sMARE: An Enhanced Query Performance Prediction Evaluation Approach

(Discussion Paper)

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

[1]*University of Padova, Padova, Italy*
[2]*RMIT University, Melbourne, Australia*

Abstract

QPP has been studied extensively in the IR community over the last two decades. Nevertheless, the Query Performance Prediction (QPP) field still lacks sound theoretical evaluation methodologies. In this work*, we re-examined 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. This, in turns, enables the use of ANalysis Of VAriance (ANOVA) models for comparative analyses, permitting deeper analyses on the QPP models performance, and allowing to measure interactions between multiple factors.

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 and journal guidelines and discussion papers [2] have led the community to appreciate the importance of a more theoretically grounded evaluation. 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 [3]. 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 effectiveness as measured using a standard IR effectiveness metric, usually Average Precision (AP). Such association is measured using a correlation coefficient: 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, by relying on repeated randomized topic sampling, we can test whether or

*This is an extended abstract of [1], best full paper award.

SEBD 2021: The 29th Italian Symposium on Advanced Database Systems, September 5-9, 2021, Pizzo Calabro (VV), Italy

not the correlation coefficients for two different QPP methods are significantly different from each other. However, it is important to note that this approach is 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.

In this work, we propose an evaluation approach (dubbed scaled Mean Absolute Rank Error (sMARE)) which 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 [4] 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 related research questions: (**RQ1**) How can statistical analysis and testing be applied to QPP evaluation exercises? (**RQ2**) What factors contribute to improving or reducing the performance of a QPP model?

2. Related Work

Retrieval performance can vary widely across different systems, even for a single query [3]. *Pre-retrieval predictors* analyze query and corpus statistics prior to retrieval [5] while *post-retrieval predictors* also analyze the retrieval results [6]. 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 methodology is based on a *point estimate* and have been shown to be unreliable when comparing multiple systems, corpora and predictors [7]. Hauff et al. [7] demonstrate that higher correlation does not necessarily attest to better prediction, and used Root Mean Square Error (RMSE) in their evaluation. When computing the Confidence Interval (CI) for Pearson's linear correlation in the evaluation using multiple previously reported pre-retrieval predictors, Hauff et al. [7] found that many of the predictors had overlapping CIs, and concluded that they were not significantly different from the best performing predictor. Also of interest, recent work using query variations for QPP [4] 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[8, 9, 10]. 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 $Model$ includes a subject component (which in IR evaluation often corresponds to the topic), one or more factors, which are the different experimental conditions (e.g, the entire system, or its components), and possibly their interactions. 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. 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 [9].

3. Experimental Analysis

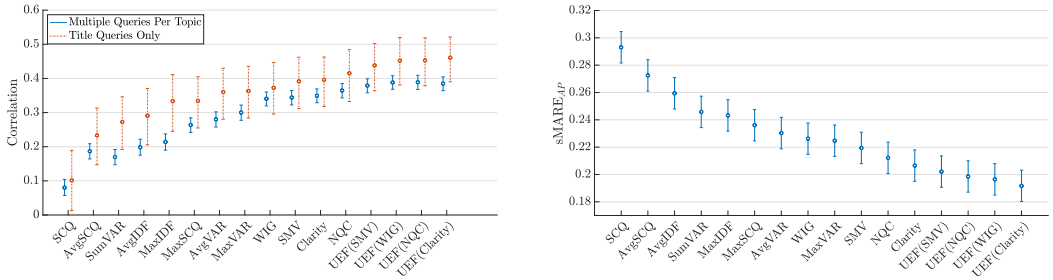
In our analyses, we use the TREC Robust 2004 (Robust04) Ad Hoc [11] collection. We enrich the set of queries for the corpus using publicly available human-curated query variants for each topic [12].[1] Our experiments use a Grid of Points (GoP) of runs, using 4 different stoplists (`atire`, `zettair`, `indri`, `lingpipe`), plus the `no stop` approach and 2 different stemmers, (`lovins`, `porter`) plus a `no stem` approach. All the runs were produced using the query-likelihood model. We experimented with the following pre-retrieval qpp models: SCQ, AvgSCQ, MaxSCQ, SumVAR, AvgVAR, MaxVAR [13], AvgIDF [14], MaxIDF [15]. Concerning post-retrieval qpp models, we considered the followings: Clarity [5], NQC [16], WIG [17], SMV [18]. We duplicate the number of post-retrieval methods, considering, for each of them, their UEF [19] counterpart. Such QPP models have been selected since they are the most well-known and representative. In total, 240 different predictor-system combinations were generated for the Robust04 collection. For predictors which required hyper-parameters, we considered those previously observed to be effective [16]. We apply Average Precision (AP) to measure the effectiveness of the different retrieval pipelines.

3.1. Traditional QPP evaluation using correlations

Prior work on QPP has relied primarily on a single evaluation paradigm: the QPP generates a candidate list where the queries are ranked by their prediction values. The correlation between such list and a list induced using a reference measure, such as AP, is then considered as measure of quality of the QPP model.

Figure 1a 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. The pairwise comparison using the data from Figure 1a (title queries only), shows that 57 pairs of predictors are found to be significantly

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



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

(b) Confidence Intervals of AP induced scaled Mean Absolute Rank Error (sMARE$_{AP}$) from MD0$_{micro}$ on the Robust04 title queries.

