Is Query Performance Prediction With Multiple Query Variations Harder Than Topic Performance Prediction?

Oleg Zendel
oleg.zendel@rmit.edu.au
RMIT University
Melbourne, VIC, Australia

J. Shane Culpepper
shane.culpepper@rmit.edu.au
RMIT University
Melbourne, VIC, Australia

Falk Scholer
falk.scholer@rmit.edu.au
RMIT University
Melbourne, VIC, Australia

ABSTRACT

Accurately estimating the retrieval effectiveness of different queries representing distinct information needs is a problem in Information Retrieval (IR) that has been studied for over 20 years. Recent work showed that the problem can be significantly harder when multiple queries representing the same information need are used in prediction. By generalizing the existing evaluation framework of Query Performance Prediction (QPP) we explore the causes of these differences in prediction quality in the two scenarios. Our empirical analysis demonstrates that for most predictors, this difference is solely an artifact of the underlying differences in the query effectiveness distributions. Our detailed analysis also demonstrates key performance distribution properties under which QPP is most and least reliable.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**.

KEYWORDS

Query Performance Prediction; Evaluation; Query Variations

ACM Reference Format:
Oleg Zendel, J. Shane Culpepper, and Falk Scholer. 2021. Is Query Performance Prediction With Multiple Query Variations Harder Than Topic Performance Prediction?. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3463039

1 INTRODUCTION

The Query Performance Prediction (QPP) task has been extensively explored over the past two decades, with new QPP methods and applications proposed over the years. Formally, the goal of QPP is to estimate the effectiveness of a search performed in response to a query in the absence of human relevance judgments [5]. In practice, most prior work performed the estimation and evaluated it based on a fixed corpus and retrieval method for a batch of pre-defined queries. Previous work [17, 19] showed that by using fixed retrieval parameters, the evaluation reduces to ranking the queries by the retrieval effectiveness for that corpus and retrieval method, but this may not be an appropriate way to evaluate quality if either of these are changed. This does not limit the value of QPP, and we should not be surprised that different rankers or collections can lead to different conclusions, but it is an important reminder that we should always make sure the evaluation we choose is appropriate for our task as properly evaluating QPP can be difficult to get right without careful planning [11].

One common application of ranking queries by effectiveness is the task of ad-hoc retrieval in a search engine. Many improvements to ad-hoc search have been made over the years. Benham et al. [4] showed how simple low-cost fusion techniques can outperform even complex State-of-the-art (SOTA) Learning to Rank (LtR) techniques. Recent advances in Natural Language Processing (NLP) modeling [10, 26] coupled with the public availability of commercial search query logs such as MS-MARCO [15] and ORCAS [7] have also led to substantial performance improvements in ranked retrieval effectiveness. An interesting new outcome of NLP modeling is the ability to induce many query variants (often shown as query suggestions in commercial search engines [16]), all of which represent a single information need [13]. While there are no guarantees which variants may be most effective, QPP might be a valuable tool to find the best candidates in the future. The importance of query variants has been studied for many years in IR, but evaluation exercises rarely consider any confounding factors they might introduce, and usually a single query is provided and treated as a unique *topic*. Most QPP evaluation exercises follow the model of "one query per topic" as well.

Given that query suggestion is now common and large test collections such as ORCAS contain many near duplicate queries which may be topic related, several recent papers have begun to explore the impact that query variants may have in various QPP applications. Using recently introduced human curated query variants for an existing TREC collection, Zendel et al. [30] showed that the relative QPP quality significantly varies when different queries are used to represent the information need. Specifically, the effectiveness of the queries that represent the topic impacts the relative prediction quality. This calls for a change in the existing evaluation framework. In order to be able to more reliably determine the effectiveness of different QPP methods, we evaluate them on a significantly larger dataset that consists of queries with varying effectiveness. In an early study of this problem, Thomas et al. [25] found that due to the conflation of queries with information needs (topics) in most of the collections used for QPP experiments, the existing predictors were in effect ranking topics, and that there were important differences in prediction quality when using several queries for one topic (variants), versus using only one query per topic, as in the previously assumed framework. Hereafter, we differentiate between estimation of queries from different topics to estimation of queries from

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463039

the same topic. The two QPP task categories are *inter-topic* QPP for ranking queries from different topics, and *intra-topic* QPP (also called query selection and Query Variation Performance Prediction (QVPP) [18]), which ranks multiple queries for a single topic.

