Query Performance Prediction and Effectiveness Evaluation Without Relevance Judgments: Two Sides of the Same Coin

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

Josiane Mothe
ESPE, Université de Toulouse, IRIT, UMR5505 CNRS
Toulouse, France
Josiane.Mothe@irit.fr

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

Md Zia Ullah
UPS, Université de Toulouse, IRIT, UMR5505 CNRS
Toulouse, France
mdzia.ullah@irit.fr

ABSTRACT

Some methods have been developed for automatic effectiveness evaluation without relevance judgments. We propose to use those methods, and their combination based on a machine learning approach, for query performance prediction. Moreover, since predicting average precision as it is usually done in query performance prediction literature is sensitive to the reference system that is chosen, we focus on predicting the average of average precision values over several systems. Results of an extensive experimental evaluation on ten TREC collections show that our proposed methods outperform state-of-the-art query performance predictors.

CCS CONCEPTS

• **Information systems** → **Test collections**; *Retrieval effectiveness*;

KEYWORDS

Query difficulty prediction, AAP, Test collections, TREC

ACM Reference Format:
Stefano Mizzaro, Josiane Mothe, Kevin Roitero, and Md Zia Ullah. 2018. Query Performance Prediction and Effectiveness Evaluation Without Relevance Judgments: Two Sides of the Same Coin. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3209978.3210146

1 INTRODUCTION

Query Performance Prediction (QPP) is about predicting the effectiveness of the system for an unknown query [3, 19] while Effectiveness Evaluation without Relevance Judgments (EEwRJ) mainly tackles the problem of the cost of human relevance judgment by considering new methodologies to assess system effectiveness [15].

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210146

We consider these problems as the two sides of the same coin and we propose to combine these two research directions that so far have been treated independently. We show by extensive experiments on ten TREC collections that EEwRJ can be exploited to obtain a more accurate QPP than state-of-the-art.

In the following, we briefly review QPP and EEwRJ in Section 2, detail how EEwRJ can be adapted to QPP in Section 3, present our experiments in Section 4, and summarize our findings and sketch future developments in Section 5.

2 BACKGROUND

Query Performance Prediction. QPP aims at estimating system effectiveness for a given query [3, 19]. Current approaches consider either individual features [4, 9, 14, 20] or a combination of them [2, 7, 13, 20] to predict query performance. QPP accuracy is evaluated by means of correlation between the predicted AP and the real AP [3, 11].

The most effective individual predictors are the post-retrieval ones, which are calculated after the query has been submitted to the search engine considering the retrieved document list and document scores [3]. Although some of these features can be quite sophisticated (e.g. Weighted Information Gain which measures the divergence between the mean of the top-retrieved document scores and the mean of the entire set of document scores [20]), they only weakly correlate with actual system effectiveness [7, 11]: Pearson correlation with actual effectiveness is about 0.5 [14].

Since using one single query feature for QPP is not fully effective, combining features looks as a reasonable alternative. Current research mainly investigated linear regression [2, 6, 13, 20]. Thanks to these types of combination, the correlation has been slightly increased but remains well below 0.6.

Evaluation Without Relevance Judgments. The objective of all the EEwRJ methods[1] is to predict system effectiveness in a TREC-like environment. The first proposal was by Soboroff et al. [15], who proposed to randomly sample documents from the pool and treated such documents as relevant; the intuition is that if a document is retrieved by many systems in the top rank positions it will be pooled and thus it is probably a relevant document. Wu and Crestani [18] used data fusion techniques to merge the ranked lists retrieved by the systems and computed a score for each system based on

[1] To avoid confusion, we speak of QQP *approaches* and of EEwRJ *methods* in this paper.

the popularity of the documents it retrieves. Aslam and Savell [1] proposed an index based on the similarity between the ranked lists of systems; their index is computed simply considering the ratio between the document intersection and the document union of the ranked lists of each pair of systems.

