Query Performance Prediction: Evaluation Contrasted with Effectiveness

Claudia Hauff[1], Leif Azzopardi[2], Djoerd Hiemstra[1], and Franciska de Jong[1]

[1] University of Twente, Enschede, The Netherlands
{c.hauff,hiemstra,f.m.g.dejong}@ewi.utwente.nl
[2] University of Glasgow, Glasgow, UK
leif@dcs.gla.ac.uk

Abstract. Query performance predictors are commonly evaluated by reporting correlation coefficients to denote how well the methods perform at predicting the retrieval performance of a set of queries. Despite the amount of research dedicated to this area, one aspect remains neglected: how strong does the correlation need to be in order to realize an improvement in retrieval effectiveness in an operational setting? We address this issue in the context of two settings: Selective Query Expansion and Meta-Search. In an empirical study, we control the quality of a predictor in order to examine how the strength of the correlation achieved, affects the effectiveness of an adaptive retrieval system. The results of this study show that many existing predictors fail to achieve a correlation strong enough to reliably improve the retrieval effectiveness in the Selective Query Expansion as well as the Meta-Search setting.

1 Introduction

Predicting the performance, i.e. the retrieval effectiveness, of queries has become a very active area of research in recent years [1,4,6,8,13,15,18,19,20,21]. Accurately predicting a query's effectiveness would enable the development of adaptive components in retrieval systems. For instance, if the performance of a query is predicted to be poor, the system may ask the user for a refinement of the query or divert it to a specialized corpus. Conversely, if the performance of a query appears sufficiently good, the query's performance can be further improved by an affirmative action such as automatic query expansion (AQE) [1].

While the perceived benefits of query performance prediction (QPP) methods are clear, current evaluations often do not consider whether those benefits are indeed realized. Commonly, the correlation between the ground truth (the retrieval effectiveness in average precision for instance) and the predicted performance of queries is reported. The subsequent application of QPP methods in an operational setting is often missing and thus it remains unclear, if QPP methods perform well enough to be useful in practice.

In this work, we attempt to bridge this knowledge gap by investigating the relationship between the correlation that QPP methods achieve, and their effect on retrieval effectiveness in two operational settings: Meta-Search (MS) [15,18]

C. Gurrin et al. (Eds.): ECIR 2010, LNCS 5993, pp. 204–216, 2010.
© Springer-Verlag Berlin Heidelberg 2010

and Selective Query Expansion (SQE) [1,5,18]. Our goal is to determine at what levels of correlation a QPP method can be considered good enough to be employable in practice. If we can determine such thresholds, we would be able to infer from a correlation-based evaluation, whether the quality of a QPP method is sufficient for an operational setting.

We conduct two empirical studies based on several TREC[1] data sets. In previous work [7], we performed preliminary experiments with respect to SQE. Here, we extend these experiments and add a second setting: Meta-Search. We will show, that the correlation a QPP method needs to achieve on average to be useful in practice, is dependent on the operational setting. In the SQE setting, moderate to high correlations result in reliable improvements. In the MS setting, low to moderate correlations are already sufficient, however for these improvements to be statistically significant, we find moderate to high correlations to be required.

The paper is organized as follows: in Sec. 2 we outline related work and the motivation for our work. The data sets and our general approach are described in Sec. 3. Then, the experiments on SQE (Sec. 4) and MS (Sec. 5) are presented, followed by the conclusions and directions for future work (Sec. 6).

2 Related Work and Motivation

First, we briefly describe the two main types (pre- and post-retrieval) of QPP methods. To give an indication of the success QPP methods have achieved in SQE and MS respectively, we also provide an overview of literature that employed QPP methods in either setting.

Query Performance Prediction. Pre-retrieval QPP methods predict the performance of a query without considering the ranked list of results returned by a retrieval system in response to a query. Approaches in this category are usually based on the query terms' corpus-statistics, such as the standard deviation of the query terms' IDF [8] or TF.IDF [19] values. Post-retrieval predictors base their predictions on the ranked list of results. The strategies employed are manifold, they include a comparison between the ranked list and the corpus [1,4], the perturbation of query terms and a subsequent comparison of the generated ranked lists [13,18,21], the perturbation of documents in the ranked list to determine the list's stability [13,20], and the reliance on different retrieval approaches to form predictions based on the diversity of the returned documents [6]. The two most commonly reported correlation coefficients in QPP evaluations are the rank correlation coefficient Kendall's Tau $\tau \in [-1, 1]$ and the linear correlation coefficient $r \in [-1, 1]$. Current state of the art QPP methods achieve up to $\tau \approx 0.55$ and $r \approx 0.65$, depending on the test corpus and query set.

