An Enhanced Evaluation Framework for Query Performance Prediction

Guglielmo Faggioli[1](✉), Oleg Zendel[2](✉), J. Shane Culpepper[2](✉), Nicola Ferro[1](✉), and Falk Scholer[2](✉)

[1] University of Padova, Padova, Italy
guglielmo.faggioli@phd.unipd.it, ferro@dei.unipd.it
[2] RMIT University, Melbourne, Australia
{oleg.zendel,shane.culpepper,falk.scholer}@rmit.edu.au

Abstract. Query Performance Prediction (QPP) has been studied extensively in the IR community over the last two decades. A by-product of this research is a methodology to evaluate the effectiveness of QPP techniques. In this paper, we re-examine the existing evaluation methodology commonly used for QPP, and propose a new approach. Our key idea is to model QPP performance as a distribution instead of relying on point estimates. Our work demonstrates important statistical implications, and overcomes key limitations imposed by the currently used correlation-based point-estimate evaluation approaches. We also explore the potential benefits of using multiple query formulations and ANalysis Of VAriance (ANOVA) modeling in order to measure interactions between multiple factors. The resulting statistical analysis combined with a novel evaluation framework demonstrates the merits of modeling QPP performance as distributions, and enables detailed statistical ANOVA models for comparative analyses to be created.

1 Introduction

The Information Retrieval (IR) community has long recognized the importance of applying statistical tests to evaluation results. Although best practices continue to evolve, conference/journal guidelines and discussion papers [20,34] have led the community to appreciate the importance of a more theoretically grounded evaluation, and practitioners in IR have been urged over the years to include sound analyses using statistical tests of significance or confidence intervals in submitted manuscripts. While this has led to higher quality analytical comparisons in many IR-related fields, not all areas have adopted the practice. An example of a common IR problem that might benefit from alternative evaluation techniques is Query Performance Prediction (QPP).

The goal of QPP is to estimate the effectiveness of a retrieval system in response to a query when no relevance judgments are available [8]. The most widely-used method for evaluating QPP approaches is based on the strength of a relationship between per-topic prediction scores, and the actual per-topic system

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12656, pp. 115–129, 2021.
https://doi.org/10.1007/978-3-030-72113-8_8

effectiveness as measured using a standard IR effectiveness metric, usually Average Precision (AP). The association is measured using a correlation coefficient, with different papers reporting the Pearson (linear) correlation, Spearman's rank correlation, or Kendall's τ. A QPP approach that achieves a higher correlation value than another is taken to be the superior approach. This evaluation method compares QPP effectiveness at a very high level, with the performance of a QPP approach over a whole set of topics being summarized just by a correlation coefficient as a *point value*.

In order to statistically validate the results two alternatives are available. First, we can test whether or not the correlation between a predictor and the retrieval results is significantly different from zero [9,11,12,14,16,23,24,27,37,48–50]. However, this validation approach just tells us how reliable the conclusions are for a single QPP method, and does not allow two or more QPP approaches to be directly compared. Second, by relying on repeated randomized topic sampling, we can test whether or not the correlation coefficients for two different QPP methods are significantly different from each other. A statistically appropriate method to test the latter would rely on Fisher's z transformation of sample correlation coefficients. In fact, this approach was previously suggested by Hauff et al. [22] and again more recently by Roitman [32] to more reliably test significant differences in QPP model performance. However, this practice has not been adopted in published QPP work to date. Instead, a Student's t-test for the difference of means of the correlated correlation coefficients is currently the preferred approach [30,46,47]. However, it is important to note that both of these approaches are fundamentally different from the pair-wise significance test used for system retrieval effectiveness, which is now common practice in IR evaluation exercises.

Motivated by these observations, we re-examine how QPP efficacy can be analyzed using a more fine-grained approach – by modeling the performance of QPP techniques as *distributions*. This approach has also previously been applied successfully in system evaluation exercises. A distribution-based model can be constructed as follows. First, an estimate of the performance for each system-topic combination is computed using a traditional performance measure, such as AP. Then, all of the topics for a collection are used to model the performance distribution. Note that this is fundamentally different from a classical QPP evaluation approach. Indeed, even when various sampling techniques (e.g., randomization/bootstrap) are currently used in QPP, this is a re-sampling of topics, and leads to a new (aggregated) *point estimate*, e.g., Kendall's τ, for that sample. The different re-samples are then used to compute an expectation and a confidence interval for the point estimate. In contrast, when randomization/bootstrap techniques are used for the evaluation of retrieval effectiveness [40], it is topics that are re-sampled; for *each* topic a performance score such as AP is computed, and a *distribution* of performance for that sample is obtained. An aggregate of this distribution, e.g., a mean or a confidence interval, is then computed, and finally, the different re-samples are used to compute a further expectation and confidence interval for the aggregate.

