Noname manuscript No.
(will be inserted by the editor)

sMARE: a New Paradigm to Evaluate and Understand Query Performance Prediction Methods*

Guglielmo Faggioli · Oleg Zendel · J. Shane Culpepper · Nicola Ferro · Falk Scholer

Submitted: July 2021 - Revised: November 2021 - Accepted: February 2022

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. To obtain such distribution, we exploit the scaled Absolute Ranking Error (sARE) measure, and its mean the scaled Mean Absolute Ranking Error (sMARE). 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.

[*] This is an extended version of Faggioli et al. (2021), awarded "Best Paper" at the European Conference on Information Retrieval, 2021.

Guglielmo Faggioli
University of Padova, Padova, Italy E-mail: guglielmo.faggioli@phd.unipd.it

Oleg Zendel
RMIT University, Melbourne, Australia E-mail: oleg.zendel@rmit.edu.au

J. Shane Culpepper
RMIT University, Melbourne, Australia

Nicola Ferro
University of Padova, Padova, Italy

Falk Scholer
RMIT University, Melbourne, Australia

Keywords Query Performance Prediction · Systems Evaluation · Analysis of Variance · Query formulations · Information Retrieval

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 including those of Fuhr (2017) and Sakai (2020) have led the community to appreciate the importance of a more theoretically grounded evaluation. 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 (Carmel and Yom-Tov 2010). 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). 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 by a single 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 (He and Ounis 2004; Zhou and Croft 2006; Cronen-Townsend et al. 2002; Zhou and Croft 2007; Chifu et al. 2018; Diaz 2007; Zhao et al. 2008; Carmel et al. 2006; Cummins 2014; Hauff et al. 2008; Mothe and Tanguy 2005; Shtok et al. 2010). 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. (2009) and again more recently by Roitman (2020) to more reliably test for 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 (Roitman 2018a; Zamani et al. 2018; Zendel et al. 2019). 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 effectiveness 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 or 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 (Smucker et al. 2007), 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. A summary 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 summary.

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. In particular, we compare the performance with the distance between the rank predicted by a QPP model for a query and the rank of the query using a given traditional performance measure. Being a measure of the rank error made by a predictor, we call the above measure scaled Absolute Ranking Error (sARE). So, we now have a measure of error for each of the topics, given a specific predictor. To have a measure of the overall quality of the predictor, we can average sARE over all topics and compute the scaled Mean Absolute Ranking Error (sMARE). We also incorporate recent work in retrieval effectiveness on query variation and reformulation for each topic (Bailey et al. 2016, 2017; Benham et al. 2019; Thomas et al. 2017; Zendel et al. 2019) into our framework, which allows a finer-grained sampling of retrieval performance, and allows us to estimate interaction between systems, topics and query formulations, which was not possible using only single pointwise estimates.

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?

This is an extended version of (Faggioli et al. 2021), awarded "Best Paper" at the European Conference on Information Retrieval, 2021 which proposed a novel evaluation framework for QPP, based on the sARE measure, and models QPP prediction performance as a distribution computed over the evaluation topics. The sARE-based approach is a statistically grounded evaluation methodology that can be used by practitioners to perform comprehensive comparative analyses of the effectiveness of new QPP prediction techniques.

Novel Contributions. In this work, we present some novel contributions with respect to Faggioli et al. (2021). We discuss the performance differences observed at the query level based on different performance characteristics ("easy" versus "hard" queries determined by system level effectiveness of AP for example) to demonstrate how predictors might be more comprehensively studied in the future. We also present, as a new contribution, further examples of the capabilities and applications of the proposed framework. In particular, we include an additional ANOVA analysis to show and compare the performance of QPP models based on multiple factors. This provides new insights and observations into how QPP algorithms behave in our original study. It also allows us to provide practitioners with new techniques to understand, debug and explain the performance of new QPP approaches. Finally, we include a detailed experimental study of ties as well as a comprehensive analysis of several alternative formulations to measure the rank error in order to ensure that we are recommending the most appropriate tie-breaking approach and rank error formulation.

2 Related Work

Retrieval performance can vary widely across different systems, even for a single query (Carmel and Yom-Tov 2010; Culpepper et al. 2021). 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 (Cronen-Townsend et al. 2002; Hauff et al. 2008; He and Ounis 2004; Mothe and Tanguy 2005; Scholer et al. 2004; Zhao et al. 2008) and *post-retrieval predictors* that also analyze the retrieval results (Aslam and Pavlu 2007; Roitman 2018b; Shtok et al. 2016; Zamani et al. 2018; Zhou and Croft 2006; Carmel et al. 2006; Cummins 2014; Diaz 2007; Amati et al. 2004). 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 (Hauff et al. 2009; Scholer and Garcia 2009). Hauff et al.

(2009) 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. (1992) 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 (Thomas et al. 2017; Zendel et al. 2019, 2021; Di Nunzio and Faggioli 2021) 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 (Maxwell and Delaney 2004; Rutherford 2011). 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 (Banks et al. 1999; Carterette 2012; Voorhees et al. 2017; Ferro and Silvello 2016; Ferro et al. 2019; Robertson and Kanoulas 2012; Tague-Sutcliffe and Blustein 1994; Faggioli and Ferro 2021). 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 of factor levels to the response variable.

3 Experimental Analysis

In this section we detail the experiments carried out to demonstrate the advantages of using the sARE measure to evaluate QPP models. In Subsection 3.1 we describe the experimental setup. Subsection 3.2 contains details on the traditional evaluation of QPP models, used as a baseline for the subsequent analyses. Subsection 3.4 contains the analysis of how the framework behaves when using several approaches to compute the error and to break ties. In Subsection 3.5 we describe how to use the sARE measure to compare systems. Finally, in Subsection 3.6, we include observations that can now be made on QPP models and query formulations when performing an evaluation using ANOVA and the sARE measure.

3.1 Experimental Setup

In our analyses, we use the TREC Robust 2004 (Robust04) Ad Hoc (Voorhees 2004) 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 original TREC relevance judgments. We enrich the set of queries for the corpus using publicly available human-curated query reformulations for each topic (Benham and Culpepper 2017).[1] Our experiments use a Grid of Points (GoP) of runs as described by Ferro and Harman (2010), using 4 different stopword lists (`atire`, `zettair`, `indri`, `lingpipe`), plus the `no stop` (not applying stopword removal) approach and 2 different stemmers, (`lovins`, `porter`) plus a `nostem` approach. The indexes are constructed from the raw postings lists created with the Apache Lucene search engine[2], and the Common Index File Format (CIFF) (Lin et al. 2020). All runs were produced using our own implementation of the query-likelihood model and use Dirichlet smoothing ($\mu = 1000$), as described originally by Zhai and Lafferty (2001). Each run was repeated 15 times. We test 16 QPP models ($12 + 4$ UEF-based methods) in our analyses, all of 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. 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

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

[2] `https://lucene.apache.org`

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

QPP model	**Description**
Pre-retrieval	
SCQ by Zhao et al. (2008)	Measures similarity based on *cf.idf* to the corpus, summed over the query terms.
AvgSCQ by Zhao et al. (2008)	SCQ normalized by the query length.
MaxSCQ by Zhao et al. (2008)	The query term with maximal SCQ score.
SumVAR by Zhao et al. (2008)	Measures the *cf.idf* variability of the query terms in the corpus.
AvgVAR by Zhao et al. (2008)	Variability normalized with the query length.
MaxVAR by Zhao et al. (2008)	The query term with maximal variability.
AvgIDF by Cronen-Townsend et al. (2004)	The mean *idf* value of the query terms.
MaxIDF by Scholer et al. (2004)	The query term with maximal *idf* value.
Post-retrieval	
Clarity by Cronen-Townsend et al. (2002)	Measures the divergence between the Language Model (LM) constructed over top documents in the result list to the LM of the entire corpus.
NQC by Shtok et al. (2012)	Measures the standard deviation of the top documents scores in the retrieval list.
WIG by Zhou and Croft (2007)	Measures the difference between the mean retrieval score of the top retrieved documents and the score of the entire corpus.
SMV by Tao and Wu (2014)	Scores the queries based on a combination of the scores standard deviation and magnitude.
UEF by Shtok et al. (2010)	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).

collection previously (Shtok et al. 2012, 2010; Tao and Wu 2014). 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 AP was the most common effectiveness metric used in prior QPP work.

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),

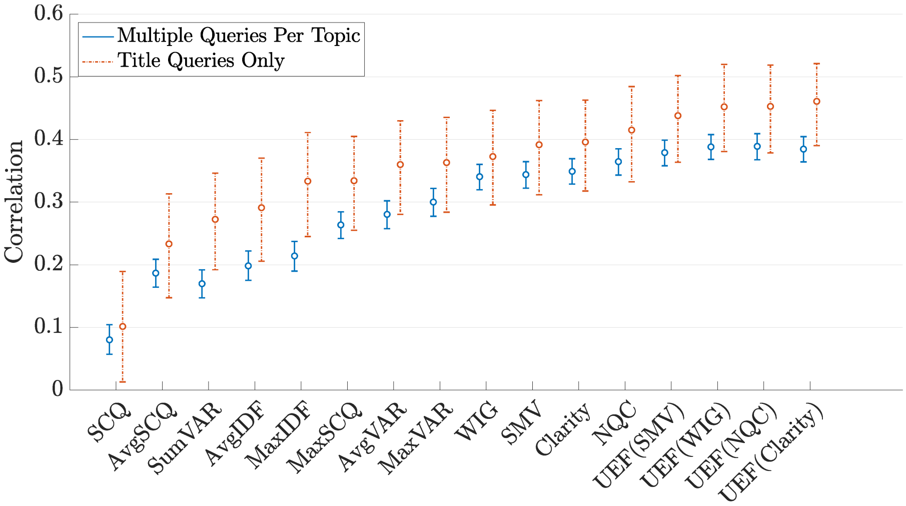


Fig. 1: Prediction quality of the selected QPP models on Robust04 (Confidence Intervals computed with Kendall's τ), using either title queries or all available formulations.