Applications of SQE. The two SQE scenarios evaluated in [18] are based on the notion that easy queries (queries with a high retrieval effectiveness) improve with the application of AQE, while difficult queries (queries with a low

[1] Text REtrieval Conference (TREC), http://trec.nist.gov/

retrieval effectiveness) degrade. The reasoning is as follows: easy queries will have relevant documents among the top ranked results, and thus an AQE algorithm [3,12,16,17], which derives additional query terms from the top ranked documents returned for the initial query, is likely to pick terms related to the information need. The ranked list retrieved for the expanded query thus further improves the quality of the results. Difficult queries have few or no relevant documents in the top ranks and an AQE algorithm will add irrelevant terms to the query, degrading the result quality. In the first scenario in [18], a support vector machine is trained on features derived from the ranked list of the original query to classify queries as either to be expanded or not. In the second scenario [18], a QPP method is used to rank the queries according to their predicted performance. The 85% best performing queries are derived from TREC topic descriptions[2] (simulating AQE), while the bottom 15% of queries are derived from TREC topic titles (simulating no AQE). In both experiments, selectively expanding the queries based on a QPP method proves slightly better than uniformly expanding all queries, with a change in Mean Average Precision (MAP) of +0.001.

An analogous scenario with a different QPP method is evaluated in [1]; the greatest improvement reported is from a MAP of 0.252 (AQE of all queries) to 0.256 (selective AQE). Better results are reported in [5], where the threshold of when (not) to expand a query is learned: in the best case, MAP increases from 0.201 (AQE of all queries) to 0.212 (selective AQE). AQE is combined with collection enrichment in [9]: depending on how the QPP method predicts a query to perform, it is either not changed, expanded based on the results of the local corpus or expanded based on the results of the external corpus. The evaluation yields mixed results, while for one data set the MAP increases from 0.220 (AQE of all queries) to 0.236 (selective AQE), no change in effectiveness is observed for a second data set. These results indicate, that successfully applying a QPP method in the SQE setting is a challenging task.

Applications of MS. In [18], the following meta-search setup is evaluated: a corpus is partitioned into four parts, each query is submitted to each partition, and the result lists of each partition are merged with weights according to their predicted performance. In this experiment, MAP increases from 0.305 (merging without weights) to 0.315 (merging with QPP based weights).

In [15], a variety of retrieval algorithms are applied on one corpus. For each query and retrieval algorithm, a result list is derived and its predicted performance score is determined. Heuristic thresholds are used to classify each result list as either poor, medium or good. The result lists are then merged with weights according to the performance classification. The best weighted data fusion method performs 2.12% better than the unweighted baseline. Finally, in [14] it is proposed to generate a number of relevance models [11] for each query and pick the model that is predicted to perform best. The results indicate the

[2] A TREC topic usually consists of a title (a small number of keywords), a description (a sentence) and a narrative (a long description of the information need).

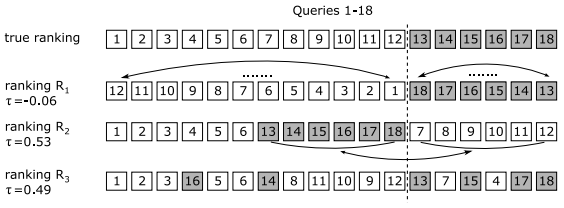


Fig. 1. Examples of predicted query rankings (R_1 to R_3) and their effect on SQE. Of the 18 queries, those ranked 1-12 (white) improve when AQE is employed, the results of queries ranked 13-18 (grey) degrade with the application of AQE.

feasibility of the approach, the QPP-based model selection strategy significantly outperforms the baseline.

Overall, few works have employed QPP in practice. Based on the results, we hypothesize, that it is easier to be successful in the MS than in the SQE setup.

Evaluation by Correlation. Evaluating QPP methods by reporting correlation coefficients is the current standard. The correlation based evaluation though can lead to problems: (i) very different predictions can lead to similar correlations, (ii) a QPP method for which a high (low) correlation is reported, may not lead to an increase (decrease) in retrieval effectiveness in an operational setting, and, (iii) following from (i) and (ii) is the observation that a single QPP method is not a reliable indicator of the level of correlation required, at which the application of a QPP method in practice is likely to lead to a consistent gain in retrieval effectiveness.

Fig. 1 contains a conrete example for Kendalls τ, the correlation coefficient we investigate in this work. The true ranking is the ranking of queries based on their retrieval effectiveness. Let us assume a SQE setup according to [18]: the queries ranked 1-12 benefit from AQE, the remaining queries do not. R_1, R_2 and R_3 are examples of predicted rankings of query performance. R_1 predicts all *ranks* incorrectly, thus $\tau_{R_1} = -0.06$ with respect to the true ranking. All AQE decisions though are correct, which leads to an optimal increase in retrieval effectiveness in SQE. The opposite case is R_2 with $\tau_{R_2} = 0.53$. The AQE decision is wrong for 12 queries, degrading the retrieval effectiveness. Based on the predicted ranking R_3, the wrong AQE decision is made for 4 queries. Although the retrieval effectiveness will be better than based on R_2, τ is similar: $\tau_{R_3} = 0.49$.