Nuray and Can [10] adapted methods from democratic election strategies to compute the popularity score of each document by treating the documents as candidates and the systems as voters; more in detail, they used the "RankPosition," "Borda," and "Condorcet" methodologies. Spoerri [16] proposed a set of trials between systems and for each trial measures the percentage of documents retrieved by a system alone, by all the systems in the trial, and a combination of the previous percentage scores.

Diaz [5] embedded the retrieved documents in a high-dimensional space and computed spatial correlation values to measure document similarity and derived a predicted retrieval performance. Diaz [5] methodology is the only one which makes use of the collection documents; we leave such technique as future work. Sakai and Lin [12] used a variation of the Condorcet method from [10] which is less computationally demanding.

3 QPP BY MEANS OF EEWRJ

While QPP focuses on individual queries, EEwRJ focuses on average over queries. By focusing on a single effectiveness measure such as Average Precision (AP), we can say that QPP aims at predicting AP, while the EEwRJ aims at predicting Mean AP (MAP) for all the runs in a given TREC edition. Usually, the EEwRJ methods are evaluated by means of correlation, like QPP. However, while QPP approaches are evaluated by the Pearson correlation between predicted and real AP, EEwRJ methods are evaluated by the correlation between predicted and real MAP.

EEwRJ methods can be taken almost off-the-shelf and, with minor adaptations, exploited "as is" for QPP. Indeed EEwRJ methods can predict (by solving some normalization issues) not only MAP but also individual AP values for each ⟨system, topic⟩ pair. Following Mizzaro and Robertson [8], we can then derive a prediction of Average AP (AAP) which is the average across systems of AP for a given query ("topic" in TREC terminology).

Considering AAP does make sense for QPP since queries (or topics) which get a low AAP are difficult queries most systems failed on, and we should pay attention to and thus predict as difficult, On the other hand, queries which get high AAP are easy queries that any system can treat. In this paper, we thus focus on AAP as the measure to predict (while most of the papers from the literature consider AP [11, 14, 20]).

Moreover, we also combine the individual EEwRJ methods. So far the EEwRJ methods have been proposed individually, without any combination. Instead, we train a Machine Learning (ML) system that, on the basis of the TREC data of the previous years, learns a model that is then applied on a subsequent year TREC test collection (previous years test collections are the training set and the new test collection is then the test set). In other terms, the combination function, or the ML model, is the one that, on the basis of historical data, provides the best prediction of real AAP values given the individual EEwRJ outcomes.

Table 1: Name, acronym, and parameters used for EEwRJ.

Citation	Acronym	Name	Pool depth
Soboroff et al. [15]	SNC	Soboroff et al.	100
Wu and Crestani [18]	WUCv0	Basic	100
	WUCv1	Version 1	100
	WUCv2	Version 2	100
	WUCv3	Version 3	100
	WUCv4	Version 4	100
Aslam and Savell [1]	AS	Aslam & Savell	100
Nuray and Can [10]	NC-NRP	Normal Rank Position	30
	NC-NB	Normal Borda	30
	NC-NC	Normal Condorcet	30
	NC-BRP	Bias Rank Position	30
	NC-BB	Bias Borda	30
	NC-BC	Bias Condorcet	30
Spoerri [16]	SPO-S	Single	100
	SPO-A	All Five	100
	SPO-SA	Single Minus All Five	100
Sakai and Lin [12]	SL	Sakai and Lin	30

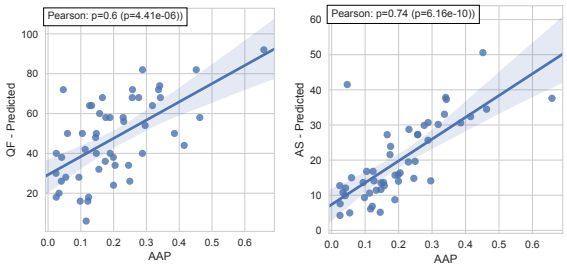
Table 2: Short description of the 10 TREC collections used.