In this work, we propose a methodology similar to the latter approach. Our evaluation approach has several appealing properties: it allows formal inferential statistics to be applied, which generalizes the results to the entire population of topics; it allows the behavior of a QPP approach to be more clearly isolated, for example through confidence intervals; and, it enables factor decomposition, which in turn allows us to measure the relative contributions to observed effectiveness systematically. We also incorporate recent work in retrieval effectiveness on query variation and reformulation of each topic [3,4,7,43,47] into our framework, which allows a more fine-grained sampling of retrieval performance, and to estimate interaction between systems, topics and query formulations, which is not possible using only a single point estimate.

Our work focuses on two closely related research questions:

- **RQ1**: How can detailed statistical analysis and testing be applied to QPP evaluation exercises?
- **RQ2**: What factors contribute to improving or reducing the performance of a QPP model?

The overall contribution of this paper is a new evaluation framework for QPP which models the performance of QPP methods as distributions of topics. Beside providing a statistically grounded evaluation procedure, our approach provides practitioners with new tools to carry out comprehensive analyses of QPP models.

2 Related Work

Retrieval performance can vary widely across different systems, even for a single query [8]. This has resulted in a large body of work on QPP, which is divided into two common approaches. *Pre-retrieval predictors* analyze query and corpus statistics prior to retrieval [12,23,24,27,36,48] and *post-retrieval predictors* that also analyze the retrieval results [1,2,9,14,16,31,38,46,49]. Predictors are typically evaluated by measuring the correlation coefficient between the AP values attained with relevance judgments and the values assigned by the predictor. Such evaluation methodologies are based on a *point estimate* and have been shown to be unreliable when comparing multiple systems, corpora and predictors [22,35]. Hauff et al. [22] demonstrate that higher correlation does not necessarily attest to better prediction, and used Root Mean Square Error (RMSE) in their evaluation. Hauff et al. applied methods from Meng et al. [26] to compare 2 or more correlation coefficients, and argued that to test the significance of differences in correlation between the predictors, Fisher's z transformation should be used and the Confidence Interval (CI) should be reported. When computing the CI for Pearson's linear correlation in the evaluation using multiple previously reported pre-retrieval predictors, they found that many of the predictors had overlapping CIs, and concluded that they were not significantly different from the best performing predictor. Hauff et al. focused on prediction of normalized scores that can be compared to AP using linear correlation as measured with a parametric statistic. In this work, we focus on ranking the queries based on the retrieval effectiveness, which is analogous to a rank-based correlation given by Kendall's τ as our reference for the existing evaluation framework, but many other alternatives are possible. We chose to use a rank-based correlation as it is a non-parametric statistical method, and hence makes no assumptions about the underlying distributions of the data.

Also of interest, recent work using query variations for QPP [43,47] has demonstrated that the relative prediction quality of predictors can vary with respect to the effectiveness of the queries used to represent the topics, and we explore such observation further using advanced statistical instrumentation. One principled approach that can be used in IR evaluation is ANOVA [25,33]. ANOVA is commonly used to assess the presence of statistically significant differences in mean performance observed when using different experimental conditions. This technique can be operationalized as a General Linear Mixed Model (GLMM), where a response variable, called $Data$, is linearly modeled into two parts: the experimental conditions (the $Model$) and the $Error$: $Data = Model + Error$. The $Error$ represents that part of the variance in the $Data$ that the $Model$ cannot account for. The ANOVA approach is particularly useful in our work as it allows us to break down the variance observed in the data, assigning it to the factors that caused it [5,10,17,19,29,41,45]. The $Model$ often includes a subject component (which in IR evaluation often corresponds to the topic), one or more factors, which are the different experimental conditions (either the entire system, or its components - e.g., the stemmer, the stoplist and the QPP model), and possibly their interactions. If all the possible combinations of factors are applied to all subjects, this is a *Factorial/Crossed Design*, and its factors are called *Crossed Factors*. Specific factors might be *nested* inside others: in the following analyses, query formulations are a nested factor of the topic, since each formulation represents a single topic and cannot be used to represent others. To compare the *effect size* of different factors, which cannot be done by looking only at the F-statistic or p-value, the Strength of Association (SOA) is reported, measured as ω^2, and is the factor significance, bounded between [0, 1]. The larger ω^2 is, the greater the impact is for factor levels to the response variable.

3 Experimental Analysis

3.1 Experimental Setup

In our analyses, we use the TREC Robust 2004 (Robust04) Ad Hoc [44] collection. The Robust04 ad hoc track consists of approximately $528K$ documents from TREC disks 4 & 5, minus the Congressional Record from the TIPSTER corpus, and contains 249 topics with at least one relevant document in the QREL file. We enrich the set of queries for the corpus using publicly available human-curated query variants for each topic [6].[1] Our experiments use a Grid of Points (GoP) of runs as described by Ferro and Harman [18], using 4 different stoplists

[1] http://culpepper.io/publications/robust-uqv.txt.gz.

