Normalized Query Commitment Revisited

Haggai Roitman
IBM Research - Haifa
haggai@il.ibm.com

ABSTRACT

We revisit the Normalized Query Commitment (NQC) query performance prediction (QPP) method. To this end, we suggest a scaled extension to a discriminative QPP framework and use it to analyze NQC. Using this analysis allows us to redesign NQC and suggest several options for improvement.

ACM Reference Format:
Haggai Roitman. 2019. Normalized Query Commitment Revisited. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331334

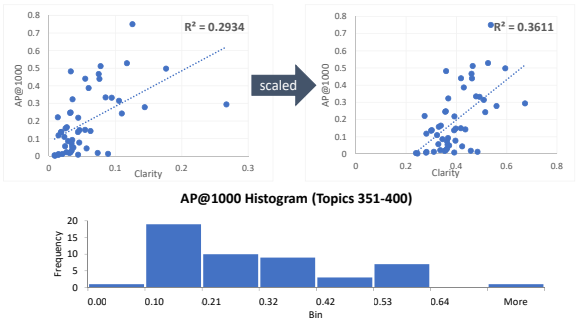

Figure 1: Example of scaling Clarity's values (TREC Robust queries 351-400, $\gamma = 0.3$)

1 INTRODUCTION

Query performance prediction (QPP) is a core IR task whose primary goal is to assess retrieval quality in the absence of relevance judgements [1]. In this work, we revisit Shtok et al.'s *Normalized Query Commitment* (NQC) QPP method [10]. NQC is a state-of-the-art post-retrieval QPP method [1], based on document retrieval scores variance analysis. Nowadays, the NQC method serves as a common competitive baseline in many QPP works [4, 6, 8, 11].

We first present the NQC method and shortly discuss the motivation behind it. We then shortly present Roitman et al.'s discriminative QPP framework [8]. Using a scaled extension to this framework, we analyze the NQC method, "deconstructing" it into its most basic parts. Based on this analysis, we revise NQC's design and suggest several options for improvement. Using an extensive evaluation over common TREC corpora, we demonstrate that, by redesigning NQC, where we extend it with more proper calibration and scaling, we are able to significantly improve its prediction accuracy.

2 BACKGROUND

2.1 Normalized Query Commitment

Let q denote a query and let D denote the respective ranked-list of top-k documents in corpus C with the highest retrieval scores. Let $s(d)$ further denote document d's ($\in D$) retrieval score with respect to q. The NQC method estimates the query's performance according to the standard-deviation of D's document retrieval scores, further normalized by the corpus score[1] $s(C)$, formally:

$$NQC(D|q) \stackrel{def}{=} \frac{\sqrt{\frac{1}{k}\sum_{d\in D}(s(d)-\hat{\mu}_D)^2}}{s(C)}, \quad (1)$$

where $\hat{\mu}_D \stackrel{def}{=} \frac{1}{k}\sum_{d\in D} s(d)$ is D's mean document score. The key idea behind NQC is the assumption that the mean score $\hat{\mu}_D$ may serve as a *pseudo-ineffective reference (score) point* of the retrieval. The more the retrieval scores deviate from this point, the less chance is assumed for a query-drift within D; and hence, a more qualitative retrieval will be predicted [10]. The corpus score $s(C)$ further serves as a query-sensitive normalization term, allowing to compare NQC values across different queries [10].

