Can we predict QPP? An approach based on multivariate outliers

No Institute Given

Abstract. Query performance prediction (QPP) aims to predict the success and failure of a search engine on a collection of queries and documents. State of the art predictors can enable this prediction with a degree of accuracy; however, it is far from being perfect. Existing studies have mainly observed QPP is a difficult task but yet have lacked in-depth qualitative analysis. In this paper, we analyze QPP from the perspective of predicting the accuracy of query performance. Our working hypothesis is that certain queries lend themselves more easy to prediction while others pose greater challenges. Moreover, by focusing on outliers, we can pinpoint queries that are particularly difficult to predict. To achieve this, we consider multivariate outlier detection. Our results show the effectiveness of this approach in identifying queries for which QPP struggles to provide accurate predictions. Furthermore, we show that by excluding these difficult to predict queries, the overall accuracy of QPP is substantially improved.

Keywords: Information Retrieval · Query performance prediction · QPP · Post-retrieval features · Multivariate outlier detection

1 Introduction

A search engine aims to process and answer any user query by retrieving relevant documents. However, the performance of a particular search engine can vary substantially depending on the specific queries it encounters.

Query performance prediction (QPP) addresses the crucial task of predicting how effective a system will be on a given query. This problem is of paramount importance for two primary reasons. Firstly, when a query is expected to be difficult, it may be necessary to adopt a specialized, albeit potentially costly, approach. Conversely, when a query is predicted to be easy, a simpler and more cost-effective method can be employed.

QPP accuracy is typically assessed by measuring the correlation between the predicted and the actual performance values. This approach is sound as long as the underlying assumptions and conditions for applying correlation measurements are respected.

Systems vary in their approach to process a given query. This variability encompasses processes such as automatic query reformulation, the choice of the weighting/matching function (it can be BM25, language model, or a LLM-based

model for examples), and the application of document re-ranking models. Consequently, different systems will perform differently on the same queries. QPP is thus considered with regard to the system it predicts the performance for. This implies that a QPP predictor should adapt its behavior to suit a particular system. This could be a reason why post-retrieval QPPs, which use the results of the search, tend to be more accurate compared to pre-retrieval ones, which rely solely on the query and the set of documents.

We found for example that *LemurTF_IDF*, a post-retrieval QPP corresponding to a Letor feature [3, 5] has a higher correlation with the actual AP[1] obtained with the LGD weighting function [6] than with the JS weighting function [1] (resp. 0.522 and 0.504 Pearson correlation).

On the other hand, two predictors will behave differently on a particular system (See Figure 1 where two QPP values and actual AP are displayed for the TREC78 collection using the same particular search engine).

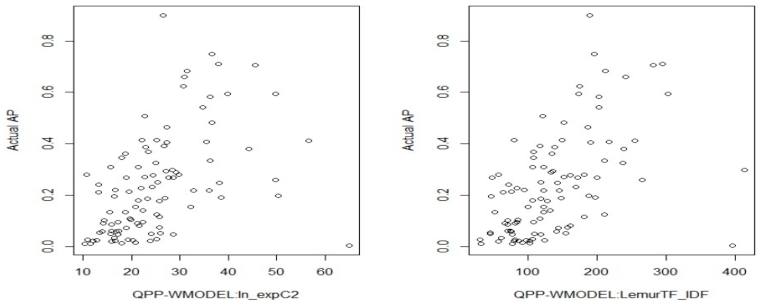

Fig. 1. QPPs behave differently on the same system and set of queries. Predicted AP (X axis) and actual AP (Y axis) obtained with LGD weighting function and their Pearson correlation. Left side In_expC2 (ρ=0.484) - Right side LemurTF_IDF QPP (ρ=0.552).

In Figure 1, we can see that the correlation is relatively weak, which is accurately reflected by the 0.484 (resp. 0.552) Pearson correlation values. Current QPP models lack accuracy. Single features, even post-retrieval ones could not demonstrate high correlation with actual performance [9, 4, 8]; even the combination of predictors, which is out of the scope of this paper, has not been very successful [7, 19, 15, 5, 12, 16].

In this paper, we aim at analyzing deeper predictor performance. Past studies showed predictors do not work well when considering a full set of queries. Our hypothesis is that the accuracy of prediction can significantly vary among queries and our objective is to identify the queries for which the predictor fails to estimate the effectiveness. In other words, we want to predict the predictability of performance and address the following research question: Is it possible to predict the queries for which a QPP can provide an accurate difficulty estimation

[1] Average precision

for a given system? To put it differently, can we concentrate on those queries that are more likely to yield accurate predictions and perhaps automatically disregard those queries that are more challenging to predict?

