Unsupervised Query Performance Prediction for Neural Models utilising Pairwise Rank Preferences

Ashutosh Singh
University of Glasgow
Glasgow, United Kingdom
smilingashutosh@gmail.com

Debasis Ganguly
University of Glasgow
Glasgow, United Kingdom
debasis.ganguly@glasgow.ac.uk

Suchana Datta
University College Dublin
Dublin, Ireland
suchana.datta@ucdconnect.ie

Craig Macdonald
University of Glasgow
Glasgow, United Kingdom
craig.macdonald@glasgow.ac.uk

ABSTRACT

A query performance prediction (QPP) method predicts the effectiveness of an IR system for a given query. While unsupervised approaches have been shown to work well for statistical IR models, it is likely that these approaches would yield limited effectiveness for neural ranking models (NRMs) because the retrieval scores of these models lie within a short range unlike their statistical counterparts. In this work, we propose to leverage a pairwise inference-based NRM's (specifically, DuoT5) output to accumulate evidences on the pairwise believes of one document ranked above the other. We hypothesize that the more consistent these pairwise likelihoods are, the higher is the likelihood of the retrieval to be of better quality, thus yielding a higher QPP score. We conduct our experiments on the TREC-DL dataset leveraging pairwise likelihoods from an auxiliary model DuoT5. Our experiments demonstrate that the proposed method called Pairwise Rank Preference-based QPP (QPP-PRP) leads to significantly better results than a number of standard unsupervised QPP baselines on several NRMs.

ACM Reference Format:
Ashutosh Singh, Debasis Ganguly, Suchana Datta, and Craig Macdonald. 2023. Unsupervised Query Performance Prediction for Neural Models utilising Pairwise Rank Preferences. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592082

1 INTRODUCTION

A query performance prediction (QPP) method seeks to estimate the retrieval quality of an IR model without actually making use of the relevance assessments. Existing QPP research has mainly been dominated by unsupervised approaches [9, 16, 18, 23, 26] because of the two key advantages: first, they are interpretable (e.g., applying intuitive heuristics such as the variance of retrieval scores, as in NQC [18]), and second, unlike supervised approaches they do not need a training set of queries with available relevance assessments. Supervised QPP approaches mainly involve learning a relative preference of the effectiveness of queries in a pairwise manner, using data such as the word embedded vectors of queries and top-retrieved documents [4, 23] or with transformer-based embeddings [1, 6]. Such supervised approaches usually require a large quantity of training data to reach adequate levels of performance. To address this limitation, our work solely focuses on unsupervised QPP approaches that do not require any training examples.

In recent years, it has been shown that supervised neural ranking models (NRMs) outperform their unsupervised counterparts [3, 7, 8, 10, 11, 21], which implies that it is important to study whether off-the-shelf QPP approaches can adequately work well for these supervised data-driven models. Existing research has shown that an off-the-shelf application of standard unsupervised QPP estimators on NRMs yields limited QPP effectiveness [5]. The authors of [5] argue that this can be attributed to the way in which the retrieval status values (RSVs) are computed in an NRM, i.e., via the application of a neural activation function, such as tanh or sigmoid, which due to their short range ($[-1, 1]$ and $[0, 1]$, respectively) restricts the RSVs to lie within short intervals. In other words, summing up these neural activation outputs over the query terms (query lengths being usually small) would yield a small number, as a result of which standard QPP methods, which leverage statistics on the top-retrieved scores (usually not bounded within such small ranges) may not be effective for NRMs.

In this paper, we propose a novel unsupervised QPP approach specifically tailored to work well for NRMs. Specifically, we propose an unsupervised method that leverages information from **pairwise rank preferences** (a way to obtain these likelihoods is via an NRM capable of pairwise inference, e.g., DuoT5 [15]). The core hypothesis is that a query for which an IR system ends up retrieving a relatively high volume of relevant content would also exhibit a higher agreement of the observed ranking of documents with the predicted probabilities of relative rank preferences inferred from an auxiliary pairwise ranking model (e.g., DuoT5). The main contribution of this work is that we show that *even unsupervised approaches can work effectively well in predicting the query performance of NRMs.*

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592082

2 PAIRED RANK PREFERENCE-BASED QPP

We now formally define our proposed unsupervised predictor, 'Paired Rank Preference-based QPP' (abbreviated as QPP-PRP).