2.2 Discriminative QPP

In this work, we build on top of Roitman et al.'s discriminative QPP framework [8]. In [8], the authors have shown that many of the previously suggested post-retrieval QPP methods (e.g., Clarity [2], WIG [12] and SMV [11]) share the following basic form:

$$WPM(D|q) \stackrel{def}{=} \frac{1}{k}\sum_{d\in D} s(d)\cdot\phi_F(d), \quad (2)$$

where $\phi_F(d) \stackrel{def}{=} \prod_j \left(f_j(d)\right)^{\alpha_j}$ is a *Weighted Product Model* (WPM) discriminative calibrator; with $f_j(d)$ represents some retrieval quality feature and $\alpha_j \geq 0$ denotes its relative importance [8]. Within this framework, $\phi_F(d)$ calibrates each document d's ($\in D$) retrieval score $s(d)$ according to the likelihood of d being a relevant response to query q [8]. To this end, $\phi_F(d)$ may encode various retrieval quality properties, such as properties of q, C, D and the document d itself. As Roitman et al. have pointed out, some of such properties may be complementing each other (e.g., query vs. corpus quality

[1] Such a score is obtained by treating the corpus as a single (large) document.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331334

effects), and therefore, tradeoffs in the design of general QPP methods should be properly modeled [8]. $\phi_F(d)$, therefore, models such tradeoffs (i.e., using the weights α_j).

2.3 Scaled Weighted Product Model

While most of existing QPP methods are designed to predict a given quality measure (with AP@1000 being the most commonly used measure [1]), the relationship between a given predictor's estimates and actual quality numbers may not be necessarily linear. As an example, the bottom part of Figure 1 depicts an histogram of the AP@1000 values obtained by a query-likelihood (QL) based-retrieval method, which was applied over TREC Robust topics 351-400. As we can observe, there is a high variability in quality values, having query difficulty non-uniformly distributed.

To address such variability during prediction, we now suggest a simple, yet effective extension to [8], which scales the calibrated-mean estimator defined in Eq. 2, as follows:

$$SWPM(D|q) \stackrel{def}{=} \left(\frac{1}{k}\sum_{d\in D} s(d)\cdot\phi_F(d)\right)^{\gamma}, \quad (3)$$

where $\gamma \geq 0$ is a scaling parameter. Note that, whenever $\gamma < 1$, higher variability in prediction values is encouraged. Going back to Figure 1, its upper part further illustrates the relationship between Clarity's [2] predicted values and actual quality numbers before and after applying scaling ($\gamma = 0.3$). As we can observe, after scaling, Clarity's prediction accuracy has significantly improved.

3 NQC REVISITED

We now show that, the NQC method can be derived as a scaled calibrated-mean estimator. To this end, we first rewrite NQC (defined in Eq. 1), as follows:

$$NQC(D|q) \stackrel{def}{=} \left[\frac{1}{k}\sum_{d\in D} s(d)\left(\frac{1}{s(C)}\right)^2\left(\frac{s(d)-\hat{\mu}_D}{\sqrt{s(d)}}\right)^2\right]^{0.5} \quad (4)$$

Using Eq. 4, it now becomes apparent that, NQC may be treated as an instance of the generic scaled estimator defined in Eq. 3. In NQC's case, we have: $f_1(d) = \frac{1}{s(C)}$ and $\alpha_1 = 2$, $f_2(d) = \frac{s(d)-\hat{\mu}_D}{\sqrt{s(d)}}$ and $\alpha_2 = 2$, and a scaling parameter $\gamma = 0.5$.

Reformulating the NQC method as a scaled calibrated-mean estimator has two main advantages. First, being represented as a discriminative QPP method, we can potentially improve its prediction accuracy by applying supervised-learning to better tune its calibration (i.e., α_1 and α_2) and scaling (i.e., γ) parameters. As we shall later demonstrate, such a more fine-tuned calibration may indeed result in a better prediction quality.

Second, and more important, similar to [8], such a representation allows to analyze the existing design of NQC's calibration features (i.e., $f_1(d)$ and $f_2(d)$), potentially even redesigning some of them. Among these two features, $f_1(d)$ was previously examined in [8], where it was shown to be utilized as a calibration feature within the SMV method [11]. We, therefore, next focus our attention on the second calibration feature of $f_2(d)$.