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 usually reported as Pearson's r for linear correlation, Kendall's τ, 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 95% confidence intervals, 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. Conducting pairwise comparisons on the data from Figure 1 (title queries only), a bootstrap hypothesis testing (Efron and Tibshirani 1994) shows that 57 pairs of predictors are statistically significantly different at significance level $\alpha = 0.05$, 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. A large number of statistical "ties" between different QPP models may be caused by one of the following two reasons: *i)* methods are in fact equal and there has been little to no improvement since Clarity was proposed by Cronen-Townsend et al. (2002); or *ii)* our current evaluation strategy is not powerful enough to measure any difference between the models. We are more inclined to believe our second hypothesis, which is inline with the observations of Hauff et al. (2009). That is, using confidence intervals can make it difficult to conclusively determine which QPP system is the best performing one. Figure 2 shows a heat-map plot of the pairwise ranking similarities between the different QPP methods. The similarity is measured with Kendall's τ correlation (Kendall 1945). Given two sorted lists of real values, the original Kendall's τ (Kendall 1938)[3] is defined as follows:

$$\tau = \frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{total number of pairs}} \tag{1}$$

Defining C as the proportion of concordant pairs, we can show that (see Appendix A):

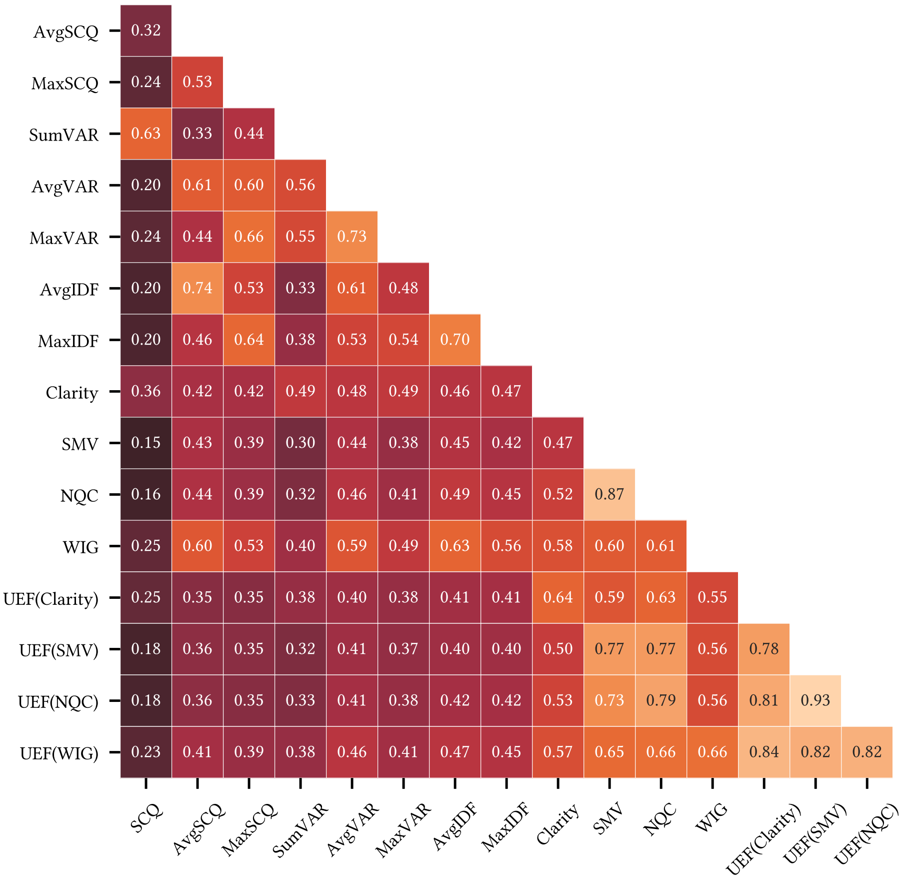
$$C = \frac{\tau + 1}{2}$$

which is an intuitive approximation of the ratio of agreement. Note there are later formulations of Kendall's τ which *do* account for ties. This distinction is discussed in greater detail in the appendix.

For example, for $\tau = 0.6$, $C = 0.8$; means that 80% of the topic pairs are ranked identically using either pair of predictors. Figure 2 further supports this result as all of the UEF based predictors show no significant differences from each other in the current setup. However, the noticeable drop in the similarity of the NQC and Clarity methods when compared to UEF(Clarity) suggests that a more powerful statistical analysis may yield a different outcome. This is a key motivation for our work and will be examined in greater detail.

In addition to using the traditional title queries, we also explore the scenario of using multiple query formulations for a topic, 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 correlation is generally lower when using multiple topic formulations (the blue bars shown in Figure 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, $p < 0.0001$). Notice that, to prevent the number of formulations for each topic from influencing the result, we randomly sample each topic using 5 different formulations. 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

[3] The original formula has no adjustments for ties in the rankings, it is mentioned here for its simplicity.



	SCQ	AvgSCQ	MaxSCQ	SumVAR	AvgVAR	MaxVAR	AvgIDF	MaxIDF	Clarity	SMV	NQC	WIG	UEF(Clarity)	UEF(SMV)	UEF(NQC)
AvgSCQ	0.32														
MaxSCQ	0.24	0.53													
SumVAR	0.63	0.33	0.44												
AvgVAR	0.20	0.61	0.60	0.56											
MaxVAR	0.24	0.44	0.66	0.55	0.73										
AvgIDF	0.20	0.74	0.53	0.33	0.61	0.48									
MaxIDF	0.20	0.46	0.64	0.38	0.53	0.54	0.70								
Clarity	0.36	0.42	0.42	0.49	0.48	0.49	0.46	0.47							
SMV	0.15	0.43	0.39	0.30	0.44	0.38	0.45	0.42	0.47						
NQC	0.16	0.44	0.39	0.32	0.46	0.41	0.49	0.45	0.52	0.87					
WIG	0.25	0.60	0.53	0.40	0.59	0.49	0.63	0.56	0.58	0.60	0.61				
UEF(Clarity)	0.25	0.35	0.35	0.38	0.40	0.38	0.41	0.41	0.64	0.59	0.63	0.55			
UEF(SMV)	0.18	0.36	0.35	0.32	0.41	0.37	0.40	0.40	0.50	0.77	0.77	0.56	0.78		
UEF(NQC)	0.18	0.36	0.35	0.33	0.41	0.38	0.42	0.42	0.53	0.73	0.79	0.56	0.81	0.93	
UEF(WIG)	0.23	0.41	0.39	0.38	0.46	0.41	0.47	0.45	0.57	0.65	0.66	0.66	0.84	0.82	0.82

Fig. 2: The Kendall's τ correlation coefficient computed between several different QPP predictors. The correlation is calculated over topics, which are represented by TREC title queries on the Robust04 *indri-porter* pipeline.

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 3 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 distance the two distributions, is 0.039, which suggests there is a high similarity between the two distributions. In summary, the distributions are similar and thus the

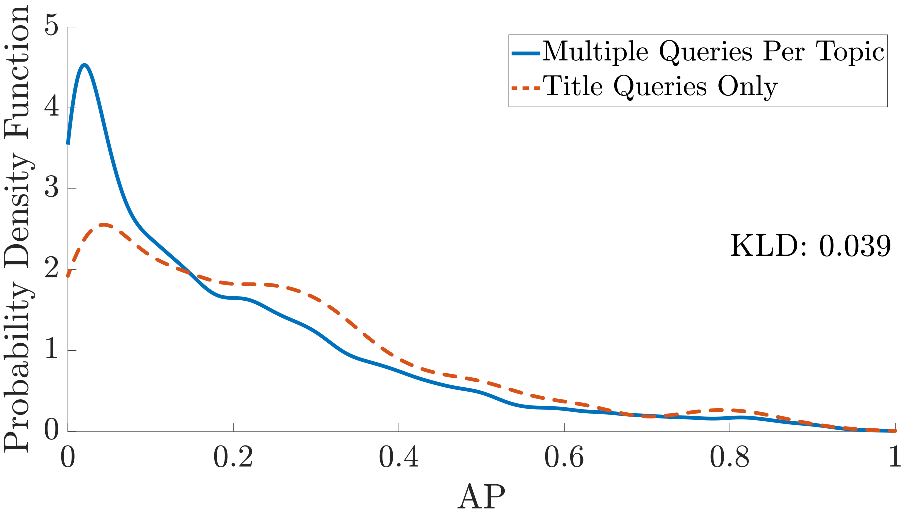


Fig. 3: A comparison of the AP score distributions of the title-only queries and multi-query topic formulations.

introduction of the multiple query formulations 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 the ranks span, as is the default in many statistical applications (Gibbons and Chakraborti 2011). Since the choice of tie breaking rule could have an impact on the results, several possible approaches are evaluated and discussed in greater detail in Subsection 3.4. Observe that this approach 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:

$$\text{sARE-AP}(q_i) := \frac{|r_i^p - r_i^e|}{|Q|}, \tag{2}$$

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 need the single

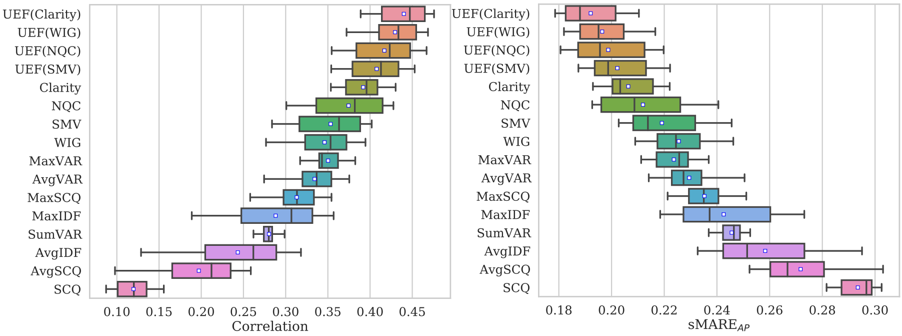


Fig. 4: Prediction quality when measuring correlation with Kendall's τ and sMARE-AP for Robust04 title-only queries and 15 different system configurations. The line inside the interquartile range (IQR) is the median, and the white square is the mean.

