conference on Information Retrieval (ECIR) from the 2nd to the 6th of April 2023 in Dublin, Ireland

✉ suchana.datta@ucdconnect.ie (S. Datta); debasis.ganguly@glasgow.ac.uk (D. Ganguly); derek.greene@ucd.ie (D. Greene); mandar@isical.ac.in (M. Mitra)
🌐 https://gdebasis.github.io/ (D. Ganguly); http://derekgreene.com/ (D. Greene); https://www.isical.ac.in/mandar-mitra (M. Mitra)
iD 0000-0001-9220-6652 (S. Datta); 0000-0003-0050-7138 (D. Ganguly); 0000-0001-8065-5418 (D. Greene); 0000-0001-9045-9971 (M. Mitra)

A major disadvantage of listwise QPP approaches is that evaluation is conducted in a relative manner, so the performance of one query is measured relative to the others. However, a downstream performance estimate of an individual query also needs to be evaluated independently of the other queries. In contrast, a pointwise approach measures the effectiveness on individual queries, and then, if required, aggregates the results over a complete set. This is analogous to measuring the retrieval effectiveness metric MAP by computing the average precision values for individual queries and then aggregating them. Pointwise evaluation also allows us to carry out a per-query analysis of a method often leading to useful insights. For instance, Buckley [6] found that, by performing an extensive per-topic retrieval analysis, they were able to identify queries where most IR systems fail to retrieve relevant documents. However, a listwise evaluation methodology is not conducive to performing this kind of detailed per-query analysis.

Another drawback of listwise methods is that they can be overly sensitive to the configuration setup used for evaluation. The two most important such configurations are: i) the target retrieval evaluation metric that induces a ground-truth ordering over the set of queries; ii) the retrieval model used to obtain the top-k set of documents for QPP estimation. Indeed, variations in these configurations can lead to both large standard deviations in the reported rank correlation measures and significant differences in the relative ranks of various QPP systems [7]. To address the limitations of listwise methods, we propose a new QPP evaluation framework, **Aggregated Pointwise Absolute Errors** (**APAE**), which is shown to not only be consistent with the existing listwise approaches, but also to be more robust to changes in QPP experimental setup.

2. A Framework for Pointwise QPP Evaluation

Correlation with listwise ground-truth Before describing our new QPP evaluation framework APAE, we begin by introducing the required notation. Formally, a QPP estimate is a function of the form $\phi(Q, M_k(Q)) \mapsto \mathbb{R}$, where $M_k(Q)$ is the set of top-k ranked documents retrieved by an IR model M for a query $Q \in \mathcal{Q}$, a benchmark set of queries.

For the purpose of listwise evaluation, for each $Q \in \mathcal{Q}$, we first compute the value of a target IR evaluation metric, $\mu(Q)$ that reflects the quality of the retrieved list $M_k(Q)$. The next step uses these $\mu(Q)$ scores to induce a *ground-truth ranking* of the set $\mathcal{Q}$, or in other words, arrange the queries by their decreasing (or increasing) $\mu(Q)$ values, i.e.,

$$\mathcal{Q}_\mu = \{Q_i \in \mathcal{Q} : \mu(Q_i) > \mu(Q_{i+1}), \forall i = 1, \ldots, |\mathcal{Q}| - 1\}\} \qquad (1)$$

Similarly, the evaluation framework also yields a *predicted ranking* of the queries, where this time the queries are sorted by the QPP estimated scores, i.e.,

$$\mathcal{Q}_\phi = \{Q_i \in \mathcal{Q} : \phi(Q_i) > \phi(Q_{i+1}), \forall i = 1, \ldots, |\mathcal{Q}| - 1\} \qquad (2)$$

A listwise evaluation framework then computes the rank correlation between these two ordered sets $\gamma(\mathcal{Q}_\mu, \mathcal{Q}_\phi)$, where $\gamma : \mathbb{R}^{|\mathcal{Q}|} \times \mathbb{R}^{|\mathcal{Q}|} \mapsto [0, 1]$ is a correlation measure, such as Kendall's τ.

Individual ground-truth In contrast to listwise evaluations, where the ground-truth takes the form of an ordered set of queries, pointwise QPP evaluation involves making $|\mathcal{Q}|$ *independent*

Table 1
QPP configurations - (QPP method, IR metric, and models) used to measure variations.