3.1 Analysis of $f_2(d)$

Using the new NQC formulation, we now try to explain the motivation behind its second calibration feature $f_2(d) = \frac{s(d)-\hat{\mu}_D}{\sqrt{s(d)}}$. To this end, we "break" $f_2(d)$ into two main parts. The first, given by $s(d) - \hat{\mu}_D$, measures the difference between a given document d's ($\in D$) score and that of the pseudo-ineffective reference (score) point, considered by NQC as the mean document score $\hat{\mu}_D$. The larger the difference is, the more will document d's own score be estimated as an informative evidence for a true view of its relevance (i.e., $s(d) > \hat{\mu}_D$) or irrelevance (i.e., $s(d) < \hat{\mu}_D$). The second part, given by $\sqrt{s(d)}$, basically serves as a scaler, allowing to measure score differences with respect to some comparable scale. In NQC's case, scaling is simply performed using the document score itself.

3.2 Redesigning $f_2(d)$

We now generalize the basic form of $f_2(d)$, and reformulate it as follows:

$$f_2(d) \stackrel{def}{=} \frac{s(d) - s_{neg}}{\sqrt{Z_d}}, \quad (5)$$

where s_{neg} represents some pseudo-ineffective reference (score) point for comparison, and Z_d denotes some scaler value. Therefore, for the original NQC method we have the configuration of $s_{neg} = \hat{\mu}_D$ and $Z_d = s(d)$.

We next suggest several other configurations, based on alternative s_{neg} and Z_d instantiations. Starting with Z_d, as a first alternative, we suggest $s_{max} \stackrel{def}{=} \max_{d\in D} s(d)$. As a second alternative, motivated by Ozdemiray et al.'s **ScoreRatio** QPP method [6], we also suggest $s_{min} \stackrel{def}{=} \min_{d\in D} s(d)$. As a third alternative, we suggest a range-sensitive scaler: $s_{max} - s_{min} + \epsilon$, where $\epsilon \stackrel{def}{=} 10^{-10}$ is used to avoid dividing by zero. Noting that both s_{max} and s_{min} may actually represent outlier values, as a fourth and final alternative, which tries to reduce such outliers, we suggest the following inter-percentile range scaler $Q_{.95} - Q_{.05} + \epsilon$. Here, $Q_{.x}$ represents the value of the x% percentile of D's observed score distribution.

Next, we suggest several alternative estimates of s_{neg}. As a first alternative, we suggest s_{min}, which represents the document with the lowest relevance score, and hence, presumably the least effective sample reference. As a second alternative, following [4] we consider s_{neg} to be the mean score of the documents at the lower ranks of D. Formally, for a given $l \leq k$, we estimate $\hat{\mu}_{D_{neg}} \stackrel{def}{=} \frac{1}{k-l}\sum_{i=l+1}^{k} s(d_i)$, where d_i denotes the document that is ranked at position i in D.

Our third and final alternative is motivated by Diaz's **Autocorrelation** QPP method [3]. Using the opposite of the logic[2] that was suggested in [3], which is based on the *Cluster-Hypothesis in IR* [5], we expect an informative document score $s(d)$ to be quite different from those scores of documents that are the most dissimilar to d. Formally, let $dist(d, d')$ denote the distance between two documents in D and let $KFN(d)$ be the set of K-farthest neighbors

[2]The Autocorrelation method treats similar documents as pseudo-effective reference points, and hence, the score of a document is actually supposed to be similar to the scores of its closest neighbors.

of d in D according to $dist(d, d')$. We then define s_{neg} as the mean score of documents in $KFN(d)$.