Generic form of a QPP Estimator. A QPP estimator is a function θ which takes as input a query Q, a list of top-k retrieved documents as ranked by an IR model ϕ - which we denote as $L_k(Q, \phi)$ and outputs a real-valued score, i.e., $\theta : (Q \times \phi \times k) \mapsto \mathbb{R}$. The output score $\theta(Q, \phi, k)$ represents an estimate of the quality of the top-k list as retrieved by an IR model ϕ.

Leveraging Pairwise Likelihoods from an Auxiliary Model. We propose to make use of the pairwise predictions from a supervised ranking model ζ that given a pair of documents (D, D') retrieved in response to a query Q outputs the probability of the event that D is observed in a ranked list above D'. A concrete example of such a pairwise inference-based ranking model is DuoT5 [15]. Note that the ranking model for the pairwise likelihood calculations, ζ, is different from the base ranking model ϕ that induces the top-k ranking, $L_k(Q, \phi)$, the quality of which is eventually to be estimated.

For each pair of documents (D, D'), where both D and $D' \in L_k(Q, \phi)$ (for brevity we denote this as L_k from hereon), the auxiliary pairwise inference model ζ outputs a parameterized probability of the form $P^{\zeta}_{D<D'}$, i.e., a probability parameterized by the pairwise model ζ indicating the likelihood of the event that D is retrieved at a better rank than D'. In practice, the output of a pairwise inference model, e.g., DuoT5, is a Sigmoid, which when fed as input a triple of the form $\langle Q, D, D' \rangle$ outputs a value less than 0.5 if D is to be ranked better than D' (a value higher than or equal to 0.5 denotes the complementary event).

In a similar manner, the model outputs a different likelihood if a triple is fed as input in the reverse order, i.e., when $\langle Q, D', D \rangle$ is used as an input, the pairwise model outputs a likelihood of D' being ranked better than D. These two likelihoods are usually not identical, and it is a common practice to take the average value of these two likelihoods as the estimated belief of D being ranked better than D' [15]. Formally,

$$P^{\zeta}_{D<D'} \stackrel{\text{def}}{=} \frac{1}{2}(P[r(D) < r(D'); \zeta] + 1 - P[r(D) > r(D')); \zeta]), \quad (1)$$

where $r(D)$ denotes the rank of a document D, and $P[X; \zeta]$ represents the estimated probability of an event X using the parameterized likelihoods obtained from the model ζ.

QPP Formulation with Pairwise Aggregates. We now describe how we use the pairwise probabilities of Equation (1) towards deriving a QPP estimate. The key idea is that a ranked list is hypothesized to be of good quality if the aggregated beliefs from the pairwise probabilities from the auxiliary model ζ agrees consistently well with an observed ranked list.

As a first step of QPP-PRP, for each document D in the top-k retrieved set, we partition the remaining documents of the top-k into two non-overlapping sets - i) documents that are ranked worse than D, denoted as D_{bottom}, and ii) those that are ranked better than D, which we denote as D_{top}. Now, for the pivot document D, the estimated log-probability of observing documents ranked better than D is given by aggregating the $P^{\zeta}_{D'<D}$ values over each D' that

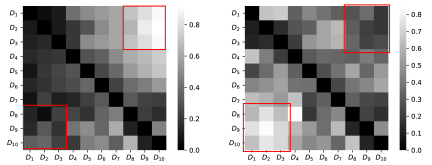


Figure 1: A visual illustration of the working principle of QPP-PRP; *Left*: a query for which the pairwise ranking preferences obtained via the auxiliary model ζ largely conforms with the observed ranking, *Right*: depicts the reverse of this situation, where the observed rank is less consistent with the pairwise rank preferences.

are actually observed to be ranked better than D, i.e.,

$$\log P(D_{\text{top}}|D; \zeta) = \sum_{D' \in D_{\text{top}}} \log(1 + P^{\zeta}_{D'<D}), \quad (2)$$

By symmetry, the likelihood for D_{bottom} is given as

$$\log P(D_{\text{bottom}}|D; \zeta) = \sum_{D' \in D_{\text{bottom}}} \log(1 + P^{\zeta}_{D'<D}). \quad (3)$$