Table 1. A summary of QPP models used in this work.

QPP model	**Description**
	Pre-retrieval
SCQ [48]	Measures similarity based on *cf.idf* to the corpus, summed over the query terms
AvgSCQ [48]	SCQ normalized by the query length
MaxSCQ [48]	The query term with maximal SCQ score
SumVAR [48]	Measures the *cf.idf* variability of the query terms in the corpus
AvgVAR [48]	Variability normalized with the query length
MaxVAR [48]	The query term with maximal variability
AvgIDF [13]	The mean *idf* value of the query terms
MaxIDF [36]	The query term with maximal *idf* value
	Post-retrieval
Clarity [12]	Measures the divergence between the Language Model (LM) constructed over top documents in the result list to the LM of the entire corpus
NQC [39]	Measures the standard deviation of the top documents scores in the retrieval list
WIG [50]	Measures the difference between the mean retrieval score of the top retrieved documents and the score of the entire corpus
SMV [42]	Scores the queries based on a combination of the scores standard deviation and magnitude
UEF [37]	Prediction framework that is based on the similarity of the initial result list with the list re-ranked using a Relevance Model (RM), scaled by an estimator of the RM quality. In this work we scale the RM with the existing post-retrieval predictors: UEF(Clarity), UEF(NQC), UEF(WIG) and UEF(SMV)

(`atire`, `zettair`, `indri`, `lingpipe`), plus the `no stop` approach and 2 different stemmers, (`lovins`, `porter`) plus a nostem approach. All the runs were produced using the query-likelihood model [28], and repeated 15 times. We test 16 QPP models (12 + 4 UEF-based methods) for our analyses, which are summarized in Table 1. Our goal was to choose representative and well known system configurations and QPP models, and the evaluation framework is not limited to any specific configuration. So it can easily be extended by others for further experiments in the future. In total, 240 different predictor-system combinations were generated for the Robust04 collection. The pre-retrieval approaches are parameter-free and do not require tuning. For the parameters of the post-retrieval predictors we used fixed settings that have been demonstrated to be effective for the Robust04 collection previously [37,39,42]. We apply Average Precision (AP) to measure the effectiveness of the different retrieval pipelines, as our primary goal is to be consistent with previous evaluation exercises, as Average Precision (AP) was the most common effectiveness metric used in prior QPP work.

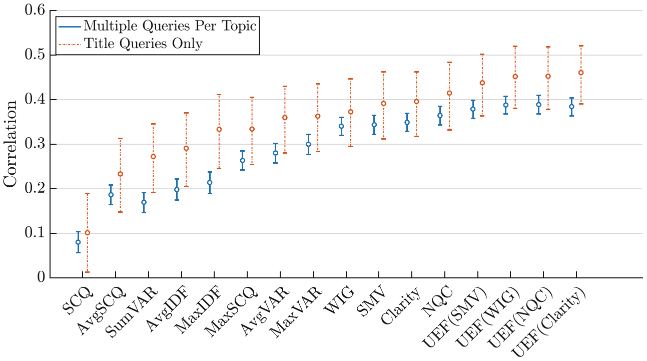

Fig. 1. Prediction quality of the selected QPP models on Robust04 (Confidence Intervals computed with Kendall's τ), using either title queries or all available formulations. (Color figure online)

3.2 Traditional QPP Evaluation Using Correlations

Prior work on QPP has relied primarily on a single evaluation paradigm. Given a set of topics (information needs), where each topic is represented by a single query, a single retrieval method, and a single document corpus, the prediction quality of the predictors is evaluated as follows:

1. Retrieval effectiveness of the queries is measured with a common IR metric, usually AP or possibly Normalized Discounted Cumulated Gain (nDCG) to induce a ranking of the queries. This ordering serves as the ground truth in the evaluation process.
2. The QPP method is applied to the queries, which generates a candidate list where the queries are ranked by their prediction values.
3. A correlation coefficient is computed between the ground truth list and the candidate list produced by the predictor.
4. The correlation coefficients of different predictors are then compared, with an underlying assumption that a higher correlation value attests to the superior quality of a predictor.

The correlation coefficient is often measured as Pearson's r for linear correlation, Kendall's τ, and/or Spearman's ρ for the monotonic rank correlation.