Thus, predictions resulting in similar correlations can have very different impacts on retrieval effectiveness, a problem which motivates our work.

3 Method and Materials

Ideally, we would like to perform the following experiment: given a large number of QPP methods, a large number of retrieval algorithms and a set of queries, let each QPP method predict the queries' quality. Evaluate the QPP methods in terms of τ, use the predictions in an operational setting and perform retrieval

experiments to derive baseline and QPP-based results. Finally, check at what level of τ the QPP-based results generally improve over the baseline. In practice, this is not feasible as state of the art QPP methods only reach up to $\tau \approx 0.55$.

Data Sets. To perform experiments on SQE and MS, test collections, sets of queries and their respective retrieval performances are required. To make our results independent of a particular retrieval approach, we utilize TREC data sets, in particular the runs submitted by the participants to different adhoc retrieval tasks. All submitted runs with a MAP greater than 0.15 are included in this study. The data sets used are as follows (in brackets the number of runs): TREC6 (49), TREC7 (77), TREC8 (103), TREC9 (53), TREC10 (59), Terabyte04 (35), Terabyte05 (50) and Terabyte06 (71).

QPP Method of Arbitrary Accuracy. As state of the art QPP methods achieve a limited τ, they are unsuitable for our investigation. Since τ is rank based, it is possible to construct predicted rankings of any level of τ, simply by randomly permutating the true performance ranking of queries. The smaller the number of permutations, the closer τ is to 1. The larger the number of permutations, the closer τ is to 0. From the full range of $\tau \in [-1, 1]$, sixteen intervals $\{c_{0.1}, ..., c_{0.85}\}$ of size 0.05 each were investigated, starting a $c_{0.1} = [0.1, 0.15)$ and ending at $c_{0.85} = [0.85, 0.9)$[3]. For each interval c_i, 1000 rankings were randomly generated with $\tau \in c_i$ with respect to the true ranking. We rely on such a large number of rankings due to the issues outlined in Fig. 1: a single predicted ranking can be misleading when applied in practice. By considering the results of 1000 rankings for each c_i, we can consider the change in retrieval effectiveness *on average*. Each predicted ranking is used in the SQE and MS experiments, which allows us to analyze the impact of varying levels of correlation against retrieval effectiveness.

4 Selective Query Expansion

We analyze the relationship between τ as evaluation measure of QPP methods and the change in retrieval performance when queries are expanded selectively in a setup analogous to [18]. The effect AQE has on retrieval effectiveness varies, depending on the AQE approach, the retrieval algorithm and the set of queries evaluated. While across the whole query set, AQE aids retrieval effectiveness (improvements range between $3 - 40\%$ [12,16,17]), not all queries benefit. The percentage of queries from a query set performing worse when AQE is applied varies between $20 - 40\%$ [1,10,12,16]. In [1,10] it is reported that the worst and the very best queries are hurt by AQE[4], whereas in [3,16] only the worst queries are reported to be hurt by AQE.

[3] This is sufficient, as negative correlations can be transformed into positive correlations by reversing the ranking and $\tau = 1$ indicates two perfectly aligned rankings.

[4] Further adding terms dilutes the results of the already highly performing queries.

4.1 Experimental Details

Let us for now assume that all well performing queries improve with AQE, while the worst performing queries all degrade with AQE. Let θ be a rank threshold. Our SQE setup is as follows: given a set of m queries, they are ranked according to their predicted performance. AQE is applied to the best $(\theta \times m - 1)$ performing queries, the remaining queries are not expanded. As this setup only requires predicted rankings, we can use our generated predicted rankings of arbitrary accuracy. To evaluate the retrieval effectiveness of SQE, we require pairs of baseline (no AQE) and AQE runs. Then, we perform SQE based on the predicted rankings and consider SQE successful if it improves over the retrieval effectiveness of the AQE run. We derive baseline and AQE run pairs from the runs in our data sets. As we are not interested in the result lists themselves, but in the effectiveness of each run on each query Q, we consider a run to consist of a list of average precision (AP) values, thus $run = (ap^{Q_1}, ap^{Q_2}, .., ap^{Q_m})$.