4 EVALUATION

4.1 Experimental setup

4.1.1 Datasets. For our evaluation we have used the following TREC corpora and topics: **TREC4** (201-250), **TREC5** (251-300), **AP** (51-150), **WSJ** (151-200), **ROBUST** (301-450, 601-700), **WT10g** (451-550) and **GOV2** (701-850). These datasets were used by many previous QPP works [1–3, 7–12]. Titles of TREC topics were used as queries, except for TREC4, where we used topic descriptions instead [8]. We used the Apache Lucene[3] open source search library for indexing and searching documents. Documents and queries were processed using Lucene's English text analysis (i.e., tokenization, lowercasing, Porter stemming and stopwords). For retrieval, we used Lucene's language-model-based cross-entropy scoring with Dirichlet smoothed document language models, where the smoothing parameter was set to 1000 [8, 10].

4.1.2 Baseline predictors. We compared the original NQC method (see again Eq. 1) and its proposed extensions to several different baseline QPP methods. As a first line of baselines, we compared with the following state-of-the-art post-retrieval QPP methods:

- **Clarity** [2]: estimates query performance relatively to the divergence between the relevance model induced from a given (retrieved) list and the corpus background model.
- **WIG** [12]: estimates query performance according to the difference between the average document retrieval score in a given list and that of the corpus $s(C)$.
- **WEG** [4]: a WIG-like alternative that uses $\hat{\mu}_{D_{neg}}$ as the pseudo-ineffective reference point instead of $s(C)$.
- **Autocorrelation** [3]: based on the *Cluster-Hypothesis in IR*, it estimates query performance according to the Pearson's-r correlation between D's original document scores and those estimated by interpolating the scores of each document d's ($\in D$) K-nearest neighbors relatively to that document's similarity with each neighbor.
- **ScoreRatio** [6]: simply estimates query performance according to the ratio $\frac{s_{\max}}{s_{\min}}$.
- **SMV** [11]: is a direct alternative to NQC that also considers score magnitude and variance, estimated as:

$$SMV(D|q) \stackrel{def}{=} \frac{1}{k \cdot s(C)} \sum_{d \in D} s(d) \left| \ln \frac{s(d)}{\hat{\mu}_D} \right|.$$

Using our derivation of NQC as a scaled calibrated QPP method (see Section 3), we further evaluated various alternatives of NQC, as follows. We first evaluated various $f_2(d)$ configurations within NQC, i.e., **NQC**(s_{neg},Z_d). To this end, we instantiated s_{neg} and Z_d according to our proposed alternatives.

Next, using the original NQC configuration (i.e., $s_{neg} = \hat{\mu_D}$ and $Z_d = s(d)$), we also evaluated a calibration-only version of NQC (**C-NQC**), where we only tuned its α_1 and α_2 parameters, while still fixing the scaling parameter to $\gamma = 0.5$. In a similar manner, we evaluated a scaled-only version of NQC (**S-NQC**), where we fixed $\alpha_1 = 2$ and $\alpha_2 = 2$ and only tuned the scaling parameter γ. We further evaluated a combined predictor (**SC-NQC**), where all the three parameters were tuned. Finally, we evaluated a pre-tuned **SC-NQC** employed with the best $f_2(d)$ configuration learned for the non-calibrated/non-scaled NQC versions, denoted **SC-NQC**(best).

4.1.3 Setup. On each setting, we predicted the performance of each query with respect to its top-1000 retrieved documents [1]. We assessed prediction over queries quality according to the correlation between the predictor's values and the actual average precision (AP@1000) values calculated using TREC's relevance judgments [1]. To this end, we report the Pearson's-r (P-r) and Kendall's-τ (K-τ) correlations, which are the most commonly used measures [1].

Most of the methods that we have evaluated required to tune some free parameters. First, following the common practice [1], for each QPP method, we tuned k – the number of documents used for prediction[4]; with $k \in \{5, 10, 20, 50, 100, 150, 200, 500, 1000\}$. For **Clarity**, we induced a relevance model using the top-m ($\in \{5, 10, 20, 50, 100\}$) documents in D and further applied clipping at the top-n terms cutoff, with $n \in \{10, 20, 50, 100\}$. For the **Autocorrelation** and **NQC**($KFN(K), \cdot$) baselines, we tuned $K \in \{3, 5, 10, 20, 30\}$, further using the *Bhattacharyya* distance between the unsmoothed language models of the documents in D as the (dis)similarity measure $dist(\cdot)$ of choice. To realize $\hat{\mu}_{D_{neg}}$ in **SMV** and **NQC**, we further tuned $l \in \{5, 10, 20, 50, 100, 200, 500\}$.