Acron.	Collections	Corpus	Size	Topics
T6	TREC6 Adhoc	ROBUST	528K	50 (301 – 350)
T7	TREC7 Adhoc	ROBUST	528K	50 (351 – 400)
T01	TREC2001 Adhoc	WT10G	1.6M	50 (501 – 550)
R04	Robust 2004	ROBUST	528K	249 (301 – 450, 601 – 700)
R05	Robust 2005	ROBUST	528K	50 (301 – 700)
Tb04	Terabyte 2004	GOV2	25M	49 (701 – 750)
Tb05	Terabyte 2005	GOV2	25M	50 (751 – 800)
Tb06	Terabyte 2006	GOV2	25M	150 (701 – 850)
W13	Web Track 2013	ClueWeb12B	52M	50 (201 – 250)
W14	Web Track 2014	ClueWeb12B	52M	50 (250 – 300)

We use six ML algorithms [17]: Linear Regression (LR), M5P model tree (M5P), Random Forest (RF), Neural Networks (NN), Support Vector Machine with Polynomial kernel (SVM_Poly), and SVM with Radial Basis Function Kernel (SVM_RBF). These, in addition to 17 state-of-the-art EEwRJ individual methods from the previous work presented in Section 2 and summarized in Table 1, sum up to 23 methods used in the following experiments.

4 EXPERIMENTS AND RESULTS

Table 2 shows the ten TREC test collections used in our experiment. For a first evaluation of the predictive power of EEwRJ features, we considered each of the systems that participated to the corresponding TREC edition, predicted its AP using any individual features and then calculated the predicted AAP (by averaging the results across system per topic). Finally, we calculated the Pearson correlation between the predicted AAP and the actual AAP.



(a) QF vs actual AAP (b) AS vs actual AAP

Figure 1: TREC7 Adhoc collection. Pearson correlation between AAP and (a) QF [20], (b) AS [1]. While dots correspond to actual and predicted AAP for individual topics, the cone represents the confidence interval.

As for comparison, we calculated also the Pearson correlation, between the actual AAP and the value obtained when using state of the art QPP. As baselines to compare with, we consider the state of the art QPP approaches such as Unnormalized Query Commitment (UQC) [14], Query Feedback (QF) [20], Weighted Information Gain (WIG) [20], and Clarity [4].

To calculate the value of the state of the art QPP post-retrieval features, we used Language modeling. Thus while EEwRJ predictors are calculated for any (topic/participant system) pairs, QPP features are calculated only once for each topic.

For comparison purposes, these QPP approaches are also combined using machine learning algorithms including the same algorithms as previously mentioned (LR, M5P, RF, and SVM_RBF); these predictors are later referred to as ML QPP. The algorithms are trained to learn AAP and thus also predict AAP.

We found that EEwRJ individual features have a higher correlation with AAP than QPP individual features. As for example, Figure 1 reports the predicted values and actual AAP we obtained for TREC7 collection (a) QF, one of the best state of the art QPP feature (correlation value 0.599) and (b) ASLAM method, one of the EEwRJ (correlation value 0.744). The plots and the correlation values confirm that the AS method is a better predictor than QF.

To turn to a more systematic and complete analysis, Table 3 reports Pearson correlation of the predicted AAP values with the actual AAP of the participants' system, for each collection.

In the first two parts, on the top of the table, we report the state of the art baseline query performance predictors when calculated as previously mentioned, and their correlation with AAP. We report first individual predictors, second their combination using machine learning algorithms with leave-one-query-out cross-validation on each collection. Leave-one-query-out cross-validation is widely used in the field as in [14, 20].

The following two parts, on the bottom of the table, report the Pearson correlation values between the predicted AAP by the EEwRJ methods and the actual AAP.[2] First, the correlation values for the EEwRJ individual features (listed in Table 1) are reported

[2]AAP is obtained averaging for each topic the AP values of all systems which participated in a given TREC collection.