Run Pairs. Each run of our data sets is considered as a baseline run run_{base} (no AQE) for which a corresponding AQE run run_{aqe} is generated. Recall, that we work with the assumption that AQE improves the effectiveness of the well performing queries, while degrading the effectiveness of poorly performing queries. Thus, for each $ap^{Q_i}_{base}$ in run_{base}, a respective $ap^{Q_i}_{aqe}$ value in run_{aqe} is generated such that $ap^{Q_i}_{aqe} > ap^{Q_i}_{base}$ when $ap^{Q_i}_{base}$ is among the top $(\theta \times m - 1)$ performing queries in run_{base}, otherwise $ap^{Q_i}_{aqe} < ap^{Q_i}_{base}$. The $ap^{Q_i}_{aqe}$ values are randomly sampled (with the outlined restrictions) from the other runs in the data sets, a strategy supported by [2], where no correlation between $ap^{Q_i}_{base}$ and the amount of improvement, i.e. $\Delta = ap^{Q_i}_{aqe} - ap^{Q_i}_{base}$, $\Delta > 0$, was found. The optimal SQE run run_{opt} is the run where the correct AQE decision is made for every query: $ap^{Q_i}_{opt} = max(ap^{Q_i}_{base}, ap^{Q_i}_{aqe})$. We only include run pairs where the MAP of run_{aqe} improves by between $15 - 30\%$ over run_{base} and run_{opt} improves by at least 3% over run_{aqe}. Due to the random component in the process, 500 run pairs are created for each setting of $\theta = \{1/2, 2/3, 3/4\}$ and $m = \{50, 150\}$. The choice of θ is based on [12,16], the settings of m are typical TREC topic set sizes.

Experiment. Given the 1000 rankings per c_i and the 500 run pairs (run_{base}/ run_{aqe}) for each setting of θ and m, SQE is thus performed $500,000$ times for each c_i. From each run pair and predicted ranking in c_i a selective AQE run run_{sqe} is formed: if according to the predicted ranking $ap^{Q_i}_{base}$ is among the top $(\theta \times m - 1)$ scores in run_{base}, then $ap^{Q_i}_{sqe} = ap^{Q_i}_{aqe}$, that is the AQE result is used. The remaining queries are not expanded and $ap^{Q_i}_{sqe} = ap^{Q_i}_{base}$. Recorded are the MAP of run_{base}, run_{aqe}, run_{opt} and run_{sqe}. We consider SQE to be successful if the MAP of run_{sqe} is higher than the MAP of run_{aqe}. Since the (run_{base}/run_{aqe}) pairs lead to different absolute changes in retrieval performance, we report a normalized value: $v_{sqe} = 100(MAP_{sqe} - MAP_{base})/(MAP_{opt} - MAP_{base})$. When the correct AQE decision is made for each query, $v_{sqe} = 100$. In contrast, $v_{sqe} < 0$ if the MAP of run_{sqe} is below the baseline's run_{base} MAP.

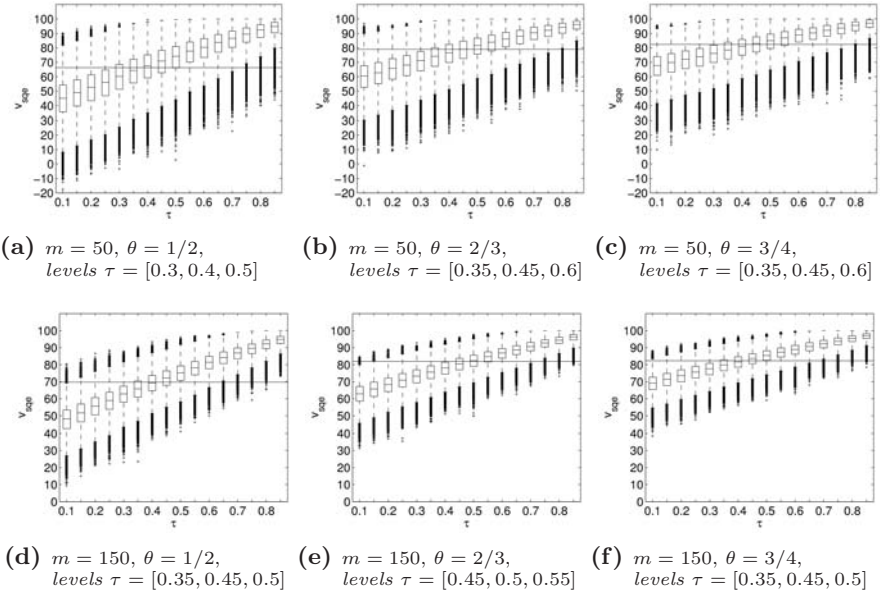
We present the results, derived for each c_i, in the form of box plots. Every box marks the lower quartile, the median and the upper quartile of $500,000$ v_{sqe} values. The whiskers mark the 1.5 inter-quartile range, the remaining separately marked points are outliers. We also plot the median normalized value of AQE runs as horizontal line, as it is the value v_{sqe} should improve upon.