QPP Methods	AvgIDF [8], Clarity [9], NQC [10], WIG [11], UEF(Clarity), UEF(NQC), UEF(WIG) [2]
IR Metrics	AP@100, nDCG@100, P@10, Recall@100
IR Models	LMJM ($\lambda = 0.6$), LMDir ($\mu = 1000$), BM25 $(k, b) = (0.7, 0.3)$

comparisons. Each comparison is made between a query Q's predicted QPP score $\phi(Q)$ and its retrieval effectiveness measure $\mu(Q)$, i.e.,

$$\eta(\mathcal{Q}, \mu, \phi) \overset{\text{def}}{=} \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \eta(\mu(Q), \phi(Q)) \tag{3}$$

Unlike the rank correlation γ, here η is a pointwise correlation function of the form $\eta : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$. It is often convenient to think of η as the inverse of a *distance* function that measures the extent to which a predicted value deviates from the corresponding true value. In contrast to ground-truth evaluation metrics, most QPP estimates (e.g., NQC, WIG etc.) are not bounded within $[0, 1]$. Therefore, to employ a distance measure, each QPP estimate $\phi(Q)$ must be normalized to the unit interval. Subsequently, η can be defined as $\eta(\mu(Q), \phi(Q)) \overset{\text{def}}{=} 1 - |\mu(Q) - \phi(Q)/\aleph|$, where $\aleph$ is a normalization constant which is sufficiently large to ensure that the denominator is positive.

Selecting an IR metric for pointwise QPP evaluation In general, an unsupervised QPP estimator will be agnostic with respect to the target IR metric μ. For instance, NQC scores can be seen as being approximations of AP@100 values, but can also be interpreted as approximating any other metric, such as nDCG@20 or P@10. Therefore, a question arises around which metric should be used to compute the individual correlations in Equation 3. Of course, the results can differ substantially for different choices of μ, e.g., AP or nDCG. This is also the case for listwise QPP evaluation, as reported in [7]. To reduce the effect of such variations, we now propose a simple yet effective solution.

Metric-agnostic pointwise QPP evaluation For a set of evaluation functions $\mu \in \mathcal{M}$ (e.g., $\mathcal{M} = \{\text{AP@100}, \text{nDCG@20}, \ldots\}$), we employ an aggregation function to compute the overall pointwise correlation (Equation 3) of a QPP estimate with respect to each metric. Formally,

$$\eta(Q, \mathcal{M}, \phi) = \Sigma_{\mu \in \mathcal{M}}(1 - |\mu(Q) - \phi(Q)/\aleph|), \tag{4}$$

where Σ denotes an aggregation function (it does not indicate summation). In particular, we use the most commonly-used such functions as choices for Σ: 'minimum', 'maximum', and 'average' – i.e., $\Sigma \in \{\text{avg}, \text{min}, \text{max}\}$. Next, we find the average over these values computed for a given set of queries $\mathcal{Q}$, i.e., we substitute $\eta(Q, \mathcal{M}, \phi)$ from Equation 4 into the summation of Equation 3.

3. Experiments

A QPP experiment context [7] involves three configuration choices: i) the **QPP method** itself that is used to predict the relative performance of queries; ii) the **IR metric** that is used to obtain a ground-truth ordering of the query performances as measured on a set of top-k ($k = 100$ in our experiments) documents retrieved by iii) a specific **IR model**. Table 1 summarizes the IR models and metrics used in our experiments, along with the relevant hyper-parameter values. The objective of our experiments is to investigate the following two key research questions:

- **RQ1**: Does APAE *agree* with the standard listwise correlation metrics?
- **RQ2**: How *robust* is APAE with respect to changes in the QPP experiment context?

An affirmative answer to **RQ1** would indicate that our proposed metric APAE is *consistent* with existing metrics used for QPP evaluation, while an affirmative answer to **RQ2** would suggest that APAE is preferable to existing methods due to its higher stability with respect to different experimental settings.

We conduct our QPP experiments on the TREC Robust dataset, which consists of 249 topics. Following the standard practice for QPP experiments [5, 12], we report results aggregated over a total of 30 randomly chosen equal-sized train-test splits of the data. The training split of each partition was used for tuning the hyper-parameters for the QPP method.

Agreement between listwise and pointwise evaluation Firstly, we investigate the consistency of APAE with respect to three standard listwise QPP evaluation metrics: Pearson's r, Spearman's ρ and Kendall's τ; and a pointwise approach, scaled Absolute Rank Error (sARE) [13]. Since sARE is an error measure, we measure correlations of APAE with $1 - \text{sARE}$ measures (which for the sake of simplicity, we refer to as sARE in Table 2). We experiment with three different instances of APAE obtained by substituting the aggregation functions – avg, min and max as Σ in Equation 4, denoted respectively as $\eta_{\text{avg}}(\mathcal{M})$, $\eta_{\text{min}}(\mathcal{M})$ and $\eta_{\text{max}}(\mathcal{M})$.

