Information Needs, Queries, and Query Performance Prediction

Oleg Zendel
olegzendel@campus.technion.ac.il
Technion

Anna Shtok
annie.shtok@gmail.com

Fiana Raiber
fiana@verizonmedia.com
Yahoo Research

Oren Kurland
kurland@ie.technion.ac.il
Technion

J. Shane Culpepper
shane.culpepper@rmit.edu.au
RMIT University

ABSTRACT

The *query performance prediction* (QPP) task is to estimate the effectiveness of a search performed in response to a query with no relevance judgments. Existing QPP methods do not account for the effectiveness of a query in representing the underlying information need. We demonstrate the far reaching implications of this reality using standard TREC-based evaluation of QPP methods: their relative prediction quality patterns vary with respect to the effectiveness of queries used to represent the information needs. Motivated by our findings, we revise the basic probabilistic formulation of the QPP task by accounting for the information need and its connection to the query. We further explore this connection by proposing a novel QPP approach that utilizes information about a set of queries representing the same information need. Predictors instantiated from our approach using a wide variety of existing QPP methods post prediction quality that substantially transcends that of applying these methods, as is standard, using a single query representing the information need. Additional in-depth empirical analysis of different aspects of our approach further attests to the crucial role of query effectiveness in QPP.

ACM Reference Format:
Oleg Zendel, Anna Shtok, Fiana Rabier, Oren Kurland, and J. Shane Culpepper. 2019. Information Needs, Queries, and Query Performance Prediction. 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, 10 pages. https://doi.org/10.1145/3331184.3331253

1 INTRODUCTION

Ad hoc (query-based) retrieval effectiveness can significantly vary across queries for a variety of retrieval methods [10]. This fact enabled a large body of work on *query performance prediction* (QPP) [10] whereby the goal is to estimate search effectiveness in the absence of human relevance judgments. There are two common approaches to this problem. *Pre-retrieval predictors* analyze the query and the corpus prior to retrieval time [14, 22, 23, 29, 38, 39, 48, 59]; e.g., queries containing terms with high IDF (inverse document frequency) values should presumably be more effective [14, 23]. *Post-retrieval predictors* use information induced from the result list of top-retrieved documents [1, 2, 5, 9, 11, 14–18, 20, 22, 25, 30, 31, 33–37, 41–44, 49, 51, 56, 57, 60, 61]. For example, some post-retrieval predictors are based on analyzing the retrieval scores of documents in the result list (e.g., [17, 35, 36, 41, 49, 61]).

A careful examination of past work in this area reveals that most current approaches rely on predicting the performance across a set of queries, where each query represents a unique information need; often on a single corpus using a single retrieval method. This is largely an artifact of the test collections commonly used to evaluate new techniques. For example, the TREC [52] initiative has historically defined an information need as a *topic* using a *title*, *description*, and *narrative*. The title is commonly viewed as a keyword query that a user might issue to a search engine, and the description and narrative more clearly define the user’s real information need. While QPP methods have been shown to be effective in this setting, many (specifically, pre-retrieval methods) were found to be ineffective in predicting the relative effectiveness of different queries that represent the same information need (a.k.a., query variations or reformulations) [48][1]. Consequently, Thomas et al. [48] postulated that QPP methods essentially predict information-need performance rather than query performance. Indeed, current QPP approaches do not explicitly account for the information need, nor for the extent to which a query effectively represents it for retrieval[2]. In fact, this is also the case for the formal fundamental probabilistic basis of most prediction methods [31, 43], where coupling an information need with the query used to represent it does not allow to account for the effectiveness of the query in representing the need for retrieval.

We show that the implications of this coupling are more far-reaching. Specifically, using the UQV datasets [3, 8], where human-generated query variations are available for TREC topics, we show that the relative prediction quality patterns of various QPP methods, when evaluated in the standard setting of different queries representing different information needs, vary with respect to the effectiveness of queries used to represent the underlying need.

Motivated by our empirical findings that attest to the importance of modeling the information need and its connection with queries used to represent it, we revise the probabilistic formalism of the QPP task [43] which is the basis for most QPP methods. Our formalism

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

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/3331184.3331253

[1] Pre-retrieval methods are also not effective in predicting the performance of different document lists retrieved for the same query using different retrieval methods [31]. Some work [50] demonstrated the limited merit of using pre-retrieval methods to select between personalized and original query variants.

[2] The main exception we are aware of is the work of Sondak et al. [44].

reflects a transition from addressing the basic question of *"What is the probability that this retrieved list is relevant (effective) for this query?"* [43] to addressing the question of *"What is the probability that this retrieved list is relevant (effective) for a (latent) information need that is represented by this query?"*.

We use our revised probabilistic formalism to address a novel prediction challenge: *query performance prediction using reference queries*. That is, how can we estimate retrieval effectiveness for a given query using knowledge derived from additional (reference) queries that represent the same information need. Our proposed approach, which can be instantiated using any existing query-performance predictor, accounts for the association between the query and the reference queries and the predicted performance for the latter. Extensive empirical evaluation attests to the clear merits of our approach. Specifically, the prediction quality is substantially better than that of using existing predictors which do not utilize reference queries.

Summary of Contributions. In this work, we show that the effectiveness of a query in representing the underlying information need has a significant impact on QPP quality. We then define a probabilistic QPP framework which encapsulates the fundamental relationship between queries and an information need, and show how it can be easily instantiated using a variety of QPP approaches to improve prediction quality. An extensive empirical evaluation shows that combining information from multiple reference queries dramatically improves prediction quality for queries regardless of their effectiveness in representing the information need, and independent of the QPP method used to instantiate our approach.

2 RELATED WORK

The prediction framework presented by Carmel et al. [11] for estimating *topic/information need difficulty* accounts for the fact that the information need can be represented by different queries. However, in contrast to our work, the model instantiated from the framework predicts performance for a single query without using information from alternative queries that represent the same information need, and without accounting for how well the query represents the information need.

Previous work exists on predicting performance for a single query by estimating the presumed extent to which it represents the underlying information need [44]. This predictor was also used by Roitman et al. [36], where the estimate is based on properties of a pseudo-feedback-based relevance model induced from the retrieved list [27]. There has also been work on ranking query reformulations [19, 37, 54] by estimating the extent to which each reformulation represents the information need. The task we pursue is different from these previous lines of work: we predict performance for a query using information from other queries representing the same information need. The estimator from Sondak et al. [44] is used as a baseline reference comparison in this work.

Other relevant recent work on using reference document lists for QPP was proposed by Shtok et al. [43]. The reference lists are retrieved by using different retrieval methods for a set of queries. Performance is then predicted based on the similarity between the retrieved list for a query and its corresponding reference lists. Our prediction model uses reference queries and assumes that the retrieval method is fixed. Some of the estimates derived in this work result in predictors that utilize reference document lists retrieved for different queries; hence, these specific predictors are similar in spirit to those proposed by Shtok et al. [43]. However, we utilize estimates (predictors) for the retrieval effectiveness of the reference queries, while the (novel) models devised and evaluated in [43] do not account for the effectiveness of the reference lists. Furthermore, the work of Shtok et al. [43] couples the query with the information need as in prior work on QPP, while our focus in this work is on explicitly de-coupling the two.

Query Variations. There is a long history of research that explores the relationship between a single query and the information needs it represents. Early work exploring the impact of query representation and effectiveness was presented by Belkin et al. [6]. Shortly thereafter, Belkin et al. [7] operationalized these insights to significantly improve retrieval effectiveness by fusing the results from multiple query variations for a single information need.