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:

$$\text{sMARE-AP}(\mathcal{P}) := \frac{1}{|Q|} \sum_{q_i \in Q} \text{sARE-AP}(q_i). \tag{3}$$

Note that sMARE-AP can be seen as a derivation of *Spearman's Footrule distance*, making it a distance metric for the full rankings instead of a correlation. Among the properties of Spearman's Footrule distance, Diaconis and Graham (1977) 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 4 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).

3.4 Computing the Measure of Deviation of an Optimal Rank Ordering

When defining a measure to accurately represent the distance between the rank of a query w.r.t the rank of all queries when sorted in decreasing order by their AP score and their associated QPP score, two choices need to be made: the tie-breaking strategy and the approach used to quantify the deviation from the optimal rank.

Tie-breaking strategies The sMARE framework is based on computing differences between the expected and observed ranks. The expected rank corresponds to the rank that the query achieves if we sort them by performance. The observed rank, on the other hand, is the rank assigned considering the prediction of a given QPP approach. Since we are considering rankings induced by scores for either observed or predicted performance, we can expect that two or more queries will obtain the same observed / predicted scores. In such cases, we must decide how to assign the value of the rank for each of the queries. Let $\mathcal{Q}_t$ be a set of queries that includes either identical QPP or AP scores, s. Given also r_k the rank of the query with the maximal score such that $s_k < s$, we can define the following tie-breaking strategies, using the list $(0.1, 0.2, 0.2, 0.3)$ as an example:

- `average` $(1, 2.5, 2.5, 4)$: the rank for all the queries in $\mathcal{Q}_t$ is the average rank in the set, equal to $(2r_k + |\mathcal{Q}_t| + 1)/2$. The main advantage of this method is that the sum of the ranks equals to the sum of the ranks when no ties exist.
- `min` $(1, 2, 2, 4)$: all the tied queries have the lowest rank in the tie set, equal to $r_k + 1$.
- `max` $(1, 3, 3, 4)$: all the tied queries have the highest rank in the tie set, equal to $r_k + |Q_t|$.
- `first` $(1, 2, 3, 4)$: ties are sorted "alphabetically" or "lexicographically", according to the order of appearance in the ranked list: where all possible values between 1 and $|\mathcal{Q}|$ are associated to a query. Note that this is similar in spirit to tie-breaking in the `trec_eval` tool which breaks ties by sorting on the document ID. However here we are sorting by query score and not scoring ranked documents.
- `dense` $(1, 2, 2, 3)$: similar to the `min` approach, the rank of all the queries in the set of ties will always be $r_k + 1$, but the rank between groups will always increase by 1. This means that, given $n \leq |\mathcal{Q}|$ unique scores associated to queries in a ranked list, every possible value between 1 to n will be assigned to at least one query.

To further highlight the importance of the analysis on the number of ties, we also report in Figure 5 the number of ties observed. The blue line shows the mean number of ties over all 13 QPP models, and the shaded area represents the 95% confidence interval. Note that, even if we consider as many as 6 digits, we still have on average more than 500 ties. Note that we have used 6 significant digits in the subsequent experiments for each raw observation, and more than

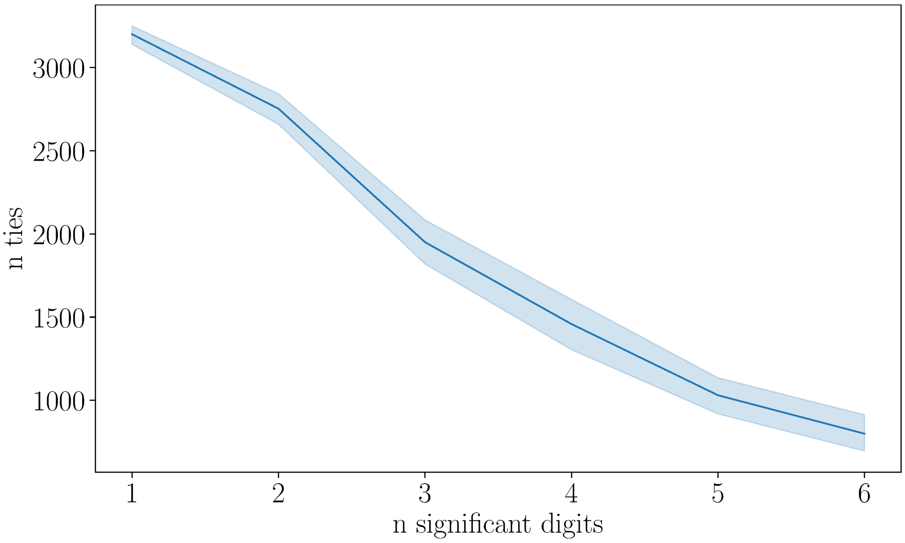


Fig. 5: The average number of ties observed between QPP methods, when the number of significant digits differs. Note that, even when using 6 significant digits, we can observe more than 500 ties on average.

is common practice, and only because it reduces the number of observed ties to a more conservative level – making them less likely to influence any of the observations being made.

Error measures Given r_i^p, the rank observed for the query i in the ranked list sorted by QPP score, and r_i^e, the rank observed in the ranked list sorted by AP, four possible measures can be defined to quantify the distance from the optimal rank:

- scaled Absolute Rank Error (sARE) , as defined in Equation 2;
- sRE (scaled Rank Error) which uses the signed distance between the two ranks, scaled by the number of queries, and is defined as

  $$\frac{r_i^p - r_i^e}{|Q|}.$$

  For the case of no ties, or using the `first` or `average` rank strategy for ties, the sum over all queries would be 0, as it is equal to

  $$\sum_Q \frac{r_i^p - r_i^e}{|Q|} = \frac{1}{|Q|} \sum_Q r_i^p - \sum_Q r_i^e.$$

  This approach is not particularly useful for our needs, but may be useful for other studies.

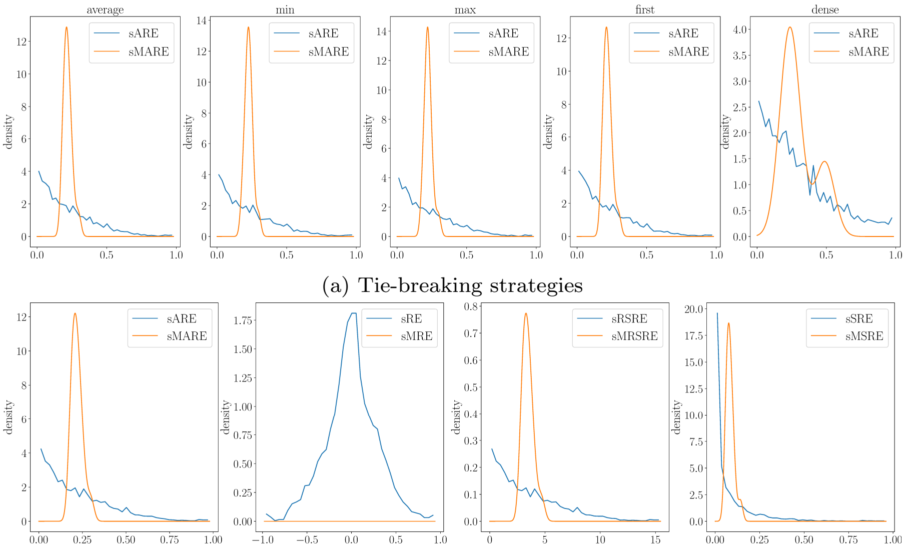
- sSRE (scaled Square Rank Error) is the square of the difference between the two ranks, normalized by the number of queries and is defined as

$$\left(\frac{r_i^p - r_i^p}{|Q|}\right)^2.$$

- sRSRE (scaled Root Square Rank Error) the root of the squared difference:

$$\sqrt{\frac{(r_i^p - r_i^e)^2}{|Q|}}.$$

As shown in Equation 3, each of these measures can be aggregated by computing the mean of all queries for each predictor, to obtain a "mean" version.



(a) Tie-breaking strategies

(b) Ranking error computations

Fig. 6: The top panel (a) shows a comparison between multiple tie-breaking strategies (average, min, max, first and dense approaches, respectively) for both sARE and sMARE. The bottom panel (b) shows the different aggregation algorithms (s(M)ARE, s(M)RE, s(M)RSRE, S(M)SRE, respectively) using average tie-breaking, in term of score density distributions.

Figure 6 compares all of the tie-breaking strategies and formulations for the ranking error of the distribution of the scores for using one possible retrieval pipeline (`indri` stoplist and `porter` stemmer). Figure 6a shows the tie-breaking comparison. Note that, we have artificially inflated the number of ties by truncating the AP and QPP scores to 2 decimal points. Using higher precision scores, the tie-breaking strategies used are all nearly identical, due to proportionally fewer ties. For our tie-breaking strategy comparison, we show only the results observed using sARE – and its averaged version, sMARE – as deviation measure. All other measures discussed exhibit similar behavior. Figure 6b shows the comparison between the different approaches of computing the deviation of the QPP prediction from the ideal rank. For this comparison, ties-breaking uses the `average` strategy since our earlier experiment shows no appreciable differences between tie-breaking strategy when using our experimental data. For each possible setting, we compute the measure of interest for each topic-predictor pair, and plot the probability density distribution of such scores (blue lines). Furthermore, we compute the mean of the scores over all topics for each predictor (orange line). This statistical measure will be used later in our ANOVA experiments when we compare the QPP predictors.

For tie-breaking strategies, we observe that the `average`, `min`, `max`, and `first` tie-breaking strategies all exhibit similar behavior with sARE, with the exception of `dense` tie-breaking which produces much more widely dispersed results. Observe, for example, the additional peak in the distribution when using the `dense` approach. This peak corresponds to the small peaks observed in the other tie breaking strategies, but is inflated in size, when compared to the others. The dense approach is strongly influenced by the number of ties present in the ranking list. This causes the results to be unpredictable since they depend on the randomly observed magnitude (the quantity of ties), which is not correlated with the magnitude of our goal – the performance of QPP. As a result, we recommend against using the `dense` approach, since it may overly inflate performance differences between systems.