The estimate of Equation (3) should be low because an effective pairwise inference model would output low values for $P^{\zeta}_{D'<D}$. This is because as per the actual observation a document $D' \in D_{\text{bottom}}$ is ranked worse than D. Intuitively, the odds-ratio of the two complementary events of Equations (2) and (3) thus should indicates a high level of agreement between the observed and the predicted rank orders as obtained with ζ. This is because in such cases, we would expect the quantity in Equation (2) to be maximised and the one in Equation (3) to be minimised. Therefore, we aggregate the beliefs over all $D \in L_k(Q)$, to compute our proposed QPP score as

$$\text{QPP-PRP}(Q) = \frac{\sum_{D \in L_k(Q)} \log P(D_{\text{top}}|D; \zeta)}{\sum_{D \in L_k(Q)} \log P(D_{\text{bottom}}|D; \zeta)}. \quad (4)$$

Relation with Uncertainty in Ranking Preferences. We also illustrate the idea visually in Figure 1, where we show the sample pairwise probability values of the top-k documents ($k = 10$) for two queries. Each is shown as a matrix, where the $(i, j)^{\text{th}}$ entry denotes the probability $P^{\zeta}_{D_i<D_j}$. The gray-scale intensity of the part of the matrix above the diagonal thus corresponds to the aggregated beliefs of the numerator of Equation (4). Similarly, the region below the diagonal represents the aggregated beliefs of the quantity in the denominator of Equation (4). The agreement of the observed ranks with the pairwise preferences can thus be visually interpreted with the relative gray-scale intensities of the regions above and below the diagonal. For instance, the left plot of Figure 1 represents an easy query (with higher agreements between the observed and the predicted ranks) because the upper part and the lower parts of the diagonal show high and low intensities, respectively. Similarly, the right plot of Figure 1 represents a more difficult query (low QPP estimate) because the lower part of the diagonal being comprised of relatively high intensity pixels increases the denominator of Equation (4).

In particular, for obtaining the pairwise probabilities of the form $P^{\zeta}_{D<D'}$, we make use of the DuoT5 model [15] as a concrete realisation of ζ. Indeed, DuoT5 has been shown to outperform other transformer based neural models, such as the DuoBERT [15], most

likely because the T5 transformer model being a generative model usually works better than a non-generative model, such as BERT.

3 EVALUATION

3.1 Experiment Setup

We investigate the following research questions:

- **RQ-1**: Are existing unsupervised QPP approaches effective for NRMs?
- **RQ-2**: Do the pairwise document ranking order probabilities as obtained from a pairwise inference model, such as DuoT5, help improve the QPP effectiveness, as per our hypothesis?
- **RQ-3**: Do we observe similar trends in QPP effectiveness for the two modes of NRMs: re-ranking documents retrieved with a sparse index, and end-to-end retrieval with a dense index?

3.1.1 Dataset and IR Models. We experiment with the MS MARCO passage ranking dataset [13], a standard IR collection commonly used to evaluate the quality of supervised NRMs. We use the TREC DL'19 and DL'20 topics for our experiments. We employ a number of different NRMs as the target IR models for QPP evaluation. In particular, we employ both sparse indexing based re-ranking and end-to-end approximate retrieval on dense indexes.

The initial top-1000 documents for the sparse index-based re-ranking methods were obtained with BM25. We denote such a sparse index-based NRM as 'BM25 + X', where X is the name of the neural approach used to re-rank the top-1000 retrieved by BM25 (an example of 'X' is ColBERT). In particular, as re-ranking models, we employed models from the following two families.

- **Neural Augmentation**: This involves augmenting the text of a document either by boosting the term frequency of important terms, such as DeepCT [2], or by including additional terms using a seq2seq transformation, such as docT5query [14]. The models used in our experiments from this family are the following - 'BM25 + DeepCT + BERT', 'BM25 + docT5query + BERT' and 'BM25 + docT5query + ColBERT' (method 'X + Aug + Y' means that 'Y' is used to re-rank the results obtained with 'X' after augmenting the index 'X' with 'Aug').
- **Dense Retrieval**: This family consists of the dense retrieval-based approaches, where a set of top-k documents is retrieved by employing approximate nearest neighbor search on dense embeddings. In this family, we employ two models - ColBERT-E2E and ANCE. ANCE [22] is a single representation-based dense retrieval model that selects the hard negatives globally from the entire corpus during training. We use the default checkpoint trained on MS MARCO training data for our experiments. Additionally, we also employ a contextualized query expansion-based model, BERT-QE [25], on our dense retrieval baseline, ColBERT-E2E.

The retrieval outputs of the neural models were obtained from the runs available at [19][1].