A more recent seminal study of query reformulation in user search sessions in a production search engine also explored the importance of subtle query variations on the overall effectiveness of search results returned to the user [24]. Subsequent studies have shown that query variations can be used in learning-to-rank models to significantly boost query performance [19, 40]. Sheldon et al. [40] and other recent studies [55] have shown that highly-effective query variations, rewrites, and suggestions can be generated offline using random walks over a bipartite graph constructed over click data extracted from large search engine query logs [13]. In the absence of access to commercial search engine logs, effective manual query variations can also be generated using crowdsourcing [3, 4, 8]. We use the UQV query collections [3, 8] as the query variations are publicly available, and hence the results are reproducible.

Using human query variations, Thomas et al. [48] showed empirically that existing pre-retrieval predictors essentially predict information-need difficulty rather than query difficulty. Their work inspires this work, where we attempt to de-couple information needs and queries in modeling the QPP task. We strengthen their findings by showing that the relative prediction quality patterns of various query-performance predictors (both pre- and post- retrieval ones) vary with respect to the effectiveness of queries used to represent the information needs. In contrast to our work, a model of the QPP task that accounts for the information need was not presented and the challenge of predicting performance using reference queries was not addressed. That is, Thomas et al. [48] show that it can be very hard to make a prediction that distinguishes the performance of queries representing a single information need, while we show that using multiple representations for an information need can significantly improve the quality of prediction for the performance of a single query for an information need.

Supervised Query Performance Prediction. It is also possible to improve prediction quality using supervised [12, 21, 26, 31, 33–35, 56] or weakly supervised [57] techniques. In general, these approaches, and others [42, 51, 61], integrate various unsupervised query performance prediction methods. In this work, we do not explore the use of query variations in supervised QPP techniques. Rather, we demonstrate the merits of our approach with many unsupervised pre- and post-retrieval predictors that are commonly the

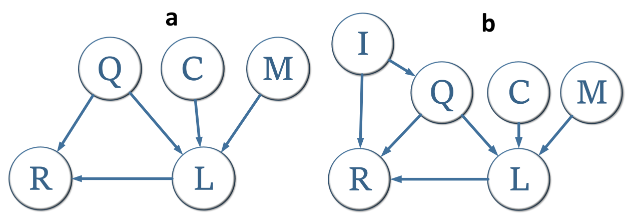

Figure 1: Graphical model representation of (a) the probabilistic model underlying existing query performance prediction methods, and (b) the extended model that accounts for the information need.

basis for these supervised approaches. Extending our new framework to a supervised learning framework is an interesting problem that is orthogonal to our own.

3 PREDICTION FRAMEWORK

The query performance prediction (QPP) task is to estimate the effectiveness of a search performed in response to a query in the absence of human relevance judgments [10]. As noted by Raiber and Kurland [31], the task can be framed as estimating the probability:

$$p(R = 1|Q, C, M), \tag{1}$$

where R, Q, C and M are random variables that take as values relevance status (1 stands for relevant), queries, corpora and retrieval methods, respectively. That is, the task is to estimate the probability for a relevance event, discussed below, given the ranked retrieval of a query over a corpus.

Previous work on QPP is based, *in spirit*, on fixing the values of C and M, and estimating Equation 1 [31]. That is, prediction is performed over different queries used for retrieval with the same retrieval method on the same corpus. Since the choice of a query, retrieval method and corpus entails a retrieved list, the QPP task pursued in past work can be framed, as recently suggested by Shtok et al. [43], as estimating:

$$p(R = 1|Q, L), \tag{2}$$

where L is a random variable that takes retrieved lists as values. Figure 1 (a) depicts the graphical model of dependence relations between the random variables used in Equations 1 and 2.

Estimating Equation 2 for a specific query and retrieved list amounts to addressing the question [43]:"*What is the probability that this retrieved list is relevant to this query?*". This question is an extension, from a single document to a ranked document list, of the question underlying probabilistic retrieval: "*What is the probability that this document is relevant to this query?*" [46]. A relevance status for a retrieved list can be *operationally* defined in terms of document relevance [43]; e.g., by thresholding a document-relevance-based evaluation measure applied to the retrieved list. It was shown that many post-retrieval QPP methods can be derived from Equation 2 [43].

3.1 A missing Piece: The Information Need

A more careful examination of past work on QPP reveals the following: evaluation of prediction quality was performed by using queries representing different information needs — specifically, TREC topics served as information needs, and each topic was represented by a title, which was treated as a query. Various predictors, including pre-retrieval methods that analyze the query and the corpus but not the retrieved list, were shown to yield high prediction quality in this evaluation setting. However, recent work shows that pre-retrieval predictors are ineffective in predicting the relative performance of different queries used to represent the same information need [48]. Now, these two prediction tasks — for queries representing the same or different information needs — cannot be differentiated at the model level using Equations 1 and 2, since the underlying information need is not explicitly accounted for.

Furthermore, the (implicit) coupling of the information need with the query in previous work on QPP has led to ignoring the fact that different queries representing the same information need can exhibit very different performance characteristics. As a case in point, consider the WIG predictor [61] which uses the (corpus normalized) mean retrieval score of the top retrieved documents as a prediction value. If the query does not effectively represent the information need, then high retrieval scores which attest to improved match between the document and the query[3] need not necessarily indicate high effectiveness. Moreover, the effectiveness of the query in representing the information need for retrieval can depend on the corpus and retrieval method. For example, a query that can be relatively ineffective for a simple bag-of-words ranking function such as Okapi BM25 [32] or query likelihood [45] might still be highly effective if a retrieval method which captures higher level term dependencies is used (e.g., [28]).

Given the observations discussed above, we revise the basic probabilistic modeling of the QPP task from Equations 1 and 2 to account for the information need. The QPP task becomes estimating:

$$p(R = 1|I, Q, C, M) = p(R = 1|I, Q, L); \tag{3}$$

I is a random variable that takes information needs as values; and, recall that the value of L is uniquely determined by the values of Q, C and M. Figure 1 (b) depicts the dependencies between the random variables. Note that an assignment to I induces a probability distribution over assignments of Q; this is the probability that a user selects a specific query to represent an information need. The retrieval effectiveness of this selection depends on the retrieval method and corpus that together with the query determine the retrieved list.

Assignments of the random variables in Equation 3 result in a novel revised fundamental question of the QPP task: "*What is the probability that this retrieved list is relevant to this (latent) information need given that the need is represented by this query?*."

3.2 QPP using Reference Queries

To further explore the importance of accounting for the connection between the information need and the query in the QPP task, we pursue the following novel challenge: predicting retrieval performance for a given query using information about additional queries that presumably represent the same information need.

[3] WIG is mainly used for ranking functions that are based on the surface-level similarity between the query and documents.

Formally, let q be the query used for retrieval ($Q = q$) and i be the information need it represents ($I = i$). Let Q_i be a set of queries, henceforth referred to as *reference queries*, that represent i; i.e., $\forall q' \in Q_i.\ p(Q = q'|I = i) > 0\ (q' \neq q)$.

We predict query performance by estimating $p(R = 1|I = i, Q = q, C = c, M = m)$ from Equation 3 using the reference queries in Q_i as disjoint proxies for the information need:

$$\hat{p}_{Ref}(R = 1|i, q, c, m) \overset{def}{=} \sum_{q' \in Q_i} \hat{p}(R = 1|i, q, q', c, m)\hat{p}(q'|i, q, c, m); \quad (4)$$

herein, $\hat{p}$ denotes an estimate for p; $\hat{p}_{Ref}$ is an estimate that utilizes reference queries; we omit random variables from formulations (except for the relevance status) for brevity.