4.2 Experimental Results

Best-Case Scenario. First, we evaluate the best-case scenario, where our assumption that AQE only hurts the worst performing queries holds for all run pairs. The results for the different settings of θ and m are presented in Fig. 2.

Independent of m and θ, for all correlation intervals c_i a number of positive outliers exist, where the MAP of run_{sqe} improves over run_{aqe}'s MAP. Thus, even if $\tau = 0.1$, a QPP method can be successful by chance. This supports the view that a single experiment and QPP method are inadequate indicators to show a QPP's method utility in practice.

When the correlation of the predicted ranking with respect to the ground truth is low, run_{sqe} may perform worse than run_{base} as observed for $\theta = 1/2$ and $m = 50$ (negative v_{sqe} values). This means that a poor QPP method may



(a) $m = 50$, $\theta = 1/2$, *levels* $\tau = [0.3, 0.4, 0.5]$ **(b)** $m = 50$, $\theta = 2/3$, *levels* $\tau = [0.35, 0.45, 0.6]$ **(c)** $m = 50$, $\theta = 3/4$, *levels* $\tau = [0.35, 0.45, 0.6]$

(d) $m = 150$, $\theta = 1/2$, *levels* $\tau = [0.35, 0.45, 0.5]$ **(e)** $m = 150$, $\theta = 2/3$, *levels* $\tau = [0.45, 0.5, 0.55]$ **(f)** $m = 150$, $\theta = 3/4$, *levels* $\tau = [0.35, 0.45, 0.5]$

Fig. 2. SQE best-case scenario. On the x-axis the starting value of each correlation interval c_i is listed, that is the results for $c_{0.2} = [0.2, 0.25)$ are shown at position 0.2. The horizontal lines mark the normalized median value of the performance of the AQE runs, which v_{sqe} must improve upon for SQE to be successful. "*levels* τ" lists the lower bound of c_i where the SQE runs outperform the AQE runs in at least 25%, 50% and 75% of all $500,000$ samples (these levels can be read from the box plots).

significantly degrade the effectiveness of a system. An increase in τ generally leads to a smaller spread in performance (the height of the boxes in the plot) of v_{sqe}, that is, outliers are rarer and the performance drop is not as sharp. Finally, the number m of queries also influences the outcome - with increased m the spread of the results decreases and the results can be considered to be more stable. This suggests that the more queries are used in the evaluation, the better the correspondence between the evaluation measure Kendalls τ and the performance in an operational setting.

Random Perturbations. We now investigate what happens when the assumption that all well performing queries improve with AQE is violated. This experiment is motivated by the divergent observations [1,10,12,16] about the change in effectiveness of the best performing queries when AQE is applied.

To simulate such violation, we perturb run_{aqe}. Given a run pair (run_{base}/run_{aqe}), we randomly select a query Q_i from the top $(\theta \times m - 1)$ performing queries of run_{aqe} and perturb its score $ap^{Q_i}_{aqe}$ to $\hat{ap}^{Q_i}_{aqe}$, which is a random value below $ap^{Q_i}_{base}$. To keep the MAP of run_{aqe} constant, the difference $(ap^{Q_i}_{aqe} - \hat{ap}^{Q_i}_{aqe})$ is randomly distributed among the other queries' ap scores. This procedure is performed for $p = \{10\%, 20\%, 30\%\}$ of the queries. The results of this experiment with fixed $\theta = 1/2$ are shown in Fig. 3. Evidently, already a small number of perturbed queries has a considerable influence on the minimum τ which is required to improve run_{sqe}

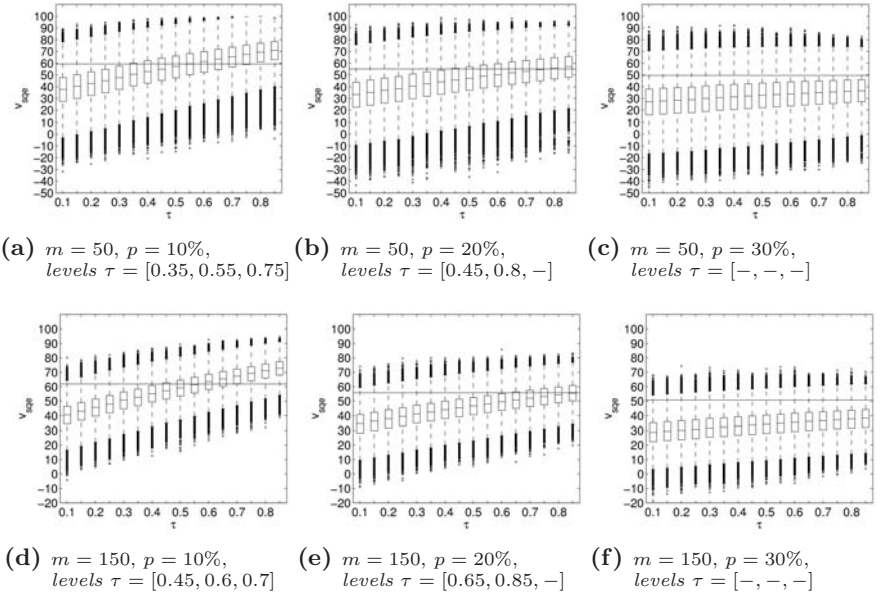


Fig. 3. SQE random perturbation scenario. $\theta = 1/2$ is fixed. "*levels* τ" lists the lower bound of c_i where the SQE runs outperform the AQE runs in at least 25%, 50% and 75% of all 500,000 samples.