Contributions. In this work, we demonstrate that previous studies of this problem had an important confounding factor in the data used, and these differences may not justify the conclusion that QPP is not effective for the intra-topic case. By re-examining both of these QPP tasks from a different angle, we show that the most important factor in QPP effectiveness is the magnitude of true effectiveness differences between queries, and not necessarily the topic they represent. When the overall distribution of score differences is the same, QPP methods tend to have similar prediction quality regardless of the underlying topic for some QPP methods, while for other methods the topic effect does still matter. The aim of this paper is to explore the circumstances for the prediction quality variance in 18 different QPP methods for intra-topic and inter-topic predictions, and to provide important insights into better understanding the conditions which lead to effective QPP.

2 RELATED WORK AND BACKGROUND

QPP methods are traditionally separated into two categories based on when they are computed: *pre-retrieval* and *post-retrieval* predictors. Pre-retrieval predictors, as the name suggests, are computed a priori to retrieval and estimate the effectiveness of a result using statistical information from the corpus and query. Post-retrieval predictors require additional information from each retrieval result, such as the retrieval scores or term distributions from the top documents in the result set. Post-retrieval predictors are generally more accurate than pre-retrieval predictors, as they have access to additional information, but are usually more expensive to compute. We perform our analysis using 18 common QPP methods that have been shown to be effective for inter-topic ranking, and are common baselines in recent QPP research.

Pre-retrieval predictors can be categorized as: (1) *Specificity* of the query terms in the collection: MaxIDF [20] and AvgIDF [8]; (2) *Similarity* between the query terms and the collection: SCQ, MaxSCQ and AvgSCQ [31]; and (3) *Coherency* with respect to the documents that contain the query terms: SumVAR, MaxVAR and AvgVAR [31]. Similarly, post-retrieval predictors can be categorized as: (1) *Clarity* based methods measure the coherence of the results with respect to the corpus. Multiple versions of Clarity [1, 6, 8, 9] have been proposed; We use the original version proposed by Cronen-Townsend et al. [8]. (2) *Robustness* of the retrieval result with respect to different perturbations, for example the QF (query feedback) method [32]. (3) *Scores distributions* of the results, such as WIG [32], NQC [22] and SMV [24], as well as all of the corresponding *utility estimation framework* (*UEF*) [21] instantiations, which measure the robustness of a result list based on underlying QPP method.

Query variations. Thomas et al. [25] used human curated query variants [2] to test existing pre-retrieval QPPs in different settings, including the representation of a topic with multiple queries. The effect differences were measured between tasks (topics), rankers and queries. Their findings concluded that the existing pre-retrieval QPPs mainly predict the difficulty of the topic, which is represented by the median effectiveness of the intra-topic queries, rather than the effectiveness of queries. The QPP methods were evaluated by *selecting queries*, that is, predicting which query is most effective in each query pair. The intra-topic prediction is identical to the query selection task, but the authors only compared the accuracy between different predictors. In this work, we propose an alternative comparison technique to inter-topic prediction. Our analysis provides several new insights into the accuracy of the query prediction task.

Scells et al. [18] presented a survey of existing QPP methods for systematic review. The authors proposed the use of the term QVPP when estimating the effectiveness of query variants for the same topic (intra-topic). Several QPP methods were tested for three different sub-tasks: ranking query variants, identifying a query variant better than a specific seed query; and identifying the best query variant for a topic. To evaluate the ranking of query variants, different correlation coefficients were calculated for each topic, and the average and standard deviation were reported. The authors concluded that the tested existing methods were insufficient for the purpose. Moreover, when comparing to inter-topic QPP, generally lower correlations were observed than in the QVPP task; but no statistical testing was reported, so it is unclear which of the outcomes were significant. In this work we propose an alternative approach to compare the two tasks. Based on our findings, we also explore several plausible explanations for the different behaviors observed for QPP performance in an inter- versus intra-topic query comparisons.