We now examine the estimates on the right hand side of Equation 4. $\hat{p}(q'|i, q, c, m)$ is an estimate for the likelihood that the reference query q' is the one selected to represent the information need i. Based on the relationships between random variables assumed in Figure 1 (b), the likelihood $p(q'|i, q, c, m)$ does not depend on the corpus, retrieval method or on other queries used to represent the need[4]. In addition, we use the query q as a signal about the (latent) information need to derive the following approximation which is inspired by the relevance-model framework [27][5]:

$$\hat{p}(q'|i, q, c, m) = \hat{p}(q'|i) \approx \hat{p}(q'|q). \quad (5)$$

Below we use inter-query association measures to derive $\hat{p}(q'|q)$.

We next examine $\hat{p}(R = 1|i, q, q', c, m)$ in Equation 4. This is an estimate for the probability of a relevance event given two queries that represent the information need i. The actual retrieved list is not specified as it can be produced in several ways; e.g., the queries can be concatenated to yield a single query used for retrieval, or the lists retrieved for the two queries can be fused to produce a single list. To devise a generic estimate which potentially applies to different approaches of fusing information about the two queries for retrieval, we make the following basic assumption: the retrieval effectiveness of using two queries representing the same information need is based, among other factors, on the extent to which each is an effective representative of the information need with respect to the corpus and retrieval method used. Specifically, the estimate we use is a linear interpolation (fusion), with a free parameter λ, of the estimates for a relevance event for the two queries:

$$\hat{p}(R = 1|i, q, q', c, m) \overset{def}{=} (1 - \lambda)\hat{p}(R = 1|i, q, c, m) + \lambda\hat{p}(R = 1|i, q', c, m); \quad (6)$$

Plugging the estimates from Equations 5 and 6 in Equation 4 and assuming that $\hat{p}(q'|q)$ is a probability distribution over $q' \in Q_i$, we arrive at:

$$\hat{p}_{Ref}(R = 1|i, q, c, m) \overset{def}{=} (1 - \lambda)\hat{p}(R = 1|i, q, c, m) + \lambda \sum_{q' \in Q_i} \hat{p}(R = 1|i, q', c, m)\hat{p}(q'|q). \quad (7)$$

Equation 7 is based on backing off from a direct estimate for a relevance event, $\hat{p}(R = 1|i, q, c, m)$, to a mixture-based estimate that uses additional queries representing the same information need. More specifically, the higher the association ($\hat{p}(q'|q)$) of the given query (q) with reference queries (q') that are effective representatives of the information need — i.e., those with high $\hat{p}(R = 1|i, q', c, m)$ — the higher the estimate for a relevance event for q.

3.3 Deriving Specific Predictors

We next derive specific predictors based on Equation 7. First, following standard practice in past work on QPP using probabilistic models [25, 33, 35, 43], we use the values $\mathcal{P}(q)$ and $\mathcal{P}(q')$, assigned by an existing performance predictor $\mathcal{P}$ to q and q', for $\hat{p}(R = 1|i, q, c, m)$ and $\hat{p}(R = 1|i, q', c, m)$, respectively. The predictor uses information induced from the query (q or q') and the corpus (c) and might also use information induced from the document list retrieved for the query using the retrieval method m. The prediction principles underlying existing predictors were derived from Equations 1 and 2 in recent work [43]. Now, if the query is coupled with the information need, as assumed in past work, then $\hat{p}(R = 1|i, q, c, m)$ and $\hat{p}(R = 1|i, q', c, m)$ become $\hat{p}(R = 1|q, c, m)$ and $\hat{p}(R = 1|q', c, m)$, respectively, and we indeed resort to Equation 2.

Our next goal is devising the estimate $\hat{p}(q'|q)$. To this end, we use inter-query association measures, $\mathcal{A}$, as described below. The resultant prediction value we use, following Equation 7, is:

$$\mathcal{P}_{Ref}(q) \overset{def}{=} (1 - \lambda)\mathcal{P}(q) + \lambda\frac{1}{|Q_i|} \sum_{q' \in Q_i} \mathcal{P}(q')\mathcal{A}(q', q). \quad (8)$$

We do not normalize the inter-query association values to form a probability estimate $\hat{p}(q'|q)$ over $q' \in Q_i$, as such normalization resulted in substantially degraded prediction quality. (Actual numbers are omitted as they convey no additional insight.) This is not a surprise: one cannot expect Q_i to include all potential queries that represent i. As a result, normalization negatively distorts the estimate for the true relative extent to which the reference queries are associated with the given query, and thereby the extent to which they represent the information need. This badly affects prediction *across* information needs — the standard prediction quality evaluation paradigm that we subscribe to in this paper. Furthermore, we show in Section 4 that the average association of the given query with the reference queries is already a descent basis for prediction. (We further discuss this in Section 3.3.1.) Hence, normalization with respect to the associations negatively affects prediction quality. On the other hand, in practice, we might have a different number of reference queries for each information need. To avoid the resulting bias, we use the $\frac{1}{|Q_i|}$ normalization factor in Equation 8.

The first inter-query association measure, $\mathcal{A}$, we consider is the Jaccard coefficient between q and q'; the resultant predictor, based on Equation 8, is **Ref-Jaccard[$\mathcal{P}$]**. The second measure is the ratio between the overlap (in terms of number of documents) at the top-k ranks of the document lists retrieved for q and q' from c using the retrieval method m; k is a free parameter; the resultant predictor is denoted **Ref-Overlap[$\mathcal{P}$]**. The third measure is the Rank-Biased-Overlap (RBO) [53] between the lists retrieved for q and q' computed at rank k with parameter p. In contrast to the overlap measure, RBO also accounts for the ranks at which

[4] For simplicity, we assume that queries are generated independently for an information need. Note that this *conditional independence* does not contradict our use of the queries as disjoint events in Equation 4. Disjoint events cannot be independent.

[5] The probability of generating a term from a relevance model is approximated by the probability to generate it given the observed query [27].

documents appear; the resulting predictor is **Ref-RBO[$\mathcal{P}$]**. We note that the fact that the overlap and RBO measures are based on information induced from the corpus and the retrieval method does not contradict the fact that, according to Figure 1, queries are generated only based on the information need. The same way users might utilize knowledge and assumptions about the corpus/retrieval method (e.g., the language and style used in the corpus) to formulate queries, the prediction method can utilize all information at hand so as to predict the use of a specific query given an example of another query representing the same information need.

3.3.1 Special-case predictors. To study the contribution to prediction quality of the two factors that govern the utilization of reference queries in Equation 8 — i.e, the predicted performance of the reference queries and their association with the query — we consider two predictors that are special cases of our prediction model. The first, named **OnlyAsso**, assumes that $\hat{p}(R = 1|i, q', c, m)$ in Equation 7 is the same constant for all $q' \in Q_i$; i.e., the reference queries are assumed to be equi-effective, and prediction is only based on the association between the given query and the reference queries in Q_i. Following Equation 8, the predicted value is:

$$\mathcal{P}_{OnlyAsso}(q) \overset{def}{=} (1-\lambda)\mathcal{P}(q) + \lambda \frac{1}{|Q_i|} \sum_{q' \in Q_i} \mathcal{A}(q', q). \quad (9)$$

The second predictor assumes that all reference queries are associated to the same extent with q and hence are equi-likely to represent the information need i. The resultant prediction value, following Equation 8, is:

$$\mathcal{P}_{OnlyRef}(q) \overset{def}{=} (1-\lambda)\mathcal{P}(q) + \lambda \frac{1}{|Q_i|} \sum_{q' \in Q_i} \mathcal{P}(q'). \quad (10)$$

This prediction method, termed **OnlyRef**, is based on the following assumption: the unweighted average of the performance-prediction values assigned to queries representing the information need is a good approximation to the performance value that should be predicted for any query (q) representing the need.

