A Unified Framework for Post-Retrieval Query-Performance Prediction

Oren Kurland[1], Anna Shtok[1], David Carmel[2], and Shay Hummel[1]

[1] Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel
kurland@ie.technion.ac.il, annabel@tx.technion.ac.il, projphoto@gmail.com
[2] IBM Research, Haifa Lab, Haifa 31905, Israel
carmel@il.ibm.com

Abstract. The query-performance prediction task is estimating the effectiveness of a search performed in response to a query in lack of relevance judgments. Post-retrieval predictors analyze the *result list* of top-retrieved documents. While many of these previously proposed predictors are supposedly based on different principles, we show that they can actually be derived from a novel unified prediction framework that we propose. The framework is based on using a pseudo effective and/or ineffective ranking as reference comparisons to the ranking at hand, the quality of which we want to predict. Empirical exploration provides support to the underlying principles, and potential merits, of our framework.

Keywords: query-performance prediction, post-retrieval prediction framework.

1 Introduction

There has been much work throughout recent years on predicting *query performance* [4]. That is, estimating the effectiveness of a search performed in response to a query in lack of relevance judgments. Pre-retrieval query-performance predictors, for example, analyze the query and may use corpus-based statistics [11,4]. Post-retrieval predictors [6,4] also utilize information induced from the *result list* of the most highly ranked documents.

We present a (simple) novel unified post-retrieval prediction framework that can be used to derive many previously proposed post-retrieval predictors that are supposedly based on completely different principles. The framework is based on using a pseudo effective and/or ineffective ranking(s) as reference comparisons to the ranking at hand, the effectiveness of which we want to predict. The more similar the given ranking to the pseudo effective ranking and dissimilar to the pseudo ineffective ranking the higher its effectiveness is presumed to be. As it turns out, many previous post-retrieval predictors simply differ by the choice of the pseudo (in)effective ranking that serves for reference, and/or the inter-ranking similarity measure used.

Experiments performed using TREC datasets provide empirical support to the underlying principles, and potential merits, of our framework. For example,

G. Amati and F. Crestani (Eds.): ICTIR 2011, LNCS 6931, pp. 15–26, 2011.

while current predictors use either a pseudo effective or a pseudo ineffective ranking, we demonstrate the potential merits of using both.

2 Related Work

Post-retrieval query-performance prediction methods are based on analyzing the result list of top-retrieved documents [4]. These methods can be classified into three categories [4]. Clarity-based approaches [6] estimate the focus of the result list with respect to the corpus. Robustness-based approaches [19,22,18,23,2] measure the stability of the result list under perturbations of the query, documents, and the retrieval method. Score-distribution-based approaches [8,23,15] utilize properties of the retrieval scores in the result list. We show that predictors representing these three categories can be derived from, and explained by, our proposed post-retrieval prediction framework.

A utility estimation framework (UEF) [16], which inspired the development of our framework, is based on estimating a relevance model and using it to induce a pseudo effective ranking. The induced ranking serves as a reference comparison in estimating the quality of a given ranking as in our framework. Yet, UEF, which we show to be a specific case of our framework, was used to derive predictors based on a specific way of inducing a pseudo effective ranking. We show that several previous predictors can be instantiated from our framework by using different approaches for inducing a pseudo effective ranking. More importantly, in contrast to our framework, UEF does not utilize a (pseudo) ineffective ranking as a reference comparison. Thus, quite a few predictors that we derive from our framework cannot be derived from UEF. Moreover, we demonstrate in Section 4 the merits of using both pseudo effective and ineffective rankings

A conceptual framework for modeling (predicting) topic difficulty [5] is based on similarities between the query, the result list, and the corpus. In contrast, our framework predicts the effectiveness of a ranking by measuring its similarity with (pseudo) effective and ineffective rankings. The corpus, which served to induce a non-relevance model in this framework [5], is utilized in our framework for inducing pseudo ineffective rankings that are used to derive several predictors.

3 Query-Performance Prediction Framework

Suppose a retrieval method $\mathcal{M}$ is employed in response to query q over a corpus of documents $\mathcal{D}$ so as to satisfy the information need expressed by q. The goal of query-performance prediction methods is to quantify the effectiveness of the resultant corpus ranking, denoted $\pi_{\mathcal{M}}(q;\mathcal{D})$, in lack of relevance judgments.