over run_{aqe}. When $p = 10\%, \tau \geq 0.55$ is required to ensure that for more than 50% of the samples run_{sqe} improves over run_{aqe}. A further increase in the number of perturbed queries leads to a situation as visible for $p = 30\%$, where independent of the accuracy of the predicted ranking, in less than 25% of all instances run_{sqe} improves over run_{aqe}.

To summarize, in the best-case scenario, we have shown that a QPP method should evaluate to at least $\tau = 0.4$, for 50% of the samples to improve with the application of SQE. This threshold increases to $\tau = 0.55$, when the assumption we make about for what queries AQE is successful, is slightly violated. These results explain the low levels of success in [1,5,9,18], where QPP methods were applied in the SQE setting. Only the best performing QPP methods reach the levels of τ required to be beneficial. Moreover, making the wrong assumption about when AQE will lead to increased system effectiveness quickly leads to a situation where it does not matter anymore, how well a QPP method performs.

5 Meta-search

The second operational setting we examine is Meta-Search where we adopt a setup analogous to [14]: given a query and a set of t runs, we select the run that is predicted to perform best for the query. We chose this setup over the data fusion setup [18], where the merging algorithm introduces an additional dimension, as it allows us to better control the experimental setting.

5.1 Experimental Details

In preliminary experiments we found two parameters influencing the retrieval effectiveness of QPP-based meta-search: (i) the number t of runs, and (ii) the percentage γ of improvement in MAP between the worst and the best performing run in the set of t runs. The experimental setup reflects these findings. We derived 500 sets of t runs from our data sets for each setting of γ: $0-5\%$, $15-20\%$, $30-35\%$ and $50-55\%$. A range $0-5\%$ means, that all t runs perform very similar in terms of MAP, while in the extreme setting of γ, the MAP of the best run is $50-55\%$ better than the MAP of the worst run. To generate each set of runs, t runs are randomly selected from the data sets. A set is valid if the maximum percentage of retrieval improvement lies in the specified interval of γ. Recall, that 1000 predicted rankings exist per correlation interval c_i. As we require t predicted rankings per set of runs, t rankings are randomly chosen from all rankings of a given c_i. To avoid result artifacts due to relying on predicted ranks instead of scores[5], the i^{th} predicted rank is replaced by the i^{th} highest ap score of the run. The meta-search run run_{meta} is then created by selecting for each query the result of the run with the highest predicted ap score. The optimal run run_{opt} is derived by $ap^{Q_i}_{opt} = max(ap^{Q_i}_{run_1}, .., ap^{Q_i}_{run_t})$.

[5] An extreme example is the case where the worst performing query of run A has a higher average precision than the best performing query of run B.

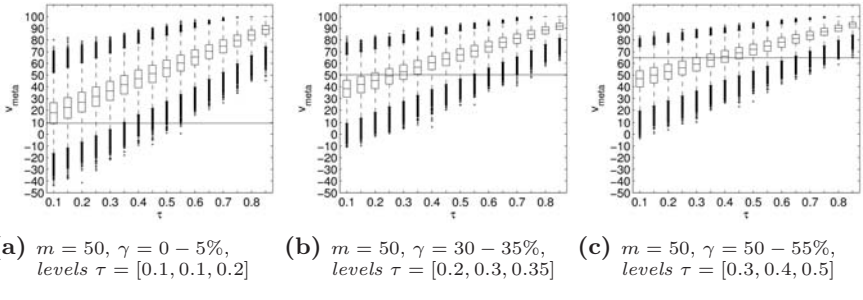
Analogous to the SQE experiments, we report the normalized performance of run_{meta}, that is, $v_{meta} = 100(MAP_{meta} - MAP_{worst})/(MAP_{opt} - MAP_{worst})$, where MAP_{worst} is the MAP of the worst of the t runs. When $v_{meta} < 0$, run_{meta} performs worse than the worst run of the set, a value of $v_{meta} = 100$ implies that run_{meta} is optimal. For MS to be successful, run_{meta} needs to perform better than the best run of the set. This threshold is indicated by the horizontal line in the plots (the normalized median of the best runs' MAP across all sets).