4 EVALUATION

4.1 Experimental Setup

Two TREC datasets were used for experiments. The first is ROBUST which is composed of 528, 155 (mainly) news articles and is associated with 249 TREC topics[6] (301-450 and 600-700). The second dataset is the Category B of the ClueWeb12 collection (CW12 hereafter), which is composed of around 50 million web pages, and is associated with 100 TREC topics (201-300). The set of queries per TREC topic includes the original topic title and additional query variations [3, 8]: human generated queries that represent the topic.[7] We removed duplicate query variations per topic and queries with out-of-vocabulary terms. On average, there are 12.75 (with standard deviation of 6.81) and 46.11 (with standard deviation of 18.66) unique variations per topic for ROBUST and CW12, respectively. We applied Krovetz stemming to all queries and documents; stopwords on the INQUERY list were removed only from queries. The Indri toolkit (www.lemurproject.org) was used for experiments.

Following common practice in work on QPP [10, 31, 35, 43, 57], we use language-model-based retrieval: the query-likelihood model [45] was used to retrieve the document lists; retrieval scores are log query likelihood; document language models were Dirichlet smoothed with the smoothing parameter set to 1000 [58].

Our proposed framework is not limited to specific predictors; we aim to demonstrate consistent patterns on a set of commonly used post- and pre-retrieval predictors. We consider six pre-retrieval predictors that were shown to be highly effective in past work [22]: AvgIDF [15], MaxIDF [15], AvgSCQ [59], MaxSCQ [59], AvgVAR [59] and MaxVAR [59]. The post-retrieval predictors considered are Clarity [14], NQC [41], WIG [61], QF (query feedback) [60] and their UEF [42] counterparts. In addition, we report the performance for SMV [47] and the recently proposed RSD [36] predictor that was instantiated in our experiments using WIG.

To measure prediction quality, we use Pearson correlation between the true AP (at cutoff 1000) values attained for queries using relevance judgments and the values assigned to them by a predictor [10]. We also used Kendall's Tau for evaluation [10] for all of the experiments shown. All trends were consistent across both correlation measures, and due to space limitations, the Kendall's Tau results are omitted as they provide no additional insight.

The free-parameter values of the predictors were set using the train-test approach used in prior QPP work [25, 31, 35, 57]. Specifically, we randomly split the topics into two equal-sized sets, where each of the two sets served in turn as the test fold. The parameter values yielding the highest Pearson correlation (see above) over the training fold were applied to the test fold. The reported prediction quality of the split is the average prediction quality of the two test folds. The partitioning procedure was repeated 30 times and we report the average prediction quality over these 30 splits. Statistically significant differences are computed using the two-tailed paired t-test at a 95% confidence level with Bonferroni correction.

The number of top-retrieved documents in all the post-retrieval predictors except for UEF was selected from $\{5, 10, 25, 50, 100, 250, 500, 1000\}$. The number of documents used to construct relevance model #1 (RM1) [27] for Clarity and QF was set to values in the same range. For UEF we set the number of top-retrieved documents to values in $\{25, 50, 100, 250, 500, 1000\}$, used Pearson correlation to measure inter-list similarity as recommended by Shtok et al. [42], and constructed RM1 using the same number of documents used for measuring the similarity. The overlap between two retrieved lists in QF was computed at the following cutoff values: $\{5, 10, 25, 50, 100\}$. For Clarity, QF, UEF and RSD, RM1 was constructed from unsmoothed document-language models and the number of terms was clipped to 100 [31]; for RSD we also experimented with an unclipped RM1 that was reported to be effective in prior work [44]. In addition, for RSD we used 100 sampled lists. For the list-based inter-query association measures used in our approach, Overlap and RBO, the cutoff k is selected from $\{5, 10, 25, 50, 100, 250, 500\}$; this range of values was also used to set the sampled-list size in RSD. The value of p in RBO is set to 0.95 following prior recommendations [43]. The value of λ was selected from $\{0, 0.1, \ldots, 1\}$.

[6] One topic was removed from the original set due to absence of relevant documents in the relevance judgments files.

[7] The variations are available at https://tinyurl.com/robustuqv and https://tinyurl.com/clue12uqv.

Table 1: Prediction quality with respect to query effectiveness: TREC title queries (Title) and query variations with the maximal (Max), median (Med) and minimum (Min) AP per topic. The top (bottom) block presents previously proposed pre-retrieval (post-retrieval) predictors. The best prediction quality per corpus and predictor is boldfaced. The highest number in each column is underlined.

	ROBUST				CW12			
	Title	Max	Med	Min	Title	Max	Med	Min
AvgIDF	.357	**.415**	.314	.133	.432	**.458**	.327	.233
AvgSCQ	.243	**.339**	.281	.220	.440	**.469**	.328	.247
AvgVar	.405	**.466**	.378	.247	.421	**.453**	.337	.214
MaxIDF	.396	**.412**	.391	.235	**.377**	.326	.369	.297
MaxSCQ	.338	**.385**	.375	.297	.400	.373	**.424**	.347
MaxVar	.420	**.456**	.441	.354	.372	**.378**	.374	.290
Clarity	.409	.384	**.460**	.417	.029	.130	**.213**	.189
NQC	.477	**.551**	.435	.166	.509	**.548**	.393	.181
WIG	.475	**.511**	.454	.391	.535	**.549**	.500	.416
SMV	.424	**.535**	.411	.159	.462	**.520**	.320	.183
RSD	.489	**.521**	.376	.185	.549	**.574**	.325	.249
QF	**.487**	.391	.356	.429	.280	**.405**	.248	.248
UEF(Clarity)	.522	.517	.541	.468	.276	**.294**	.263	.292
UEF(NQC)	.523	**.558**	.444	.236	**.438**	.435	.355	.288
UEF(WIG)	**.509**	.385	.367	.352	**.441**	.387	.333	.348
UEF(QF)	**.495**	.444	.435	.451	**.345**	.298	.263	.324

Table 2: Main result: Prediction quality of Ref-RBO when predicting performance for queries which are TREC's topic titles. The baseline is applying $\mathcal{P}$ to the title query as is standard. '*' marks statistically significant differences with the baseline. The best result in a column is underlined.

	ROBUST		CW12	
$\mathcal{P}$	baseline	Ref-RBO	baseline	Ref-RBO
AvgIDF	.357	.603*	.432	.669*
AvgSCQ	.243	.595*	.440	.696*
AvgVar	.405	.604*	.421	.673*
MaxIDF	.396	.611*	.377	.627*
MaxSCQ	.338	.601*	.400	.675*
MaxVar	.420	.606*	.372	.637*
Clarity	.409	.598*	.029	.613*
NQC	.477	.586*	.509	.664*
WIG	.475	.590*	.535	.691*
SMV	.424	.584*	.462	.646*
RSD	.489	.568*	.549	.650*
QF	.487	.588*	.280	.662*
UEF(Clarity)	.522	.603*	.276	.578*
UEF(NQC)	.523	.579*	.438	.658*
UEF(WIG)	.509	.589*	.441	.658*
UEF(QF)	.495	.575*	.345	.594*

4.2 Experimental Results

4.2.1 Query effectiveness. The classic QPP task is to predict retrieval effectiveness for queries, where each represents a different information need (topic). In most work to date, a TREC topic title was used as the representative query for a topic. In Table 1 we study the prediction quality when queries of different effectiveness are used to represent each topic: the query variation with the highest AP (Max), median[8] AP (Med) and lowest AP (Min) per topic[9].