Now, let $\pi_{opt}(q;\mathcal{D})$ be the optimal corpus ranking with respect to the information need expressed by q as defined by the probability ranking principle [13]; that is, a ranking that corresponds to the "true" degrees (probabilities) of documents' relevance. Naturally, the more "similar" the given ranking $\pi_{\mathcal{M}}(q;\mathcal{D})$ is to the optimal ranking $\pi_{opt}(q;\mathcal{D})$, the more effective it is:

$$Q(\pi_{\mathcal{M}}(q;\mathcal{D})) \overset{def}{=} Sim(\pi_{\mathcal{M}}(q;\mathcal{D}), \pi_{opt}(q;\mathcal{D})) \; ; \tag{1}$$

$Q(\pi_{\mathcal{M}}(q;\mathcal{D}))$ is the quality (effectiveness) of $\pi_{\mathcal{M}}(q;\mathcal{D})$ that we aim to predict; and, $Sim(\cdot,\cdot)$ is an inter-ranking similarity measure discussed below.

One way to derive a prediction method using Eq. 1 is to try to approximate the optimal ranking. This is the task addressed, for example, by probabilistic retrieval methods that estimate the probability of a document being relevant. Now, if we have a retrieval approach that is known, in general, to be quite effective, we could use it to induce a *pseudo effective* (PE) corpus ranking $\pi_{PE}(q;\mathcal{D})$. Then, the PE ranking can be used in Eq. 1, instead of the optimal ranking, as a reference comparison in estimating (predicting) $\mathcal{M}$'s ranking effectiveness:

$$\hat{Q}_{PE}(\pi_{\mathcal{M}}(q;\mathcal{D})) \stackrel{def}{=} Sim(\pi_{\mathcal{M}}(q;\mathcal{D}),\pi_{PE}(q;\mathcal{D})) \ . \tag{2}$$

Clearly, the quality of predictors derived from Eq. 2 depends on the actual effectiveness of $\pi_{PE}(q;\mathcal{D})$, and on the inter-ranking similarity measure used. To potentially improve the ranking-quality estimate in Eq. 2, we use the dissimilarity between $\mathcal{M}$'s ranking and a *pseudo ineffective* (PIE) ranking as a means of regularization:

$$\begin{aligned}&\hat{Q}_{PE;PIE}(\pi_{\mathcal{M}}(q;\mathcal{D})) \stackrel{def}{=} \\ &\quad \alpha(q)Sim(\pi_{\mathcal{M}}(q;\mathcal{D}),\pi_{PE}(q;\mathcal{D})) - \beta(q)Sim(\pi_{\mathcal{M}}(q;\mathcal{D}),\pi_{PIE}(q;\mathcal{D})) \ ;\end{aligned} \tag{3}$$

$\alpha(q)$ and $\beta(q)$ are (query-dependent) weights. This approach is conceptually reminiscent of Rocchio's retrieval method [14] that is based on using interpolation of prototypes of relevant and non-relevant documents for query refinement.

Retrieval effectiveness measures such as mean average precision (MAP) and precision@k attribute much more importance to documents at high ranks than to those at low ranks. Consequently, post-retrieval query-performance predictors [4] analyze the *result list* of the documents most highly ranked rather than the entire corpus ranking. Along the same lines, we approximate the quality of the given corpus ranking, $\pi_{\mathcal{M}}(q;\mathcal{D})$, by focusing on the highest ranks. Formally, let $L_x^{[k]}$ denote the result list of the k highest ranked documents in x's ranking. The ranking-quality estimate from Eq. 3 is approximated using an estimate for the quality of the result list $L_{\mathcal{M}}^{[k]}$, which is in turn estimated based on the similarity of $L_{\mathcal{M}}^{[k]}$ with the result lists of the PE and PIE rankings:

$$\hat{Q}_{PE;PIE}(\pi_{\mathcal{M}}(q;\mathcal{D})) \approx \alpha(q)Sim(L_{\mathcal{M}}^{[k]},L_{PE}^{[k]}) - \beta(q)Sim(L_{\mathcal{M}}^{[k]},L_{PIE}^{[k]}) \ . \tag{4}$$

Various inter-ranking (list) similarity measures ($Sim(\cdot,\cdot)$) can be used. For example, if both lists are (different) rankings of the same document set, then Kendall's-τ, which uses rank information, or Pearson's correlation coefficient computed based on retrieval scores in the lists, can be applied. Document content can also be used to induce inter-ranking (list) similarity as we discuss below.

3.1 Deriving Previously Proposed Predictors

We next show that several previously proposed post-retrieval predictors can be instantiated from the framework described above (Eq. 4). Specifically, either a pseudo effective or ineffective result list is used as a reference comparison to the given result list ($L_{\mathcal{M}}^{[k]}$), and some inter-list similarity measure is used.