Figure 1 shows the performance of 16 different QPP models when using this common evaluation approach – Kendall's τ correlation in this case – with 95% confidence intervals shown as well. In this example, the results are generated for a specific retrieval pipeline, using the `indri` stoplist and `porter` stemmer. To compute the confidence intervals (at significance level $\alpha = 0.05$), we used a bias-corrected and accelerated bootstrap procedure with 10,000 samples. Observe that when using title queries only (orange bars), there is a large degree of overlap between the different QPP approaches. Similar results were observed when using all of the other pipelines described in this work. The pairwise comparison

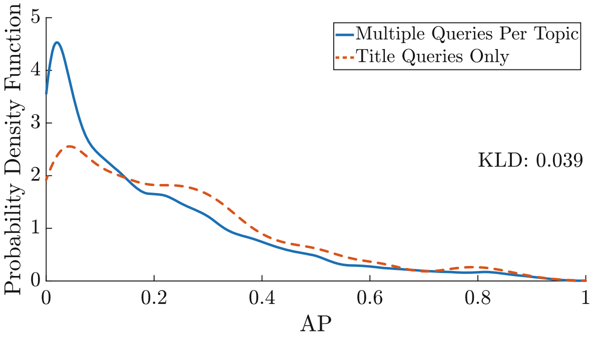

Fig. 2. Comparison between the AP score distributions of title-only queries and multi-query topic formulations. (Color figure online)

using the data from Fig. 1 (title queries only, p-values omitted due to space constraints), shows that 57 pairs of predictors are found to be statistically significantly different, out of 120 total pairs of QPP models (47.5%). In particular, among the best performing predictors, UEF(Clarity) is not statistically different from UEF(WIG), UEF(NQC), UEF(SMV), Clarity and NQC. This suggests that using confidence intervals does indeed make it difficult to decide which QPP system is the best performing one, as suggested by Hauff et al. [22].

In addition to using the traditional title queries, we also explore the scenario of using multiple formulations, which allows us to produce replicas for the same experimental conditions (i.e., the retrieval system or the QPP model used) on the same subject (i.e., the topic). While the performance is generally lower when using multiple topic formulations (the blue bars shown in Fig. 1), there is a high degree of similarity between the ordering of the QPP models for multiple query formulations to the ordering for title-only (Kendall's tau correlation between using title-only versus multiple queries per topic is 0.98, $p < 0.0001$). Overall, the statistically induced bootstrap intervals are substantially larger if a traditional title-only evaluation approach is used, which makes it less suitable for determining if any single system is a clear winner, while using multiple queries does induce smaller intervals and better discriminative power between the QPP approaches. Even if, as shown, using query variants does not dramatically impact the ranking of QPP models, it is nevertheless important to consider whether adding variants has an impact on the distribution of the raw AP scores. The Mean Average Precision (MAP) values are 0.211 and 0.254 for the set of all query formulations and title queries only, respectively, and thus are quite consistent. Figure 2 shows the Probability Density Function (PDF) for the AP scores for the two scenarios – title-only (red line) and multiple queries per topic (blue line). The Kullback-Leibler Divergence (KLD), a measure of the similarity between the two distributions, is 0.039. In summary, the distributions are similar and thus the introduction of the multiple formulations for each topic does not appear to skew the overall AP score distribution.

3.3 ANOVA Modeling and Analysis of QPP

To support a more detailed analysis of QPP methods and associated factors, we now explore the use of ANOVA, which can be achieved by modifying steps 3 and 4 of the traditional QPP evaluation process shown above. Instead of computing the correlations between the complete lists, we measure the difference, for each query, in the rank position assigned by a QPP method and the ground truth rank position assigned by AP. Ties in ranks are broken using the average of tie rank spans, as is the default in many statistical applications [21]. Other tie breaking rules were also considered but initial investigation led to consistent final results, so are not reported here. Observe that this transitions us from *point estimates* of a single correlation value for the two lists over a whole set of topics to a *distribution* of the rank differences between the two lists for each query in the set. In order to scale the scores to the range $[0, 1]$ we divide them by the number of samples. The error, labeled as AP induced scaled Absolute Rank Error (sARE$_{AP}$), for each query is:

$$\text{sARE}_{AP}(q_i) := \frac{|r_i^p - r_i^e|}{|Q|}, \tag{1}$$

where r_i^p and r_i^e are the ranks assigned by the predictor and the evaluation metric respectively for query i; Q is the set of queries. If we still require the single point estimate of the prediction quality for each predictor $\mathcal{P}$, we can calculate the AP induced scaled Mean Absolute Rank Error (sMARE$_{AP}$) as follows:

$$\text{sMARE}_{AP}(\mathcal{P}) := \frac{1}{|Q|} \sum_{q_i \in Q} \text{sARE}_{AP}(q_i). \tag{2}$$

