Query-Performance Prediction Using Minimal Relevance Feedback

Olga Butman
Intel Corporation, Israel
olga.krz@gmail.com

Anna Shtok
Faculty of Industrial Engineering and Management, Technion, Israel
annabel@tx.technion.ac.il

Oren Kurland
Faculty of Industrial Engineering and Management, Technion, Israel
kurland@ie.technion.ac.il

David Carmel
Yahoo! Research, Haifa 31905, Israel
dcarmel@yahoo-inc.com

ABSTRACT

There has been much work on devising query-performance prediction approaches that estimate search effectiveness without relevance judgments (i.e., zero feedback). Specifically, post-retrieval predictors analyze the result list of top-retrieved documents. Departing from the zero-feedback approach, in this paper we show that relevance feedback for even very few top ranked documents can be exploited to dramatically improve prediction quality. Specifically, applying state-of-the-art zero-feedback-based predictors to only a very few relevant documents, rather than to the entire result list as originally designed, substantially improves prediction quality. This novel form of prediction is based on quantifying properties of relevant documents that can attest to query performance. We also show that integrating prediction based on relevant documents with zero-feedback-based prediction is highly effective; specifically, with respect to utilizing state-of-the-art direct estimates of retrieval effectiveness when minimal feedback is available.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

Keywords: query-performance prediction

1. INTRODUCTION

There is a large body of work on query-performance prediction [5]; that is, estimating the effectiveness of a search performed in response to a query with no relevance judgements. We take a step further and address a novel question: how can prediction quality be improved when a minimal amount of relevance feedback (e.g., for the highest ranked document) is available? The resultant prediction challenge is still very difficult. For example, the fact that the highest ranked document is relevant (or not) does not necessarily attest to the average precision (AP) at cutoff 1000.

In addition to the theoretical interest, the challenge we address is important for several applications that can benefit from improved query performance prediction and for which some relevance feedback might be available. For example, in interactive retrieval systems, the user's interaction with the search results is used to repeatedly adjust the query and the retrieval strategy [16]. Hence, using query-performance predictors that exploit user feedback can potentially help to guide these adjustments, e.g., by predicting which query formulations will result in improved performance [2, 7, 11].

Our approach to query-performance prediction with minimal relevance feedback relies on our following key finding. Some state-of-the-art (post-retrieval) predictors, which were devised to analyze the result list of top-retrieved documents with zero feedback, turn out to be highly effective in predicting performance when employed over *only* very few relevant documents in the result list. In this case, several of these predictors could be viewed as estimating the "inherent" difficulty of the query as manifested by characteristics of relevant documents. Accordingly, we devise a method that integrates prediction based on relevant documents with zero-feedback-based prediction; the relative impact of each is based on the amount of positive feedback available.

We empirically show that even when feedback is available only for the highest ranked document, our method's prediction quality substantially transcends that of state-of-the-art zero-feedback-based prediction. Furthermore, when minimal relevance feedback is available, the prediction quality is much better than that of state-of-the-art direct estimates of retrieval effectiveness that also use this feedback [4, 17].

2. RELATED WORK

Our approach that integrates prediction based on relevant documents with zero-feedback-based prediction is conceptually reminiscent of integrating true and pseudo feedback to improve retrieval effectiveness [10, 19]. Zero-feedback-based predictors were used to improve retrieval effectiveness when using positive relevance feedback for retrieval [11]. In con-

ICTIR'13, September 29-October 02, 2013, Copenhagen, Denmark.

http://dx.doi.org/10.1145/2499178.2499201.

trast, we aim to improve prediction quality using relevance feedback, whether positive, negative, or both. Clarity was the only predictor employed over relevant documents in this work [11]. Yet, it is much less effective, in this capacity, than other predictors we study.

Work on evaluation of retrieval systems using minimal relevance judgments has focused on methods for (i) selecting documents to be judged for relevance (e.g., pool sampling methods), and (ii) approximating evaluation measures [12]. Our goal here is to evaluate the effectiveness of a single retrieved list when relevance feedback is provided for a few top ranks. In contrast to these previously proposed approximation methods [12], our prediction method is based on integrating measures for properties of relevant documents and those of unjudged documents. We show that the query-performance prediction quality of our method substantially outperforms that of (state-of-the-art) approximation measures — bpref [4] and infAP [17] — that use the judged documents for direct estimation of retrieval effectiveness.

3. PREDICTION FRAMEWORK

Let $\mathcal{M}$ be a retrieval method that ranks documents in corpus $\mathcal{D}$ in response to a *given* query q. Our goal here is to quantify the effectiveness of the induced ranking. We use $\mathcal{D}_q^{[n]}$ to denote the list of n most highly ranked documents, often referred to as the *result list*.

If relevance judgments, such as those of TREC, are available, various evaluation measures can be used to estimate ranking effectiveness; for example, average precision (AP) computed at cutoff 1000:

$$AP(q) \stackrel{def}{=} \frac{1}{|R(q,\mathcal{D})|} \sum_{i=1}^{1000} \frac{\delta[d_i \in R(q,\mathcal{D})]}{i} \sum_{j=1}^{i} \delta[d_j \in R(q,\mathcal{D})]; \quad (1)$$

$R(q,\mathcal{D})$ is the set of documents in the corpus that were judged relevant to q; d_i is the document at rank i; $\delta[s]$ is 1 if the statement s holds, and 0 otherwise. Accordingly, $\frac{1}{i}\sum_{j=1}^{i} \delta[d_j \in R(q,\mathcal{D})]$ is the precision at i ($p@i(q)$).