Using a Pseudo Ineffective (PIE) Result List

Clarity. The clarity predictor estimates the focus of the given result list, $L_{\mathcal{M}}^{[k]}$, with respect to the corpus by measuring the (KL) divergence between their induced language models [6]. The assumption is that the more distant the models are, the more focused the result list; therefore, the higher the quality of $\pi_{\mathcal{M}}(q;\mathcal{D})$.

Clarity can be explained as a specific instance of the prediction framework described above. Let $\alpha(q) = 0$ and $\beta(q) = 1$; i.e., only a pseudo ineffective (PIE) result list $L_{PIE}^{[k]}$ is used. The PIE list is composed of k instances of the corpus that represents a general (average) non-relevant document. (The documents in the corpus can be concatenated to yield one long document to be used.) Let $p(\cdot|L)$ denote a language model induced from the document list L; and, let $Sim(L_1, L_2) \overset{def}{=} -KL\big(p(\cdot|L_1)||p(\cdot|L_2)\big)$ be an inter-list similarity measure that is based on the KL divergence between the lists' language models. Indeed, the clarity of $L_{\mathcal{M}}^{[k]}$ is defined as $-Sim(L_{\mathcal{M}}^{[k]}, L_{PIE}^{[k]})$; $p(\cdot|L_{\mathcal{M}}^{[k]})$ is a relevance language model [12] induced from $L_{\mathcal{M}}^{[k]}$; and, $p(\cdot|L_{PIE}^{[k]})$ is the corpus language model, as the corpus is the only (pseudo) document that appears (k times) in $L_{PIE}^{[k]}$.

Weighted information gain (WIG). The WIG predictor is based on measuring the amount of information in the given result list $L_{\mathcal{M}}^{[k]}$ with respect to that in a result list that is created using the corpus as an average non-relevant document [23]. In practice, WIG is computed by the average divergence of retrieval scores of documents in $L_{\mathcal{M}}^{[k]}$ from that of the corpus. When retrieval scores reflect surface-level document-query similarities, the higher the divergence, the higher the query-similarity documents in the list exhibit with respect to that of the corpus; consequently, the more effective $L_{\mathcal{M}}^{[k]}$ is presumed to be.

As with clarity, to derive WIG from our framework we set $\alpha(q) = 0$, $\beta(q) = 1$, and $L_{PIE}^{[k]}$ to k copies of the corpus, which serves for a non-relevant document. The (average) $L1$ distance between retrieval scores serves for an inter-list similarity measure; that is, $Sim(L_{\mathcal{M}}^{[k]}, L_{PIE}^{[k]}) \overset{def}{=} \frac{1}{k}(\sum_{i=1\ldots k} Score(L_{\mathcal{M}}^{[k]}(i)) - Score(L_{PIE}^{[k]}(i))$, where $L(i)$ is the document at rank i of list L and $Score(L(i))$ is its retrieval score in the list. (Recall that for $i \in \{1,\ldots,k\}$ $L_{PIE}^{[k]}(i) \overset{def}{=} \mathcal{D}$.)[1]

Thus, the difference between WIG and clarity, as instantiated from our framework, is the measure used to compute the (dis)similarity between the given result list and a result list composed of k copies of the corpus that serves for a non-relevant document.[2]

[1] In implementation, the retrieval scores used by WIG are further normalized so as to ensure inter-query compatibility.

[2] See Zhou [21] for an alternative view of the connection between WIG and clarity.

NQC. The NQC predictor [15] measures the standard deviation of retrieval scores in the result list. It was shown that the mean retrieval score in the list corresponds to the retrieval score of a centroid-based representation of the documents in the list [15] for some retrieval methods for which retrieval scores represent document-query similarities. Furthermore, the list centroid was argued to manifest query drift, and hence, could be thought of as a pseudo non-relevant document that exhibits relatively high query similarity. Accordingly, high divergence of retrieval scores from that of the centroid, measured by the standard deviation, was argued, and empirically shown, to imply high quality of the result list.

Hence, if we (i) set $\alpha(q) = 0$ and $\beta(q) = 1$, (ii) use k instances of the centroid-based representation of $L_{\mathcal{M}}^{[k]}$ (denoted $Cent(L_{\mathcal{M}}^{[k]})$) to create a pseudo ineffective list ($L_{PIE}^{[k]}$), and (iii) use $Sim(L_{\mathcal{M}}^{[k]}, L_{PIE}^{[k]}) \overset{def}{=} -\sqrt{\frac{1}{k}\sum_{i=1\ldots k}\left(Score(L_{\mathcal{M}}^{[k]}(i)) - Score(L_{PIE}^{[k]}(i))\right)^2}$ for an inter-list similarity measure (note that $L_{PIE}^{[k]}(i) \overset{def}{=} Cent(L_{\mathcal{M}}^{[k]})$), we derive NQC from our framework[3].