To learn the calibration feature weights (i.e., α_1 and α_2) and scaling parameter (i.e., γ) of the NQC variants, following [7–9], we used a *Coordinate Ascent* approach. To this end, we selected the feature weights over the grid of $[0, 5] \times [0, 5] \times [0, 1]$, in steps of 0.1 per dimension. Following [7–10], we trained and tested all methods using a holdout (2-fold cross validation) approach. On each dataset, we generated 30 random splits of the query set; each split had two folds. We used the first fold as the (query) train set. We kept the second fold for testing. We recorded the average prediction quality over the 30 splits. Finally, we measured statistical significant differences of prediction quality using a two-tailed paired t-test with $p < 0.05$ computed over all 30 splits.

4.2 Results

We report our evaluation results in Table 1. As a "stand-alone" QPP method, even the original NQC method (i.e., **NQC**($\hat{\mu}_D$,$s(d)$)) already provides highly competitive prediction quality results compared to the other state-of-the-art QPP methods. We, therefore, next evaluate the impact of our proposed NQC extensions.

4.2.1 Evaluation of various $f_2(d)$ configurations. We start with a **qualitative** examination of the relative contribution of each of the two parts of $f_2(d)$ (i.e., s_{neg} and Z_d). To this end, we count the relative number of cases, per dataset and quality measure (i.e., P-r or K-τ), in which using a specific configuration part has resulted in a better prediction accuracy than that of the original NQC's configuration. Overall, in 136 out of the $14 \times 19 = 266$ possible cases, utilizing one of the alternative configurations has resulted in a better prediction.

Among the three alternative s_{neg} options, the relative preference was: $KFN(d)$ (45/70), $\hat{\mu}_{D_{neg}}$ (38/70) and $s_{\min}$ (29/70). This

[3] http://lucene.apache.org

[4] All NQC variants were tuned with the **same** value of k.

Table 1: Evaluation results. "greenish" and "reddish" colored values represent an improvement or a decline in prediction accuracy compared to the original NQC method. A statistical significant difference between a given NQC variant and the original NQC are marked with $*$ ($p < .05$).