On the other hand, there is much work on predicting query (ranking) performance *in the absence* of relevance judgments [5]. A common approach to quantifying prediction quality [5], which we adopt in this paper, is measuring the correlation between the values assigned by a predictor to queries and the "true" AP values for these queries; the latter are computed based on relevance judgments using Equation 1.

The novel question that we pursue below is whether using minimal relevance feedback, e.g., for the highest ranked document, can improve query-performance-prediction quality. More generally, we assume a relevant/non-relevant feedback for the top-k documents in the result list $\mathcal{D}_q^{[n]}$, where $k << n$; that is, we are provided with binary relevance judgments for $\mathcal{D}_q^{[k]}$. We use this feedback, along with information induced from the query (q), the result list ($\mathcal{D}_q^{[n]}$), and the corpus ($\mathcal{D}$), to devise query-performance predictors.

3.1 Direct estimates of retrieval effectiveness

The first approach that we consider is using the feedback to directly predict AP. However, with a very small number of judgments, estimating the number of relevant documents in the entire corpus and their ranks is a very hard task. Hence, we only use the positioning of relevant documents in $\mathcal{D}_q^{[k]}$ following Equation 1. Let $R(q,\mathcal{D}_q^{[k]})$ denote the set of relevant documents among the k judged, then

$$\widehat{AP}(q) \stackrel{def}{=} \sum_{i=1}^{k} \frac{\delta[d_i \in R(q,\mathcal{D}_q^{[k]})]}{i} \sum_{j=1}^{i} \delta[d_j \in R(q,\mathcal{D}_q^{[k]})]. \quad (2)$$

Note that $0 \leq \widehat{AP}(q) \leq k$. The more relevant documents are among the k judged (i.e., the higher recall is at cutoff k), the better query-performance is presumed. In Section 4 we demonstrate the merits of $\widehat{AP}$ with respect to bpref [4] and infAP [17] in terms of correlation with the true AP.

Another predictor that we consider is the precision in the set of judged documents, i.e., precision at k of the result list:

$$p@k(q) \stackrel{def}{=} \frac{r}{k}; \quad (3)$$

$r = |R(q,\mathcal{D}_q^{[k]})|$ is the number of relevant documents in $\mathcal{D}_q^{[k]}$. In contrast to $\widehat{AP}$, $p@k$ does not consider the positioning of relevant documents. Hence, $p@k$ is expected to post weaker correlation to the true AP than $\widehat{AP}$ does.

3.2 Our prediction approach

The $\widehat{AP}$ and $p@k$ predictors only consider the relevant/non-relevant signal. For $k = 1$, for example, both predictors assign either 1 or 0 based on whether the highest ranked document is relevant or not. However, ranking effectiveness also depends on documents at lower ranks. Furthermore, the characteristics of relevant and non-relevant documents can potentially be exploited to improve prediction quality.

To address these issues, we assume a post-retrieval query-performance predictor $\mathcal{P}$ that operates with no relevance judgments, as is standard [5]. Formally, given query q, and the result list $\mathcal{D}_q^{[n]}$, $\mathcal{P}(q;\mathcal{D}_q^{[n]},\mathcal{D})$ is the query-performance-prediction value assigned to q; the value can depend on corpus-based statistics. To simplify notation, we define:

$$\mathcal{P}_{Res}(q) \stackrel{def}{=} \mathcal{P}(q,\mathcal{D}_q^{[n]},\mathcal{D}). \quad (4)$$

$\mathcal{P}_{Res}$ quantifies properties of the result list ($\mathcal{D}_q^{[n]}$) that we evaluate. This quantification is important as we empirically show in Section 4, because only a small number of relevance judgements is available. Yet, $\mathcal{P}_{Res}$ does not exploit this relevance feedback. As it turns out, some state-of-the-art post-retrieval predictors, when employed *only* over relevant documents in the result list, yield quite effective query-performance prediction. For example, some of these predictors can leverage characteristics of the relevant documents that can attest to the "inherent" difficulty of q; and, consequently, to the potential ranking effectiveness. We argue for why this is the case for three such predictors in Section 3.2.1, and present empirical support in Section 4. Accordingly, the $\mathcal{P}_{RF}$ predictor (RF stands for relevance feedback) utilizes information induced only from relevant documents if there are any among the k judged, and assigns a 0 prediction value otherwise:

$$\mathcal{P}_{RF}(q) \stackrel{def}{=} \begin{cases} \mathcal{P}(q,R(q,\mathcal{D}_q^{[k]}),\mathcal{D}) & \text{r}> 0; \\ 0 & \text{r}=0. \end{cases} \quad (5)$$

Integrating zero-feedback and relevance-feedback -based prediction. We integrate the zero-feedback-based predictor, $\mathcal{P}_{Res}$, with the predictor based on relevance feedback,

$\mathcal{P}_{RF}$, as follows. The more relevant documents there are among those judged, the more weight is put on $\mathcal{P}_{RF}$:

$$\mathcal{P}_{Res:RF}(q) \stackrel{def}{=} (1-\lambda)(1-\frac{r}{k})\mathcal{P}_{Res}(q) + \lambda\frac{r}{k}\mathcal{P}_{RF}(q); \quad (6)$$

λ is a regularization parameter, the value of which is set using a train set of queries.