Some queries can be considered as outliers (*abnormally* easy or *abnormally* difficult); similarly, predictions may have abnormal values. Our hypothesis is that the queries with abnormal prediction values are difficult to predict or get non-accurate prediction.

In this paper, we consider multivariate outlier detection as a mean to identify these hard to predict queries. Since a given QPP may behave differently to predict the effectiveness of a system, the idea is to consider multiples QPPs in the query identification phase.

We consider several effectiveness measures and several benchmark collections. We show that our hypothesis could pave the way for a new research direction on QPP on the topic of the accuracy predictability.

2 Multivariate outliers to identify difficult to predict queries

Here, we hypothesize that some predictions are outliers because the queries are difficult to predict. And we want to identify those queries for which we anticipate the QPP will not be accurate.

Johnson defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data [11]. Univariate methods consider each variable independently, so only observations that appear odd for that variable are detected, while in the case of multivariate outlier detection, the interactions among different variables are compared. Multivariate outliers are a combination of unusual scores on several variables [14] and the idea is to detect the observations that are located relatively far from the center of the data distribution [2]. To identify outliers, we thus consider here multivariate outliers and identify the queries for which predictions are abnormal for different QPPs.

Mahalanobis distance is a common criterion for multivariate outlier detection. Applied to QPP, the Mahalanobis distance for a qiven query q_i, from a set of n queries $Q = \{q_1, q_2, \ldots, q_n\}$, can be defined as follows:

$$D_M(q_i, Q) = \sqrt{(q_i - \overline{q})^T V_n^{-1} (q_i - \overline{q})}, \tag{1}$$

where the query vector q_i is composed of m ($n >> m$) predictor values $q_i = \left(p_1^i, p_2^i, \ldots, p_m^i\right)$, $\overline{q}$ is the mean vector of the queries, and V_n^{-1} is the inverse of the covariance matrix of the queries. The superscript T denotes the transpose of the vector.

3 Data

In this study, we use 2 TREC collections (see details in Table 1). We use Average Precision (AP@1000) [17] and normalized Discounted Cumulative Gain

Table 1. Statistics of the collections used.

Collection	#Docs	Queries	Avg. Rel.	Avg. Irrel.
TREC78	528K	100 (351 – 450)	92.02	1577.73
WT10G	1.692M	100 (451 – 550)	59.80	1344.90

(nDCG@20) [10] as measures to evaluate the search engine performance which are common measures in adhoc retrieval. Like previous studies on QPP [7, 19, 4, 18, 15, 5], we use Pearson correlation as a measure to evaluate the accuracy of the QPP, in addition to plots as recommended in statistics.

We consider a series of systems that treat the same set of queries over the same set of documents. In this study, we construct systems using Terrier [13]. These systems differ on several factors: the scoring function employed, the variant of the query reformulation module utilized if any, and the number of documents and terms incorporated in the process of query reformulation. We report the results on the best system according to the considered collection and performance measure; best in terms of the average effectiveness over the set of queries.

In this study, we consider four Letor features which have been shown as among the most accurate for single feature QPP [3, 5]. Letor features have been initially used for retrieved document re-ranking [3]. Letor features are associated to each (query, retrieved document) pair [2]. To obtain a single value for each query for a given letor feature, the values are aggregated over the documents. We used maximum as the aggregation function which has been shown as the most accurate for QPP [5]. The four features we kept are LemurTF_IDF, In_expC2, InB2 and InL2, aggregated using max function.

4 Results

Outlier queries are different according to the considered QPP, which shows the importance of using multivariate outlier detection (See Figure 2). For example, query 403 is a clear outlier for InL2 QPP (top right of the right-side sub-figure) but not for the other QPPs. This implies that, had we used a univariate method such as LemurTF_IDF in isolation, we would not have identified this query as one that is difficult to predict, even though it truly is (See Figure 3). Similarly, consider Query 369 (shown at the top left in the left-side sub-figure). It clearly stands out as an outlier when evaluated using the LemurTF_IDF QPP, but does not exhibit the same outlier behavior when assessed by the other three QPPs.

These two queries (360 and 403) are indeed not well predicted (See Figure 3). In addition, we can see that query 356 has also been accurately identified as an outlier, as well as some other queries (See Figure 3). If these queries were not considered when calculating the correlation, the correlation would be higher.