3 EVALUATION METHODOLOGY

The standard evaluation framework for QPP uses correlation coefficients to measure the strength of the relationship between retrieval effectiveness scores and the predicted scores, for each query. Since this can be seen as a ranking task, consider the traditional Kendall's τ_b [12] correlation coefficient, which is a commonly reported measure for QPP evaluation. The measure can be reduced to comparing a distribution of scores over all query pairs. As such, Vigna [27] shows that the numerator of Kendall's τ can be reformulated as:

$$\sum_{i<j} \operatorname{sgn}\big(AP(q_i) - AP(q_j)\big) \operatorname{sgn}\big(\mathcal{P}(q_i) - \mathcal{P}(q_j)\big). \tag{1}$$

In our context, $\mathcal{P}(q_i)$ and $AP(q_i)$ are the prediction and Average Precision (AP) values for query i, respectively; and

$$\operatorname{sgn}(x) \coloneqq \begin{cases} 1 & \text{if } x > 0; \\ 0 & \text{if } x = 0; \\ -1 & \text{if } x < 0. \end{cases}$$

In the simple case with no ties, the denominator would be $\binom{N}{2}$ where N is the total number of queries in our context. The correlation over pairwise comparisons is shifted to an accuracy score by treating each pair of queries as a single sample. A binary score is defined for each pair of queries in the set $\{(q_i, q_j) \mid i < j\}$ as follows:

$$S(q_i, q_j) \coloneqq \begin{cases} 1 & \text{if } \operatorname{sgn}\big(AP(q_i) - AP(q_j)\big) = \operatorname{sgn}\big(\mathcal{P}(q_i) - \mathcal{P}(q_j)\big), \\ 0 & \text{otherwise;} \end{cases} \tag{2}$$

and the pairwise accuracy is defined as the arithmetic mean of the samples (in other words, the proportion of pairs that the ordering of the predicted scores agrees with the order of the actual AP scores):

$$\text{Pairwise Accuracy} \coloneqq \frac{2}{N(N-1)} \sum_{(q_i, q_j)} S(q_i, q_j). \tag{3}$$

Note that this treatment of ties is different from Kendall [12] as we allow ties in the ground truth ranking, and expect a good prediction method to have the same result. So QPP evaluation corresponds to measuring concordance between judges, rather than measuring an objective ordering. If both rankings are fully tied, the pairwise accuracy in Eq. 3 is 1. Though different methods treat ties differently, we leave this exploration for future work.

Next, we construct a new set of pairwise combinations with the score from Eq. 2, the result is $\binom{N}{2}$ samples, and separate the set $\{(q_i,q_j) \mid i<j\}$ of pairwise samples into two sets: inter-topic, for samples with queries from different topics; and, intra-topic for pairs from the same topic. The final result is 5,074,518 and 24,703 possible inter-topic and intra-topic pairs in the test collection we describe in the next section. This process is repeated for each predictor.

4 EXPERIMENTS

Experimental Setup. The TREC Robust-2004 (Robust04) [29] Ad Hoc collection was used for our analysis. The Robust04 ad-hoc track is composed of 528,155 documents, mostly news articles, and is associated with 249 TREC topics with at least one relevant document in the relevance judgments file. The set of queries for each TREC topic consists of the original TREC topic title, and additional human-generated query variants [3].[1] After removing duplicate queries for each topic, there are 12.83($\pm$6.85) distinct query variants per topic on average; and a total of 3194 queries.

To comply with common practice in prior work on QPP the retrieval results were produced using a Language Model (LM) based approach, the query-likelihood model [23] with Dirichlet smoothing ($\mu=1000$). Also aligned with prior work on QPP, AP was used as the ground truth effectiveness measure. The parameters of the QPP post-retrieval methods were fixed to values that were previously reported as effective for the Robust04 collection [21, 22].

Inter-Topic vs Intra-Topic Prediction. We now compare inter-topic (QPP) and intra-topic (QVPP) results for each predictor. As expected, we observe that the pairwise accuracy is higher for the inter-topic pairs for all 18 predictors that were tested.

We apply a two-way ANalysis Of VAriance (ANOVA) model to analyze the equality of the pairwise accuracy among the methods.[2] The factors are predictor, task-type and their interaction. The results were significant for all the factors, and specifically the cross factor *predictor:task-type* ($F = 1248.043, p < 0.0001$). We then run Tukey's HSD post hoc test to determine pairwise differences. For each predictor of all the inter-/intra- pairwise comparisons show highly significant differences – corroborating previous results [25].