An important role of λ is in the r=0 case, where all k judged documents are not relevant. Instead of backing off completely to zero-feedback-based prediction ($\mathcal{P}_{Res}(q)$), we discount this prediction by using $(1-\lambda)\mathcal{P}_{Res}(q)$, since the top-$k$ documents are not relevant — a strong evidence for low effectiveness. Similarly, if all judged documents are relevant ($r = k$), we use a discounted relevance-feedback-based prediction ($\lambda\mathcal{P}_{RF}(q)$). As shown in Section 4, the discount often becomes smaller as k increases, i.e., the optimal value of λ increases when k grows. Indeed, more confidence should be attributed to prediction based on relevant documents as their number increases, which is the case *on average* when increasing k.

Another observation about the $\mathcal{P}_{Res:RF}$ predictor is that once k is fixed (and so is λ), larger number of relevant documents (r) does not necessarily imply higher prediction value, in contrast to $p@k$ and $\widehat{AP}$. This is because the result-list-based prediction could potentially dominate the relevance-feedback-based prediction. In a related vein, $\mathcal{P}_{Res:RF}$ does not use information about the positioning of relevant documents in contrast to $\widehat{AP}$. Thus, integrating these two predictors can yield some merit as we show in Section 4.

3.2.1 Instantiating specific predictors

We next present three (state-of-the-art) post-retrieval predictors $\mathcal{P}$ that do not utilize relevance feedback ($\mathcal{P}_{Res}$). We then argue for why, and how, these predictors can be used upon *only* relevant documents ($\mathcal{P}_{RF}$ from Equation 5) to predict query performance. Using Equation 6 we can then instantiate the $\mathcal{P}_{Res:RF}$ predictor.

As the three predictors operate in the language modeling framework, we set as a goal to predict the performance of standard language-model-based retrieval, namely, the query likelihood (QL) approach [14]. Let $p(w|x)$ denote the probability assigned to term w by a (smoothed) unigram language model induced from text (or text collection) x. (Details regarding language model induction are provided in Section 4.1.) The QL score of x with respect to q ($= \{q_i\}$) is

$$Score_{QL}(q;x) \stackrel{def}{=} \log p(q|x) \stackrel{def}{=} \log \prod_{q_i \in q} p(q_i|x). \quad (7)$$

Our goal is to predict the effectiveness of the ranking induced by the QL scores of documents in the corpus.

Two of the predictors we analyze use relevance models [9, 1]. Let R^{Res} be a relevance model constructed from the result list $\mathcal{D}_q^{[n]}$:[1]

$$p(w|R^{Res}) \stackrel{def}{=} \sum_{d \in \mathcal{D}_q^{[n]}} p(w|d)p(d|q); \quad (8)$$

$p(d|q) \stackrel{def}{=} \frac{p(q|d)}{\sum_{d' \in \mathcal{D}_q^{[n]}} p(q|d')}$ is d's normalized query likelihood.

[1]This is RM1 [9] that was reported to be more effective than RM3 [1] for the predictors we present [13].

To construct a relevance model, R^{Rel}, only from the relevant documents (i.e., from $R(q, \mathcal{D}_q^{[k]})$), we set $p(d|q) \stackrel{def}{=} \frac{1}{|R(q,\mathcal{D}_q^{[k]})|} = \frac{1}{r}$ following previous recommendations [9]:

$$p(w|R^{Rel}) \stackrel{def}{=} \frac{1}{r} \sum_{d \in R(q,\mathcal{D}_q^{[k]})} p(w|d). \quad (9)$$

To rank documents using a relevance model R, the minus cross entropy, $\sum_w p(w|R)\log p(w|d)$, is used [9].

Clarity. The Clarity predictor is based on measuring the KL divergence between a relevance language model induced from the result list $\mathcal{D}_q^{[n]}$ (see Equation 8), and a language model induced from the entire corpus [6]:

$$Clarity_{Res}(q) \stackrel{def}{=} \sum_w p(w|R^{Res}) \log \frac{p(w|R^{Res})}{p(w|\mathcal{D})}.$$

The larger the divergence, the more focused $\mathcal{D}_q^{[n]}$ is assumed to be; thereby, improved performance is presumed.

The clarity of the relevant-documents set ($R(q, \mathcal{D}_q^{[k]})$) is computed using the relevance model from Equation 9:

$$Clarity_{RF}(q) \stackrel{def}{=} \sum_w p(w|R^{Rel}) \log \frac{p(w|R^{Rel})}{p(w|\mathcal{D})}.$$

The rationale for using the $Clarity_{RF}$ predictor is based on the following premise. The more coherent the set of relevant documents is with respect to the corpus, the easier we assume it is to satisfy the underlying information need. Therefore, the more likely the retrieval is to be of high effectiveness. Conversely, a non-focused set of relevant documents potentially attests to a query spanning over different aspects. Such queries are known to be difficult [3].