3.1.2 Baseline QPP Methods. As baselines, we employ two standard unsupervised QPP approaches that have been reported to work well in the literature, namely - (i) **NQC** [18, 24], and (ii) **UEF** [17], with NQC as the base estimator. While NQC estimates the retrieval quality by measuring how skewed is the distribution of the RSVs at the very top-ranks, UEF involves sampling various sub-sequences of the top-ranked documents followed by estimating the robustness of each by computing the perturbations in the rank orders before and after relevance feedback (e.g., by RLM [12]).

3.1.3 Ablations of QPP-PRP. Our proposed method hypothesises that leveraging pairwise ranking preference information should improve QPP estimation. We use one of the most effective pairwise models, DuoT5 [15], for calculating the pairwise ranking preferences. In doing so, we implicitly measure the agreement between DuoT5 and the input ranking. Hence, as an ablation of our proposed method, we first obtain a list with an NRM ϕ, say $L_k(\phi)$, and then re-rank this list by applying DuoT5 (a pairwise re-ranker) to obtain $L_k(\zeta)$. In particular, we employ two methods, namely Kendall's τ and RBO [20], to compute the QPP estimate as a rank correlation between $L_k(\phi)$ and $L_k(\zeta)$ yielding two ablations, which we respectively, denote as **DuoT5-τ** and **DuoT5-RBO**.

It is worth mentioning that this ablation cannot be applied on the output of the auxiliary pairwise model ζ itself, because the rank correlation between $L_k(\phi)$ and $L_k(\zeta) = 1$ for all queries if ϕ is identical to ζ. Consequently, no ablation effectiveness is reported on the Duo-T5 results of Table 1 (the gray cells).

Furthermore, since as per the visual example of Figure 1, an anti-symmetric matrix of pairwise rank preferences indicates high retrieval quality, as a different ablation we simply compute how symmetric a matrix is by employing the following standard measure of anti-symmetry of a matrix: $\text{AS} = (|A+A^T|-|A-A^T|)/(|A+A^T|+|A-A^T|)$. This value if smaller indicates a better retrieval quality. We call this method **AS**; this is an ablation of QPP-PRP because it only checks for the lack of symmetric properties of the pairwise preference matrix without taking into account which part of the matrix (the part above or below the diagonal) is comprised of larger values as checked by our method (Equation (4)).

3.1.4 Parameters and Evaluation Measures. We follow the standard setup for QPP experiments of conducting a 30-fold evaluation on 50:50 splits [18, 23, 24]. For each fold, the training set was used to tune the hyper-parameter common to all the QPP methods - the number of top-documents. This was tuned via grid search on the set $\{10, 20, 30, 40, 50\}$. To evaluate QPP, we use the Pearson's-r and Kendall's-τ, denoted as P-r and K-τ, respectively, as prescribed in the literature [18, 23]. The QPP evaluation considers AP@50 for deriving the ground-truth ordering of the queries[2].

3.2 Results and Discussion

Main observations. Table 1 presents a comparison between the different QPP methods for the different NRMs investigated. In relation to **RQ-1**, it is observed that off-the-shelf QPP approaches (baselines NQC and UEF), despite performing well on BM25, do not perform well on NRMs (except for BM25+MonoT5+DuoT5). This observation is consistent with those reported in [5].

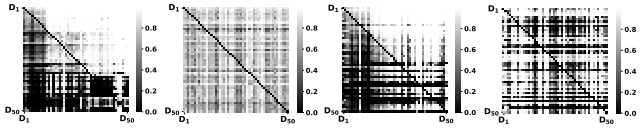
Second, in relation to **RQ-2**, it is observed that QPP-PRP significantly outperforms the unsupervised baselines NQC and UEF, which shows that our hypothesis of making use of the pairwise rank preferences is indeed beneficial. Moreover, the rank perturbation based ablation methods DuoT5-τ and DuoT5-RBO, although

[1] https://github.com/Xiao0728/ColBERT-PRF-VirtualAppendix

[2] Implementation available at: https://github.com/smilingashutosh/QPP-PRP

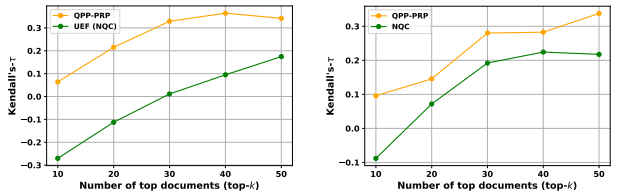
Table 1: QPP effectiveness for both sparse-reranked (top-half) and dense (bottom-half) neural models on TREC-DL ('19 and '20) topics. The best results in each group are bold-faced, and the best results across the groups are underlined. A '*' alongside a baseline or an ablation indicates that QPP-PRP (ours), is significantly better than the corresponding method (t-test with 95% confidence).