Recall from above that WIG uses the $L1$ distance between retrieval scores in $L_{\mathcal{M}}^{[k]}$ and those in a PIE list composed of k copies of the corpus, which serves as a *general non-relevant document.* In comparison, NQC uses the $L2$ distance between the retrieval scores in $L_{\mathcal{M}}^{[k]}$ and those in a PIE list composed of k copies of a *pseudo non-relevant document* that exhibits high surface-level query similarity (i.e., $Cent(L_{\mathcal{M}}^{[k]})$).

Query-independent vs. query-dependent ranking. Another approach for producing a pseudo ineffective result list, $L_{PIE}^{[k]}$, is based on re-ranking the given result list, $L_{\mathcal{M}}^{[k]}$, using non-query-dependent information; e.g., based on documents' PageRank [3]. The idea is that the higher the divergence between $L_{\mathcal{M}}^{[k]}$'s original ranking and its query-independent re-ranked version, the higher the quality of $L_{\mathcal{M}}^{[k]}$; Kendall's-tau, for example, can serve for an inter-ranking similarity measure [3]. Thus, this approach is another instance of our framework when using only a PIE list (i.e., the query-independent ranked version of $L_{\mathcal{M}}^{[k]}$) with $\beta(q) = 1$.

Using a Pseudo Effective (PE) Result List

Query feedback. In the query feedback (QF) predictor [23], a query model is induced from $L_{\mathcal{M}}^{[k]}$ and is used to rank the entire corpus. Then, the overlap (i.e., number of shared documents) between the n_{QF} highly ranked documents by this retrieval, and the n_{QF} highly ranked documents by the given ranking $\pi_{\mathcal{M}}(q; \mathcal{D})$ (n_{QF} is a free parameter), presumably indicates the effectiveness of the latter. That is, the higher the overlap, the less non-query-related noise there is in $L_{\mathcal{M}}^{[k]}$ from which the query model was induced; hence, $L_{\mathcal{M}}^{[k]}$, and the $\pi_{\mathcal{M}}(q; \mathcal{D})$ ranking from which it was derived, are considered of higher quality.

[3] To ensure inter-query compatibility of prediction values, documents' retrieval scores are scaled using that of the corpus.

The retrieval performed over the corpus using the query model induced from $L_{\mathcal{M}}^{[k]}$ is essentially pseudo-feedback-based query-expansion retrieval. As is known, such retrieval outperforms, on average, that of using only the original query. Thus, the result list of k highest ranked documents produced by using the induced query model could be considered as pseudo effective (PE) on average; let $L_{PE}^{[k]}$ denote this list. Accordingly, the overlap at cutoff n_{QF} between $L_{\mathcal{M}}^{[k]}$ and $L_{PE}^{[k]}$ serves as the inter-list similarity measure. Setting $\alpha(q) = 1$ and $\beta(q) = 0$, i.e., using only the similarity with the pseudo effective ranking just mentioned, we get that QF is a specific instance of our framework.

Utility Estimation Framework (UEF). The basic idea underlying UEF [16] is to devise a supposedly effective representation of the underlying information need (specifically, using a relevance model approach [12]). This representation is used to re-rank the given result list $L_{\mathcal{M}}^{[k]}$. The resultant re-ranked version of $L_{\mathcal{M}}^{[k]}$ is presumably of relatively high quality, and is thereby denoted here $L_{PE}^{[k]}$. The similarity between $L_{\mathcal{M}}^{[k]}$ and $L_{PE}^{[k]}$ ($Sim(L_{\mathcal{M}}^{[k]}, L_{PE}^{[k]})$) is measured using Kendall's-τ, Pearson's coefficient, or Spearman's-ρ. The similarity value is scaled by an estimate for the quality of the information need representation. The motivation is to model the confidence in the ability to derive an effective representation of the information need, and use the level of confidence so as to adjust the prediction value. Thus, UEF is a specific instance of our proposed framework wherein $\beta(q) = 0$ (i.e., no pseudo ineffective result list is used), and $\alpha(q)$ is the estimate for the quality of the information need representation.