5.2 Experimental Results

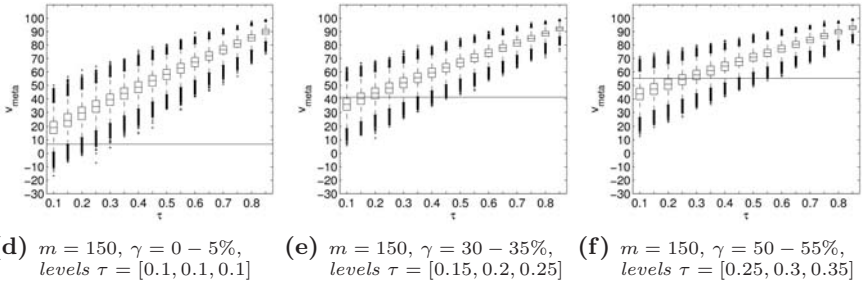
We experimented with $t = \{2, 3, 4, 5\}$. Due to space constraints though, in Fig. 4 we only report the results for $t = 4$ and $\gamma = \{0 - 5\%, 30 - 35\%, 50 - 55\%\}$.

Evidently, low correlations are sufficient to improve over the effectiveness of the best run. Note though, that for low correlations outliers exist that perform worse than the worst run (negative v_{meta}). As the performance difference between the best and the worst run increases (γ), higher correlations are required to improve run_{meta} over the best individual run. For $m = 50$ and $\gamma = 50 - 55\%$ for example, $\tau \geq 0.4$ is required to improve 50% of the samples. As in the case of the SQE experiments, an increase in m leads to more stable results.

The table in Fig. 5 contains basic statistics of the MAP values of the worst and best run of the sets of t runs as well as run_{opt}. The results were averaged over the 500 sets of runs for each m and γ. A different view on the sets of runs offers the percentage of queries in run_{opt} that are drawn from the worst performing



(a) $m = 50$, $\gamma = 0 - 5\%$, *levels* $\tau = [0.1, 0.1, 0.2]$ **(b)** $m = 50$, $\gamma = 30 - 35\%$, *levels* $\tau = [0.2, 0.3, 0.35]$ **(c)** $m = 50$, $\gamma = 50 - 55\%$, *levels* $\tau = [0.3, 0.4, 0.5]$



(d) $m = 150$, $\gamma = 0 - 5\%$, *levels* $\tau = [0.1, 0.1, 0.1]$ **(e)** $m = 150$, $\gamma = 30 - 35\%$, *levels* $\tau = [0.15, 0.2, 0.25]$ **(f)** $m = 150$, $\gamma = 50 - 55\%$, *levels* $\tau = [0.25, 0.3, 0.35]$

Fig. 4. MS experiments with $t = 4$ systems, m queries and varying γ

m	γ	% worst in opt.	MAP worst	MAP best	MAP opt
50	$0-5\%$	23.4%	0.236	0.245	0.328
	$15-20\%$	18.4%	0.220	0.259	0.334
	$30-35\%$	14.5%	0.209	0.277	0.344
	$50-55\%$	11.5%	0.196	0.298	0.356
150	$0-5\%$	23.5%	0.222	0.231	0.345
	$15-20\%$	20.0%	0.211	0.249	0.347
	$30-35\%$	17.5%	0.195	0.259	0.348
	$50-55\%$	14.5%	0.182	0.277	0.354

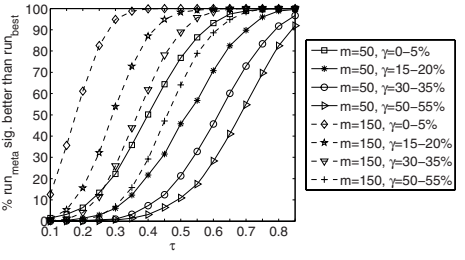


Fig. 5. The table contains the average MAP values over the 500 sets of $t = 4$ runs for each setting of m and γ. Listed are the MAP of the worst, best and optimal meta-search run.

The figure shows the development of the percentage of samples in which run_{meta} statistically significantly outperforms the best individual run (run_{best}) across the correlation intervals c_i.

run (column **% worst in opt.**). At $t = 4$, if the run selection is random, we expect 25% of the results to come from each run. As expected, with increasing γ, less results from the worst performing run are utilized in the optimal run. Though even for large differences in retrieval effectiveness ($\gamma = 50-55\%$), the worst performing run still contributes towards the optimal run.

We also investigate if the improvements of the MAP of run_{meta} are statistically significant[6]. We performed a paired t-test with significance level 0.05. Shown in Fig. 5 is the percentage of samples in which run_{meta} significantly outperforms the best individual run in the set of t runs. At $m = 50$, the threshold of τ to improve 50% of the samples significantly, ranges from 0.4 ($\gamma = 0-5\%$) to 0.7 ($\gamma = 50-55\%$). The thresholds are lower for $m = 150$: $\tau = 0.2$ ($\gamma = 0-5\%$) and $\tau = 0.5$ ($\gamma = 50-55\%$) respectively.