			QPP Evaluation											
			Baselines				Ablations						Ours	
			NQC		UEF		AS		DuoT5-τ		DuoT5-RBO		QPP-PRP	
	IR Model	AP@50	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ
Sparse Re-ranked	BM25	0.2571	**0.4669**	**0.2528**	0.3789	0.2223	-0.1938*	-0.1469*	0.2253*	0.2349*	0.2623*	0.2259	0.2896	0.2135
	+BERT	0.3591	0.3462*	0.1282*	0.3746*	0.1245*	0.1930*	0.1159*	0.4356*	0.3118	0.3962*	0.2723*	**0.5213**	**0.3160**
	+MonoT5	0.3871	0.2985*	0.0766*	0.3432*	0.1173*	0.3011*	0.2103*	0.4297*	0.2736*	0.3188*	0.1793*	**0.5256**	**0.3790**
	+DeepCT+BERT	0.3585	0.3543*	0.0916*	0.3700*	0.0776*	0.2426*	0.1600*	0.4400*	0.3085*	0.3760*	0.2515*	**0.4745**	**0.3170**
	+docT5query+ColBERT	0.3797	0.4259*	0.2520*	0.3437*	0.1556*	0.2144*	0.1525*	0.4886*	0.3495	0.5099*	0.3402*	**0.5205**	**0.3560**
	+docT5query+BERT	0.3651	0.2382*	-0.0266*	0.3028*	-0.0167*	0.2515*	0.1759*	0.4581*	0.3345*	0.4326*	0.2927*	**0.6143**	**0.4160**
	+MonoT5+DuoT5	0.3621	**0.4084**	**0.2253**	0.3747	0.1828*	0.2243*	0.1491*					0.3256	0.2157
Dense	ColBERT-E2E	0.3497	0.2492*	0.2111*	0.1989*	0.1770*	0.2129*	0.1132*	0.3467*	0.2922*	0.3120*	0.2356*	**0.4383**	**0.3026**
	ANCE	0.2963	0.1837*	0.1622*	0.1523*	0.1454*	0.1190*	0.0893*	**0.3563**	**0.2839**	0.3264*	0.2320*	0.3381	0.2411
	ColBERT-E2E+BERT-QE	0.3613	0.2455*	0.2070*	0.1796*	0.1505*	0.2298*	0.1196*	0.2953*	0.2554*	0.3051*	0.2331*	**0.4677**	**0.3171**



(a) AP=0.9751 **(b) AP=0.0596** **(c) AP=0.9118** **(d) AP=0.0018**

Figure 2: Pairwise probabilities ($P^{\zeta}_{D<D'}$) of the top 50 documents for four sample queries both with high and low AP@50. *(a)* and *(b)*: the QPP-PRP matrix on BM25+MonoT5 (sparse); *(c)* and *(d)*: the pairwise probabilities for a dense NRM (ColBERT-E2E).



(a) BM25+MonoT5 **(b) ColBERT-E2E**

Figure 3: Effect of variations in k (the number of top documents) on QPP-PRP and the best performing baselines (Table 1).

worse than QPP-PRP is, in fact, mostly better than the baselines, NQC and UEF. This also shows that the ranking order of DuoT5 with respect to a target IR model can itself be a good indicator of QPP effectiveness. The poor performance of the ablation method AS is attributed to its inability to discern whether it is the upper part of the pairwise probability matrix that is comprised of high values. This is, in fact, addressed by our method QPP-PRP, via the odds-ratio of Equation (4).

Interestingly, the effect of QPP-PRP on DuoT5 itself is not that effective because it is likely that the method of relative rank preferences can only reliably estimate the robustness of any other retrieval model, i.e., *relative to* DuoT5. In relation to **RQ-3**, we observe that the trends are similar even for end-to-end dense NRMs.