	TREC4		TREC5		AP		WSJ		ROBUST		WT10g		GOV2	
	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ	P-r	K-τ
Autocorrelation	.456	.366	.188	.136	.348	.226	.586	.496	.385	.321	.299	.198	.247	.198
ScoreRatio	.420	.444	.344	.204	.231	.166	.620	.411	.405	.370	.226	.275	.295	.246
Clarity	.401	.357	.314	.208	.413	.265	.500	.355	.404	.329	.289	.203	.206	.164
WIG	**.533**	**.502**	.347	.252	.613	.417	.685	.463	.560	.399	.221	.323	.498	.352
WEG	.349	.454	.296	.201	.499	.357	.696	.482	.562	.400	.462	**.383**	.419	.337
SMV	.458	.386	.414	.332	.536	.379	.713	.484	.534	.391	.466	.279	.419	.310
NQC($\hat{\mu}_D$,$s(d)$) (Shtok et al. [10])	.491	.383	.454	.275	.619	.405	.722	.510	.580	.422	.522	.330	.378	.253
NQC($\hat{\mu}_D$,max)	.484	.371	.455	.277	.628	.413	.723	.494	.582	**.424**	.518	.324	.386*	.256
NQC($\hat{\mu}_D$,min)	.486	.403	.468	.265	.592*	.390*	.689*	.486*	.540	.425	.538*	.337*	.350*	.240*
NQC($\hat{\mu}_D$,max − min)	.436*	.325*	.417*	.301*	.645*	.414*	.706*	.460*	**.585***	.405*	.407*	.237*	.460*	.320*
NQC($\hat{\mu}_D$,$Q.95 - Q.05$)	.451*	.335*	.406*	.297*	.643*	.386*	.715	.496	.440*	.409*	.384*	.274*	.421*	.294*
NQC($\hat{\mu}_{D_{neg}}$,$s(d)$)	.482	.393	.503*	.304*	.625*	.423*	.729*	.504	.567*	.412*	.523	.333	.416*	.292*
NQC($\hat{\mu}_{D_{neg}}$,max)	.474*	.381	.495*	.314	.632*	.422*	.734*	.513	.568*	.409*	.516	.325	.427*	.300*
NQC($\hat{\mu}_{D_{neg}}$,min)	.483	.412*	.518*	.306*	.615	.407	.705*	.499*	.541*	.420*	.539*	.334*	.384*	.280*
NQC($\hat{\mu}_{D_{neg}}$,max − min)	.415*	.320*	.427*	.299*	.635*	.409	.734*	.496*	.551*	.377*	.399*	.233*	.507*	.355*
NQC($\hat{\mu}_{D_{neg}}$,$Q.95 - Q.05$)	.443*	.344*	.438*	**.336***	.678*	.406	.736*	.502*	.465*	.394*	.383*	.273*	.504*	.363*
NQC(min,$s(d)$)	.466*	.393	.506*	.312*	.602*	.423*	.715	.496	.538*	.393*	.406*	.277*	.445*	.306*
NQC(min,max)	.459*	.374	.498*	.312*	.617	.430*	.723	.494	.536*	.388*	.391*	.265*	.456*	.318*
NQC(min,min)	.484	.425*	.522*	.317*	.591*	.424*	.714	.500	.529*	.408*	.429*	.291*	.412*	.296*
NQC(min,max − min)	.380*	.311*	.430*	.304*	.630*	.412*	.706*	.460*	.493*	.329*	.245*	.176*	.528*	.358*
NQC(min,$Q.95 - Q.05$)	.407*	.338*	.442*	.330*	.674*	.412*	.715	.496	.457*	.353*	.255*	.205*	.532*	.367*
NQC($KFN(d)$,$s(d)$)	.475*	.422*	.524*	.297*	.638*	.445*	.730*	.561*	.580	.413*	.506*	.333*	.432*	.327*
NQC($KFN(d)$,max)	.470*	.418*	.519*	.306*	.643*	.449*	.739*	.561*	.583	.414	.500*	.324*	.443*	.327*
NQC($KFN(d)$,min)	.488	.434*	.530*	.286*	.635*	.428*	.691*	.523*	.556*	.422	.528*	.351*	.400*	.310*
NQC($KFN(d)$,max − min)	.415*	.349	.474*	.317*	.646*	.452*	.760*	.558*	.572*	.394*	.385*	.253*	.515*	.374*
NQC($KFN(d)$,Q.95-Q.05)	.448*	.371	.479*	.330*	.688*	.453*	.744*	**.582***	.504*	.403*	.380*	.272*	.509*	.373*
C-NQC($\hat{\mu}_D$,$s(d)$)	.492	.412*	.455	.310*	.622*	.412*	.737*	.510	.580	.412	.559*	.377*	.507*	.360*
S-NQC($\hat{\mu}_D$,$s(d)$)	.496	.383	.473*	.275	.620	.405	.728*	.502*	.580	.422	.522	.330	.398*	.253
SC-NQC($\hat{\mu}_D$,$s(d)$)	.504*	.412*	.473*	.310*	.623*	.412*	.737*	.510	.580	.422	.559*	.377*	.527*	.360*
SC-NQC(best)	.510*	.457*	**.560***	.317*	**.691***	**.454***	**.770***	.567*	**.585***	.405*	**.561***	.380*	**.566***	**.401***