The results presented in Table 2 answer **RQ1** in the affirmative. Each reported value here corresponds to the rank correlation (Kendall's τ) between the relative ranks of the QPP systems ordered by their effectiveness as computed via one of the standard metrics (one of r, ρ, τ or sARE) and APAE, i.e., one of $\eta_{\text{avg}}(\mathcal{M})$, $\eta_{\text{min}}(\mathcal{M})$ and $\eta_{\text{max}}(\mathcal{M})$). The high correlation values between the standard listwise and the proposed pointwise metrics show that APAE can be used as a substitute for the standard listwise evaluation. Notably, we see that the average aggregate function yields the best results, and hence for the subsequent experiments we use $\eta_{\text{avg}}(\mathcal{M})$ as the pointwise evaluation metric.

Table 2
The correlation of our proposed pointwise evaluation metric APAE with the standard listwise metrics - Pearson's r, Spearman's ρ, Kendall's τ and sARE. The rank correlations between each pair of QPP system ranks (evaluated with a listwise measure and a pointwise measure) were computed with Kendall's τ. The high values indicate that the pointwise measurement can effectively *substitute* a standard list-based measure, since they lead to a fairly similar relative ordering between the effectiveness of different QPP methods.

	$\eta_{\text{avg}}(\mathcal{M})$				$\eta_{\text{min}}(\mathcal{M})$				$\eta_{\text{max}}(\mathcal{M})$			
	r	ρ	τ	sARE	r	ρ	τ	sARE	r	ρ	τ	sARE
BM25	0.810	0.810	0.905	0.887	0.778	0.778	0.794	0.813	0.802	0.810	0.794	0.794
LMDir	**0.905**	**0.810**	**0.905**	**0.887**	0.778	0.794	0.794	0.810	0.769	0.782	0.794	0.796
LMJM	0.810	0.810	0.810	0.846	0.794	0.794	0.782	0.786	0.794	0.769	0.810	0.846

Table 3
Stability of the proposed pointwise QPP metric APAE with respect to listwise approach, across different pairs of IR metrics and IR models. Red cells indicate the lowest value in each group, while the lowest values along each column are bold-faced.

Model	Metric	AP@100	R@10	R@100	nDCG@10	nDCG@100
LMJM	AP@10	0.497	0.813	0.429	0.783	0.429
BM25		0.897	0.722	0.722	0.793	0.793
LMDir		0.897	0.786	0.786	0.823	0.905
LMJM	AP@100		**0.328**	0.811	0.363	0.783
BM25			0.783	0.784	0.714	0.642
LMDir			0.823	0.901	0.834	0.789
LMJM	R@10			0.624	0.893	0.503
BM25				0.803	0.982	0.894
LMDir				0.903	0.864	0.864
LMJM	R@100				0.852	0.804
BM25					0.786	0.890
LMDir					0.738	0.738
LMJM	nDCG@10					0.537
BM25						0.904
LMDir						0.868

(a) Correlations between the relative ranks of 7 different QPP systems across different pairs of IR target metrics. QPP systems were evaluated with the baseline listwise metric - Kendall's τ.

Model	Metric	AP@100	R@10	R@100	nDCG@10	nDCG@100
LMJM	AP@10	0.904	1.000	0.715	1.000	0.792
BM25		1.000	1.000	1.000	1.000	1.000
LMDir		1.000	1.000	1.000	1.000	1.000
LMJM	AP@100		0.905	0.811	0.669	1.000
BM25			1.000	1.000	1.000	1.000
LMDir			1.000	1.000	1.000	1.000
LMJM	R@10			0.603	0.905	**0.542**
BM25				1.000	1.000	1.000
LMDir				1.000	1.000	1.000
LMJM	R@100				0.654	1.000
BM25					1.000	1.000
LMDir					1.000	1.000
LMJM	nDCG@10					0.649
BM25						1.000
LMDir						1.000

(b) Similar to Table 3a, except QPP performance was evaluated with the pointwise approach APAE. A comparison with Table 3a indicates a better consistency in the relative ranks of QPP systems for variations in the IR metrics.

Metric	Model	LMJM	BM25	BM25	LMDir	LMDir
		(0.6)	$(0.7, 0.3)$	$(0.3, 0.7)$	(500)	(1000)
AP@100	LMJM (0.3)	0.826	0.904	0.819	0.714	0.895
nDCG@100		0.780	0.694	0.695	0.759	0.759
R@100		0.824	0.769	0.782	0.904	0.904
AP@100	LMJM (0.6)		0.703	0.712	0.904	0.823
nDCG@100			0.781	0.827	0.811	0.811
R@100			0.813	0.725	0.731	**0.675**
AP@100	BM25 $(0.7, 0.3)$			0.903	0.785	0.785
nDCG@100				0.897	0.786	0.786
R@100				0.812	0.752	0.779
AP@100	BM25 $(0.3, 0.7)$				0.887	0.882
nDCG@100					0.901	0.895
R@100					0.889	0.901
AP@100	LMDir (500)					0.901
nDCG@100						0.893
R@100						0.903

(c) Here rank correlations between the relative ranks of QPP systems are measured across IR model pairs. As in Table 3a, QPP systems were evaluated with τ. The numbers alongside the IR models denote their respective parameters.