When considering the two reference collections of this study, we can observe that the correlation when considering the outliers only is very weak (See Figure 4,

[2] Terrier implements these scores see `http://terrier.org/docs/v5.1/learning.html`

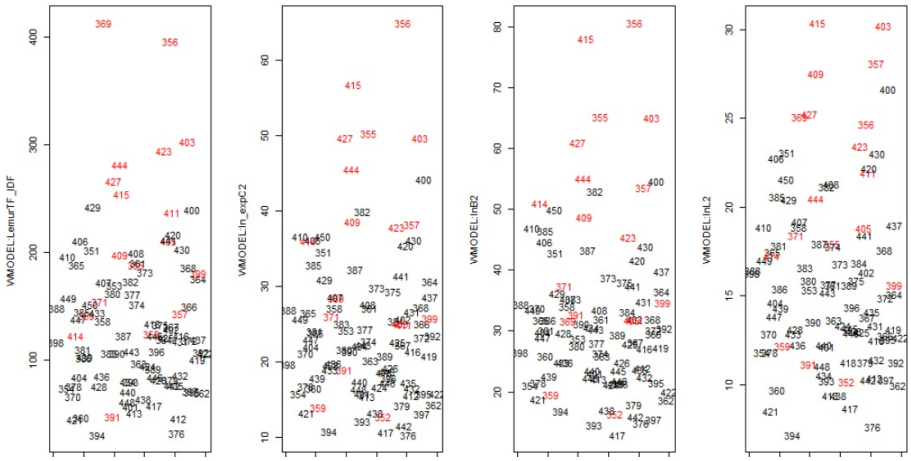

Fig. 2. Outlier queries are different according to the QPP. Queries that have been detected as outliers by our method are displayed in red. The four plots correspond to the four QPPs. The X-axis is not meaningful, it is just used to spread the queries so that we can read their numbers. Y-axis corresponds to the predicted value by the considered QPP. Here the QQP values are calculated for the TREC78 collection and a model based on the LGD weighting function.

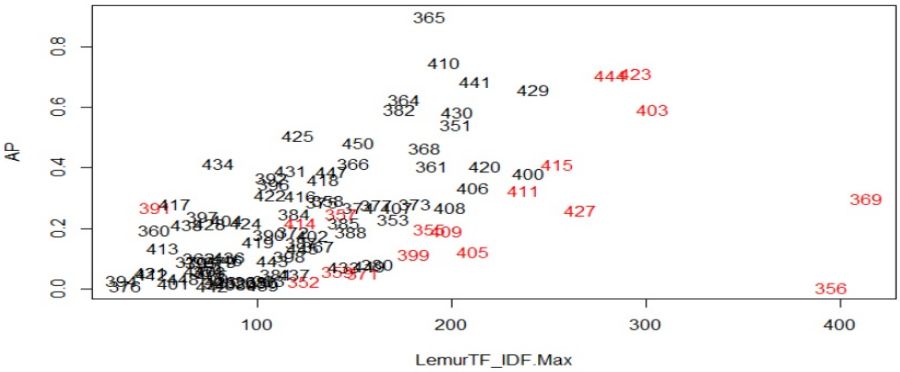

Fig. 3. Outlier queries are correctly identified. Queries that have been detected as outliers by our method are displayed in red. X-axis represents the QPP values and Y-axis is the actual AP. Here the values are calculated for LemurTF_IDF on the TREC78 collection and a model based on the LGD weighting function.

"LemurTF_IDF - Outliers only" row. This result was expected since we want to remove these difficult to predict queries. On the contrary, when the outlier queries are removed, the correlation between the QPPs and the actual effectiveness measures is much higher (see Figure 4 "No Outliers" rows). Note that if we use univariate outlier detection using LemurTF_IDF predictor only, the correlation decreases (0.658 compared to 0.700 here) (See also Figure 4 "Univariate" rows).

Results are consistent across the QPPs for TREC78 collection. This consistency remains valid for the best predictor LemurTF_IDF also for WT10G. We also evaluated the results considering other reference systems (although we keep including LLM-based model for future work) and the results were also consistent (e.g., using In_expB2 weighting function).