Overall, we see that the best prediction quality is in most cases attained when the variation with the maximal AP (Max) represents the topic. (Refer to the boldfaced numbers.) The lowest prediction quality is almost always observed when the chosen query is the one with the minimal AP (Min). More generally, we see that prediction quality varies considerably depending on the effectiveness of the queries used. For example, for ROBUST, the best predictor for title queries is UEF(NQC) (0.523) whereas for Min queries it is UEF(Clarity) (0.468); i.e., the decision about the best predictor to use also appears to depend on the effectiveness of the query used to represent the information need (topic). To shed some light on this phenomenon, in Figure 2 we present the retrieval effectiveness (AP@1000) attained for the title queries, and all additional query

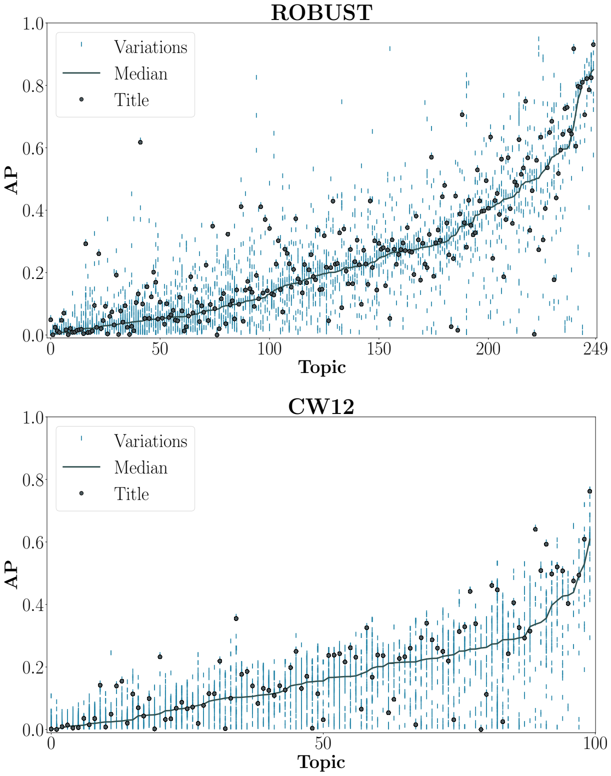

Figure 2: The retrieval effectiveness (AP@1000) of query variations and TREC title queries for each topic. Each point on the x-axis is a different topic; the topics are ordered on the x-axis by the median effectiveness — represented by the curves — of all known variants for the topic.

[8] For topics with an even number of query variations, the larger of the two middle values was chosen.

[9] Query variations with a zero AP were omitted for this analysis; there are 71 such variations in CW12 representing 31 topics and 34 in ROBUST representing 11 topics.

Table 3: Prediction quality when using different inter-query association measures in our approach. '*b*' and '*r*': statistically significant differences with baseline and Ref-RBO in a column (except for OnlyAsso), respectively. '*o*': statistically significant differences with OnlyAsso in a row. Bold: best result in a column per dataset.

	ROBUST				
	MaxIDF	AvgSCQ	WIG	UEF(Clarity)	OnlyAsso
baseline	.396	.243	.475	.522	-
Ref-Jaccard	$.484^{ro}_{b}$	$.364^{ro}_{b}$	$.506^{ro}_{b}$	$.553^{ro}_{b}$	.269
Ref-Overlap	$\mathbf{.613}^{o}_{b}$	$.591^{o}_{b}$	$.588_{b}$	$.599^{o}_{b}$	.582
Ref-RBO	$.611^{o}_{b}$	$\mathbf{.595}^{o}_{b}$	$\mathbf{.590}_{b}$	$\mathbf{.603}^{o}_{b}$	**.588**
Ref-Geo	$.395^{ro}_{b}$	$.243^{ro}$	$.475^{ro}$	$.521^{ro}_{b}$	.187
OnlyRef	$.393^{r}$	$.290^{r}_{b}$	$.471^{r}_{b}$	$.532^{r}$	-
	CW12				
	MaxIDF	AvgSCQ	WIG	UEF(Clarity)	OnlyAsso
baseline	.377	.440	.535	.276	-
Ref-Jaccard	$.527^{ro}_{b}$	$.608^{ro}_{b}$	$.613^{ro}_{b}$	$.551^{ro}_{b}$	.504
Ref-Overlap	$\mathbf{.670}^{ro}_{b}$	$\mathbf{.724}^{r}_{b}$	$\mathbf{.703}^{ro}_{b}$	$\mathbf{.622}^{ro}_{b}$	**.720**
Ref-RBO	$.627^{o}_{b}$	$.696^{o}_{b}$	$.691_{b}$	$.578^{o}_{b}$	.685
Ref-Geo	$.373^{ro}$	$.438^{ro}$	$.534^{ro}$	$.275^{ro}$	.071
OnlyRef	$.383^{r}$	$.490^{r}_{b}$	$.537^{r}$	$.447^{r}_{b}$	-

variations per topic. We see that the difference in retrieval effectiveness among query variations per topic can be quite striking. While for some queries the AP is nearly 0, most topics have at least one variant with an AP higher than 0.3. In addition, we see that using the TREC title queries for retrieval does not necessarily yield median retrieval effectiveness per topic. For some topics, the performance can be much better or much worse than the median. In other words, the relative effectiveness of the title queries in representing the underlying information need (topic) varies across topics.

4.2.2 Main result: Using reference queries for prediction. We now turn to study the merits of using reference queries to predict retrieval effectiveness. In this section, and in Sections 4.2.3-4.2.6, prediction quality is evaluated using standard practice in work on QPP [10]; that is, prediction is performed for a set of queries, each of which is the title of a different TREC topic. In Section 4.2.7 we evaluate prediction quality when each topic is represented by a query variation which is not necessarily the topic title.

Table 2 presents the prediction quality of Ref-RBO, which was derived from Equation 8 using RBO as the inter-query association measure. We show in Section 4.2.3 that RBO is one of the most effective inter-query association measures among those considered. Hence, Ref-RBO is the main instantiation of our approach that we focus on throughout this section. We hasten to point out that the prediction quality patterns we report for Ref-RBO are consistent with those attained when using the other inter-query association measures, as exemplified in Section 4.2.3.

We can see in Table 2 that for all the considered predictors $\mathcal{P}$, for both datasets, Ref-RBO, which relies on reference queries, substantially and statistically significantly outperforms the **baseline**: applying the predictor $\mathcal{P}$ to the title query, as is standard, without using reference queries. Note that this baseline is a specific case of our approach when setting $\lambda = 0$ in Equation 8. The best prediction quality attained for Ref-RBO (refer to the underlined numbers) surpasses the best results attained by any baseline by a large margin. We can clearly conclude that there is much merit in using our QPP approach that utilizes reference queries.

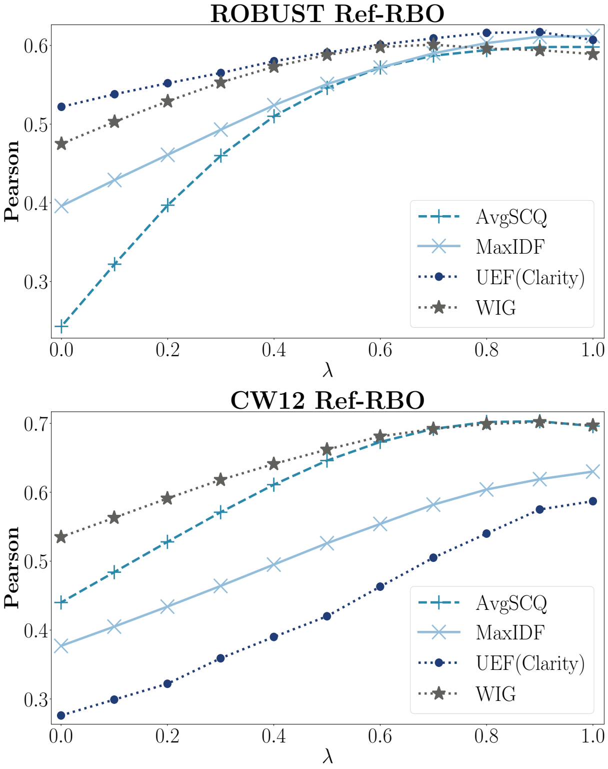

Figure 3: The effect of λ on prediction quality (Equation 8).