Autocorrelation. Applying score regularization — specifically, adjusting the retrieval score of a document using information induced from similar documents — upon the given result list $L_{\mathcal{M}}^{[k]}$ so that the resultant retrieval scores "respect" the cluster hypothesis is another way to produce a pseudo effective result list [8]. The (Pearson) correlation between the retrieval scores in $L_{\mathcal{M}}^{[k]}$ and $L_{PE}^{[k]}$ serves for an inter-list similarity measure. Hence, this (spatial) *autocorrelation* approach [8] is also an instance of our framework (with $\alpha(q) = 1$ and $\beta(q) = 0$).

Utilizing fusion. All predictors discussed above are based on a single retrieval (if at all) used to create a pseudo (in)effective ranking. Alternatively, fusion of multiple rankings can be used to produce a pseudo effective ranking [8]. Indeed, the merits of fusion, in terms of retrieval effectiveness, have been acknowledged [9]. Pearson's correlation between the given result list and that produced by fusion served for query-performance prediction [8]. Clearly, this prediction approach is a specific instance of our framework (with $\alpha(q) = 1$ and $\beta(q) = 0$).

Intermediate summary. As was shown above, various post-retrieval predictors can be derived from Eq. 4. The predictors use either a pseudo effective ranking or a pseudo ineffective ranking but not both. The pseudo effective rankings were induced using pseudo-feedback-based retrieval [23,16], score regularization [8], and fusion [8]. Pseudo-ineffective rankings were induced using the corpus [6,23], a centroid of the result list [15], and a query-independent retrieval method [3]. The inter-ranking similarity measures used were based on (i) the $L1$

[23] and $L2$ [15] distances of retrieval scores and their Pearson correlation [8,16], (ii) the KL divergence between induced language models [6], (iii) Kendall's-τ [3,16] and the document overlap [23] between result lists.

4 Experiments

We next present an empirical study of the potential merits of our framework. In Section 4.2 we explore the basic premise underlying the framework, the utilization of both pseudo effective and pseudo ineffective rankings, and a use case demonstrating the intricacies of utilizing pseudo ineffective rankings.

Parts of the study are based on utilizing (little) relevance feedback to control the effectiveness of reference rankings. Although feedback is often not available for query-performance prediction, the exploration using it shows that the ability to devise effective reference rankings to be used in our framework yields prediction quality that substantially transcends state-of-the-art.

4.1 Experimental Setup

We used the following TREC collections and queries for experiments:

Collection	Data	Num Docs	Topics
TREC4	Disks 2&3	567,529	201-250
TREC5	Disks 2&4	524,929	251-300
WT10G	WT10g	1,692,096	451-550
ROBUST	Disk 4&5-CR	528,155	301-450,601-700

Topics' titles serve for queries; for TREC4 topics' descriptions are used as titles are not available. Porter stemming and stopword removal (using INQUERY's list) were applied using the Lemur toolkit (www.lemurproject.org), which was also used for retrieval.

To measure prediction quality, we follow common practice [4] and compute Pearson's correlation between the values assigned by a predictor to queries, and the "true" average precision (AP, computed at cutoff 1000) values for these queries determined based on TREC's relevance judgments.

Language modeling framework. The goal of the predictors we study is predicting the effectiveness of rankings induced in response to the queries specified above by the *query likelihood* (QL) retrieval method [17]. Let $p(w|x)$ denote the probability assigned to term w by a (smoothed) unigram language model induced from text (collection) x. (Specific language-model induction details are provided below.) The (log) query likelihood score of document d with respect to query q ($= \{q_i\}$), which is used for ranking the corpus, is $Score_{QL}(q;d) \overset{def}{=} \log p(q|d) \overset{def}{=} \log \prod_{q_i \in q} p(q_i|d)$. The result list of k highest ranked documents is denoted $L^{[k]}_{q;QL}$.

Some of the predictors we explore utilize *relevance language models* [12]. Let R^S be a relevance model[4] constructed from a document set S: $p(w|R^S) \overset{def}{=}$

[4] We use the RM1 relevance model. While for retrieval purposes, RM3 [1], which interpolates RM1 with the query model is more effective, RM1 is more effective for performance prediction with the predictors we study as previously reported [16].

$\sum_{d \in S} p(w|d) wt(d)$; $wt(d)$ is d's weight ($\sum_{d \in S} wt(d) = 1$). To score document d with respect to R^S, so as to induce ranking, the minus cross entropy between R^S and d's language model is used: $Score_{CE}(R; d) \stackrel{def}{=} \sum_w p(w|R) \log p(w|d)$.