Turning our attention to the `first` tie-breaking approach, even though it has a roughly similar distribution to the other strategies, it also introduces a bias as the queries are sorted in an arbitrary order. Such an order does not depend on the actual performance. This problem is particularly relevant when we have large number of ties. In general, if we have many queries and small groups of ties, then the bias does not heavily impact sARE. Nevertheless, we recommend against using it, in order to minimize any possible corner cases. Based on these experimental results, in the remainder of our experiments we will use the `average` tie-breaking method, as it is the most common method, and was the best performing method in our experimental analysis.

With respect to ranking error, we observe that both sARE (scaled Absolute Rank Error) and sRSRE (scaled Root Square Rank Error) have similar density distributions, but sARE is in the [0, 1] interval. Similarly, sSRE is bound by a [0, 1] interval. Overall, the shape of the distribution is quite similar to sARE for our collection, but has two distinguishing differences: it has lower values on average, and it has a smaller range of values. Since differences are squared, sSRE tends to be higher when there are large differences between predicted and observed ranks. Conversely, sARE is larger when there are many errors, even when many of them are small. The smaller values of sSRE when compared against sARE suggests that the QPP models tested tend to make many small errors, and not too many large errors. That is, sSRE is less *discriminative.*

To investigate this further, we compare the two approaches using sensitivity. Using the paired bootstrap test described by Sakai (2006), the Achieved Significance Level (ASL) is computed for each pair of QPP methods using the title queries and the bootstrap with $10,000$ samples. The outcome of our pairwise comparisons is presented in Figure 7. While in general the patterns are similar, sARE does appear to be more sensitive, identifying 74/120 statistically significantly different pairs (61.7%), compared with 68/120 (56.7%) for sSRE. Note that when using this approach, both methods identify more pairs of predictors which are significantly different (where the significance level is $\alpha = 0.05$) than when using the Kendall's τ correlation measured with bootstrap resampling. Both lead to the SMV predictor being added to the cluster of best performing methods. As discussed previously, sMARE can be associated with Spearman's footrule distance, sMSRE (scaled Mean Squared Rank Error) on the other hand can be associated with Spearman's coefficient of association ρ. While both sARE and sSRE have valuable statistical properties (Diaconis and Graham 1977), sARE appears to be more sensitive, and is more useful in our ANOVA analysis, as we want to perform a detailed comparative analysis of methods. The sRE (scaled Rank Error), despite being on a larger interval scale ([-1, 1]), is not useful when computing a mean, here called sMRE (scaled Mean Ranked Error), and is always equal to 0. This is easily explainable since the sum over the ranking errors (using the `average` and `first` tie-breaking strategies) will always be equal to 0. So, based on our desiderata, we have adopted the use of sARE /sMARE since: *i)* they are bounded between 0 and 1;[4] and *ii)* sMARE is not always equal to 0.

3.5 Comparing Systems Using sMARE-AP

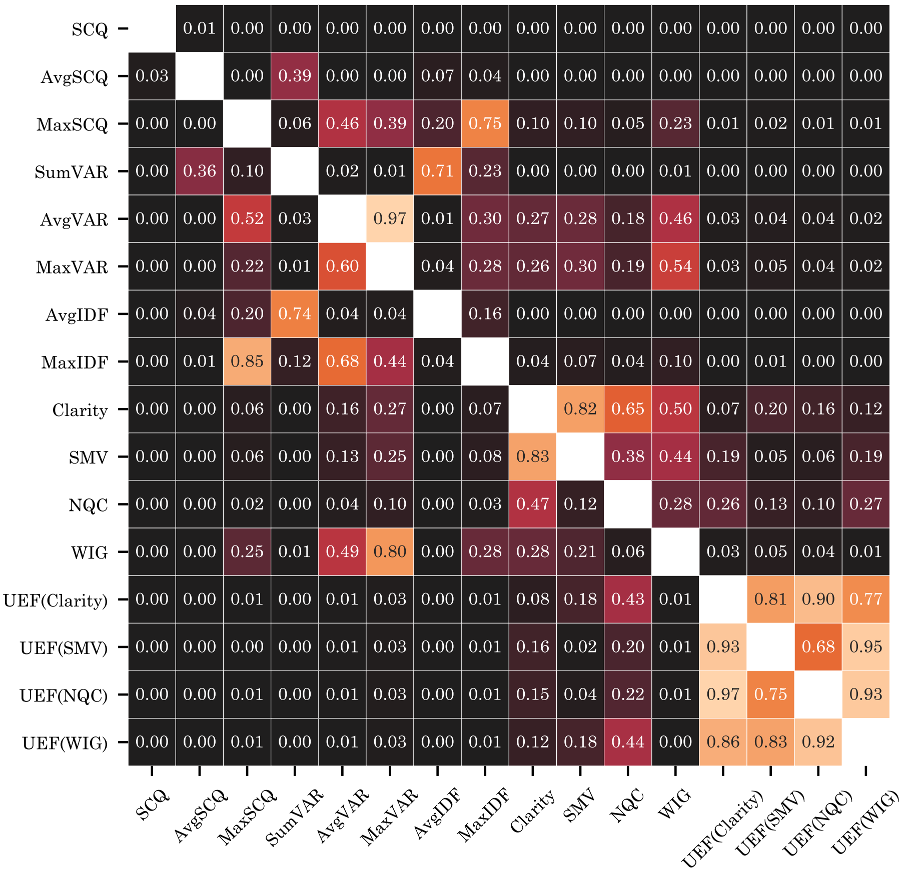
We are now in a position to introduce our first ANOVA model which will enable a more comprehensive experimental analysis of the results:

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

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 effect size, while the topic effect is large, indicating that most of the performance of the QPP depends on the chosen topic.

Based on these results, we next ran a Tukey's Honestly Significant Difference (HSD) post-hoc analysis to test for pairwise comparisons. Figure 8 shows the Tukey's HSD confidence intervals for sMARE-AP over the different QPP

[4] The values of sSRE are bounded as well, and sMSRE $\in [0, \frac{1}{3})$, or $[0, \frac{1}{\sqrt{3}})$ if the squared root is applied on the mean.



	SCQ	AvgSCQ	MaxSCQ	SumVAR	AvgVAR	MaxVAR	AvgIDF	MaxIDF	Clarity	SMV	NQC	WIG	UEF(Clarity)	UEF(SMV)	UEF(NQC)	UEF(WIG)
SCQ		0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AvgSCQ	0.03		0.00	0.39	0.00	0.00	0.07	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaxSCQ	0.00	0.00		0.06	0.46	0.39	0.20	0.75	0.10	0.10	0.05	0.23	0.01	0.02	0.01	0.01
SumVAR	0.00	0.36	0.10		0.02	0.01	0.71	0.23	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
AvgVAR	0.00	0.00	0.52	0.03		0.97	0.01	0.30	0.27	0.28	0.18	0.46	0.03	0.04	0.04	0.02
MaxVAR	0.00	0.00	0.22	0.01	0.60		0.04	0.28	0.26	0.30	0.19	0.54	0.03	0.05	0.04	0.02
AvgIDF	0.00	0.04	0.20	0.74	0.04	0.04		0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MaxIDF	0.00	0.01	0.85	0.12	0.68	0.44	0.04		0.04	0.07	0.04	0.10	0.00	0.01	0.00	0.00
Clarity	0.00	0.00	0.06	0.00	0.16	0.27	0.00	0.07		0.82	0.65	0.50	0.07	0.20	0.16	0.12
SMV	0.00	0.00	0.06	0.00	0.13	0.25	0.00	0.08	0.83		0.38	0.44	0.19	0.05	0.06	0.19
NQC	0.00	0.00	0.02	0.00	0.04	0.10	0.00	0.03	0.47	0.12		0.28	0.26	0.13	0.10	0.27
WIG	0.00	0.00	0.25	0.01	0.49	0.80	0.00	0.28	0.28	0.21	0.06		0.03	0.05	0.04	0.01
UEF(Clarity)	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.01	0.08	0.18	0.43	0.01		0.81	0.90	0.77
UEF(SMV)	0.00	0.00	0.00	0.00	0.01	0.03	0.00	0.01	0.16	0.02	0.20	0.01	0.93		0.68	0.95
UEF(NQC)	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.01	0.15	0.04	0.22	0.01	0.97	0.75		0.93
UEF(WIG)	0.00	0.00	0.01	0.00	0.01	0.03	0.00	0.01	0.12	0.18	0.44	0.00	0.86	0.83	0.92	

Fig. 7: ASL value comparison showing the sensitivity of the sSRE and sARE deviation measures. Values above the diagonal show ASL values for sSRE and the ASL values for sARE are below the diagonal. Computing the sARE pairs result yields $ASL < 0.05 : 74/120$ (61.7%) and the sSRE pairs yield $ASL < 0.05 : 68/120$ (56.7%)

models. Comparing Figure 1 (orange bars) and Figure 8, we can observe that there is less overlap between the CIs, in particular computing the p-values for the pairwise comparisons, out of 120 pairs of predictors, 96 of them are significantly different (80.0%). The outcomes observed when using the bootstrap-based approach resulted in 68.4% [5] more statistically significant differences between predictor pairs when compared against the original data, and the top performing cluster consists of UEF(WIG), UEF(SMV), UEF(NQC), and UEF(Clarity).

[5] $(96 - 57)/57 = 0.684$, where 96 is the number of statistically significantly different pairs found now, and 57 pairs were found using the bootstrap based approach.