QF. The query feedback (QF) predictor, QF_{Res}, measures ranking robustness [21]. The result list, $\mathcal{D}_q^{[n]}$, is considered robust if it contains relatively small amounts of non-query-related noise. The potential amount of noise is assumed to be negatively correlated with the number of documents shared (i.e., overlap) between the (i) n_{QF} highest ranked in $\mathcal{D}_q^{[n]}$, and (ii) n_{QF} highest ranked by a search performed over the corpus using R^{Res}; n_{QF} is a free parameter.[2] The idea is that a relevance model constructed from an effective result list, which does not contain much noise, would not yield a ranking that drifts much from the original ranking.

Using R^{Rel} instead of R^{Res} (i.e., using only the relevant documents rather than the entire result list to construct a relevance model) to rank the corpus, and computing the document overlap described above, we get the QF_{RF} predictor. If the number of relevant documents, r, is not extremely small, R^{Rel} can be considered as faithfully representing the information need; specifically, to a larger extent than R^{Res} does. (Note that R^{Res} is based on pseudo feedback, while R^{Rel} is based on true feedback.) Furthermore, the ranking induced over the corpus using R^{Rel} is of high effectiveness as is known in work on utilizing relevance feedback for retrieval. Hence, the overlap between the highest ranked documents

[2]For consistency with the clarity-based predictors, we use the relevance model in Equation 8, which is somewhat different than the query model used originally [21]. Yet, the resultant prediction quality is state-of-the-art [13].

Collection	Data	Num Docs	Topics
WT10G	WT10g	1,692,096	451-550
ROBUST	Disk 4&5-CR	528,155	301-450, 601-700
GOV2	GOV2	25,205,179	701-850
TREC123	Disks 1&2	741,856	51-200
TREC4	Disks 2&3	567,529	201-250
TREC5	Disks 2&4	524,929	251-300

Table 1: Test collections and topics.

by R^{Rel}, and those at the highest ranks of the result list $\mathcal{D}_q^{[n]}$, can attest to the effectiveness of $\mathcal{D}_q^{[n]}$ as was also noted in some previous work [8].

WIG. The WIG (Weighted Information Gain) predictor [21] is based on the premise that high retrieval scores with respect to that of the corpus attest to improved performance:[3]

$$WIG(q;S) \stackrel{def}{=} \frac{1}{\sqrt{|q|}} \frac{1}{|S|} \sum_{d \in S} (Score_{QL}(q;d) - Score_{QL}(q;\mathcal{D}));$$

S is a set of documents; the query-length ($|q|$) normalization is employed for inter-query compatibility.

Setting $S \stackrel{def}{=} \mathcal{D}_q^{[n_{WIG}]}$ we get the standard WIG_{Res} predictor that operates on the n_{WIG} highest ranked documents in the result list; n_{WIG} is a free parameter. Setting $S \stackrel{def}{=} R(q, \mathcal{D}_q^{[k]})$ results in the WIG_{RF} predictor that measures the divergence between retrieval scores of the relevant documents and that of the corpus. Recall that QL retrieval scores reflect the language-model-based surface-level similarity to the query. Hence, the higher the similarity is for relevant documents with respect to that for the corpus, the less potential for *vocabulary mismatch* between relevant documents and the query. Consequently, the more effective the QL-based ranking is presumed to be.

4. EVALUATION

4.1 Experimental setup

We conducted experiments with the TREC corpora specified in Table 1. These were also used in work on query-performance prediction with zero feedback [5].

Titles of TREC topics serve for queries, except for TREC4 for which topic descriptions are used as titles are not available. We applied Porter stemming and stopword removal (using the INQUERY list) to queries and documents using the Lemur/Indri toolkit (www.lemurproject.org), which was used for experiments.

As noted in Section 3.2.1, we set as a goal to predict the performance of the query likelihood (QL) retrieval approach [14]. The k documents most highly ranked for a query are those for which we have feedback. A document is considered relevant if there is such judgment in TREC's qrels file. A document is considered non-relevant if it was judged as such, or if it does not have a judgment.

Following common practice [5], prediction quality is measured as follows. We compute Pearson's correlation between the values assigned by a predictor to queries, and the "true" AP values for these queries determined based on TREC's relevance judgments (see Equation 1).[4]

The zero-feedback-based predictors ($\mathcal{P}_{Res}$) Clarity and QF depend on the result list size, n; QF also incorporates an overlap cutoff parameter, n_{QF}; and, WIG depends on n_{WIG}, the number of highest-ranked documents considered. The prediction quality of the predictors is sensitive to the values of these parameters [7, 21, 20]. When employed over only relevant documents ($\mathcal{P}_{RF}$), QF is the only predictor among the three that still depends on a free parameter (n_{QF}); in the $QF_{Res:RF}$ predictor, the same value of n_{QF} is used for QF_{Res} and QF_{RF}. Furthermore, the $\mathcal{P}_{Res:RF}$ predictors use the parameter λ that controls the reliance on feedback.

To have a robust evaluation with respect to effects of free-parameter values, we take a train-test approach. The query set for a corpus is randomly split in half. One half serves as the training set; that is, the predictor's free parameters are set to values that optimize prediction quality (Pearson correlation) over this set. The second half of the query set serves as the test set on which we apply the predictor with the learned free-parameter values. We repeat this procedure 30 times and report the average prediction quality and its standard deviation over the 30 test sets. The standard deviation of prediction quality serves for studying its robustness.