Note that sMARE$_{AP}$ can be seen as a derivation of *Spearman's Footrule distance*, making it a metric for the full rankings instead of a correlation. Among the properties of Spearman's Footrule distance, Diaconis and Graham [15] list that it is bounded between $[0, \lfloor 0.5n^2 \rfloor]$, where n is the length of the ranking. Since both sARE$_{AP}$ and sMARE$_{AP}$ are normalized by the number of queries, sMARE$_{AP}$ is bounded between [0, 0.5]. To demonstrate the agreement between the proposed evaluation method with existing evaluation practices from a high-level (point estimate) perspective, we use the QPP methods over the Robust04 title queries. Figure 3 plots the ranking of the predictors based on the median of the point estimates for each predictor for all 15 system configurations which is simply the median of the Kendall's τ correlation for the traditional evaluation approach and the median of sMARE$_{AP}$ for our evaluation approach. Each predictor consists of 15 values that represent the prediction quality. Though the directionality of the two approaches is inverted, the ranking of the predictors clearly agrees on the overall rank ordering. The corresponding box-plots also demonstrate the similarity of the variance estimate. In order to validate the agreement we computed the Pearson's correlation coefficient over the point estimates for the predictors for each of the 15 system configurations. The resulting correlations coefficients were all -0.99 or higher ($p < 0.0001$ for each).

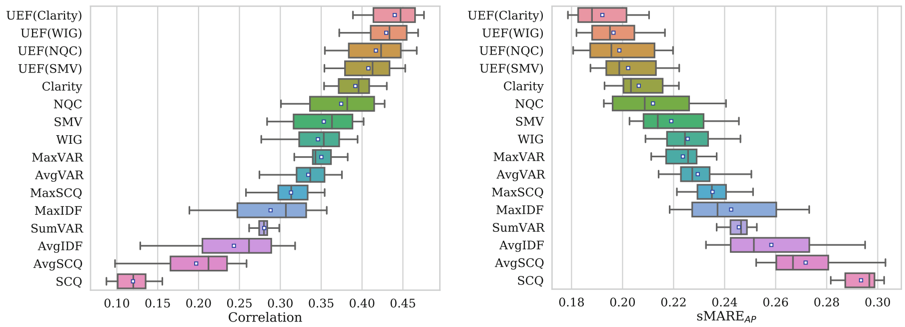

Fig. 3. Prediction quality when measuring correlation with Kendall's τ and sMARE$_{AP}$ for Robust04 title-only queries and 15 different system configurations. The line inside the interquartile range (IQR) is the median, and the white square is the mean.

Table 2. MD0$_{micro}$ ANOVA on the Robust04 collection. Topics are represented with the title queries. SS: Sum of Squares; DF: Degrees of Freedom; MS: Mean Square; F: F statistics.

Source	**SS**	**DF**	**MS**	**F**	**p-value**	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	876.524	248	3.534	168.136	<0.001	0.410
Stoplist	1.185	4	0.296	14.095	<0.001	0.001
Stemmer	5.218	2	2.609	124.108	<0.001	0.004
QPP model	46.569	15	3.105	147.691	<0.001	0.036
Error	1250.538	59490	0.021			
Total	2180.034	59759				

We are in a position to introduce our first ANOVA model which will enable a more comprehensive experimental analysis of the results.

$$y_{iqrs} = \mu + \tau_i + \gamma_q + \delta_r + \zeta_s + \varepsilon_{iqrs} \qquad (\text{MD0}_{micro})$$

where: $y_{i...}$ is the performance (sARE$_{AP}$) on the i-th topic (using the specified QPP pipeline); μ is the *grand mean*; τ_i is the effect of the i-th topic (represented with the title query formulation); γ_q, δ_r, and ζ_s are the effect of the q-th stoplist, the r-th stemmer, and the s-th QPP model; ε_{iqrs} is the error component. Table 2 summarizes the ANOVA results of our first experiment. It can be seen that the stoplist, the stemmer, and the QPP model have a small size effect, while the topic effect is large (indicating that most of the performance of the QPP depends on the chosen topic). Based on the results of this analysis, we also ran a Tukey's Honestly Significant Difference (HSD) post-hoc analysis to test for pairwise differences. Figure 4 shows the Tukey's HSD confidence intervals for sMARE$_{AP}$ over the different QPP models.

When comparing Fig. 1 (orange bars) and Fig. 4, we can observe that there is less overlap between the CIs, in particular, we observe that, by computing the

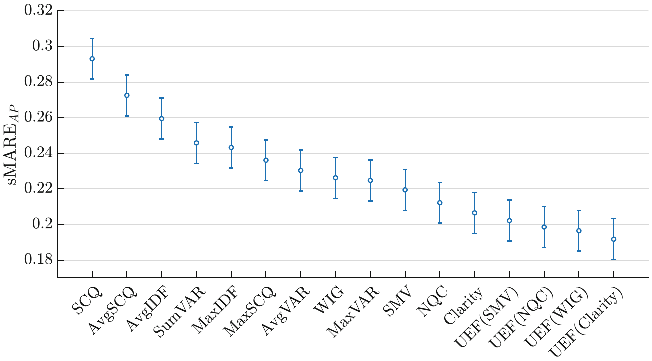

Fig. 4. Confidence Intervals of sMARE$_{AP}$ from MD0$_{micro}$ on the Robust04 title queries.

p-values for the pairwise comparisons, out of 120 pairs of predictors, 96 of them are significantly different (80.0%). Thus, compared to the results observed for the bootstrap-based approach, we are able to differentiate between 68.4% more pairs of predictors. In this case, the top performing cluster includes UEF(WIG), UEF(SMV), UEF(NQC), and UEF(Clarity).