Thus, in the MS setting, QPP methods that result in low correlations can already lead to a gain in retrieval effectiveness. For these improvements to be statistically significant though, QPP methods with moderate to high correlations are needed.

6 Conclusions

We have investigated the relationship of one standard evaluation measure of QPP methods (Kendalls τ) and the change in retrieval effectiveness when QPP methods are employed in two operational settings: Selective Query Expansion and Meta-Search. We aimed to give a first answer to the question: when are QPP methods good enough to be usable in practice? To this end, we performed a comprehensive evaluation based on TREC data sets.

We found that the required level of τ depends on the particular setting a QPP method is employed in. In the case of SQE, in the best-case scenario, $\tau \geq 0.4$

[6] This analysis was not required in the SQE setup, due to the run construction.

was found to be the minimum level of τ for the SQE runs to outperform the AQE runs in 50% of the samples. In a second experiment we showed the danger of assuming AQE to behave in a certain way - slightly violating an assumption already requires QPP methods with $\tau \geq 0.75$ for them to be viable in practice.

The outcome is different for the MS experiments. Here, the level of τ is dependent on the performance differences of the participating runs. If the participating runs are similar, a QPP method with $\tau = 0.1$ is sufficient to improve 50% of the runs over the baseline. If the performance differences are great, $\tau = 0.3$ is required. To achieve statistically significant improvements for 50% of runs under large system differences, $\tau = 0.7$ ($m = 50$) and $\tau = 0.5$ ($m = 150$) are required.

These results indicate that (i) QPP methods need further improvement to become viable in practice, in particular for the SQE setting and (ii) that with increasing query set sizes m the evaluation in terms of τ relates better to the change in effectiveness in an operational setting.

We have restricted ourselves to an analysis of Kendalls τ as it is widely used in QPP evaluations. We plan to perform a similar analysis for the linear correlation coefficient r. In contrast to the rank-based τ, r is based on raw scores, which adds another dimension - the distribution of raw scores - to the experiments.

References

1. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness and selective application of query expansion. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004)
2. Billerbeck, B., Zobel, J.: When query expansion fails. In: SIGIR 2003, pp. 387–388 (2003)
3. Carmel, D., Farchi, E., Petruschka, Y., Soffer, A.: Automatic query refinement using lexical affinities with maximal information gain. In: SIGIR 2002, pp. 283–290 (2002)
4. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: SIGIR 2002, pp. 299–306 (2002)
5. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A framework for selective query expansion. In: CIKM 2004, pp. 236–237 (2004)
6. Diaz, F.: Performance prediction using spatial autocorrelation. In: SIGIR 2007, pp. 583–590 (2007)
7. Hauff, C., Azzopardi, L.: When is query performance prediction effective? In: SIGIR 2009, pp. 829–830 (2009)
8. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 43–54. Springer, Heidelberg (2004)
9. He, B., Ounis, I.: Combining fields for query expansion and adaptive query expansion. Inf. Process. Manag. 43(5), 1294–1307 (2007)
10. Kwok, K.: An attempt to identify weakest and strongest queries. In: SIGIR 2005 Query Prediction Workshop (2005)
11. Lavrenko, V., Croft, W.B.: Relevance based language models. In: SIGIR 2001, pp. 120–127 (2001)
12. Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: SIGIR 1998, pp. 206–214 (1998)

13. Vinay, V., Cox, I.J., Milic-Frayling, N., Wood, K.: On ranking the effectiveness of searches. In: SIGIR 2006, pp. 398–404 (2006)
14. Winaver, M., Kurland, O., Domshlak, C.: Towards robust query expansion: model selection in the language modeling framework. In: SIGIR 2007, pp. 729–730 (2007)
15. Wu, S., Crestani, F.: Data fusion with estimated weights. In: CIKM 2002, pp. 648–651 (2002)
16. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: SIGIR 1996, pp. 4–11 (1996)
17. Xu, J., Croft, W.B.: Improving the effectiveness of information retrieval with local context analysis. ACM Trans. Inf. Syst. 18(1), 79–112 (2000)
18. Yom-Tov, E., Fine, S., Carmel, D., Darlow, A.: Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In: SIGIR 2005, pp. 512–519 (2005)
19. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 52–64. Springer, Heidelberg (2008)
20. Zhou, Y., Croft, W.B.: Ranking robustness: a novel framework to predict query performance. In: CIKM 2006, pp. 567–574 (2006)
21. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: SIGIR 2007, pp. 543–550 (2007)