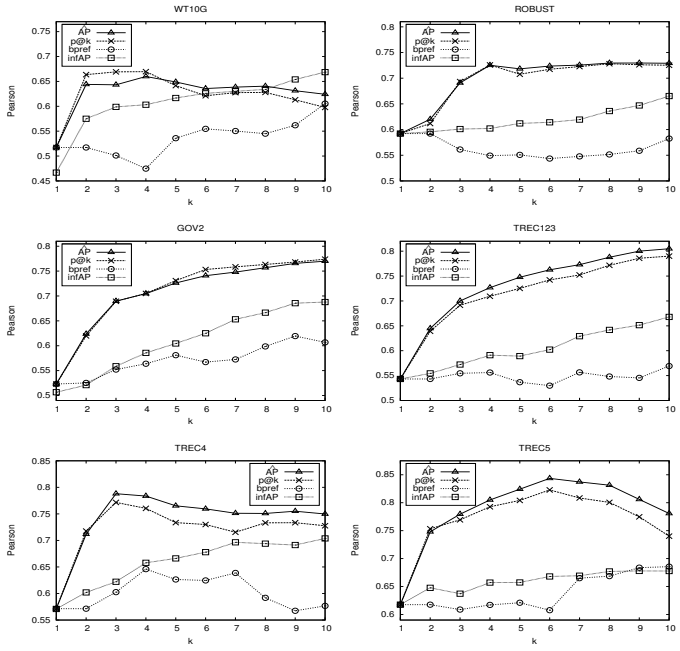

Figure 1: Prediction quality of direct estimates of retrieval effectiveness with relevance feedback available for the top-k documents. Prediction quality is measured *only here* by Pearson correlation with *all* queries per corpus, rather than using the train-test split. Note: figures are not to the same scale.

[3]While WIG was originally proposed in the Markov Random Field framework [21], it is highly effective in measuring the query-performance of the query-likelihood method where term dependencies are not considered [20, 13].

[4]We found that measuring prediction quality by computing Kendall's-τ between the ranking of queries based on their "true" AP and that based on assigned prediction values [15] yields the same relative prediction quality patterns as those for Pearson's correlation. Hence, Kendall's-τ numbers are omitted as they convey no additional insight.

The free-parameter values are selected from the following ranges: $n \in \{150, 300, 500, 700, 1000\}$; $n_{QF} \in \{5, 10, 20, 30, 50, 70, 100\}$; $n_{WIG} \in \{1, 2, 3, 4, 5, 10, 15, 20, 30, 50, 70, 100\}$; and, $\lambda \in \{0, 0.1, \ldots, 1\}$. We use the same number of feedback documents, k, for all queries in an experimental setup; the effect of k on prediction quality is studied below.

We use Dirichlet smoothed unigram language models with the smoothing parameter set to 1000 [18]. Following previous recommendations [13], the number of terms used by the relevance model utilized by Clarity and QF, in the various predictors, is 100; and, the language models of documents used to construct the relevance model are unsmoothed maximum likelihood estimates.

4.2 Experimental results

4.2.1 Direct estimates of retrieval effectiveness

We first compare the prediction quality of the direct estimates of retrieval effectiveness, $\widehat{AP}$ and $p@k$. As reference comparisons we use the bpref [4] and the state-of-the-art infAP [17] measures. These estimates do not include free parameters. Hence, in contrast to the train-test-based evaluation we use below, here we use all queries per corpus for computing prediction quality. Figure 1 presents the prediction quality with feedback available for the top-k documents.

We see in Figure 1 that both $\widehat{AP}$ and $p@k$ almost always (specifically, for $k > 1$) outperform bpref and infAP; often, by quite a large margin. This finding can be attributed to the fact that bpref and infAP were designed to cope with incomplete relevance judgments when evaluating the relative average effectiveness of different retrieval systems; specifically, addressing cases wherein judged and unjudged documents are interleaved in the result list. In contrast, the prediction task we address here is estimating the relative effectiveness of a single retrieval method over a set of queries; and, relevance feedback is provided for the top-k documents.

For $k > 1$, $\widehat{AP}$ outperforms $p@k$ in most cases. This is because $\widehat{AP}$ accounts for the positioning of relevant documents and $p@k$ does not. For $k = 1$, $\widehat{AP}$ and $p@k$ are equivalent: the prediction value is 1 if the document is relevant, and 0 otherwise. Since $\widehat{AP}$, which is novel to this study, is the most effective direct estimate among those considered, it will be used below as a baseline to our prediction methods.

4.2.2 Applying predictors

Table 2 reports the prediction quality of our approach for two extreme cases: having (i) the minimal possible feedback (i.e., for the highest ranked document; $k = 1$); and, much feedback, specifically, for the top $k = 10$ documents. Later we discuss the range between these two extremes.

Feedback for a single document ($k = 1$). A striking observation with regard to Table 2 is the prediction-quality increase attained by moving from zero feedback to a single feedback. For example, $\widehat{AP}$, which directly predicts AP, outperforms all zero-feedback-based predictors $\mathcal{P}_{Res}$ for four out of the six collections. In other words, knowing whether the highest ranked document is relevant or not, and assigning a prediction value of 1 and 0, respectively, yields prediction quality that often transcends that of state-of-the-art zero-feedback-based predictors. Furthermore, our $\mathcal{P}_{RF}$ predictor (see Equation 5) consistently improves over $\widehat{AP}$ and $\mathcal{P}_{Res}$; substantially so over the latter. Thus, we see that applying zero-feedback-based predictors upon a single relevant document can be of much merit.