Collection	WT10G	WT10G	TREC78	TREC78
Measure	NDCG	AP	NDCG	AP
Best system	0.444	0.236	0.524	0.238
Ouliers	16	16	19	18
LemurTF_IDF - Univariate	0.337	0.342	0.544	0.658
LemurTF_IDF - Outliers only	0.206	0.292	0.095	0.350
LemurTF_IDF - No Ouliers	**0.438**	**0.468**	**0.601**	**0.700**
LemurTF_IDF - All	0.365	0.393	0.381	0.522
In_expC2 - - Univariate	0.423	0.368	0.607	0.631
In_expC2 - No Outliers	0.391	0.350	**0.607**	**0.635**
In_expC2 - All	0.425	0.371	0.418	0.484
InB2 - Univariate	0.329	0.286	0.542	0.536
InB2 - No Outliers	0.286	0.214	0.530	**0.543**
InB2 - All	0.336	0.274	0.372	0.416
InL2 - Univariate	0.264	0.341	0.380	0.426
InL2 - No Outliers	0.258	0.347	**0.458**	**0.491**
InL2 - All	0.340	**0.353**	0.398	0.446

Fig. 4. Pearson correlation is consistently better when queries our method detects as difficult to predict. Pearson correlation is calculated between actual effectiveness (either AP or ndcg) on the two TREC reference collections (TREC78 and WT10G). We report the number of outlier queries detected as well as the correlation when outliers have been removed using univariate and multivariate methods as well as and when all the queries are considered for the 4 Letor features. There is no statistical test to be applied here.

5 Conclusion

Studies on QPP generally focus on prediction accuracy [7, 19, 4, 18, 15, 5, 8], but seldom on the difficulty of that prediction. QPP is clearly a difficult task, since current predictors are not very accurate.

In this study we show that difficult to predict queries can be detected. We used multivariate outlier detection for that; it has the advantage to consider different QPPs to detect the queries for which the prediction may not be accurate. We also show that removing these automatically detected queries, we have a higher accuracy of the predictor. That means that we know that the predictor is not accurate for some queries that we can identify. This result pave the way to a new research direction: the prediction of the accuracy of the prediction. Some predictor may be accurate for certain queries and not for others. In future work, we will re-examine other QPPs from the literature such as Normalized Query Commitment (NQC) [18], Unnormalized Query Commitment (UCQ) [18], Query Feedback (QF) [19], Weighted Information Gain (WIG) [19] as well as other benchmark collections and reference retrieval systems.

References

1. Amati, G.: Frequentist and bayesian approach to information retrieval. In: European Conference on Information Retrieval. pp. 13–24. Springer (2006)
2. Ben-Gal, I.: Outlier detection. Data mining and knowledge discovery handbook pp. 131–146 (2005)
3. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: ICML. pp. 129–136 (2007)
4. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Morgan & Claypool Publishers (2010)
5. Chifu, A.G., Laporte, L., Mothe, J., Ullah, M.Z.: Query performance prediction focused on summarized letor features. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1177–1180 (2018)
6. Clinchant, S., Gaussier, E.: Information-based models for ad hoc ir. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. pp. 234–241 (2010)
7. De, R., Grivolla, J., Jourlin, P., de Mori, R.: Automatic classification of queries by expected retrieval performance. In: ACM SIGIR 2005 Workshop on Predicting Query Difficulty–Methods and Applications (2005) (2005)
8. Faggioli, G., Formal, T., Marchesin, S., Clinchant, S., Ferro, N., Piwowarski, B.: Query performance prediction for neural ir: Are we there yet? In: European Conference on Information Retrieval. pp. 232–248. Springer (2023)
9. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: ACM CIKM. pp. 1419–1420 (2008)
10. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: ACM SIGIR Forum. vol. 51, pp. 243–250. ACM, New York, NY, USA (2017)
11. Johnson, R.A., Wichern, D.W., et al.: Applied multivariate statistical analysis (2002)
12. Mizzaro, S., Mothe, J., Roitero, K., Ullah, M.Z.: Query performance prediction and effectiveness evaluation without relevance judgments: Two sides of the same coin. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 1233–1236 (2018)
13. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27. pp. 517–519. Springer (2005)
14. Peña, D., Prieto, F.J.: Multivariate outlier detection and robust covariance matrix estimation. Technometrics **43**(3), 286–310 (2001)
15. Raiber, F., Kurland, O.: Query-performance prediction: setting the expectations straight. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 13–22 (2014)
16. Roy, D., Ganguly, D., Mitra, M., Jones, G.J.: Estimating gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. Information Processing & Management **56**(3), 1026–1045 (2019). https://doi.org/https://doi.org/10.1016/j.ipm.2018.10.009, `https://www.sciencedirect.com/science/article/pii/S0306457318302437`
17. Sakai, T.: On the reliability of information retrieval metrics based on graded relevance. Information processing & management **43**(2), 531–548 (2007)

18. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting query performance by query-drift estimation. ACM Transactions on Information Systems (TOIS) **30**(2), 11 (2012)
19. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: ACM SIGIR. pp. 543–550 (2007)