Fig. 1 shows a comparison between the distributions of AP differences for inter-topic (blue) and intra-topic (orange) query pairs. The eCDF plot clearly demonstrates that the inter-topic distribution of AP differences stochastically dominates the intra-topic distribution. For example, the proportion of inter-topic pairs with an AP difference greater than 0.2 is around 0.4, while the proportion of intra-topic pairs with a similar AP difference is less than 0.2. The KL divergence of the intra-topic to inter-topic distributions is 0.353. We also applied a Kolmogorov–Smirnov (K–S) test to test the null hypothesis that the two samples are from the same distribution. The K-S statistic is relatively easy to interpret, being the supremum of the distances between the two eCDFs; for our data, the test shows statistical significance ($D=0.351, p<0.0001$). This is an important insight into the distribution of score differences in our intra- and inter-topic comparison: there are many more small AP

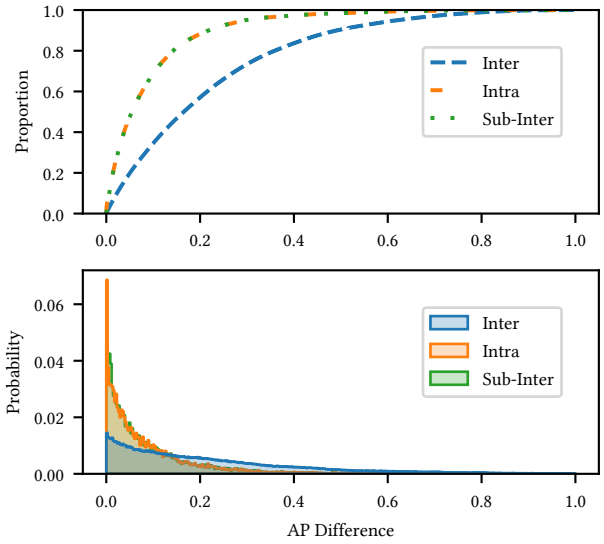

Figure 1: Comparison of pairwise AP differences distributions between all inter-topic (blue), intra-topic (orange) and sub-sampled inter-topic (green) pairs. Upper: empirical Cumulative Distribution Function (eCDF) plot; Lower: Histogram plot.

Table 1: Tukey's Honestly Significant Difference (HSD) post hoc test for pairwise differences. ‡ indicates significant result with significance level 0.01.

Predictor	Intra-	Inter-	Conf. Intvl.	p-adj
SCQ ‡	0.477	0.526	[−0.068,−0.030]	0.001
AvgSCQ ‡	0.500	0.545	[−0.064,−0.026]	0.001
MaxSCQ ‡	0.261	0.562	[−0.320,−0.282]	0.001
SumVAR ‡	0.479	0.546	[−0.086,−0.049]	0.001
AvgVAR ‡	0.533	0.572	[−0.058,−0.021]	0.001
MaxVAR ‡	0.256	0.574	[−0.337,−0.300]	0.001
AvgIDF‡	0.491	0.537	[−0.064,−0.027]	0.001
MaxIDF‡	0.275	0.534	[−0.278,−0.240]	0.001
Clarity ‡	0.557	0.591	[−0.052,−0.015]	0.001
SMV	0.586	0.580	[−0.013, 0.025]	0.900
NQC	0.588	0.584	[−0.015, 0.022]	0.900
WIG	0.581	0.594	[−0.031, 0.007]	0.643
QF	0.565	0.563	[−0.016, 0.022]	0.900
UEF(Clarity)	0.581	0.598	[−0.036, 0.002]	0.054
UEF(SMV)	0.596	0.595	[−0.017, 0.020]	0.900
UEF(NQC)	**0.599**	0.597	[−0.017, 0.020]	0.900
UEF(WIG)	0.598	**0.601**	[−0.021, 0.016]	0.900
UEF(QF)	0.598	0.590	[−0.011, 0.027]	0.900

[1] Variants are publicly available at http://culpepper.io/publications/robust-uqv.txt.gz.
[2] ANOVA was shown to be appropriate with a dichotomous dependent variable [14] for large samples.

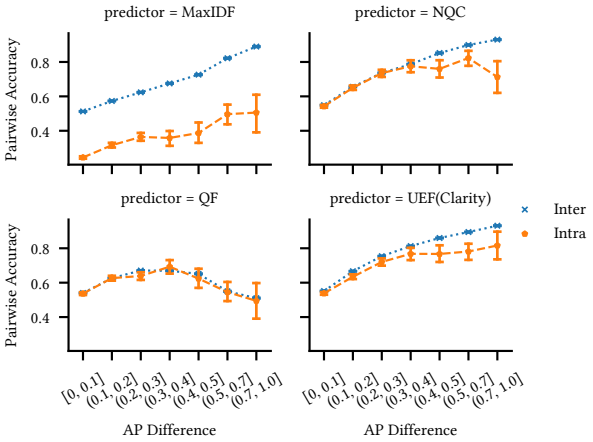

Figure 2: Point plots estimating the pairwise accuracy of different predictors, with 95% bootstrap confidence intervals.

differences observed in the intra-topic set than in an inter-topic set. Our hypothesis is that this difference is why previous studies have found that intra-topic appears to be harder than inter-topic QPP.