The $\mathcal{P}_{Res:RF}$ predictors integrate zero-feedback-based prediction ($\mathcal{P}_{Res}$) with prediction based on the single feedback document ($\mathcal{P}_{RF}$). This integration turns out to be very effective; specifically, $\mathcal{P}_{Res:RF}$ outperforms $\mathcal{P}_{Res}$ and $\widehat{AP}$ to a substantial extent in almost all relevant comparisons; furthermore, $\mathcal{P}_{Res:RF}$ outperforms $\mathcal{P}_{RF}$ in most relevant comparisons. Most cases for which $\mathcal{P}_{Res:RF}$ is outperformed by $\mathcal{P}_{RF}$ are for QF, which is highly effective when employed over a single relevant document as we show below. The learned value of λ, which controls the reliance on the single feedback document, may not be high enough in these cases.

Feedback for ten documents ($k = 10$). It is evident in Table 2 that, as expected, the prediction-quality improvement over zero-feedback-based prediction when utilizing the 10 feedback documents can be huge. (Compare, for example, $\mathcal{P}_{Res:RF}$ with $\mathcal{P}_{Res}$.) The notable exception is for $Clarity_{RF}$ that is sometimes outperformed by $Clarity_{Res}$, the zero-feedback-based predictor. As we further show and discuss below, computing clarity over potentially very few (≤ 10) relevant documents can result in much less effective prediction than that of clarity computed over a result list containing a few dozen documents.[5] Nevertheless, the integration of clarity computed over relevant documents ($Clarity_{RF}$) with clarity computed over the entire result list ($Clarity_{Res}$), i.e., the $Clarity_{Res:RF}$ predictor, much improves over using each alone. More generally, $\mathcal{P}_{Res:RF}$ substantially improves over $\mathcal{P}_{Res}$ and $\mathcal{P}_{RF}$ in almost all cases. Thus, as was the case for a single feedback, we see that integrating zero-feedback-based prediction with prediction based only on relevant documents is beneficial. This conclusion is further supported by the observation that $\mathcal{P}_{Res:RF}$, which does not utilize the positioning of relevant documents, often outperforms $\widehat{AP}$, which accounts for the positioning.

We see in Table 2 that while Clarity, QF, and WIG are effective basis for deriving predictors that utilize the 10 feedback documents, QF stands out in several cases, especially, when using only relevant documents ($\mathcal{P}_{RF}$); e.g., QF-based predictors post the best prediction quality for three collections. These findings can be explained as follows. The relevance model constructed from the relevant documents, even if there are only a few of these, induces effective ranking over the corpus. This ranking essentially serves as a reference comparison in predicting the effectiveness of the given result list, as their overlap at highest ranks is QF's prediction value [8]. On the other hand, WIG and Clarity, when computed over relevant documents, quantify properties of them that can serve as proxies for the difficulty of the query at hand, as mentioned in Section 3.2.1. While this quantification is effective for prediction, as the numbers in Table 2 attest, it addresses the quality of the given result list in a less "direct" way than that of QF.

We can also see in Table 2 that for both $k = 1$ and $k = 10$, the standard deviation of prediction quality (computed over the 30 test sets) for $\mathcal{P}_{Res:RF}$ is often much lower than that

[5]For a single feedback the case was different as for many queries the judged document is non-relevant and therefore $\mathcal{P}_{RF}(q)$=0. Here, there are few queries for which there is no relevant document among the 10 judged; thus, the prediction also relies on relevant documents, which are often very few.