The "Topic" factor, as Table 2 suggests, is responsible for the largest part of the variance; this is in line with results from IR effectiveness evaluation (see for example Tague-Sutcliffe and Blustein [41]). Thus, the estimation of the performance for a specific QPP model can vary significantly as it is dependent on properties of the underlying collection (performance differences in topics/queries). By removing the contribution of the topics from the global variance, ANOVA removes any volatility in the underlying experimental data allowing the relative performance of predictors to be compared more precisely. When using only correlations aggregated across all topics, such information is lost, while an ANOVA analysis facilitates more discriminative performance comparisons between systems by systematically accounting for each factor separately.

3.4 ANOVA Modeling of Multiple Queries and Interactions

One of the most interesting aspects of our framework is the capability to compute the effect sizes of interactions between factors. This is achieved using MD1$_{micro}$

$$\begin{aligned} y_{ijqrs} = \mu + \tau_i + \nu_{j(i)} + \gamma_q + \delta_r + \zeta_s + (\tau\gamma)_{iq} + (\tau\delta)_{ir} + (\tau\zeta)_{is} \\ + (\nu\gamma)_{j(i)q} + (\nu\delta)_{j(i)r} + (\nu\zeta)_{j(i)s} + (\gamma\delta)_{qr} + (\gamma\zeta)_{qs} + (\delta\zeta)_{rs} + \varepsilon_{ijqrs} \end{aligned} \tag{MD1$_{micro}$}$$

which extends MD0$_{micro}$ to include $\nu_{j(i)}$ to represent the effect of the j-th query formulation for the i-th topic. Moreover, this model considers all of the possible two-way interactions which are now computable using the replicates provided by the multi-query topic formulations.

Table 3. MD1$_{micro}$ ANOVA applied on ROBUST04 collection. ω^2 for non-significant factors is ill-defined and thus not reported.

Source	**SS**	**DF**	**MS**	**F**	**p-value**	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	1840.082	248	7.420	1293.936	<0.001	0.518
Formulation (Topic)	1746.213	996	1.753	305.749	<0.001	0.504
Stoplist	1.179	4	0.295	51.402	<0.001	0.001
Stemmer	10.622	2	5.311	926.188	<0.001	0.006
QPP model	305.796	15	20.386	3555.233	<0.001	0.151
Topic*Stoplist	40.224	992	0.041	7.071	<0.001	0.020
Topic*Stemmer	154.200	496	0.311	54.216	<0.001	0.081
Topic*QPP model	2051.688	3720	0.552	96.182	<0.001	0.542
Frm.*Stoplist	87.110	3984	0.022	3.813	<0.001	0.036
Frm.*Stemmer	312.955	1992	0.157	27.398	<0.001	0.150
Frm.*QPP model	3348.894	14940	0.224	39.091	<0.001	0.656
Stoplist*Stemmer	0.059	8	0.007	1.288	0.2444	–
Stoplist*QPP model	0.901	60	0.015	2.618	<0.001	<0.001
Stemmer*QPP model	4.850	30	0.162	28.195	<0.001	0.003
Error	1555.757	271312	0.006			
Total	11460.530	298799				

Table 3 presents the ANOVA summary statistics for Ex. MD1$_{micro}$. In this analysis we add the query formulations as a nested factor for each topic, in this case we randomly chose 5 for each topic.[2] The table empirically shows that the largest differences in QPP performance are due to the topics, and their formulations. While this is a well-known phenomenon, our model is able to explicitly quantify the magnitude of this effect. The effect for the QPP factor is medium-sized. It is important to note that the dimension of the effect is due to the wide variety of QPP models (and their performance) taken into account. For example, a practitioner wishing to evaluate new QPP models may observe a smaller ω^2 for the QPP model factor if the relative performance differences between the models being compared is less substantial.

We have also ran similar experiments using alternative models with fewer factors, but found that including all of the possible interactions is the most informative. For example, the effect size of stoplists and stemmers are both small but still significant. This suggests that stemmers and stoplists may affect overall prediction quality, and practitioners should consider all possible factors when comparing and contrasting QPP performance for a corpus.