Metric	Model	LMJM	BM25	BM25	LMDir	LMDir
		(0.6)	$(0.7, 0.3)$	$(0.3, 0.7)$	(500)	(1000)
AP@100	LMJM (0.3)	1.000	1.000	1.000	1.000	1.000
nDCG@100		1.000	0.864	1.000	0.843	0.864
R@100		1.000	0.864	1.000	1.000	1.000
AP@100	LMJM (0.6)		1.000	1.000	1.000	1.000
nDCG@100			0.914	1.000	**0.813**	0.914
R@100			1.000	1.000	1.000	1.000
AP@100	BM25 $(0.7, 0.3)$			1.000	1.000	1.000
nDCG@100				1.000	1.000	1.000
R@100				0.812	0.905	1.000
AP@100	BM25 $(0.3, 0.7)$				1.000	1.000
nDCG@100					1.000	1.000
R@100					1.000	1.000
AP@100	LMDir (500)					1.000
nDCG@100						1.000
R@100						1.000

(d) Unlike Table 3c, here the QPP outcomes were evaluated by APAE (instead of τ).

Variances in relative effectiveness of QPP methods To investigate **RQ2**, we consider the relative stability of QPP system ranks for variations in QPP contexts (i.e., different IR models and target metrics), comparing both listwise and pointwise approaches (see Table 3). To clarify with an example, if working with three QPP methods, say AvgIDF, NQC, WIG, we observe that τ(NQC) $>$ τ(WIG) $>$ τ(AvgIDF) for LMDir as measured relative to AP@100. We expect to observe a similar ordering for a different choice of the IR model and target IR metric, say BM25 with nDCG@100. As in our previous experiments, here we measure the rank correlations between a total of seven QPP systems (see Table 1) via Kendall's τ.

4. Concluding Remarks

Unlike the standard listwise QPP evaluation mechanism of measuring an overall rank correlation with respect to a reference ranking of the queries (in terms of retrieval effectiveness), we have proposed a pointwise evaluation method that computes the relative difference between a normalized QPP score and a true IR evaluation measure (e.g., AP@100 or nDCG@20). Our experiments demonstrated that the proposed metric exhibits a high correlation with standard listwise approaches and is more robust to changes in QPP experimental setup than listwise evaluation measures. Using this metric, it should thus be possible to evaluate the effectiveness of different QPP methods on downstream tasks on a per-query basis.

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

References

[1] Y. Zhou, W. B. Croft, Ranking robustness: A novel framework to predict query performance, in: Proc. of CIKM '06, 2006, p. 567–574.

[2] A. Shtok, O. Kurland, D. Carmel, Using statistical decision theory and relevance models for query-performance prediction, in: Proc. of SIGIR '10, 2010, p. 259–266.

[3] Y. Lv, C. Zhai, Adaptive relevance feedback in information retrieval, in: Proc. of CIKM '09, 2009, p. 255–264.

[4] R. Cummins, Document score distribution models for query performance inference and prediction, ACM Transactions on Information Systems 32 (2014) 2:1–2:28.

[5] H. Zamani, W. B. Croft, J. S. Culpepper, Neural query performance prediction using weak supervision from multiple signals, in: Proc. of SIGIR '18, 2018, p. 105–114.

[6] C. Buckley, Why current IR engines fail, in: Proc. of SIGIR'04, 2004, p. 584–585.

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

[8] C. Hauff, D. Hiemstra, F. de Jong, A survey of pre-retrieval query performance predictors, in: Proc. of CIKM '08, 2008, p. 1419–1420.

[9] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting query performance, in: Proc. of SIGIR '02, 2002, p. 299–306.

[10] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits, Predicting query performance by query-drift estimation, ACM Transactions on Information Systems 30 (2012).

[11] Y. Zhou, W. B. Croft, Query performance prediction in web search environments, in: Proc. of SIGIR '07, 2007, p. 543–550.

[12] O. Zendel, A. Shtok, F. Raiber, O. Kurland, J. S. Culpepper, Information needs, queries, and query performance prediction, in: Proc. of SIGIR '19, 2019, pp. 395–404.

[13] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: Advances in Information Retrieval, 2021, pp. 115–129.