The standard pseudo-feedback-based relevance model, denoted R^{Res}, is constructed from the result list ($S \stackrel{def}{=} L^{[k]}_{q;QL}$); $wt(d) \stackrel{def}{=} p(d|q) \stackrel{def}{=} \frac{p(q|d)}{\sum_{d' \in L^{[k]}_{q;QL}} p(q|d')}$ [12]. To control the effectiveness of some reference rankings, we also use a relevance model, R^{Rel}, that is constructed from a set S of r relevant documents that are the highest ranked by QL (r is a free parameter); $wt(d) \stackrel{def}{=} \frac{1}{r}$.

Implementation. We use three state-of-the-art predictors that were shown above to be specific instances of our framework. The first is a (conceptually) generalized version of the QF method [23]: the overlap at top (n_{QF}) ranks between the given result list, $L^{[k]}_{q;QL}$, and a result list created from the corpus using a relevance model constructed from documents in $L^{[k]}_{q;QL}$ serves for prediction. Changing the relevance model enables to study the effect of using reference rankings of varying effectiveness. The other two predictors are clarity [6] and NQC [15].

We use Dirichlet-smoothed unigram document language models with the smoothing parameter set to 1000 [20]. For constructing a relevance model, a non-smoothed maximum likelihood estimate is used for document language models [16]; and, all relevance models use 100 terms [16]. For the QF and clarity predictors, the QL result list size, k, is set to 100, which yields high quality prediction [16]; for NQC, the effect of k is studied.

4.2 Experimental Results

Using Effective Rankings as Reference Comparisons. In Fig. 1 we present the effect on QF's prediction quality of using reference rankings of varying effectiveness. Specifically, we construct a relevance model R^{Rel} from r (≥ 1) relevant documents. We then depict the MAP performance of using R^{Rel} for retrieval over the corpus; and, the resultant prediction quality of QF when using the corpus ranking induced by R^{Rel} for a reference ranking. We set the overlap cutoff parameter, n_{QF}, to 10; the patterns observed in the graphs are quite similar for $n_{QF} \in \{10, 25, 50, 100\}$. For $r = 0$, we use the result-list-based relevance model, R^{Res}, which corresponds to the original QF [23].

We can see in Fig. 1 that, as is known, the retrieval effectiveness of the relevance model increases when increasing the number of relevant documents from which it is constructed. Accordingly, the resultant prediction quality of QF increases when increasing the effectiveness of the ranking induced by the relevance model; specifically, the prediction quality becomes much better than that of using the result-list-based relevance model ($r = 0$), which is the current state-of-the-art QF approach.

Hence, we see that using reference rankings of higher effectiveness, which are induced here by using relevance models of higher quality, results in improved query-performance prediction. This finding provides support to the underlying premise of our framework. That is, high quality query-performance prediction

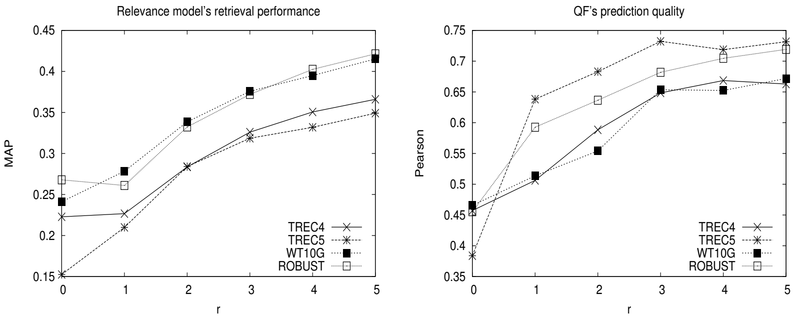

Fig. 1. Using a relevance model, R^{Rel}, constructed from r (≥ 1) relevant documents in QF; for $r = 0$, the (pseudo feedback) result-list-based relevance model, R^{Res}, is used. The left figure presents the MAP performance of using the relevance model for retrieval over the corpus. (The list of top-retrieved documents serves for reference in QF.) The right figure presents QF's resultant prediction quality.

can be attained by using an estimate of the "optimal" ranking as a reference comparison in estimating the effectiveness of the given ranking.

Using Both Effective and Ineffective Rankings. As the predictors discussed in Section 3 use either a (pseudo) effective or ineffective reference rankings, but not both, we now study the potential merits of using both.