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				

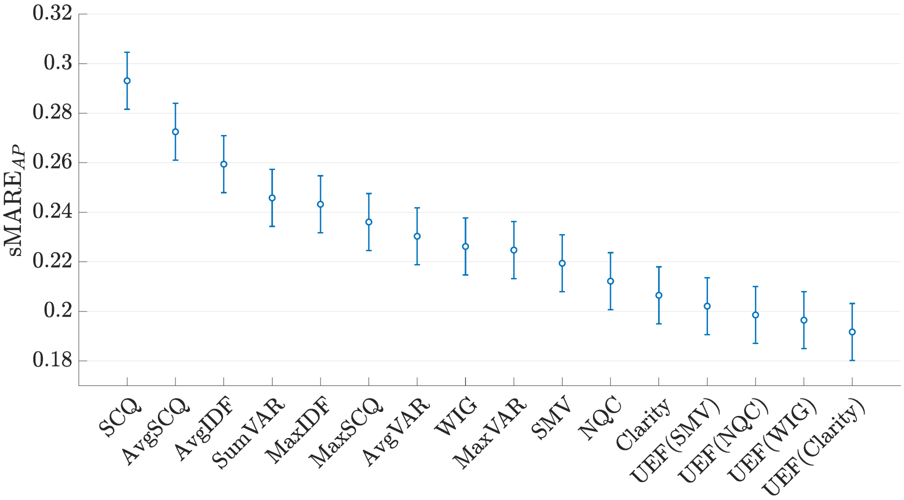
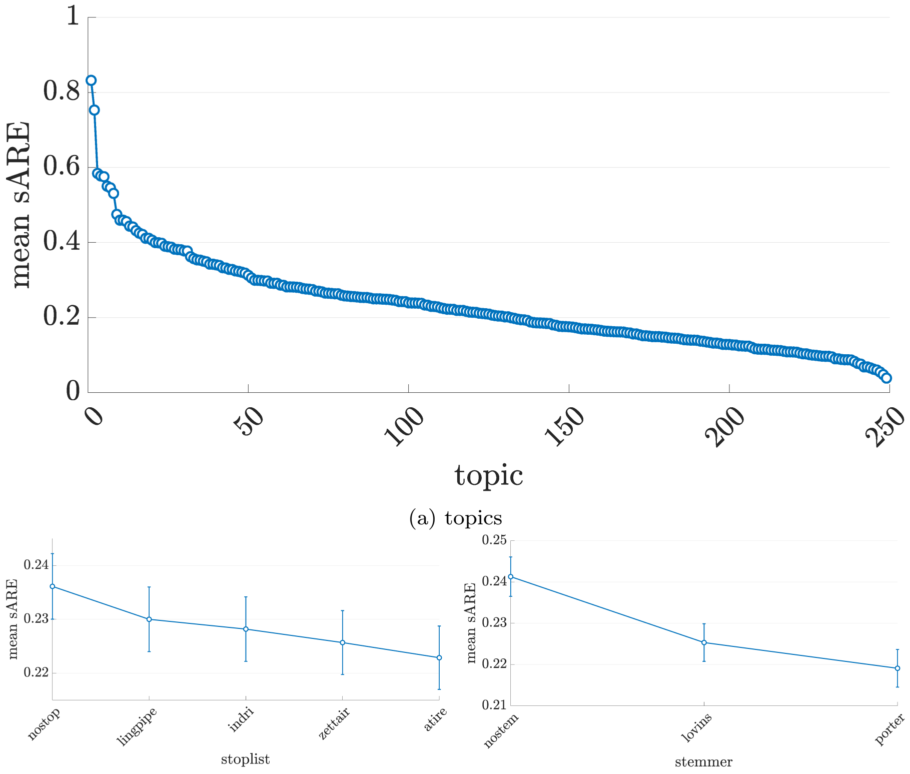


Fig. 8: Confidence Intervals of sMARE-AP from MD0$_{micro}$ for the QPP models on the ROBUST04 title queries.

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 (1994)). Thus, the estimate 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, therefore 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.

Figure 9 shows the main effects observed for different factors and levels when using ANOVA with MD0$_{micro}$. From Figure 9a we can see that, in line



(a) topics

(b) stoplists

(c) stemmers

Fig. 9: Main effects for topics, stoplists and stemmers of sMARE-AP from MD0$_{micro}$ for the QPP models on the Robust04 title queries. We also report the confidence interval for stoplists and stemmers. We do not report CI for the topics, for the sake of image readability.

with Table 2, the topic factor exhibits a very large variance. The topics `‘356’` and `‘679’` present a very large sMARE (0.832 and 0.753 respectively). The title formulation for topic `‘356’` is “`postmenopausal estrogen Britain`”, while for topic `‘679’` it is “`opening adoption records`”.

Figure 9b shows the main effect for the different stoplists included in our analysis. It is interesting to observe that the variance over the different stoplists is very small – changing from the best stoplist (`atire`) to the worst (`nostop`) only leads to an increase of approximately 1.5%. Furthermore, post-hoc analysis shows that `atire` and `zettair` are not statistically significanlty different, while `indri`, `lingpipe` and `nostop` are statistically significantly worse then `atire`. Furthermore, all the stoplists help QPP models in predicting performance more accurately.

Figure 9c highlights the main effect for the stemmer component. Note that the stemmer selected has a bigger impact than the stoplist. Using the

Table 3: Summary table for ANOVA using model MD0$_{micro}$ and representing topics with multiple formulations.

Source	**SS**	**DF**	**MS**	**F**	**p-value**	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	1653.019	248	6.665	214.777	<0.001	0.151
Stoplist	0.405	4	0.101	3.266	0.0110	<0.001
Stemmer	12.726	2	6.363	205.028	<0.001	0.001
QPP model	349.503	15	23.300	750.795	<0.001	0.036
Error	9264.609	298530	0.031			
Total	11280.263	298799				

best stemmer allows us to predict the performance of the queries more easily. In more detail, we observe that Porter's stemmer performs best, followed by Lovins's stemmer. The worst approach is to not use stemming. All pairs of stemmers show a statistically significant difference in performance.

3.6 ANOVA Modeling of Multiple Queries and Interactions

To more fully explore the impact of the query formulations on the performance of QPP predictor, we use the ANOVA model MD0$_{micro}$ in a multiple query formulation setting. We randomly sample 5 formulations[6] to represent the topics. In total, 1,245 different queries were used. Then, we compute the sARE score for each query-predictor pairing. The ANOVA summary table computed using the model MD0$_{micro}$ of multiple formulations of topics is shown in Table 3. Comparing Table 3 to Table 2, we can see that the introduction of multiple query formulations and model MD0$_{micro}$ results in a reduced topic effect size, with the originally observed large-size effect becoming a medium-to-large sized effect. The introduction of the multiple formulations increases the variance of each topic, so the possible score differences between the topics tend to be smaller, smoothing the effect size. The QPP model factor effect is similar for both models. The large Sum of Squares (SS) for the Error component indicates that this model is not suitable if we wish to study/explain any variance in the data. To do this, the model complexity must be increased in order to fit the data more tightly.

To help the model more fit the data more closely, one possible solution is to include a query Formulation factor in the ANOVA. This allows the partial modeling of the additional variance due to the multiple formulations for each topic. Therefore, we now propose another alternative as an ANOVA model:

$$y_{ijqrs} = \mu + \tau_i + \nu_{j(i)} + \gamma_q + \delta_r + \zeta_s + \varepsilon_{ijqrs} \qquad (\text{MD0f}_{micro})$$

The model MD0f$_{micro}$ extends model MD0$_{micro}$ by including $\nu_{j(i)}$, the effect of the j-th formulation of the i-th topic. Note that Topic, Stoplist, Stemmer, and

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

Table 4: Summary table for ANOVA using model MD0f$_{micro}$ and representing topics with multiple formulations.

Source	**SS**	**DF**	**MS**	**F**	**p-value**	$\hat{\omega}^2_{\langle fact \rangle}$
Topic	1653.019	248	6.665	260.704	<0.001	0.177
Formulation(Topic)	1657.578	996	1.664	65.093	<0.001	0.176
Stoplist	0.405	4	0.101	3.965	0.0032	0.000
Stemmer	12.726	2	6.363	248.871	<0.001	0.002
QPP model	349.503	15	23.300	911.343	<0.001	0.044
Error	7607.031	297534	0.026			
Total	11280.263	298799				

QPP model are crossed since each of them can be used in combination with all the others. This is not the case for the multiple formulations of a topic. A formulation can represent only the topic used to create it. Therefore, we cannot treat the formulation as a *crossed* factor, and so query formulations are *nested* for each Topic factor. This ensures the variance produced by different query formulations contribute only to the variance of the topic they represent.

Table 4 presents the results of the ANOVA when using model MD0f$_{micro}$. In order to differentiate the case where a Formulation factor is nested in a Topic from our previous models, we use the term "Formulation (Topic)". When examining Table 4, observe that the effect of the performance of both the Topic and Formulation is a large-sized effect. For formulations of a topic, a good formulation can dramatically change the performance of a QPP model. The effect of the QPP model observed in Table 4 is still small-sized, but has a relative increase of 22.2% when compared against Table 3. Such observations highlight the importance of introducing query formulations into the analysis, both in our data and in the ANOVA model, allowing us to learn more about a predictor. Note that the model MD0f$_{micro}$ still results in a high SS error. This indicates that the model may benefit from further modification to model the data more tightly. However, this will require additional efficiency improvements to be made for running multi-factor ANOVA algorithms on large data collections. The current techniques used to compute the models in this work already require substantial memory and computational resources, and successfully increasing the complexity of the model, or using additional data, is unlikely using any of the currently available hardware and software at either of our universities. We run the above-describe ANOVA via Matlab (version 2017b) on a server with 72 Intel(R) Xeon(R) Gold 6140M CPU 2.30GHz. The largest analysis occupied 250GB of RAM and it required approximately 200 hours to fit the whole model.

One of the most interesting aspects of our framework is the ability to compute the effect size for the interactions between factors. This is possible since the relative performance of a QPP model for each topic can be computed using sARE , and multiple query formulations were introduced as a nested factor. The resulting ANOVA model MD1$_{micro}$ includes component level interactions,

Table 5: MD1$_{micro}$ ANOVA applied on Robust04 collection. ω^2 for non-significant factors is ill-defined and thus not reported. When compared against Faggioli et al. (2021), a different set of random formulations for the topics is used: which leads small differences in the results – the Sum of the Squares being the largest. Nevertheless, the magnitude of the effects and the p-values, which are the focus in an ANOVA, are the same as those in Faggioli et al. (2021).