Figure 1: Quality and Confidence Intervals of the selected QPP models using the traditional approach (left) and the sMARE$_{AP}$ measure (right).

different, out of 120 total pairs of QPP models (47.5%). This suggests that using confidence intervals does indeed make it difficult to decide which QPP system is the best performing, as suggested by [7].

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 Figure 1a), 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 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.

3.2. 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. 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. 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:

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

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. For each predictor $\mathcal{P}$, we can calculate the sMARE$_{AP}$ as follows:

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

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 [20] 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 $[0, 0.5]$. The proposed metric presents a Kendall's τ correlation coefficient with the original metric higher than 0.99 ($p < 0.0001$) for all configurations.

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

$$y_{iqrs} = \mu + \tau_i + \gamma_q + \delta_r + \zeta_s + \varepsilon_{iqrs} \tag{MD0$_{micro}$}$$

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. Such ANOVA shows that all factors are significants. Furthermore, we observe a large size ω^2 effect for the topic ($\omega^2_{<topic>} = 0.410$). Both the stoplist and the stemmer factors are significant, but with a negligible effect ($\omega^2_{<stoplist>} = 0.001$ and $\omega^2_{<stemmer>} = 0.004$), while the QPP model displays a small effect ($\omega^2_{<qpp\ model>} = 0.036$). 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 1b shows the Tukey's HSD confidence intervals for sMARE$_{AP}$ over the different QPP models.

When comparing Figure 1a (orange bars) and Figure 1b, we can observe that there is less overlap between the CIs, in particular, we observe that, by computing the 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. The "Topic" factor 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 [21]). 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.

Table 1
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				

3.3. 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}

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

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 1 presents the ANOVA summary statistics for 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. The effect for the QPP factor is medium-sized. The significance of the stoplist and stemmer factors suggests that they affect the 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 single topic. Finding the QPP model where interactions are smallest is valuable in practice as this corresponds

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

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.

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. The two-way ANOVA described in MD0_{micro} is sufficient to determine if QPP models are significantly different, and has the added benefit of relying on a statistically-sound 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.

References

[1] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An Enhanced Evaluation Framework for Query Performance Prediction, in: Proc. ECIR, 2021, pp. 115–129.

[2] N. Fuhr, Some Common Mistakes In IR Evaluation, And How They Can Be Avoided, SIGIR Forum 51 (2017) 32–41.

[3] D. Carmel, E. Yom-Tov, Estimating the Query Difficulty for Information Retrieval, Morgan & Claypool Publishers, USA, 2010.

[4] O. Zendel, A. Shtok, F. Raiber, O. Kurland, J. S. Culpepper, Information Needs, Queries, and Query Performance Prediction, in: Proc. SIGIR, 2019, p. 395–404.

[5] S. Cronen-Townsend, Y. Zhou, W. B. Croft, Predicting Query Performance, in: Proc. SIGIR, 2002, pp. 299–306.

[6] J. A. Aslam, V. Pavlu, Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions, in: Proc. ECIR, 2007, pp. 198–209.

[7] C. Hauff, L. Azzopardi, D. Hiemstra, The Combination and Evaluation of Query Performance Prediction Methods, in: Proc. ECIR, 2009, pp. 301–312.

[8] D. Banks, P. Over, N.-F. Zhang, Blind Men and Elephants: Six Approaches to TREC data, Information Retrieval 1 (1999) 7–34.
[9] N. Ferro, G. Silvello, A General Linear Mixed Models Approach to Study System Component Effects, in: Proc. SIGIR, 2016, pp. 25–34.
[10] G. Faggioli, N. Ferro, System effect estimation by sharding: A comparison between anova approaches to detect significant differences (2021).
[11] E. M. Voorhees, Overview of the TREC 2004 Robust Track, in: Proc. TREC, 2004.
[12] R. Benham, J. S. Culpepper, Risk-Reward Trade-offs in Rank Fusion, in: Proc. ADCS, 2017, pp. 1:1–1:8.
[13] Y. Zhao, F. Scholer, Y. Tsegay, Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence, in: Proc. ECIR, 2008, pp. 52–64.
[14] S. Cronen-Townsend, Y. Zhou, W. B. Croft, A Language Modeling Framework for Selective Query Expansion, Technical Report, 2004.
[15] F. Scholer, H. E. Williams, A. Turpin, Query Association Surrogates for Web Search, J. Assoc. Inf. Sci. Technol. 55 (2004) 637–650.
[16] A. Shtok, O. Kurland, D. Carmel, F. Raiber, G. Markovits, Predicting Query Performance by Query-Drift Estimation, ACM Trans. Inf. Syst 30 (2012) 1–35.
[17] Y. Zhou, W. B. Croft, Query Performance Prediction in Web Search Environments, in: Proc. SIGIR, 2007, p. 543–550.
[18] Y. Tao, S. Wu, Query Performance Prediction By Considering Score Magnitude and Variance Together, in: Proc. CIKM, 2014, p. 1891–1894.
[19] A. Shtok, O. Kurland, D. Carmel, Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction, in: Proc. SIGIR, 2010, pp. 259–266.
[20] P. Diaconis, R. L. Graham, Spearman's Footrule as a Measure of Disarray, J. Royal Stat. Soc. 39 (1977) 262–268.
[21] J. M. Tague-Sutcliffe, J. Blustein, A Statistical Analysis of the TREC-3 Data, in: Proc. TREC, 1994, pp. 385–398.