	Predictor	WT10G	ROBUST	GOV2	TREC123	TREC4	TREC5
k=1	$\widehat{AP}$	0.514 (0.07)	0.600 (0.03)	0.538 (0.06)	0.552 (0.05)	0.576 (0.08)	0.615 (0.08)
	$Clarity_{Res}$	0.457 (0.09)	0.493 (0.06)	0.428 (0.12)	0.465 (0.09)	0.466 (0.12)	0.392 (0.15)
	$Clarity_{RF}$	0.526 (0.09)	0.628 (0.03)	0.517 (0.15)	0.561 (0.11)	**0.580** (0.08)	0.652 (0.09)
	$Clarity_{Res:RF}$	**0.528** (0.07)	**0.640** (0.03)	**0.542** (0.06)	**0.564** (0.06)	0.575 (0.09)	**0.661** (0.08)
	QF_{Res}	0.464 (0.13)	0.462 (0.12)	0.497 (0.10)	0.648 (0.15)	**0.625** (0.12)	0.556 (0.16)
	QF_{RF}	0.669 (0.09)	**0.651** (0.07)	0.577 (0.09)	0.773 (0.08)	0.616 (0.10)	**0.703** (0.10)
	$QF_{Res:RF}$	**0.683** (0.08)	0.648 (0.03)	**0.591** (0.09)	**0.791** (0.04)	0.607 (0.11)	0.695 (0.11)
	WIG_{Res}	0.402 (0.11)	0.522 (0.07)	0.485 (0.14)	0.684 (0.08)	0.532 (0.14)	0.250 (0.18)
	WIG_{RF}	0.573 (0.07)	0.680 (0.04)	0.605 (0.08)	0.652 (0.07)	**0.660** (0.07)	**0.653** (0.09)
	$WIG_{Res:RF}$	**0.604** (0.05)	**0.688** (0.04)	**0.622** (0.05)	**0.690** (0.06)	**0.660** (0.08)	0.646 (0.09)
k=10	$\widehat{AP}$	0.610 (0.07)	0.735 (0.03)	0.765 (0.04)	0.816 (0.03)	0.759 (0.10)	0.781 (0.06)
	$Clarity_{Res}$	0.457 (0.09)	0.493 (0.06)	0.428 (0.12)	0.465 (0.09)	0.466 (0.12)	0.392 (0.15)
	$Clarity_{RF}$	0.447 (0.05)	0.400 (0.05)	0.299 (0.09)	0.363 (0.20)	0.536 (0.06)	0.454 (0.08)
	$Clarity_{Res:RF}$	**0.629** (0.08)	**0.768** (0.03)	**0.785** (0.03)	**0.804** (0.03)	**0.725** (0.08)	**0.779** (0.07)
	QF_{Res}	0.464 (0.13)	0.462 (0.12)	0.497 (0.10)	0.648 (0.15)	0.625 (0.12)	0.556 (0.16)
	QF_{RF}	**0.657** (0.10)	0.675 (0.04)	0.775 (0.07)	0.818 (0.07)	0.816 (0.04)	0.687 (0.12)
	$QF_{Res:RF}$	0.594 (0.09)	**0.703** (0.04)	**0.812** (0.06)	**0.916** (0.02)	**0.824** (0.08)	**0.749** (0.06)
	WIG_{Res}	0.402 (0.11)	0.522 (0.07)	0.485 (0.14)	0.684 (0.08)	0.532 (0.14)	0.250 (0.18)
	WIG_{RF}	0.600 (0.05)	0.579 (0.05)	0.589 (0.11)	0.631 (0.04)	0.636 (0.10)	0.504 (0.09)
	$WIG_{Res:RF}$	**0.678** (0.09)	**0.783** (0.03)	**0.807** (0.04)	**0.869** (0.03)	**0.756** (0.11)	**0.722** (0.07)

Table 2: Prediction quality when using relevance feedback for the k ($\in \{1, 10\}$) highest ranked documents. Best result per k, corpus and family of predictors is boldfaced. Best result per k and corpus is underlined. Numbers in parentheses indicate the standard deviation of prediction quality over the 30 test sets.

	Predictor	WT10G	ROBUST	GOV2	TREC123	TREC4	TREC5
k=1	$\widehat{AP}$	0.514 (0.07)	0.600 (0.03)	0.538 (0.06)	0.552 (0.05)	0.576 (0.08)	0.615 (0.08)
	$Clarity_{Res:RF}$	**0.528** (0.07)	**0.640** (0.03)	0.542 (0.06)	**0.564** (0.06)	0.575 (0.09)	**0.661** (0.08)
	$Clarity_{Res:RF:\widehat{AP}}$	0.526 (0.07)	0.628 (0.03)	**0.550** (0.06)	0.561 (0.06)	**0.580** (0.08)	0.652 (0.08)
	$QF_{Res:RF}$	**0.683** (0.08)	0.648 (0.03)	**0.591** (0.09)	**0.791** (0.04)	0.607 (0.11)	0.695 (0.11)
	$QF_{Res:RF:\widehat{AP}}$	0.670 (0.08)	**0.651** (0.03)	0.577 (0.09)	0.774 (0.05)	**0.616** (0.10)	**0.703** (0.10)
	$WIG_{Res:RF}$	**0.604** (0.05)	**0.688** (0.04)	**0.622** (0.05)	**0.690** (0.06)	**0.660** (0.08)	0.646 (0.09)
	$WIG_{Res:RF:\widehat{AP}}$	0.573 (0.06)	0.680 (0.04)	0.614 (0.05)	0.652 (0.06)	**0.660** (0.07)	**0.653** (0.09)
k=10	$\widehat{AP}$	0.610 (0.07)	0.735 (0.03)	0.765 (0.04)	0.816 (0.03)	0.759 (0.10)	0.781 (0.06)
	$Clarity_{Res:RF}$	0.629 (0.08)	0.768 (0.03)	0.785 (0.03)	0.804 (0.03)	**0.725** (0.08)	0.779 (0.07)
	$Clarity_{Res:RF:\widehat{AP}}$	**0.684** (0.07)	**0.794** (0.03)	**0.792** (0.03)	**0.820** (0.03)	0.712 (0.10)	**0.787** (0.09)
	$QF_{Res:RF}$	0.594 (0.09)	0.703 (0.04)	**0.812** (0.06)	**0.916** (0.02)	0.824 (0.08)	0.749 (0.06)
	$QF_{Res:RF:\widehat{AP}}$	**0.655** (0.08)	**0.726** (0.04)	0.802 (0.05)	0.912 (0.02)	**0.853** (0.08)	**0.786** (0.08)
	$WIG_{Res:RF}$	0.678 (0.09)	0.783 (0.03)	**0.807** (0.04)	**0.869** (0.03)	**0.756** (0.11)	0.722 (0.07)
	$WIG_{Res:RF:\widehat{AP}}$	**0.699** (0.07)	**0.803** (0.03)	0.797 (0.03)	0.868 (0.03)	0.742 (0.07)	**0.741** (0.06)