We are now in a position to observe the interaction between topics (and their query formulations) and the predictors, which is large, indicating that important differences between QPP model performance exists within reformulations of a

[2] The topic with the minimal number of query formulations had 5 formulations.

single topic. Finding the QPP model where interactions are smallest is valuable in practice as this corresponds to be choosing a model that is most robust to query reformulation. Additionally, this enables a series of additional analyses, such as a failure analysis for topics with the largest interaction with a QPP model. There are many additional factors that could influence the performance of various QPP approaches, beyond the ones included in our model. For example, alternative ranking functions or evaluation metrics can also be used with sMARE, and may provide additional experimental evidence and insights into performance differences between various QPP models in the future.

4 Conclusion

We have presented a novel evaluation framework for QPP. The framework estimates the performance of QPP on every topic as the distance between its predicted rank - computed using the QPP – and the expected one – measured through AP (or any other traditional IR measure). This allows us to obtain a distribution of performance for the QPP over the different topics. Furthermore, our framework makes use of multiple query formulations for each topic to enhance the power of our analyses. Together, the use of multiple query formulations and the distributional representation of the performance enables carrying out more accurate studies. In particular, we showed that it is possible to rely on the statistical properties of ANOVA and corresponding post hoc procedures to better identify pairs of QPP approaches that are statistically significantly different. The newly proposed framework also enables the analysis of interaction effects for QPP models and topics, allowing failure analyses and a deeper understanding into how a QPP model works. Our framework can be extended and adapted to different investigation needs. For example, in an academic setting, you may add further factors to the model such as tokenizers, query expansion components, or ranking functions to deepen the investigation into the factors that influence QPP performance. In industrial deployment settings, comparisons between competing QPP techniques may require an ANOVA model consisting of only two factors: topics and QPP approaches. This simple two-way ANOVA is sufficient to determine if QPP models are significantly different, and has the added benefit of relying on a statistically-sound and easy to deploy framework. In future work, we plan to study additional components of the evaluation framework, such as the impact of the ranking methods which are used to establish "ground truth" performance; new factors that influence QPP systems such as the ranking approach used in the post-retrieval QPP; and the effects of using multiple corpora, in order to more comprehensively model and understand corpus and QPP interactions. In order to aid reproducibility of our results, the code for our proposed evaluation framework is publicly available.[3]

[3] https://github.com/Zendelo/QPP-EnhancedEval.

Acknowledgments. We thank the reviewers for their comments. The work is partially funded by the DAta BenchmarK for Keyword-based Access and Retrieval (DAKKAR) Starting Grants project sponsored by University of Padua and Fondazione Cassa di Risparmio di Padova e di Rovigo. This work was also partially supported by the Australian Research Council's Discovery Projects Scheme (DP190101113).

References

1. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: McDonald, S., Tait, J. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24752-4_10
2. Aslam, J.A., Pavlu, V.: Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 198–209. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71496-5_20
3. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: UQV100: a test collection with query variability. In: Proceedings SIGIR, pp. 725–728 (2016)
4. Bailey, P., Moffat, A., Scholer, F., Thomas, P.: Retrieval consistency in the presence of query variations. In: Proceedings of the SIGIR, pp. 395–404 (2017)
5. Banks, D., Over, P., Zhang, N.F.: Blind men and elephants: six approaches to TREC data. Inf. Retrieval **1**(1–2), 7–34 (1999)
6. Benham, R., Culpepper, J.S.: Risk-reward trade-offs in rank fusion. In: Proceedings ADCS, pp. 1:1–1:8 (2017)
7. Benham, R., Mackenzie, J., Moffat, A., Culpepper, J.S.: Boosting search performance using query variations. ACM Trans. Inf. Syst. **37**(4), 41:1–41:25 (2019)
8. Carmel, D., Yom-Tov, E.: Estimating the Query Difficulty for Information Retrieval. Morgan & Claypool Publishers, San Rafael (2010)
9. Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: Proceedings of the SIGIR, pp. 390–397 (2006)
10. Carterette, B.A.: Multiple testing in statistical analysis of systems-based information retrieval experiments. ACM Trans. Inf. Syst. **30**(1), 4:1–4:34 (2012)
11. Chifu, A.G., Laporte, L.é., Mothe, J., Ullah, M.Z.: Query performance prediction focused on summarized letor features. In: Proceedings of the SIGIR, pp. 1177–1180 (2018)
12. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the SIGIR, pp. 299–306 (2002)
13. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A language modeling framework for selective query expansion. Technical report, Center for Intelligent Information Retrieval, University of Massachusetts (2004)
14. Cummins, R.: Document score distribution models for query performance inference and prediction. ACM Trans. Inf. Syst. **32**(1), 2:1–2:28 (2014)
15. Diaconis, P., Graham, R.L.: Spearman's footrule as a measure of disarray. J. R. Stat. Soc. **39**(2), 262–268 (1977)
16. Diaz, F.: Performance prediction using spatial autocorrelation. In: Proceedings of the SIGIR, pp. 583–590 (2007)
17. Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Soboroff, I.: CENTRE@CLEF 2019. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11438, pp. 283–290. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_38

18. Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF pilot track overview. In: Peters, C., et al. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 552–565. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15754-7_68
19. Ferro, N., Silvello, G.: A general linear mixed models approach to study system component effects. In: Proceedings of the SIGIR, pp. 25–34 (2016)
20. Fuhr, N.: Some common mistakes in IR evaluation, and how they can be avoided. SIGIR Forum **51**(3), 32–41 (2017)
21. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference, 5th edn. Chapman & Hall/CRC, Taylor and Francis Group, Boca Raton (2011)
22. Hauff, C., Azzopardi, L., Hiemstra, D.: The combination and evaluation of query performance prediction methods. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 301–312. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_28
23. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proceedings of the CIKM, pp. 1419–1420 (2008)
24. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 43–54. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30213-1_5
25. Maxwell, S., Delaney, H.D.: Designing Experiments and Analyzing Data. A Model Comparison Perspective, 2nd edn. Lawrence Erlbaum Associates, Mahwah (2004)
26. Meng, X.L., Rosenthal, R., Rubin, D.B.: Comparing correlated correlation coefficients. Psychol. Bull. **111**(1), 172–175 (1992)
27. Mothe, J., Tanguy, L.: Linguistic features to predict query difficulty. In: Proceedings of the SIGIR, pp. 7–10 (2005)
28. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the SIGIR, pp. 275–281 (1998)
29. Robertson, S.E., Kanoulas, E.: On per-topic variance in IR evaluation. In: Proceedings of the SIGIR, pp. 891–900 (2012)
30. Roitman, H.: An extended query performance prediction framework utilizing passage-level information. In: Proceedings of the SIGIR, pp. 35–42 (2018)
31. Roitman, H.: Query performance prediction using passage information. In: Proceedings of the SIGIR, pp. 893–896 (2018)
32. Roitman, H.: ICTIR tutorial: modern query performance prediction: theory and practice. In: Proceedings of the SIGIR, pp. 195–196 (2020)
33. Rutherford, A.: ANOVA and ANCOVA. A GLM Approach, 2nd edn. Wiley, New York (2011)
34. Sakai, T.: Topic set size design. Inf. Retrieval J. **19**(3), 256–283 (2016)
35. Scholer, F., Garcia, S.: A case for improved evaluation of query difficulty prediction. In: Proceedings of the SIGIR, pp. 640–641 (2009)
36. Scholer, F., Williams, H.E., Turpin, A.: Query association surrogates for web search. J. Assoc. Inf. Sci. Technol. **55**(7), 637–650 (2004)
37. Shtok, A., Kurland, O., Carmel, D.: Using statistical decision theory and relevance models for query-performance prediction. In: Proceedings of the SIGIR, pp. 259–266 (2010)
38. Shtok, A., Kurland, O., Carmel, D.: Query performance prediction using reference lists. ACM Trans. Inf. Syst. **34**(4), 19:1–19:34 (2016)
39. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting query performance by query-drift estimation. ACM Trans. Inf. Syst. **30**(2), 1–35 (2012)
40. Smucker, M.D., Allan, J., Carterette, B.A.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the CIKM, pp. 623–632 (2007)

41. Tague-Sutcliffe, J.M., Blustein, J.: A statistical analysis of the TREC-3 data. In: Proceedings of the TREC, pp. 385–398 (1994)
42. Tao, Y., Wu, S.: Query performance prediction by considering score magnitude and variance together. In: Proceedings of the CIKM, pp. 1891–1894 (2014)
43. Thomas, P., Scholer, F., Bailey, P., Moffat, A.: Tasks, queries, and rankers in pre-retrieval performance prediction. In: Proceedings of the ADCS (2017)
44. Voorhees, E.M.: Overview of the TREC 2004 robust track. In: Proceedings of the TREC (2004)
45. Voorhees, E.M., Samarov, D., Soboroff, I.: Using replicates in information retrieval evaluation. ACM Trans. Inf. Syst. **36**(2), 12:1–12:21 (2017)
46. Zamani, H., Croft, W.B., Culpepper, J.S.: Neural query performance prediction using weak supervision from multiple signals. In: Proceedings of the SIGIR, pp. 105–114 (2018)
47. Zendel, O., Shtok, A., Raiber, F., Kurland, O., Culpepper, J.S.: Information needs, queries, and query performance prediction. In: Proceedings of the SIGIR, pp. 395–404 (2019)
48. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 52–64. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78646-7_8
49. Zhou, Y., Croft, W.B.: Ranking robustness: a novel framework to predict query performance. In: Proceedings of the CIKM, pp. 567–574 (2006)
50. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of the SIGIR, pp. 543–550 (2007)