Table 3: Pearson correlation, over the ten TREC collections, between the actual AAP and the predicted AAP by the individual QPP, ML QPP, individual EEwRJ, and ML EEwRJ predictors. "‡", "†", and "*" stand for p-value < 0.001, < 0.01, and < 0.05, respectively. Values in bold are the largest in each part of the table for each collection.

Method	T6	T7	T01	R04	R05	Tb04	Tb05	Tb06	W13	W14
QPP										
UQC	**.606**‡	.493‡	.214	**.521**‡	.208	.161	.299*	.296‡	.102	.342*
WIG	.435†	.281*	.197	.356‡	.142	.223	.311*	.285‡	**.487**‡	**.422**†
QF	.368†	**.599**‡	.107	.409‡	**.268**	**.454**†	**.337***	**.392**‡	.009	-.121
Clarity	.415†	.587‡	**.316***	.476‡	.164	.251	.121	.136	-.430†	-.221
ML QPP										
LR	.490‡	.538‡	**.152**	**.569**‡	-.051	.251	.162	.382‡	.517‡	.490‡
M5P	**.529**‡	.578‡	-.077	.548‡	.049	**.327***	.155	.351‡	**.538**‡	**.605**‡
RF	.519‡	.597‡	.000	.549‡	**.076**	.195	.227	.312‡	.423†	.281*
SVM-RBF	.453‡	**.671**‡	.003	.501‡	.060	.268	**.285***	**.289**‡	.308*	.082
EEwRJ										
SNC	.268	.269	.253	.134*	.210	.590‡	.405†	.488‡	.460‡	**.656**‡
WUCv0	.156	-.068	.317*	.150*	.275	.673‡	.465‡	.474‡	.364†	.263
WUCV1	.180	-.039	.331*	.163*	.287*	.687‡	.477‡	.492‡	.372†	.293*
WUCV2	.175	-.023	.321*	.160*	.277	.665‡	.481‡	.478‡	.374†	.280*
WUCV3	.205	.020	.332*	.176†	.290*	.681‡	.493‡	.495‡	.381†	.317*
WUCV4	.159	.182	-.066	.0430	-.001	-.178	.238	-.069	-.010	-.089
AS	**.671**‡	**.744**‡	**.647**‡	**.683**‡	**.466**‡	.460‡	.476‡	.474‡	**.460**‡	.601‡
NC-NRP	-.314*	-.0160	.395†	.018	.261	.464‡	.282*	.139	.366†	.261
NC-NB	.384†	.311*	.484‡	.398‡	.379†	**.765**‡	.610‡	**.592**‡	.430†	.542‡
NC-NC	.399†	.341*	.453‡	.402‡	.373†	.761‡	.603‡	.585‡	.420†	.590‡
NC-BRP	.358*	.287*	.448†	.415‡	.352*	.761‡	.581‡	.533‡	.411†	.608‡
NC-BB	.344*	.281*	.473‡	.389‡	.357*	.761‡	.585‡	.534‡	.437†	.590‡
NC-BC	.513‡	.597‡	.584‡	.566‡	.464‡	.636‡	.550‡	.525‡	.446†	.657‡
SPO-S	.244	.112	.360*	.181†	.259	.625‡	.509‡	.501‡	.369†	.346*
SPO-A	.288*	.243	.308*	.231‡	.336*	.744‡	.504‡	.540‡	.245	.366†
SPO-SA	.265	.186	.332*	.219‡	.313*	.697‡	.518‡	.534‡	.316*	.361†
SL	.504‡	.475‡	.516‡	.502‡	.41†	.712‡	**.620**‡	.58‡	.419†	.629‡
ML EEwRJ										
LR	**.668**‡	**.761**‡	**.636**‡	.607‡	.457‡	**.644**‡	**.555**‡	.568‡	.439†	.201
M5P	.648‡	.693‡	.605‡	.565‡	.474‡	.602‡	.407†	.585‡	.429†	**.327***
NET	.582‡	.738‡	.634‡	.509‡	.479‡	.557‡	.367†	**.596**‡	.440†	.159
RF	.621‡	.685‡	.634‡	**.651**‡	**.519**‡	.578‡	.509‡	.589‡	.408†	.187
SVM_Poly	.656‡	.760‡	.620‡	.614‡	.465‡	.662‡	.549‡	.583‡	**.448**†	.204
SVM_RBF	.642‡	.752‡	.626‡	.613‡	.465‡	.680‡	.546‡	.575‡	.441†	.207

then the ones obtained when using the six ML based combination methods.