Source	**SS**	**DF**	**MS**	**F**	**p-value**	$\hat{\omega}^2_{\langle fact\rangle}$
Topic	1653.019	248	6.665	1186.233	<0.001	0.496
Formulation(Topic)	1657.578	996	1.664	296.182	<0.001	0.496
Stoplist	0.405	4	0.101	18.041	<0.001	0.001
Stemmer	12.726	2	6.363	1132.393	<0.001	0.008
QPP model	349.503	15	23.300	4146.715	<0.001	0.172
Topic*Stoplist	39.333	992	0.040	7.057	<0.001	0.020
Topic*Stemmer	147.087	496	0.297	52.776	<0.001	0.079
Topic*QPP model	2297.031	3720	0.617	109.892	<0.001	0.575
Frm.*Stoplist	85.596	3984	0.021	3.824	<0.001	0.036
Frm.*Stemmer	292.736	1992	0.147	26.154	<0.001	0.144
Frm.*QPP model	3215.366	14940	0.215	38.302	<0.001	0.651
Stoplist*Stemmer	0.041	8	0.005	0.918	0.5000	—
Stoplist*QPP model	0.840	60	0.014	2.492	<0.001	<0.001
Stemmer*QPP model	4.509	30	0.150	26.749	<0.001	0.003
Error	1524.492	271312	0.006			
Total	11280.263	298799				

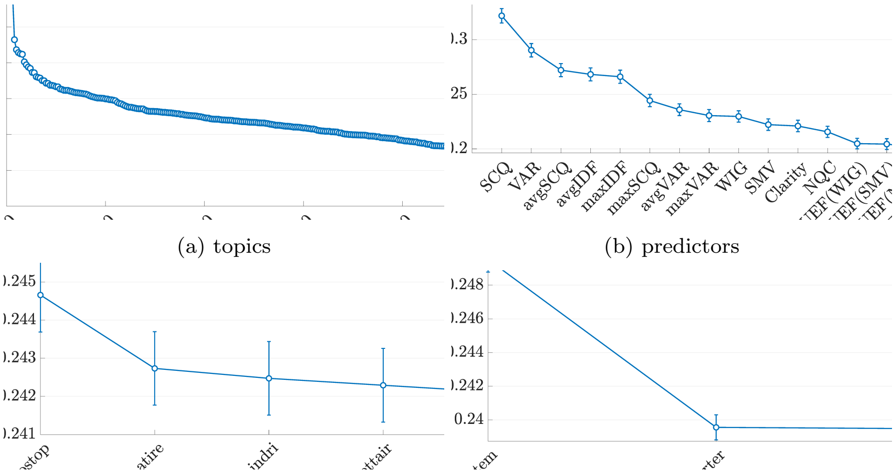
and is defined as:

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

This model extends MD0f$_{micro}$ to include all possible two-way interactions.

Table 5 presents the ANOVA summary statistics for the model MD1$_{micro}$. The table empirically shows that the largest differences in QPP performance are due to the topics, and their formulations. While the importance of topics 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 (medium-sized effects are associated with ω^2 between 6% and 14%). It is important to note that the dimension of the effect is due to the wide variety of QPP models (and their performance) that are 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.

The effect sizes of different 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.



(a) topics

(b) predictors

(c) stoplists

(d) stemmers

Fig. 10: Main effects observed using model MD1$_{micro}$ with multiple topic formulations.

We are now in a position to explore the interaction between topics (and their query formulations) and the predictors. The large effect size indicates that important differences between QPP model performance exist within reformulations of a single topic. Identifying the QPP model where interactions are smallest is valuable in practice, as this corresponds to be choosing a model that is the most robust to query reformulation. Additionally, this approach enables a series of additional analyses, such as a failure analysis for topics to determine which QPP model has the largest interactions with another factor.

There are many additional factors that can influence the performance of the QPP method, beyond the ones tested in the current model. For example, other ranking algorithms or evaluation measures can also be used with sMARE, and could provide new experimental evidence and insights into performance differences between various QPP models in the future.

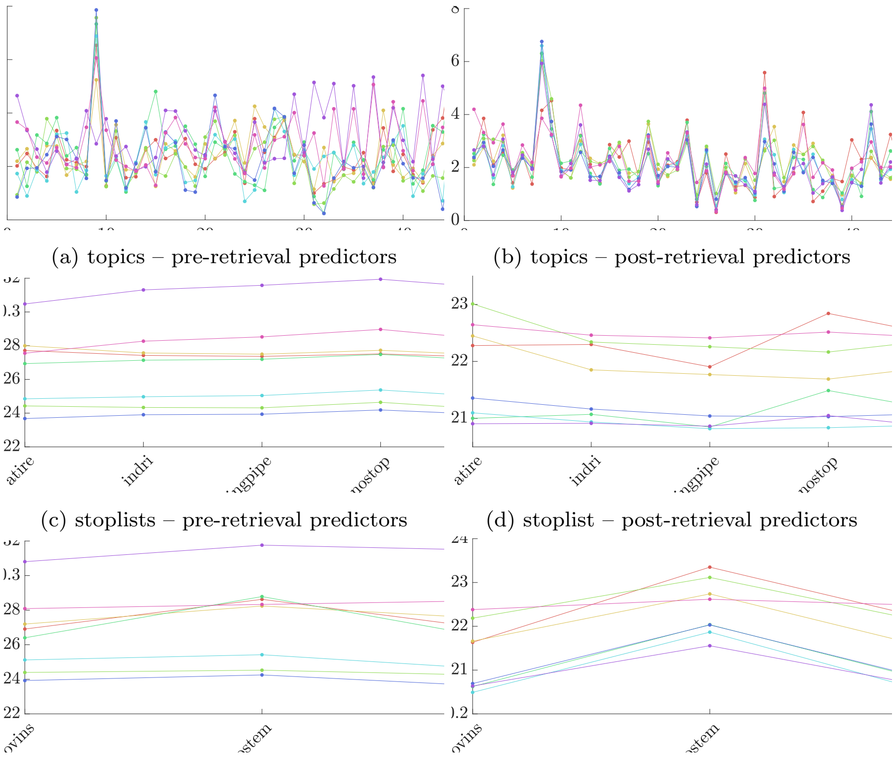
Figure 10 shows the main effects observed using the multiple formulations. Comparing the plot with Figure 9, we can see that overall, the results tend to be more uniform when multiple formulations are included. Formulations tend to have large performance variability: such variability is responsible for the flattening of relative predictor performance. Nevertheless, they give additional power to statistical techniques, allowing the obtaining of more precise results that better generalize to reality. Comparing Figure 10a to Figure 9a we observe that the main effects for the topics tend to be more stable, with a smaller variance. We still have two outliers – the biggest outlier is `'356'`, which also had the biggest effect in MD0$_{micro}$. The second is `'344'`, with the title formulation "`Abuses of E-Mail`". These two topics have sMARE of 0.6463 and 0.6016, respectively. Observing that the topic `'356'` remains particularly complex suggests that the problem is likely linked to the semantic gap between the topic formulations and the relevant documents for that topic. In contrast to what was observed in Figure 9a, the topic `'679'` is not an outlier anymore. This corroborates what was observed in Table 5, showing the importance of the query formulations: different formulations might help the predictors to estimate the query difficulty.

Figure 10b shows an interesting pattern when compared to Figure 8. In particular, we observe that the distribution of the main effects contains much more evident steps if we include multiple query formulations. While the overall order of predictors is close to the one that we observed previously, using multiple formulations we are better able to distinguish between clusters of systems. In particular, SCQ performance suggests that it belongs to its own cluster of quality. VAR, avgSCQ, avgIDF and maxIDF form a distinct cluster, and so do maxSCQ, avgVAR and maxVAR. We then have two clusters of post-retrieval predictors: the first includes the original form of all the predictors, while the second includes the UEF version.

Figures 10c and 10d show the main effect for stoplists and stemmers respectively, when multiple formulations are included in the analysis. The post-hoc analysis shows that all the stoplists are statistically significantly different from the no-stop approach, indicating the importance of applying a stoplist in the QPP scenario. Nevertheless, they are all in the same equivalence class. This empirically suggests that what makes the difference in the QPP setting is either removing stopwords or not, but the stoplists are overall equivalent. Similar conclusions can be drawn for stemmers: in Figure 10d) both stemmers (Porter's and Lovins') are statistically better than the no-stemming approach, but the two stemming approaches do not differ statistically significantly.

Figure 11 shows the interaction plots for the model MD1$_{micro}$. We report the interaction between the predictors and topic, stoplist and stemmer factors. The predictors are further separated into pre- and post-retrieval approaches, shown one the left and right, respectively. Figures 11a and 11b describe the interaction between topics and predictors. Note that, to ease the readability, we report the interaction of the systems with 50 randomly sampled topics. Similar results where observed with different topic samplings. Both plots exemplify the strong interaction between the predictors and the topics, showing several cross-overs between lines and lines tending not to be parallel. This in general confirms what was observed in Table 5. Nevertheless, we observe that lines for the post-retrieval predictors (Figure 11b) are more stable (a similar conclusion can be reached also by looking at Figure 10b). This means that *i)* different post-retrieval predictors tend to perform more similarly than pre-retrieval predictors; and *ii)* the interaction between topics and post-retrieval predictors is lower compared to that between pre-retrieval predictors and topics.