Per-query analysis. We now illustrate the matrix of pairwise rank preferences on sample queries from the TREC-DL dataset (a schematic was earlier shown in Figure 1). Figures 2a and 2b plot the pairwise rank preference on the top-50 results obtained with BM25+MonoT5 (sparse index) for two sample queries - one for which MonoT5 yields good retrieval quality, and the other for which the retrieval quality is not effective (see the AP@50 values). A visual inspection of Figure 2a reveals that the top-right part of the matrix is brighter than the bottom-left part, which shows that a query for which a target IR model leads to a consistent matrix of rank preferences also leads to higher retrieval effectiveness. On the other hand, in Figure 2b, no such emergent pattern is seen and QPP-PRP estimator in this case yields a low value. This means that the QPP estimate, in this case, is also accurate because the AP value measured for this query is indeed low. In Figures 2c and 2d, a similar trend is observed for the end-to-end NRM, ColBERT-E2E.

Sensitivity analysis. We now investigate the sensitivity of the QPP-PRP model with respect to the number of top documents (k) used to compute the QPP estimate. Figure 3 shows that our proposed QPP-PRP performs consistently better than the best performing baseline (as per Table 1) for a range of different values of k for both the best performing sparse-reranking and dense NRMs.

4 CONCLUDING REMARKS

In this paper, we proposed an unsupervised QPP approach specifically targeted to act effectively on neural models, where the similarity scores are distributed over a short range. Our proposed model alleviates this limitation by leveraging the pairwise rank preference probabilities obtained from a pairwise inference model, such as DuoT5. Our experiments demonstrated that our proposed QPP method significantly outperforms other state-of-the-art unsupervised QPP approaches.

In future, we may investigate ways of using the information from pairwise rank preference likelihoods as a part of a supervised model to further improve results.

REFERENCES

[1] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized Pre-Trained Transformers for Query Performance Prediction. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2857–2861.

[2] Zhuyun Dai and Jamie Callan. 2020. Context-Aware Document Term Weighting for Ad-Hoc Search. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1897–1907.

[3] Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional Neural Networks for Soft-Matching N-Grams in Ad-Hoc Search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 126–134.

[4] Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. 2022. Deep-QPP: A Pairwise Interaction-Based Deep Learning Model for Supervised Query Performance Prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 201–209.

[5] Suchana Datta, Debasis Ganguly, Mandar Mitra, and Derek Greene. 2022. A Relative Information Gain-Based Query Performance Prediction Framework with Generated Query Variants. *ACM Trans. Inf. Syst.* 41, 2, Article 38 (Dec 2022), 31 pages.

[6] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A 'Pointwise-Query, Listwise-Document' based Query Performance Prediction Approach. In *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. ACM, New York, NY, USA, 2148–2153.

[7] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 65–74.

[8] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-Hoc Retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*. ACM, New York, NY, USA, 55–64.

[9] Claudia Hauff. 2010. Predicting the Effectiveness of Queries and Retrieval Systems. *SIGIR Forum* 44, 1 (Aug. 2010), 88.

[10] Yongyu Jiang, Peng Zhang, Hui Gao, and Dawei Song. 2020. A Quantum Interference Inspired Neural Matching Model for Ad-Hoc Retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 19–28.

[11] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48.

[12] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proc. of SIGIR '01* (New Orleans, Louisiana, USA). ACM, New York, NY, USA, 120–127.

[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *CoRR* abs/1611.09268 (2016).

[14] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *CoRR* abs/1904.08375 (2019).

[15] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. *CoRR* abs/2101.05667 (2021).

[16] Dwaipayan Roy, Debasis Ganguly, Mandar Mitra, and Gareth J. F. Jones. 2019. Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Inf. Process. Manag.* 56, 3 (2019), 1026–1045.

[17] Anna Shtok, Oren Kurland, and David Carmel. 2010. 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 (SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 259–266.

[18] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (2012), 35 pages.

[19] Xiao Wang, Craig MacDonald, Nicola Tonellotto, and Iadh Ounis. 2023. ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval. *ACM Trans. Web* 17, 1, Article 3 (Jan 2023), 39 pages.

[20] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (Nov 2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[21] Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. 2017. End-to-End Neural Ad-Hoc Ranking with Kernel Pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 55–64.

[22] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *CoRR* abs/2007.00808 (2020).

[23] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. ACM, New York, NY, USA, 105–114.

[24] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. In *Proc. of SIGIR '19* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 395–404.

[25] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4718–4728.

[26] Yun Zhou and W. Bruce Croft. 2007. Query Performance Prediction in Web Search Environments. In *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 543–550.