To help to understand these data, Figure 2 shows a graphical comparison of the Pearson correlation. The figure contains a series of box-plots for the individual EEwRJ methods and for the ML combination of the EEwRJ methods, as well as a series of point-plots for the individual QPP features, and for the ML combination of the QPP features, i.e., our baselines. We can draw several conclusions from this figure and the data from the Table 3. When comparing individual and combined baseline (QPP) predictors in the top part of the Table 3, we can see that in many cases the combination is better than individual features, although for some collections (e.g. TREC2001) the combination fails and single features are better.

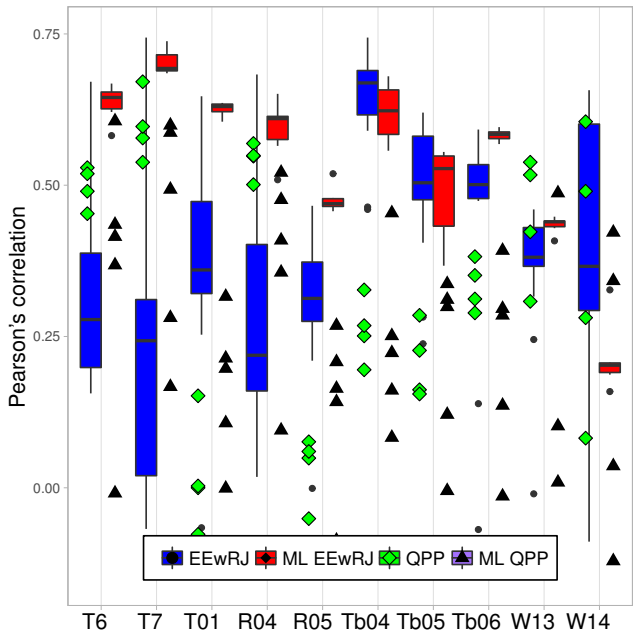


Figure 2: Comparison over the 10 collections of Pearson correlation between the actual AAP and the predicted AAP by the individual QPP, ML QPP, individual EEwRJ, and ML EEwRJ predictors. Boxplots are for EEwRJ while dots are for QPP.

When considering individual features, EEwRJ (specifically AS method) outperforms QPP baseline. When combining features, the method we propose based on EEwRJ also outperforms the combination of QPP. For example, the best correlation when combining EEwRJ methods is obtained for TREC7 where our combined methods get a correlation from .685 to .761 while ML QPP correlations are from .538 to .671, depending on the ML algorithm.

Turning to comparing individual and combined EEwRJ methods, we can clearly see that overall the combination of EEwRJ methods (ML EEwRJ) is better than the EEwRJ considered individually, although there are a few individual methods that outperform the ML combinations.

For all but one collection (W13) the best EEwRJ individual method outperforms all the baselines, and in all but two cases (W13 and W14) the EEwRJ ML methods outperform all the baselines as well. We also can see that apart for W14 collection, all the correlations are statistically significant which is not the case for QPP approaches that correlate only for some of the collections. Finally, the correlation values are much higher than any reported correlation when considering AP to be predicted [6, 11, 14].