Concerning the stoplist and stemmer components, Figures 11c and 11e illustrate how much they interact with the pre-retrieval predictors. In both cases, the interaction between the component and the predictor is light, with parallel lines overall. The only exception to this is avgIDF and maxIDF, which show a swift drop in performance when used in combination with the nos-



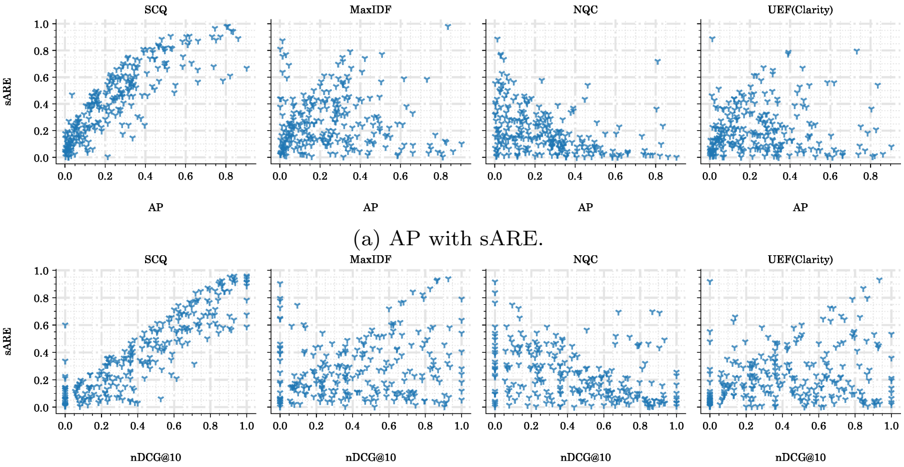
(a) topics – pre-retrieval predictors (b) topics – post-retrieval predictors

(c) stoplists – pre-retrieval predictors (d) stoplist – post-retrieval predictors

(e) stemmers – pre-retrieval predictors (f) stemmers – post-retrieval predictors

Fig. 11: Interaction effects observed using model MD1$_{micro}$ with multiple topic formulations. We report interactions between pre-retrieval (left) and post-retrieval (right) models, with topics (top), stoplists (center) and stemmers (bottom).

tem approach. Figure 11d reports the interaction between the different post-retrieval QPP models with the stoplist component. The choice of stoplist does not interact particularly with the post-retrieval methods, as also shown in Table 5. The QPP approach most affected by the different stoplists is Clarity, both in its traditional and UEF versions. This indicates that, if the practitioner intends to use Clarity, it is important to validate its performance over different stoplists. On the other hand, the WIG model (both traditional and UEF versions) is the most stable. Concerning the choice of stemmer, Figure 11f shows that the stemmer interacts slightly with predictor performance, similar to what was observed for pre-retrieval QPP approaches. All of the QPP models appear to be overall stable across different stemmers, with small interaction with the stemmer used. In particular, most QPP models suffer when query terms are not stemmed. The traditional version of WIG is the most stable QPP approach for this: it does not benefit if the stemmer is used or not.



(a) AP with sARE.

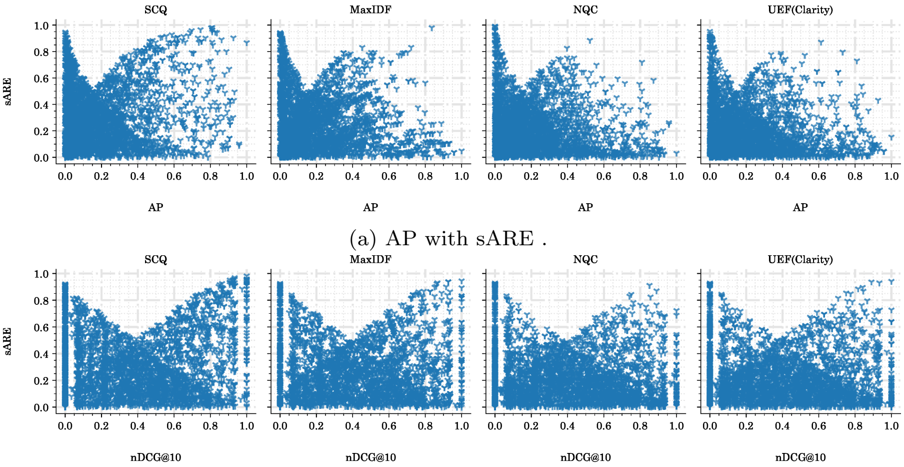
(b) nDCG@10 with sARE.

Fig. 12: Scatter plot of error (sARE) versus effectiveness (AP and nDCG@10), using only title queries.

4 Potential Applications

The evaluation approach presented in this paper supports a range of new possible performance analyses. As an example of what might be done in the future, we conduct a preliminary analysis to explore the relationship between query effectiveness and the quality of prediction. For several years, there has been a well-known problem observed in the QPP community where a deep measure such as AP tends to be amenable to high quality predictions which early precision measures, such as nDCG or Expected Reciprocal Rank (ERR), with a cutoff of, say, 5 or 10, tend to have worse overall performance when compared directly in a single collection. Indeed, we can see a somewhat similar trend when looking at performance per query. Figures 12a and 12b present several scatter plots for sARE when using AP and nDCG@10, respectively. Using only the title queries, we can observe different trends across different QPP methods, indicating that the effectiveness of the query has a varying effect on different QPP methods. Specifically, the SCQ method which has the worst prediction quality, tends to make greater errors on queries with higher effectiveness (measured both by AP and nDCG@10). On the other hand NQC, which has a significantly better prediction quality, shows the opposite trend, tending to make smaller mistakes on queries with high effectiveness.

Figure 13 shows the corresponding scatter plots when all query variants are used. Overall, the observed trends are slightly smaller for AP, and almost completely erased for nDCG@10. However, we can observe an interesting difference in the distribution of the points in the two evaluation measures. Interestingly, a visual comparison of these plots side-by-side for multiple predictors consistently exhibit the performance difference trend between AP and nDCG@10



(a) AP with sARE .

(b) nDCG@10 with sARE .

Fig. 13: Scatter plot of the error (sARE) versus the effectiveness (AP and nDCG@10).

alluded to previously. The "dips" in the graph w.r.t the evaluation measure are where the median score for all the queries occur. Since this is the mid-point of the rank ordering, the mid error can never be greater than 0.5 at this point. That is, a rank ordering error can only be 1.0 for queries at the top or the bottom of the ranking, as they are the ones that can be inverted (i.e., an AP query with a score of 1.0 might be predicted to be the worst query in the set, and such cases are where the error is the greatest). Consider the following concrete example: For (a,b,c) the errors for each item would be: $a \in \{1/3, 2/3\}, b \in \{1/3\}, c \in \{1/3, 2/3\}$ in that case b is the median query and the maximal error it can achieve is $0.5 * maximal\_error$. The median nDCG@10 is 0.3718 and AP is 0.1491.

When studying the AP plots, we can see a strong trend where the performance gets better as AP increases. However, for nDCG@10, this is not as consistent. So, we can see that for nDCG@10 the predictors can have very poor performance for high and low performing queries more often. In fact there are a number of reasons such a trend might exist related to differences in gain function in the measures. Overall, while we cannot say why such behavior exists in current predictors, it is a valuable start in a comprehensive failure analysis in QPP prediction behavior. What we can say is that such performance differences warrant further study, and we intend to explore this problem in greater detail in the future. Here, we wanted one small example of what is possible using the framework introduced in this paper.

The new methodology also allows us to determine which topics are "harder" for QPP methods. In Figure 14 we plot the topics that had the "worst" query variants, as measured by sARE. In order to visualize this, we extract 1% of all queries that have the highest absolute error (sARE) for each predictor (if the

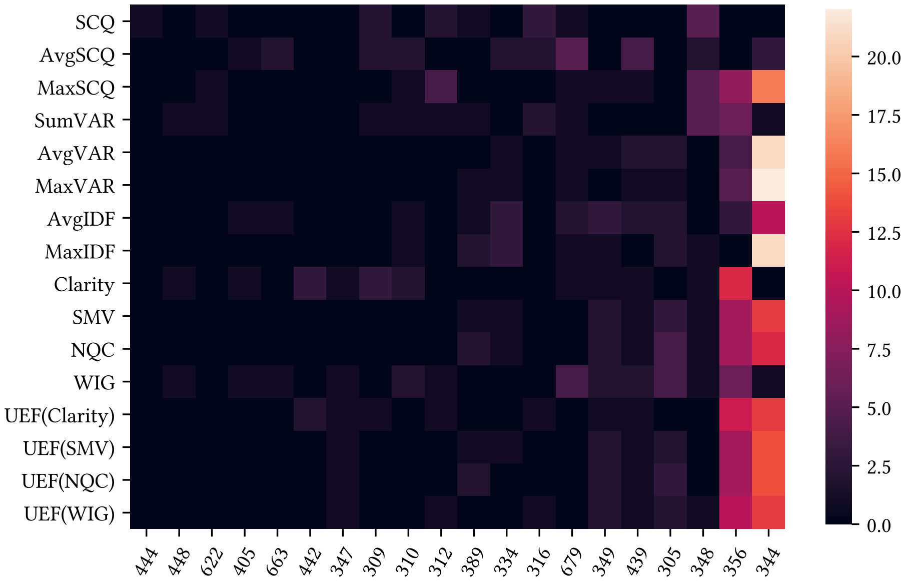


Fig. 14: Heat map of the hardest 1% of all query variants (by sARE-AP value), grouped by topic. For readability, the plot is constrained to show the 1/3 of the full set of topics that had the highest number of *hard* queries.

boundary query was tied, then all queries with the same tied error level were included). So, for each predictor we show at least 32 queries with the largest error. Next, the queries are grouped by topic, and for each topic we count the number of the queries that appear in the worst 1% list. Finally, the topics are ordered by the total number of queries for all of the predictors.

In Figure 14, we can see that some topics immediately stand out – which represent the topics with many hard queries for all predictors. This indicates that some topics are indeed harder than others to accurately predict. This further corroborates the results of our ANOVA analysis, where the topic factor was found to have a large effect size. Similar to previous observations, the two topics that stand out as the "hardest" for the majority of the QPP methods are topic number 344 and topic 356. These trends will be explored further in future work.

5 Conclusion

In this paper 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 a particular QPP approach) and the expected rank (measured using AP, or any other traditional IR effectiveness measure). Such approaches allow us to obtain a distribution of performance for the QPP over the different topics. Furthermore, our framework can leverage multiple query formulations for each topic to enhance the power of the analysis. Together, the use of multiple query formulations and the distributional representation of the performance enables carrying out more precise studies. In particular, we show that it is possible to rely on the statistical properties of ANOVA and additional post-hoc procedures such as Tukey's HSD test to better identify statistically significant differences between QPP approaches. The proposed framework also enables the analysis of interaction effects for QPP models and topics, supporting failure analyses and a deeper understanding of how a QPP model works. Our framework can be extended and adapted to different investigation needs. For example, in an academic setting, it may be useful to add further factors to the model such as tokenizers, query expansion components, or ranking functions, to further 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 using different ranking methods to establish "ground truth" performance; new factors that influence QPP systems such as the ranking approach used in the post-retrieval QPP approaches; and the effects of using multiple corpora, in order to more comprehensively model and understand corpus and QPP interactions. To support the reproducibility of our results, the code for our evaluation framework has been made publicly available.[7]

Acknowledgments

The work is partially funded by University of Padova Strategic Research Infrastructure Grant 2017: "CAPRI: Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione". This work was also partially supported by the Australian Research Council's Discovery Projects Scheme (DP190101113).