4.2.3 Inter-query association measures. Table 3 presents the prediction quality of our approach with the various inter-query association measures proposed in Section 3.3. Hereafter, due to space limitations, we only report the results for two pre-retrieval and two post-retrieval predictors which yield the best prediction quality (per collection) in Table 2 when used in our approach.

In addition, we present the results for the two special cases of our approach: OnlyAsso, which assumes that reference queries are effective to the same extent (Equation 9), and OnlyRef, which assumes that reference queries are uniformly distributed (Equation 10). We also experimented with a previously proposed estimate for $p(q'|i)$ [44], termed Geo. We report the prediction quality of using Equation 8, where Geo is used instead of the inter-query association measure; the resulting predictor is denoted Ref-Geo.

We can see in Table 3 that our approach statistically significantly outperforms the baseline in the vast majority of cases regardless of the inter-query association measure employed. The best prediction quality is almost always attained for ROBUST by Ref-RBO and for CW12 by Ref-Overlap. The lowest numbers are observed for Ref-Geo; this is presumably because Geo is not effective in predicting

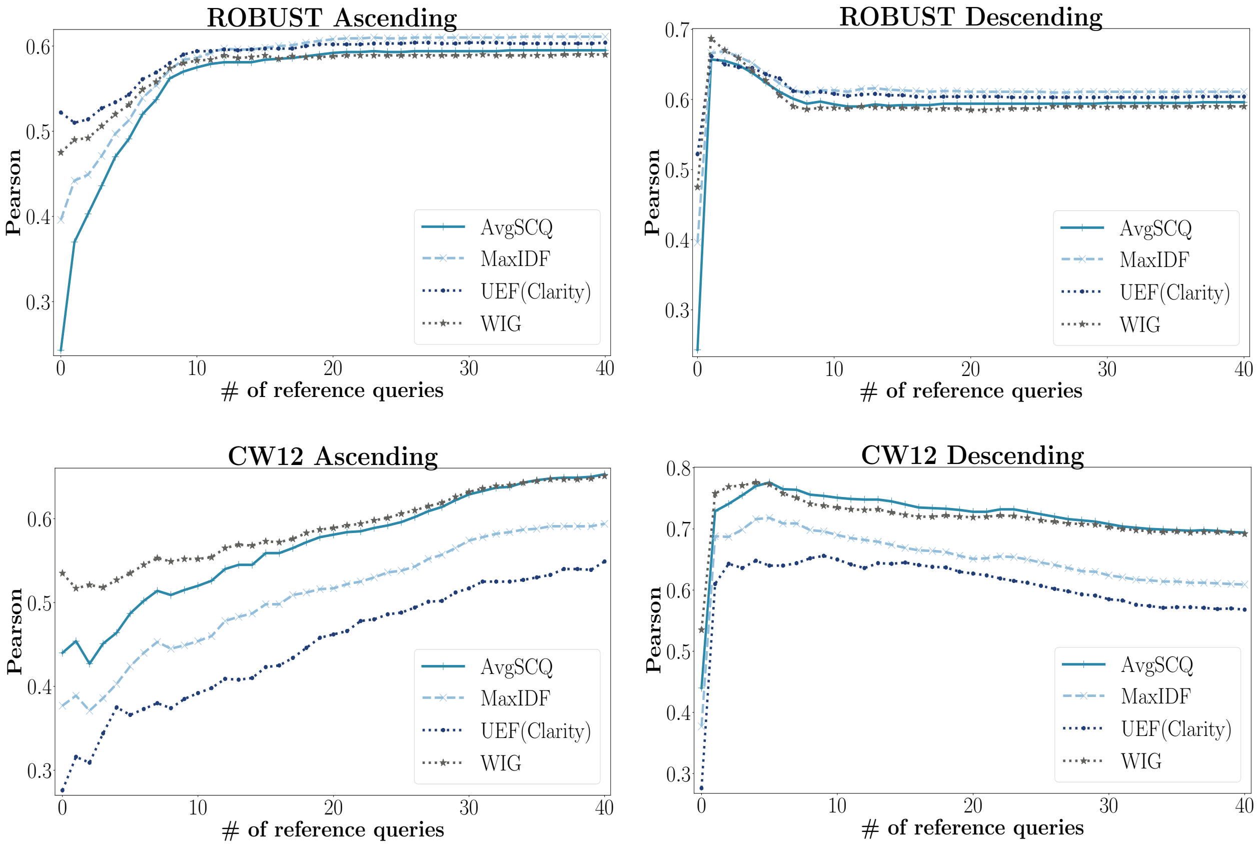

Figure 4: The effect of the number of reference queries on the prediction quality of Ref-RBO (Equation 8) when reference queries are added in Ascending or Descending order of AP effectiveness.

performance for queries representing the same information need in our approach — i.e., the reference queries we use.

We also see in Table 3 that even when all reference queries are assumed to be associated with the query to the same extent (OnlyRef), or when we do not use prediction for the reference queries in our approach (OnlyAsso), the prediction quality of our approach surpasses that of the baseline in the vast majority of the cases; nonetheless, using both estimates in our approach results in better prediction quality in the vast majority of cases.

As already noted, pre-retrieval predictors are typically more efficient than post-retrieval predictors because they can be computed before the retrieval is performed. All the predictors instantiated in our framework are post-retrieval. The only exception is when Ref-Jaccard is used together with a pre-retrieval predictor $\mathcal{P}$. Interestingly, this combination results in very high prediction quality. A case in point, for CW12, the prediction quality of Ref-Jaccard[AvgSCQ] (.608) surpasses that of all the baseline pre- and post-retrieval predictors presented in Table 2.

4.2.4 The effect of λ on prediction quality. In Figure 3 we study the effect of λ on the prediction quality of Ref-RBO. (Recall that $\lambda = 0$ amounts to the baseline: applying an existing predictor directly to the query as is standard.) Similar patterns were observed for the other inter-query association measures considered. We see that the best prediction quality is always attained for $\lambda \geq 0.7$, i.e., when a high weight is given to the reference queries; yet, in most cases the optimal λ is (slightly) smaller than 1, attesting to the additional contribution to prediction quality of also accounting for the prediction performed directly to the query.

4.2.5 The effectiveness of the reference queries. Thus far, all the available query variations served as the reference queries in our framework regardless of their retrieval effectiveness. In what follows, we divide the variations into two halves: the queries with the highest (High) and lowest (Low) AP values per topic. In Table 4 we study the merits of using each of the two sets (in comparison to using all variations) as reference queries. We observe the following: (i) using High yields better prediction quality than using Low or using all the variations; the differences are statistically significant in the vast majority of cases; and, (ii) using each of the sets (including Low) is superior to not using reference queries at all (i.e., the baseline); that is, using even poor variations as reference queries can be beneficial for prediction using our approach.

