When is Query Performance Prediction Effective?

Claudia Hauff
University of Twente
Enschede, the Netherlands
c.hauff@ewi.utwente.nl

Leif Azzopardi
University of Glasgow
Glasgow, United Kingdom
leif@dcs.gla.ac.uk

ABSTRACT

The utility of Query Performance Prediction (QPP) methods is commonly evaluated by reporting correlation coefficients to denote how well the methods perform at predicting the retrieval performance of a set of queries. However, a quintessential question remains unexplored: how strong does the correlation need to be in order to realize an increase in retrieval performance? In this work, we address this question in the context of Selective Query Expansion (SQE) and perform a large-scale experiment. The results show that to consistently and predictably improve retrieval effectiveness in the ideal SQE setting, a Kendall's Tau correlation of $\tau \geq 0.5$ is required, a threshold which most existing query performance prediction methods fail to reach.

Categories and Subject Descriptors: H.3.4 Information Storage and Retrieval : Information Search and Retrieval

General Terms: Experimentation

Keywords: Evaluation, Query Performance Prediction

1. INTRODUCTION

Predicting the retrieval performance or the degree of difficulty of a query is a challenging research area which has attracted a significant amount of attention in recent years [2, 4, 6]. A reliable and accurate prediction mechanism would enable the development of adaptive components within retrieval systems. For instance, if the performance of a query is considered to be poor, remedial action can be taken by the system to try and ensure that the user's information needs are satisfied. This may be performed through asking for refinement of the query, or some subsequent automatic disambiguation process. On the other hand, if the performance of a query appears sufficiently good, the query can be further improved by some affirmative action such as query expansion.

While these are the perceived benefits of developing QPP methods, current evaluations seldom consider whether a QPP method actually realizes these presumed benefits. The focus of QPP evaluations has been on producing methods which increase the correlation[1] between the retrieval performance and the predicted performance. In this work we investigate the correspondence between the often reported rank correlation coefficient τ of a QPP method and the change in retrieval performance in the operational setting of SQE when QPP is applied. The idea of SQE is to utilize pseudo-relevance feedback in a query-adaptive manner instead of applying it uniformly to all queries. The rational being, that queries which perform well are also likely to benefit from automatic query expansion (AQE), while queries that perform poorly are likely to have their result's quality further decreased due to query drift [1]. Thus, if we can predict the performance of a query, we can selectively expand only the queries that are predicted to perform well according to the QPP method. Since, between one fifth and one third of queries perform worse when AQE is applied [3, 5], SQE aims to reduce any loss in the application of AQE. So our goal, here, is to estimate a lower bound for the strength of the correlation coefficient τ that a QPP method needs to achieve in order be likely to obtain increases in retrieval effectiveness when performing SQE. As a single QPP method and its SQE application is insufficient to draw conclusions about the relationship between the evaluation measure and retrieval performance, we rely on a large number of generated QPP and retrieval results to perform a large-scale experiment.

[1]Reported in the literature are usually Kendall's τ, Spearman's ρ and/or Pearson's r.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

2. EXPERIMENTAL SETUP

Ideally, we would like to perform the following experiment: given a large number of QPP methods, a large number of retrieval approaches and a set of queries, (1) let each QPP method predict the queries' performance and determine the method's performance in terms of τ, (2) use the predictions to determine which queries (not) to expand, (3) perform retrieval experiments with and without AQE and finally (4) determine at what level of τ the retrieval approaches generally show improvements on their selectively expanded results in comparison to the uniformly expanded results. In practice though, this approach is not feasible for two main reasons. Most importantly, existing QPP methods only reach correlations of $\tau \leq 0.5$ [4], which would not allow us to investigate the change in retrieval performance at higher correlations. Furthermore, not all retrieval approaches may strictly adhere to the SQE assumption of top performing queries improving when applying AQE, a noise factor which needs to be taken into account and controlled.

For these reasons, we relied on a "theoretical" QPP method to derive predictions across a wide range of $\tau \in [-1, 1]$. As τ is based on ranks, we can construct predicted query performance rankings by randomly permutating the true performance ranking of a set of queries. From the full range of

τ, we investigated sixteen τ-intervals of size 0.05, starting at $c_{0.1} = [0.1, 0.15)$ and ending at $c_{0.85} = [0.85, 0.9)$. For each τ-interval, 1000 different predicted rankings were generated.

In order to make our results generalizable and less dependent on a particular retrieval approach, we utilize the runs submitted to different TREC ad hoc retrieval tasks[2]. As we are not interested in the document rankings themselves, but in the performance of each run on each query, here, we consider a *run* to consist of a list of average precision (AP) scores. Let θ be the percentage of top performing queries of a set of queries that perform better when applying AQE and let m be the query set size. Based on the assumptions and observations made in the literature about AQE, we created 500 pairs of unexpanded (run_{base}) and AQE runs (run_{qe}) from the pool of available TREC runs for each setting of $\theta = \{50, 66, 75\}\%$ and $m = \{50, 150\}$, the latter being typical sizes of TREC query sets.