We create an effective corpus ranking using a relevance model, R^{Rel}, constructed from 5 relevant documents. To measure the similarity between the corpus ranking and the QL ranking, the quality of which we want to predict, we use the **drift** method [7]. That is, we construct a relevance language model, denoted R_{QL}, from the QL result list ($L^{[k]}_{q;QL}$; $k = 100$); and, from the top-100 documents retrieved from the corpus using R^{Rel}, denoted $R_{R^{Rel}}$; uniform weights ($wt(d) \overset{def}{=} \frac{1}{100}$) are used, and R_{QL} and $R_{R^{Rel}}$ use 100 terms; $R_{R^{Rel}}$ is Jelinek-Mercer smoothed using a smoothing weight of 0.1. The minus KL divergence, $-KL\big(p(\cdot|R_{QL})||p(\cdot|R_{R^{Rel}})\big)$, serves for inter-list similarity measure. The resultant drift-based predictor is a variant of QF that used document overlap for inter-list similarity. We use this variant to have proper interpolation in Eq. 4 with the dissimilarity to an ineffective corpus-based ranking used by clarity.

Recall from Section 3 that clarity is defined as $KL\big(p(\cdot|R^{Res})||p(\cdot|L^{[k]}_{PIE})\big)$; $L^{[k]}_{PIE}$ is an ineffective list composed of k ($= 100$) copies of the corpus. We set $\alpha(q) \overset{def}{=} \lambda$ and $\beta(q) \overset{def}{=} (1 - \lambda)$ in Eq. 4 ($\lambda \in \{0, 0.1, \ldots, 1\}$) and derive the (novel) **drift+clarity** predictor, the quality of which is reported in Table 1. To study the potential prediction quality of utilizing both effective and ineffective lists, λ is set to a value that yields optimal prediction quality per corpus: 0.5, 0.3, 0.5, and 0.3 for TREC4, TREC5, WT10G and ROBUST, respectively.

It is important to conceptually differentiate the drift+clarity predictor just presented from the general case of linear interpolation of *prediction values.* Such interpolation can be based on the output of predictors that can use, for example, different inter-ranking similarity measures [23,10,16]. In contrast, drift+clarity is derived as a single predictor from Eq. 4, wherein the similarity of the given result list with an effective reference list (created using R^{Rel}), and dissimilarity with an ineffective reference list (created from the corpus) are interpolated; the (minus) KL divergence between lists' language models serves for inter-list similarity measure. In implementation, however, drift+clarity amounts to interpolating the prediction values of drift and clarity.

We see in Table 1 that although drift is much inferior to clarity, drift+clarity is much superior to clarity. This finding supports the potential merits of using both effective and ineffective reference rankings for performance prediction.

On Using Ineffective Rankings as Reference Comparisons. The NQC predictor [15] turns out to be an interesting example for demonstrating the merits, and intricacies, of using a pseudo ineffective reference ranking. NQC measures the standard deviation of retrieval scores in the result list ($L^{[k]}_{q;QL}$). As noted above, the mean retrieval score was shown to be the retrieval score of a centroid-based representation of the list; and, the centroid was argued to serve as a pseudo non-relevant document that exhibits high query similarity [15]. We showed above that NQC can be derived from our framework using a pseudo ineffective list that is composed of multiple copies of the centroid. In Fig. 2 we present the effect on NQC's prediction quality of varying the result list size, k. Below we argue that varying k affects the usefulness, in terms of resultant query-performance prediction, of the pseudo ineffective list created from the centroid.

We see in Fig. 2 that NQC's prediction quality monotonically improves when increasing k up till a point from which it monotonically decreases. Indeed, with very few documents in the result list (small k), the centroid is much affected by highly ranked relevant documents; thereby, it is not a very good basis for a useful ineffective reference list. Having more documents in the list when increasing k towards its optimal value, results in considering more non-relevant query-similar documents; thus, the centroid's usefulness for constructing ineffective reference ranking grows, and accordingly, prediction is improved.

Increasing k beyond its optimal value results in the centroid being much affected by non-relevant documents that exhibit low query similarity. Consequently, the centroid gradually becomes a "general" non-relevant document (as the corpus), rather than a query-similar non-relevant one. Now, the centroid's high query similarity was argued to be an important factor in NQC's high quality prediction [15]. Accordingly, further increasing k makes the centroid-based list less informative as a reference comparison thus decreasing prediction quality.