4.2.6 Varying the number of reference queries. We next study the effect of the number of reference queries on prediction quality. In Figure 4 we show the prediction quality of Ref-RBO as a function of the number of reference queries used; the queries are added one by

Table 4: Prediction quality with respect to the effectiveness of reference queries. Low and High: using the set of queries with the highest and lowest AP values, respectively. All: using all queries. '*b*' and '*a*' mark statistically significant difference with the baseline and All, respectively. '*l*' marks statistically significant differences between Low and High.

		ROBUST		CW12	
$\mathcal{P}$	Quantile	baseline	Ref-RBO	baseline	Ref-RBO
AvgSCQ	Low	.243	$.495^{a}_{b}$	.440	$.555^{a}_{b}$
	High	.243	$.634^{al}_{b}$	.440	$.730^{al}_{b}$
	All	.243	$.595_{b}$	.440	$.696_{b}$
MaxIDF	Low	.396	$.525^{a}_{b}$	.377	$.492^{a}_{b}$
	High	.396	$.647^{al}_{b}$	.377	$.671^{al}_{b}$
	All	.396	$.611_{b}$	.377	$.627_{b}$
UEF(Clarity)	Low	.522	$.562^{a}_{b}$	.276	$.413^{a}_{b}$
	High	.522	$.639^{al}_{b}$	.276	$.636^{al}_{b}$
	All	.522	$.603_{b}$	.276	$.578_{b}$
WIG	Low	.475	$.533^{a}_{b}$	.535	$.577^{a}_{b}$
	High	.475	$.652^{al}_{b}$	.535	$.721^{al}_{b}$
	All	.475	$.590_{b}$	.535	$.691_{b}$

one either in ascending or descending order of AP effectiveness[10]. We can see that the highest prediction quality is attained when using a relatively small number of highly effective queries: only one query is needed in ROBUST and about five queries are required in CW12. However, when using less effective reference queries, a much larger number of queries is needed to reach the highest prediction quality. For CW12, prediction quality gradually improves as more queries are added. For ROBUST, a plateau is attained after adding about twenty queries.

4.2.7 Putting it all together. Thus far, we demonstrated the clear merits of our approach when predicting performance for titles of TREC topics which served for queries. In Table 1 we showed that prediction quality of existing predictors varies considerably depending on the effectiveness of the query for which performance is predicted. Obviously, the TREC title query might be the best or the worst in terms of representing the topic (information need). Indeed, Figure 2 showed that in some cases, the effectiveness of the title query can be quite different than the median effectiveness of variants representing the topic. So, an important question we consider next is whether our approach is effective in predicting performance for queries of varying effectiveness in terms of representation of the underlying information need. It is important to differentiate this question from the one we explored in Section 4.2.5: the impact on prediction quality of using reference queries of different effectiveness to predict the performance of title queries.

Table 5 present the results for Ref-RBO when prediction is performed for the queries with the highest (Max), median (Med) and lowest (Min) AP per topic. We present for reference the prediction quality of predicting performance for title queries. All variations available for a topic are used as reference queries. We can conclusively see that our approach substantially outperforms the baseline regardless of the effectiveness of the query for which prediction is performed. This finding attests to the robustness of our approach with respect to existing predictors.

Table 5: Prediction quality of Ref-RBO for queries with the maximal (Max), median (Med) and minimal (Min) AP, and for the title queries. '∗' marks statistically significant differences with the baseline: applying $\mathcal{P}$ directly to the query as is standard. Best result in a column in a block is boldfaced.

		ROBUST		CW12	
$\mathcal{P}$		baseline	Ref-RBO	baseline	Ref-RBO
AvgSCQ	Max	**.339**	$.583^{*}$	**.469**	**$.747^{*}$**
	Med	.281	$.631^{*}$	.328	$.660^{*}$
	Min	.220	**$.724^{*}$**	.247	$.695^{*}$
	Title	.243	$.595^{*}$	.440	$.696^{*}$
MaxIDF	Max	**.412**	$.588^{*}$	.326	**$.697^{*}$**
	Med	.391	$.627^{*}$	.369	$.614^{*}$
	Min	.235	**$.694^{*}$**	.297	$.688^{*}$
	Title	.396	$.611^{*}$	**.377**	$.627^{*}$
UEF(Clarity)	Max	.517	$.583^{*}$	**.294**	$.659^{*}$
	Med	**.541**	$.644^{*}$	.263	$.557^{*}$
	Min	.468	**$.698^{*}$**	.292	**$.725^{*}$**
	Title	.522	$.603^{*}$	.276	$.578^{*}$
WIG	Max	**.511**	$.595^{*}$	**.549**	**$.774^{*}$**
	Med	.454	$.644^{*}$	.500	$.698^{*}$
	Min	.391	**$.714^{*}$**	.416	$.688^{*}$
	Title	.475	$.590^{*}$	.535	$.691^{*}$

5 CONCLUSIONS AND FUTURE WORK

We demonstrated important connections between a query, an information need, and the prediction quality achievable with many commonly used query performance predictors. Specifically, we showed that the relative prediction quality patterns of existing predictors can substantially vary with respect to the effectiveness of the queries for which performance is predicted.

Accordingly, we reformulated the probabilistic foundation of the query-performance-prediction (QPP) task by explicitly accounting for the underlying information need and its connection to queries used to represent it. We then presented a novel QPP approach that incorporates additional information from the information need, in the form of queries that can represent it. The approach, which can be instantiated using any existing performance predictor, dramatically improves prediction quality irrespective of the effectiveness of the query for which prediction is performed, or of the QPP method used to instantiate the approach.

We intend to continue our exploration of the relationship between predicting performance for queries representing different information needs and predicting performance for queries representing the same information need, the latter of which remains as a grand challenge for the IR community.

[10] If the number of variations (on the x-axis) exceeds the number of variations for a specific topic, all variations available for this topic are used as reference queries.

Acknowledgements. We thank the reviewers for their comments. This work was partially supported by the Israel Science Foundation (grant no. 1136/17), the Australian Research Council's *Discovery Projects* Scheme (DP170102231), a Google Faculty Award, and an Amazon Research Award.

REFERENCES

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

[2] J. A. Aslam and V. Pavlu. 2007. Query Hardness Estimation Using Jensen-Shannon Divergence Among Multiple Scoring Functions. In *Proc. of ECIR*. 198–209.

[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proc. of SIGIR*. 725–728.

[4] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proc. of SIGIR*. 395–404.

[5] N. Balasubramanian and J. Allan. 2010. Learning to select rankers. In *Proc. of SIGIR*. 855–856.

[6] N. J. Belkin, C. C., W. B. Croft, and J. P. Callan. 1993. The effect of multiple query representations on information retrieval system performance. In *Proc. of SIGIR*. 339–346.

[7] N. J. Belkin, P. Kantor, E.A. Fox, and J.A. Shaw. 1995. Combining evidence of multiple query representation for information retrieval. *Information Processing and Management* 31, 3 (1995), 431–448.

[8] R. Benham and J. S. Culpepper. 2017. Risk-Reward Trade-offs in Rank Fusion. In *Proc. of ADCS*. 1–8.

[9] Y. Bernstein, B. Billerbeck, S. Garcia, N. Lester, F. Scholer, and J. Zobel. 2005. RMIT University at TREC 2005: Terabyte and Robust Track. In *Proc. of TREC-14*.

[10] D. Carmel and E. Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers.

[11] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg. 2006. What makes a query difficult?. In *Proc. of SIGIR*. 390–397.

[12] A.-G. Chifu, L. Laporte, J. Mothe, and Md Z. Ullah. 2018. Query Performance Prediction Focused on Summarized Letor Features. In *Proc. of SIGIR*. 1177–1180.

[13] N. Craswell and M. Szummer. 2007. Random walks on the click graph. In *Proc. of SIGIR*. 239–246.