Each run_{base}/run_{qe} pair was derived by sampling AP values from the pool of TREC runs available for each query. In order to ensure that the generated run pairs adhere to the SQE assumptions, a valid run pair has to fulfill three restrictions: (i) the AP scores of the θ% top performing queries of run_{qe} outperform run_{base} and vice versa for the remaining queries, (ii) the mean average precision (MAP) of run_{qe} improves over run_{base} by between $15 - 30\%$ and (iii) the optimal SQE run's performance, where for each query the better of the two runs is chosen, increases by at least 3% over the MAP of run_{qe}.

Given the 1000 predicted rankings in each τ-interval and the 500 run pairs, SQE is thus performed $500,000$ times per τ-interval. For each run_{base}/run_{qe} pair and predicted ranking, the queries that are predicted to be among the top θ% are expanded (the run_{qe} result is picked), while the remaining queries are not (the run_{base} result is picked), resulting in QPP-based SQE runs (run_{psqe}).

In a second experiment, we consider the case of slightly violating the SQE assumption; the run_{base}/run_{qe} run pairs are "perturbed" such that $p = 10\%$ of the top θ% performing queries of run_{qe} are randomly assigned AP values which are lower than those of the respective queries of run_{base}. To keep the MAP constant, the difference between the old and newly assigned AP is randomly redistributed among the other queries of run_{qe}.

3. RESULTS

The application of QPP-based SQE can be considered successful if the MAP of run_{psqe} is higher than of run_{qe}. The results are summarized in Table 1. Reported are the minimum τ-intervals where the MAP of run_{psqe} is higher than of run_{qe} in $\{25, 50, 75\}\%$ of the $500,000$ cases. For instance, in the case of 50 queries with the expansion threshold θ set to 50%, the first τ-interval for which at least a quarter of the QPP-based SQE runs have a higher retrieval effectiveness than the uniformly expanded runs, is $c_{0.3} = [0.3, 0.35)$. Though not shown, we observed that for $\tau \leq 0.3$ and $m = 50$, run_{psqe} can lead to a MAP *below* run_{base}, which by way of construction performs at least 15% worse than run_{qe}. This changes as m increases, the results can be considered to be more stable and the outlying instances are less extreme.

The value of m has little influence though on the big picture; in order to improve 50% of all instances when applying QPP-based SQE, $\tau \geq 0.4$; this threshold increases to $\tau \geq 0.5$ when using the more reliable mark of 75% improving cases. Finally, the results of the perturbed setup in Table 1 show that even a small number of perturbed queries already has great influence on the usability of a QPP method in the SQE setting. If $p = 10\%$ of the top ranked queries are perturbed, QPP methods can still lead to improvements in retrieval effectiveness. However, the minimum τ-interval where 75% of the instances improve is considerably higher: $\tau \geq 0.7$. When further increasing the percentage of perturbed queries, no more consistent improvements in retrieval effectiveness can be observed in the SQE setting, independent of the quality of the predicted rankings.

[2]TREC-{6,7,8,9}, TREC-{2001,2004,2005,2006}

m		τ where X% SQE runs improve		
#Queries	θ	**25%**	**50%**	**75%**
50	50%	0.30	0.40	0.50
	66%	0.35	0.45	0.60
	75%	0.35	0.45	0.60
150	50%	0.35	0.45	0.50
	66%	0.45	0.50	0.55
	75%	0.35	0.45	0.50
Perturbations (10%)				
50	50%	0.35	0.55	0.75
150	50%	0.45	0.60	0.70

Table 1: Minimum τ-interval where the MAP of run_{psqe} is higher than of run_{qe} in $\{25, 50, 75\}\%$ of all instances.

4. CONCLUSIONS

In this work, we investigated the relationship between τ and the retrieval performance when applying QPP in the SQE setting. It was shown, that moderate to high τ coefficients are required to obtain reliable improvements in retrieval performance. However, if the assumptions behind the application of SQE are even slightly violated, the level of τ required increases considerably. Given that current state-of-the-art QPP methods only obtain low to moderate correlations, it is unlikely that they are able to realize any tangible and consistent increases in retrieval effectiveness. Furthermore, we need to emphasize, that our results can only serve as an estimate of the lower bound for the level of τ, as for instance we assumed θ to be known. Nonetheless, these findings show that realizing the potential of QPP is difficult to achieve in practice and requires considerable further research. As a next step, a similar analysis will be performed on other tasks that use QPP, along with considering other QPP evaluation measures (such as Spearman's ρ).

5. REFERENCES

[1] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness and selective application of query expansion. In *ECIR'04*, pages 127–137, 2004.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02*, pages 299–306, 2002.

[3] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *SIGIR '98*, pages 206–214, 1998.

[4] V. Vinay, I. J. Cox, N. Milic-Frayling, and K. Wood. On ranking the effectiveness of searches. In *SIGIR '06*, pages 398–404, 2006.

[5] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *SIGIR '96*, pages 4–11, 1996.

[6] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *SIGIR '07*, pages 543–550, 2007.