References

Amati G, Carpineto C, Romano G (2004) Query Difficulty, Robustness, and Selective Application of Query Expansion. In: Proc. ECIR, pp 127–137

Aslam JA, Pavlu V (2007) Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions. In: Proc. ECIR, pp 198–209

Bailey P, Moffat A, Scholer F, Thomas P (2016) UQV100: A Test Collection with Query Variability. In: Proc. SIGIR, pp 725–728

Bailey P, Moffat A, Scholer F, Thomas P (2017) Retrieval Consistency in the Presence of Query Variations. In: Proc. SIGIR, pp 395–404

Banks D, Over P, Zhang NF (1999) Blind Men and Elephants: Six Approaches to TREC data. Information Retrieval 1(1-2):7–34

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

Benham R, Culpepper JS (2017) Risk-Reward Trade-offs in Rank Fusion. In: Proc. ADCS, pp 1:1–1:8

Benham R, Mackenzie J, Moffat A, Culpepper JS (2019) Boosting Search Performance Using Query Variations. ACM Trans Inf Syst 37(4)

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

Carmel D, Yom-Tov E, Darlow A, Pelleg D (2006) What Makes a Query Difficult? In: Proc. SIGIR, p 390–397

Carterette BA (2012) Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. ACM Trans Inf Syst 30(1):4:1–4:34

Chifu AG, Laporte Léa, Mothe J, Ullah MZ (2018) Query Performance Prediction Focused on Summarized Letor Features. In: Proc. SIGIR, pp 1177–1180

Cronen-Townsend S, Zhou Y, Croft WB (2002) Predicting Query Performance. In: Proc. SIGIR, pp 299–306

Cronen-Townsend S, Zhou Y, Croft WB (2004) A Language Modeling Framework for Selective Query Expansion. Tech. rep., Center for Intelligent Information Retrieval, University of Massachusetts

Culpepper JS, Faggioli G, Ferro N, Kurland O (2021) Topic difficulty: Collection and query formulation effects. ACM Trans Inf Syst 40(1)

Cummins R (2014) Document Score Distribution Models for Query Performance Inference and Prediction. ACM Trans Inf Syst 32(1):2:1–2:28

Di Nunzio GM, Faggioli G (2021) A study of a gain based approach for query aspects in recall oriented tasks. Applied Sciences 11(19)

Diaconis P, Graham RL (1977) Spearman's Footrule as a Measure of Disarray. J Royal Stat Soc 39(2):262–268

Diaz F (2007) Performance Prediction Using Spatial Autocorrelation. In: Proc. SIGIR, p 583–590

Efron B, Tibshirani RJ (1994) An Introduction to the Bootstrap. Chapman and Hall/CRC, USA

Faggioli G, Ferro N (2021) System Effect Estimation by Sharding: A Comparison Between ANOVA Approaches to Detect Significant Differences. In: Proc. of ECIR, pp 33–46

Faggioli G, Zendel O, Culpepper JS, Ferro N, Scholer F (2021) An Enhanced Evaluation Framework for Query Performance Prediction. In: Proc. of ECIR, pp 115–129

Ferro N, Harman D (2010) CLEF 2009: Grid@CLEF Pilot Track Overview. In: Proc. CLEF, pp 552–565

Ferro N, Silvello G (2016) A General Linear Mixed Models Approach to Study System Component Effects. In: Proc. SIGIR, pp 25–34

Ferro N, Fuhr N, Maistro M, Sakai T, Soboroff I (2019) CENTRE@CLEF 2019. In: Proc. ECIR, pp 283–290

Fuhr N (2017) Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. SIGIR Forum 51(3):32–41

Gibbons JD, Chakraborti S (2011) Nonparametric Statistical Inference, 5th edn. Chapman & Hall/CRC, Taylor and Francis Group, Boca Raton (FL), USA

Hauff C, Hiemstra D, de Jong F (2008) A Survey of Pre-Retrieval Query Performance Predictors. In: Proc. CIKM, pp 1419–1420

Hauff C, Azzopardi L, Hiemstra D (2009) The Combination and Evaluation of Query Performance Prediction Methods. In: Proc. ECIR, pp 301–312

He B, Ounis I (2004) Inferring Query Performance Using Pre-retrieval Predictors. In: Proc. SPIRE, pp 43–54

Kendall MG (1938) A New Measure of Rank Correlation. Biometrika 30(1/2):81–93

Kendall MG (1945) The Treatment of Ties in Ranking Problems. Biometrika 33(3):239–251

Lin J, Mackenzie J, Kamphuis C, Macdonald C, Mallia A, Siedlaczek Michał Trotman A, de Vries A (2020) Supporting Interoperability Between Open-Source Search Engines with the Common Index File Format. In: Proc. SIGIR, pp 2149–2152

Maxwell S, Delaney HD (2004) Designing Experiments and Analyzing Data. A Model Comparison Perspective, 2nd edn. Lawrence Erlbaum Associates, Mahwah (NJ), USA

Meng XL, Rosenthal R, Rubin DB (1992) Comparing Correlated Correlation Coefficients. Psychological Bulletin 111(1):172–175

Mothe J, Tanguy L (2005) Linguistic Features to Predict Query Difficulty. In: Proc. SIGIR, pp 7–10

Robertson SE, Kanoulas E (2012) On Per-topic Variance in IR Evaluation. In: Proc. SIGIR, pp 891–900

Roitman H (2018a) An Extended Query Performance Prediction Framework Utilizing Passage-Level Information. In: Proc. SIGIR, p 35–42

Roitman H (2018b) Query Performance Prediction using Passage Information. In: Proc. SIGIR, pp 893–896

Roitman H (2020) ICTIR Tutorial: Modern Query Performance Prediction: Theory and Practice. In: Proc. SIGIR, pp 195–196

Rutherford A (2011) ANOVA and ANCOVA. A GLM Approach, 2nd edn. John Wiley & Sons, New York, USA

Sakai T (2006) Evaluating Evaluation Metrics based on the Bootstrap. In: Proc. SIGIR, pp 525–532

Sakai T (2020) On Fuhr's Guideline for IR Evaluation. Proc SIGIR 54(1):1–8

Scholer F, Garcia S (2009) A Case for Improved Evaluation of Query Difficulty Prediction. In: Proc. SIGIR, p 640–641

Scholer F, Williams HE, Turpin A (2004) Query Association Surrogates for Web Search. J Assoc Inf Sci Technol 55(7):637–650

Shtok A, Kurland O, Carmel D (2010) Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction. In: Proc. SIGIR, pp 259–266

Shtok A, Kurland O, Carmel D, Raiber F, Markovits G (2012) Predicting Query Performance by Query-Drift Estimation. ACM Trans Inf Syst 30(2):1–35

Shtok A, Kurland O, Carmel D (2016) Query Performance Prediction Using Reference Lists. ACM Trans Inf Syst 34(4):19:1–19:34

Smucker MD, Allan J, Carterette BA (2007) A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In: Proc. CIKM, pp 623–632

Tague-Sutcliffe JM, Blustein J (1994) A Statistical Analysis of the TREC-3 Data. In: Proc. TREC, pp 385–398

Tao Y, Wu S (2014) Query Performance Prediction By Considering Score Magnitude and Variance Together. In: Proc. CIKM, p 1891–1894

Thomas P, Scholer F, Bailey P, Moffat A (2017) Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. In: Proc. ADCS

Voorhees EM (2004) Overview of the TREC 2004 Robust Track. In: Proc. TREC

Voorhees EM, Samarov D, Soboroff I (2017) Using Replicates in Information Retrieval Evaluation. ACM Trans Inf Syst 36(2):12:1–12:21

Zamani H, Croft WB, Culpepper JS (2018) Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In: Proc. SIGIR, pp 105–114

Zendel O, Shtok A, Raiber F, Kurland O, Culpepper JS (2019) Information Needs, Queries, and Query Performance Prediction. In: Proc. SIGIR, p 395–404

Zendel O, Culpepper JS, Scholer F (2021) Is Query Performance Prediction With Multiple Query Variations Harder Than Topic Performance Prediction?, Association for Computing Machinery, New York, NY, USA, p 1713–1717

Zhai C, Lafferty J (2001) A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In: Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., New York, NY, USA, SIGIR '01, pp 334–342

Zhao Y, Scholer F, Tsegay Y (2008) Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In: Proc. ECIR, pp 52–64

Zhou Y, Croft WB (2006) Ranking Robustness: A Novel Framework to Predict Query Performance. In: Proc. CIKM, p 567–574

Zhou Y, Croft WB (2007) Query Performance Prediction in Web Search Environments. In: Proc. SIGIR, p 543–550

A Kendall's τ formulation derivation

Given the formulation of Kendall's τ as defined in 1, if we define C and D as:

$$C = \frac{\text{number of concordant pairs}}{\text{total number of pairs}},$$
$$D = \frac{\text{number of discordant pairs}}{\text{total number of pairs}};$$

The general Kendall's τ formula (as defined in eq. 1) becomes:

$$\tau = \frac{\text{number of concordant pairs}}{\text{total number of pairs}} - \frac{\text{number of discordant pairs}}{\text{total number of pairs}}$$

And therefore:

$$\tau = C - D,$$
$$C + D = 1;$$

We can observe that:

$$C = \tau + D,$$
$$D = 1 - C;$$

Thus $C = \tau + (1 - C) = \tau + 1 - C$. Therefore, $2C = \tau + 1$ and thus $C = \frac{\tau+1}{2}$

Note that this is the original version of Kendall's τ (Kendall 1938), the actual formula applied in the correlation calculations throughout the paper is a later version, which is commonly known as τ_b (τ_s in the original paper) (Kendall 1945). The correlation coefficient τ_b is extending the original formula to treat ties.