[14] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. 2002. Predicting query performance. In *Proc. of SIGIR*. 299–306.

[15] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. 2004. *A Language Modeling Framework for Selective Query Expansion*. Technical Report IR-338. Center for Intelligent Information Retrieval, University of Massachusetts.

[16] R. Cummins. 2011. Predicting Query Performance Directly from Score Distributions. In *Proc. of AIRS*. 315–326.

[17] R. Cummins. 2014. Document Score Distribution Models for Query Performance Inference and Prediction. *ACM Transactions on Information Systems* 32, 1 (2014), 2.

[18] R. Cummins, J. M. Jose, and C. O'Riordan. 2011. Improved query performance prediction using standard deviation. In *Proc. of SIGIR*. 1089–1090.

[19] V. Dang, M. Bendersky, and W. B. Croft. 2010. Learning to rank query reformulations. In *In Proc. of SIGIR*. 807–808.

[20] F. Diaz. 2007. Performance prediction using spatial autocorrelation. In *Proc. of SIGIR*. 583–590.

[21] C. Hauff, L. Azzopardi, and D. Hiemstra. 2009. The Combination and Evaluation of Query Performance Prediction Methods. In *Proc. of ECIR*. 301–312.

[22] C. Hauff, D. Hiemstra, and F. de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proc. of CIKM*. 1419–1420.

[23] B. He and I. Ounis. 2004. Inferring Query Performance Using Pre-retrieval Predictors. In *Proc. of SPIRE*. 43–54.

[24] R. Jones, B. Rey, O. Madani, and W. Greiner. 2006. Generating query substitutions. In *Proc. of WWW*. 387–396.

[25] O. Kurland, A. Shtok, S. Hummel, F. Raiber, D. Carmel, and O. Rom. 2012. Back to the Roots: A Probabilistic Framework for Query-performance Prediction. In *Proc. of CIKM*. 823–832.

[26] K. Kwok, L. Grunfeld, H. Sun, P. Deng, and N. Dinstl. 2004. TREC 2004 Robust Track Experiments using PIRCS. In *Proc. of TREC-13*.

[27] V. Lavrenko and W. B. Croft. 2001. Relevance-Based Language Models. In *Proc. of SIGIR*. 120–127.

[28] D. Metzler and W. B. Croft. 2005. A Markov random field model for term dependencies. In *Proc. of SIGIR*. 472–479.

[29] J. Mothe and L. Tanguy. 2005. Linguistic features to predict query difficulty. In *ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications*. http://www.haifa.il.ibm.com/sigir05-qp/papers/Mothe.pdf

[30] J. Pérez-Iglesias and L. Araujo. 2010. Standard Deviation as a Query Hardness Estimator. In *Proc. of SPIRE*. 207–212.

[31] F. Raiber and O. Kurland. 2014. Query-performance prediction: Setting the expectations straight. In *Proc. of SIGIR*. 13–22.

[32] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *Proc. of TREC-3*.

[33] H. Roitman. 2018. An Extended Query Performance Prediction Framework Utilizing Passage-Level Information. In *Proc. of ICTIR*. 35–42.

[34] H. Roitman. 2018. Query Performance Prediction using Passage Information. In *Proc. of SIGIR*. 893–896.

[35] H. Roitman, S. Erera, O. S. Shalom, and B. Weiner. 2017. Enhanced Mean Retrieval Score Estimation for Query Performance Prediction. In *Proc. of ICTIR*. 35–42.

[36] H. Roitman, S. Erera, and B. Weiner. 2017. Robust Standard Deviation Estimation for Query Performance Prediction. In *Proc. of ICTIR*. 245–248.

[37] H. Scells, L. Azzopardi, G. Zuccon, and B. Koopman. 2018. Query Variation Performance Prediction for Systematic Reviews. In *Proc. of SIGIR*. 1089–1092.

[38] F. Scholer and S. Garcia. 2009. A case for improved evaluation of query difficulty prediction. In *Proc. of SIGIR*. 640–641.

[39] F. Scholer, H. E. Williams, and A. Turpin. 2004. Query association surrogates for Web search. *JASIST* 55, 7 (2004), 637–650.

[40] D. Sheldon, M. Shokouhi, M. Szummer, and N. Craswell. 2011. LambdaMerge: merging the results of query reformulations. In *Proc. of WSDM*. 795–804.

[41] A. Shtok, O. Kurland, and D. Carmel. 2009. Predicting query performance by query-drift estimation. In *Proc. of ICTIR*. 305–312.

[42] A. Shtok, O. Kurland, and D. Carmel. 2010. Using statistical decision theory and relevance models for query-performance prediction. In *Proccedings of SIGIR*. 259–266.

[43] A. Shtok, O. Kurland, and D. Carmel. 2016. Query Performance Prediction Using Reference Lists. *ACM Trans. Inf. Syst.* 34, 4 (2016), 19:1–19:34.

[44] M. Sondak, A. Shtok, and O. Kurland. 2013. Estimating query representativeness for query-performance prediction. In *Proc. of SIGIR*. 853–856.

[45] F. Song and W. B. Croft. 1999. A general language model for information retrieval. In *Proc. of SIGIR*. 279–280.

[46] K. Sparck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments - Part 1. *Information Processing and Management* 36, 6 (2000), 779–808.

[47] Y. Tao and S. Wu. 2014. Query Performance Prediction By Considering Score Magnitude and Variance Together. In *Proc. of CIKM*. 1891–1894.

[48] P. Thomas, F. Scholer, P. Bailey, and A. Moffat. 2017. Tasks, Queries, and Rankers in Pre-Retrieval Performance Prediction. In *Proc. of ADCS*. 11:1–11:4.

[49] S. Tomlinson. 2004. Robust, Web and Terabyte Retrieval with Hummingbird Search Server at TREC 2004. In *Proc. of TREC-13*.

[50] Eduardo Vicente-López, Luis M. Campos, Juan M. Fernández-Luna, and Juan F. Huete. 2018. Predicting IR Personalization Performance Using Pre-retrieval Query Predictors. *J. Intell. Inf. Syst.* 51, 3 (2018), 597–620.

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

[52] E. M. Voorhees and D. K. Harman. 2005. *TREC: Experiments and evaluation in information retrieval*. The MIT Press.

[53] W. Webber, A. Moffat, and J. Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (Nov. 2010), 38 pages.

[54] M. Winaver, O. Kurland, and C. Domshlak. 2007. Towards robust query expansion: Model selection in the language model framework to retrieval. In *Proc. of SIGIR*. 729–730.

[55] D. Yin, Y. Hu, J. Tang, T. Daly, M. Zhou, H. Ouyang, J. Chen, C. Kang, H. Deng, C. Nobata, J.-M. Langlois, and Y. Chang. 2016. Ranking relevance in yahoo search. In *Proc. of KDD*. 323–332.

[56] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. 2005. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proc. of SIGIR*. 512–519.

[57] H. Zamani, W. B. Croft, and J. S. Culpepper. 2018. Neural Query Performance Prediction using Weak Supervision from Multiple Signals. In *Proc. of SIGIR*. 105–114.

[58] C.-X. Zhai and J. D. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proc. of SIGIR*. 334–342.

[59] Y. Zhao, F. Scholer, and Y. Tsegay. 2008. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In *Proc. of ECIR*. 52–64.

[60] Y. Zhou and W. B. Croft. 2006. Ranking robustness: a novel framework to predict query performance. In *Proc. of CIKM*. 567–574.

[61] Y. Zhou and W. B. Croft. 2007. Query performance prediction in web search environments. In *Proc. of SIGIR*. 543–550.