In order to control the effect of AP differences of the prediction accuracy, we created a sub-sample of inter-topic query pairs using stratified random sampling. Specifically, we sample pairs based on AP differences from inter-topic query pairs (the larger set) to obtain the same proportion of items in each stratum to match intra-topic query pairs (the smaller set). The results are shown in Fig. 1, where the new distribution of sub-sampled inter-topic pairs (green) is virtually identical to the intra-topic distribution. Correspondingly, this results in a much lower KL divergence of 0.006, and a lower K-S test statistic ($D = 0.024, p < 0.0001$).[3]

Controlling Score Differences. Next, we reran the ANOVA experiment using the sub-sample, and once more the results are significant for all factors, with the cross factor *predictor:task-type* ($F = 597.775, p < 0.0001$). However, the HSD test result is different, and is shown in Table 1. After correcting for the distribution of AP gaps, only half of the predictors show significantly different accuracy scores. In particular, all post-retrieval QPP methods, with the exception of Clarity, did not show significant differences. This suggests that for half of the tested methods, the AP difference is a confounding factor, which was our initial hypothesis. To further explore the effect of AP differences on the ability of the QPP methods to distinguish between a pair of queries, we conducted an analysis controlling for these differences. Query pairs were partitioned into intervals based on their AP difference. Given that the distribution of AP gaps is skewed towards lower values, the two largest intervals are much larger then the rest. To reliably estimate the differences between the two types of prediction, we then computed bootstrap confidence intervals using random sampling. We computed 95% Confidence Interval (CI) using the bootstrap with 1,000 iterations in each interval. The largest interval (0.7,1], which had the least number of query pairs consists of only 87 intra-topic pairs, hence the CIs are wider.

[3]While the p-value remains significant, this is likely due to the large samples sizes [28].

In Fig. 2 we present the details of four representative predictors (others omitted due to space constraints). In this analysis the pre-retrieval QPP MaxIDF, MaxSCQ and MaxVAR methods stood out from the rest, and had substantially larger differences. These methods were examined by Thomas et al. [25], and contributed to the conclusion that there are significant differences between the ability of QPP methods to distinguish between intra-topic and inter-topic queries. The "Max" methods all estimate the effectiveness of a query based only on the single query term with the greatest (measured) value. Since queries that represent the same topic tend to have overlapping terms, particularly topically distinctive terms, this is a plausible explanation for these predictors to exhibit poorer accuracy for the set of intra-topic query pairs.

Another interesting trend that stands out is that all of the QPPs show an increasing accuracy with respect to the size of the AP difference. The only exceptions were the QF and UEF(QF) predictors, which seem to reach their highest accuracy around an AP difference of 0.3. However, these two approaches do exhibit lower accuracy than the other post-retrieval methods.

Broadly speaking, most predictors show an accuracy as low as 50% for query pairs with small AP differences (towards the left of each plot). However, when the AP differences increase, some predictors achieve an accuracy as high as 90%. Furthermore, for some of the methods even a difference of 0.3 is enough to achieve an 80% accuracy, which is a promising result.

It is noteworthy that even for QPP methods with the best performance – the UEF based predictors – there are still significant differences between the intra-topic and inter-topic predictions in the highest score intervals. This implies that additional confounding factors beyond the size of AP differences do exist, and should be taken into account using a more reliable evaluation process.

5 CONCLUSIONS AND FUTURE WORK

This work presents an evaluation framework for QPP which enables a fair comparison of inter- and intra-topic tasks with statistical significance testing. We demonstrate that while significant differences between the intra-topic (QVPP) and inter-topic (QPP) tasks may occur, effectiveness score (AP) difference is usually the dominant factor. Pairwise effectiveness differences can have a significant effect on the prediction quality of QPP methods, particularly for those with high overall accuracy. Our experiments warrant further study of how best to account for the magnitude of performance differences in predictor-query pairs when evaluating new QPP techniques. We also show that the AP difference only explains part of the differences we observed in the two scenarios, as even when this factor is controlled for, some prediction methods still show significant performance differences based on the query type and/or AP score difference. In future work we intend to explore other factors that might affect prediction quality, such as collection or retrieval method properties. To support the reproducibility of our experiments, the code and data are publicly available.[4]

Acknowledgements. We thank the reviewers for their comments. The work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP190101113).