demonstrates that, a better choice of a pseudo-ineffective reference point within NQC should be one that is more sensitive to the document in mind d (i.e., $KFN(d)$). This in comparison to a point that is more generally estimated (i.e., $\hat{\mu}_{D_{neg}}$ and $s_{\min}$). Moreover, among the latter two alternatives, considering more than one sample from the lower ranks of D is a better choice. In a similar manner, among the four alternative Z_d options, the relative preference was: $s_{\max}$ (34/56), $s_{\min}$ (29/56), $s_{\max} - s_{\min}$ (25/56) and lastly $Q_{.95} - Q_{.05}$ (23/56). This further demonstrates that, a better choice for a scaler is one that depends on a single point (having $s_{\max}$ a better choice than $s_{\min}$), rather than a range of values. Finally, **quantitatively**, by more proper configuration of $f_2(d)$, NQC's prediction accuracy has improved in all datasets, up to 38% and 46% improvement in P-r and K-τ, respectively. This demonstrates the usefulness of analyzing NQC using our scaled extension to Roitman et al.'s discriminative QPP framework [8].

4.2.2 Effect of calibration and scaling. The four bottom rows in Table 1 further report the effect of NQC's calibration and scaling. First, we observe that, by better tuning of NQC's calibration features within **C-NQC**($\hat{\mu}_D$,$s(d)$), its prediction accuracy has improved in most cases (up to 42% better). Next, scaling (i.e., **S-NQC**($\hat{\mu}_D$,$s(d)$)) by itself also improves NQC's accuracy (up to 5%). Combining both calibration and scaling (i.e., **SC-NQC**($\hat{\mu}_D$,$s(d)$)), has resulted in most cases in a further improved accuracy (up to 42%). Finally, further using the best $f_2(d)$ configuration together with calibration and scaling (i.e., **SC-NQC**(best)) provides the best prediction strategy for NQC (up to 60% improvement).

5 CONCLUSIONS

We introduced a simple, yet highly effective, extension to Roitman et al.'s discriminative QPP framework [8]. Our main focus was on the NQC method, where using our proposed extension, we were able to redesign it and suggest several options for improvement.

REFERENCES

[1] David Carmel and Oren Kurland. Query performance prediction for ir. In *Proceedings of SIGIR '12*.
[2] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of SIGIR '02*.
[3] Fernando Diaz. Performance prediction using spatial autocorrelation. In *Proceedings of SIGIR '07*.
[4] Ahmad Khwileh, Andy Way, and Gareth J. F. Jones. Improving the reliability of query expansion for user-generated speech retrieval using query performance prediction. In *CLEF*, 2017.
[5] Oren Kurland. The cluster hypothesis in information retrieval. In *Advances in Information Retrieval*, pages 823–826. Springer International Publishing, 2014.
[6] A. M. Ozdemiray and Ismail S. Altingovde. Query performance prediction for aspect weighting in search result diversification. *Proceedings of CIKM '14*.
[7] Haggai Roitman. An extended query performance prediction framework utilizing passage-level information. In *Proceedings of ICTIR*, pages 35–42, New York, NY, USA, 2018. ACM.
[8] Haggai Roitman, Shai Erera, Oren Sar-Shalom, and Bar Weiner. Enhanced mean retrieval score estimation for query performance prediction. In *Proceedings of ICTIR '17*.
[9] Haggai Roitman, Shai Erera, and Bar Weiner. Robust standard deviation estimation for query performance prediction. In *Proceedings of ICTIR '17*.
[10] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2):11:1–11:35, May 2012.
[11] Yongquan Tao and Shengli Wu. Query performance prediction by considering score magnitude and variance together. In *Proceedings of CIKM '14*.
[12] Yun Zhou and W. Bruce Croft. Query performance prediction in web search environments. In *Proceedings of SIGIR '07*.