Table 1. Using both effective (drift) and ineffective (clarity) reference rankings for prediction (drift+clarity). Boldface marks the best result per column

	TREC4	TREC5	WT10G	ROBUST
drift	0.406	0.081	0.317	0.130
clarity	0.448	0.426	0.330	0.508
drift+clarity	**0.588**	**0.461**	**0.412**	**0.521**

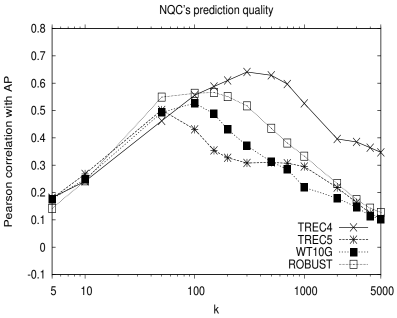

Fig. 2. NQC's prediction quality as a function of the result list size, *k*

We conclude that it is not only the ineffectiveness, in terms of retrieval performance, of the reference list that is important for successful performance prediction, but also the characteristics of the documents in it.

5 Summary and Future Work

We presented a novel unified framework for post-retrieval query-performance prediction which we used for deriving previously proposed predictors that are supposedly based on completely different principles. The framework uses (pseudo) effective and/or ineffective rankings as reference comparisons in estimating the effectiveness of a given ranking. Empirical exploration, based in part on exploiting little relevance feedback to induce effective reference rankings, provided support to the underlying principles, and potential merits, of the framework. Devising improved pseudo (in)effective reference rankings *for a given ranking* with zero feedback, and applying the framework to devise new post-retrieval predictors, is a future venue.

Acknowledgments. We thank the reviewers for their comments. This paper is based upon work supported in part by the Israel Science Foundation under grant no. 557/09, and by IBM's SUR award. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

References

1. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: UMASS at TREC 2004 — novelty and hard. In: Proceedings of TREC-13, pp. 715–725 (2004)
2. Aslam, J.A., Pavlu, V.: Query hardness estimation using jensen-shannon divergence among multiple scoring functions. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECiR 2007. LNCS, vol. 4425, pp. 198–209. Springer, Heidelberg (2007)
3. Bernstein, Y., Billerbeck, B., Garcia, S., Lester, N., Scholer, F., Zobel, J.: RMIT university at trec 2005: Terabyte and robust track. In: Proceedings of TREC-14 (2005)

4. Carmel, D., Yom-Tov, E.: Estimating the Query Difficulty for Information Retrieval. In: Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, San Francisco (2010)
5. Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: Proceedings of SIGIR, pp. 390–397 (2006)
6. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of SIGIR, pp. 299–306 (2002)
7. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: A language modeling framework for selective query expansion. Tech. Rep. IR-338, Center for Intelligent Information Retrieval, University of Massachusetts (2004)
8. Diaz, F.: Performance prediction using spatial autocorrelation. In: Proceedings of SIGIR, pp. 583–590 (2007)
9. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: Proceedings of TREC-2 (1994)
10. Hauff, C., Azzopardi, L., Hiemstra, D.: The combination and evaluation of query performance prediction methods. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 301–312. Springer, Heidelberg (2009)
11. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) SPIRE 2004. LNCS, vol. 3246, pp. 43–54. Springer, Heidelberg (2004)
12. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of SIGIR, pp. 120–127 (2001)
13. Robertson, S.E.: The probability ranking principle in IR. Journal of Documentation, 294–304 (1977)
14. Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall, Englewood Cliffs (1971)
15. Shtok, A., Kurland, O., Carmel, D.: Predicting query performance by query-drift estimation. In: Azzopardi, L., Kazai, G., Robertson, S., Rüger, S., Shokouhi, M., Song, D., Yilmaz, E. (eds.) ICTIR 2009. LNCS, vol. 5766, pp. 305–312. Springer, Heidelberg (2009)
16. Shtok, A., Kurland, O., Carmel, D.: Using statistical decision theory and relevance models for query-performance prediction. In: Proceedings of SIGIR, pp. 259–266 (2010)
17. Song, F., Croft, W.B.: A general language model for information retrieval (poster abstract). In: Proceedings of SIGIR, pp. 279–280 (1999)
18. Vinay, V., Cox, I.J., Milic-Frayling, N., Wood, K.R.: On ranking the effectiveness of searches. In: Proceedings of SIGIR, pp. 398–404 (2006)
19. Yom-Tov, E., Fine, S., Carmel, D., Darlow, A.: Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In: Proceedings of SIGIR, pp. 512–519 (2005)
20. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of SIGIR, pp. 334–342 (2001)
21. Zhou, Y.: Retrieval Performance Prediction and Document Quality. PhD thesis, University of Massachusetts (September 2007)
22. Zhou, Y., Croft, W.B.: Ranking robustness: A novel framework to predict query performance. In: Proceedings of CIKM, pp. 567–574 (2006)
23. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of SIGIR, pp. 543–550 (2007)