[4]https://github.com/Zendelo/qpp-inter-intra-eval.

REFERENCES

[1] Giambattista Amati, Claudio Carpineto, and Giovanni Romano. 2004. Query Difficulty, Robustness, and Selective Application of Query Expansion. In *Advances in Information Retrieval*. 127–137.

[2] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proc. SIGIR*. 725–728.

[3] Rodger Benham and J. Shane Culpepper. 2017. Risk-reward trade-offs in rank fusion. In *ACM International Conference Proceeding Series*. 1–8.

[4] Rodger Benham, Joel Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting Search Performance Using Query Variations. *ACM Trans. Inf. Syst.* 37, 4 (2019), 41:1–41:25.

[5] David Carmel and Elad Yom-Tov. 2010. Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 2, 1 (2010), 1–89.

[6] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. 2006. What Makes a Query Difficult?. In *Proc. SIGIR*. 390–397.

[7] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo von Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. (6 2020). https://doi.org/10.1145/1122445.1122456

[8] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting Query Performance. In *Proc. SIGIR*. 299–306.

[9] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2006. Precision prediction based on ranked list coherence. *Information Retrieval* 9, 6 (2006), 723–755.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[11] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2021. An Enhanced Evaluation Framework for Query Performance Prediction. In *Proc. ECIR*. 115–129.

[12] M. G. Kendall. 1945. The Treatment of Ties in Ranking Problems. *Biometrika* 33, 3 (1945), 239–251.

[13] Binsheng Liu, Nick Craswell, Xiaolu Lu, Oren Kurland, and J. Shane Culpepper. 2019. A Comparative Analysis of Human and Automatic Query Variants. In *Proc. SIGIR*. 47–50.

[14] Gerald H Lunney. 1970. Using Analysis of Variance with a Dichotomous Dependent Variable: An Empirical Study. *J. Educ. Meas.* 7, 4 (1970), 263–269.

[15] 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. In *Proc. NeurIPS*.

[16] Xi Niu and Diane Kelly. 2014. The use of query suggestions during information search. *Information Processing & Management* 50, 1 (2014), 218–234.

[17] Fiana Raiber and Oren Kurland. 2014. Query-Performance Prediction: Setting the Expectations Straight. In *Proc. SIGIR*. 13–22.

[18] Harrisen Scells, Leif Azzopardi, Guido Zuccon, and Bevan Koopman. 2018. Query Variation Performance Prediction for Systematic Reviews. In *Proc. SIGIR*. 1089–1092.

[19] Falk Scholer and Steven Garcia. 2009. A Case for Improved Evaluation of Query Difficulty Prediction. In *Proc. SIGIR*. 640–641.

[20] Falk Scholer, Hugh E. Williams, and Andrew Turpin. 2004. Query association surrogates for Web search. *J. Assoc. Inf. Sci. Technol.* 55, 7 (2004), 637–650.

[21] Anna Shtok, Oren Kurland, and David Carmel. 2010. Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction. In *Proc. SIGIR*. 259–266.

[22] 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 (2012), 1–35.

[23] Fei Song and W Bruce Croft. 1999. A General Language Model for Information Retrieval. In *Proc. CIKM*. 316–321.

[24] Yongquan Tao and Shengli Wu. 2014. Query Performance Prediction By Considering Score Magnitude and Variance Together. In *Proc. CIKM*. 1891–1894.

[25] Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. 2017. Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. In *Proc. ADCS*.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.

[27] Sebastiano Vigna. 2015. A Weighted Correlation Index for Rankings with Ties. In *Proc. WWW*. 1166–1176.

[28] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272.

[29] Ellen M Voorhees. 2004. Overview of the TREC 2004 Robust Retrieval Track. In *Proc. TREC*.

[30] Oleg Zendel, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. In *Proc. SIGIR*. 395–404.

[31] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In *Proc. ECIR*. 52–64.

[32] Yun Zhou and W Bruce Croft. 2007. Query Performance Prediction in Web Search Environments. In *Proc. SIGIR*. 543–550.