Clearly, EEwRJ is an effective method for QPP: both as an individual predictor and when combined our method outperforms state of the art.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed to apply the methods to effectiveness evaluation without relevance judgments (EEwRJ) to the problem of query performance prediction (QPP). Our results clearly show that EEwRJ is an effective approach to QPP. If the AAP of a TREC topic is a reliable measure of query ease/difficulty, as it seems reasonable to assume, then it is possible to find specific EEwRJ methods (both individual and combined by means of ML) that outperform state-of-the-art query performance predictors.

In the future we plan to add more test collections to the analysis, for generality and also for a better understanding of the variation across datasets (e.g., W13 and W14 look different from the other collections). We will also take into account different correlation measures and effectiveness metrics. More in general, we believe that QPP and EEwRJ are "two sides of the same coin". Our approach clearly shows that they are related, and we plan in the future to explore and exploit their relationships in a complete way.

REFERENCES

[1] Javed A. Aslam and Robert Savell. 2003. On the Effectiveness of Evaluating Retrieval Systems in the Absence of Relevance Judgments. In *Proceedings of 26th ACM SIGIR*. 361–362.

[2] Shariq Bashir. 2014. Combining Pre-retrieval Query Quality Predictors Using Genetic Programming. *Applied Intelligence* 40, 3 (April 2014), 525–535.

[3] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval* (1st ed.). Morgan and Claypool Publishers.

[4] Steve Cronen-Townsend and W. Bruce Croft. 2002. Quantifying Query Ambiguity. In *Conference on Human Language Technology Research*. 104–109.

[5] Fernando Diaz. 2007. Performance Prediction Using Spatial Autocorrelation. In *Proceedings of 30th ACM SIGIR*. 583–590.

[6] Claudia Hauff. 2010. Predicting the effectiveness of queries and retrieval systems. *SIGIR Forum* 44, 1 (2010), 88. https://doi.org/10.1145/1842890.1842906

[7] Claudia Hauff, Djoerd Hiemstra, Leif Azzopardi, and Franciska de Jong. 2010. A Case for Automatic System Evaluation. In *Proceedings of ECIR (LNCS)*, Vol. 5993. 153–165.

[8] Stefano Mizzaro and Stephen Robertson. 2007. HITS hits TREC: exploring IR evaluation results with network analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 479–486.

[9] Josiane Mothe and Ludovic Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM SIGIR, Predicting query difficulty-methods and applications workshop*. 7–10.

[10] Rabia Nuray and Fazli Can. 2006. Automatic ranking of information retrieval systems using data fusion. *Information Processing & Management* 42, 3 (May 2006), 595–614.

[11] Fiana Raiber and Oren Kurland. 2014. Query-performance prediction: setting the expectations straight. In *ACM SIGIR*. ACM, 13–22.

[12] Tetsuya Sakai and Chin-Yew Lin. 2010. Ranking Retrieval Systems without Relevance Assessments — Revisited. In *Proceeding of 3rd EVIA — A Satellite Workshop of NTCIR-8*. National Institute of Informatics, Tokyo, Japan, 25–33.

[13] Anna Shtok, Oren Kurland, and David Carmel. 2010. Using statistical decision theory and relevance models for query-performance prediction. In *ACM SIGIR*. 259–266.

[14] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting Query Performance by Query-Drift Estimation. *ACM Trans. Inf. Syst.* 30, 2, Article 11 (May 2012), 35 pages.

[15] Ian Soboroff, Charles Nicholas, and Patrick Cahan. 2001. Ranking Retrieval Systems Without Relevance Judgments. In *Proceedings of 24th ACM SIGIR*. 66–73.

[16] Anselm Spoerri. 2007. Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Processing & Management* 43, 4 (2007), 1059 – 1070.

[17] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[18] Shengli Wu and Fabio Crestani. 2003. Methods for Ranking Information Retrieval Systems Without Relevance Judgments. In *Proceedings of the 2003 ACM Symposium on Applied Computing*. 811–816.

[19] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In *Proceedings of 30th ECIR*. 52–64.

[20] Yun Zhou and W Bruce Croft. 2007. Query performance prediction in web search environments. In *ACM SIGIR*. 543–550.