Table 3: Integrating $\mathcal{P}_{Res:RF}$ and $\widehat{AP}$ ($\mathcal{P}_{Res:RF:\widehat{AP}}$) when using k feedback documents. Boldface marks the better performing between $\mathcal{P}_{Res:RF}$ and $\mathcal{P}_{Res:RF:\widehat{AP}}$ per corpus and k; underline marks the best result per corpus and k. Numbers in parentheses indicate the standard deviation of prediction quality.

for $\mathcal{P}_{Res}$ and $\mathcal{P}_{RF}$ that it integrates. Furthermore, $\mathcal{P}_{Res:RF}$'s standard deviation is on par with that of $\widehat{AP}$, which does not incorporate free parameters. These findings attest to the prediction-quality *robustness* of integrating zero-feedback-based prediction with that based on the relevance feedback.

4.2.3 Effect of the amount of feedback

Figure 2 presents the effect of the number of feedback documents (k) on the prediction quality of $\widehat{AP}$ and $\mathcal{P}_{Res:RF}$ for additional values of k .[6] For $k = 0$, we present the quality of $\mathcal{P}_{Res}$, the zero-feedback-based predictors.

With a few exceptions, the general picture in Figure 2 is that with increased amount of feedback the prediction quality of the various predictors rises, and then might level off. In almost all cases, the prediction is much better than that of zero-feedback-based prediction. We can also see that for a small number of feedback documents ($k \leq 3$) the $\mathcal{P}_{Res:RF}$ predictors often outperform $\widehat{AP}$. Indeed, $\widehat{AP}$, as any other direct estimate of retrieval effectiveness, is based on little statistics in these cases. Furthermore, recall from Figure 1 that $\widehat{AP}$ substantially outperforms bpref and infAP for $k \geq 2$, and is as effective as bpref and infAP for $k = 1$. These findings further support the motivation for integrating zero-feedback-based prediction with that based on relevant documents when minimal feedback is available. With a larger number of feedback documents, the prediction quality of $\widehat{AP}$ naturally often rises, and in some cases transcends that of some of the $\mathcal{P}_{Res:RF}$ predictors; however, it is still almost always inferior to that of the best performing $\mathcal{P}_{Res:RF}$.

4.2.4 Integrating predictors

The $\mathcal{P}_{Res:RF}$ predictors, which were shown above to be highly effective, do not use information about the positioning of relevant documents. Hence, we study the potential benefit of integrating them with $\widehat{AP}$, which relies only on the number of relevant documents and their positioning. To this end, we study the $\mathcal{P}_{Res:RF:\widehat{AP}}$ predictor:

$$\mathcal{P}_{Res:RF:\widehat{AP}}(q) \stackrel{def}{=} (\widehat{AP}(q) + \epsilon)\mathcal{P}_{Res:RF}(q); \qquad (10)$$

ϵ is set to 0.0001 to avoid zero multiplication.

[6]One important difference between Figure 1 and Figure 2 is that the former is based on using all queries per corpus as the test set and the latter is based on the train-test split.

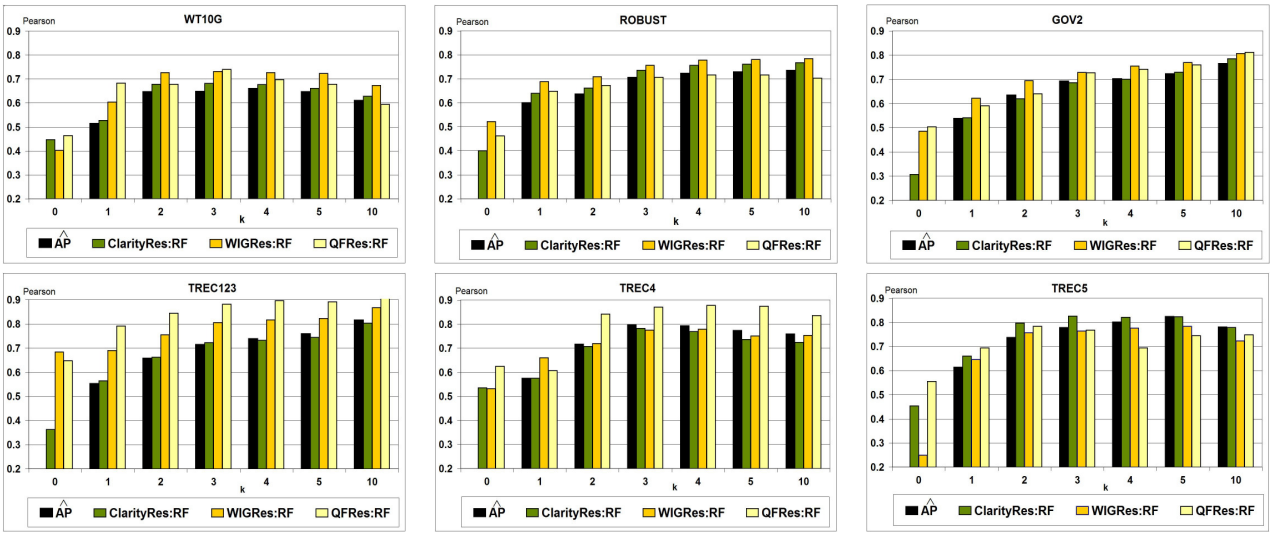

Figure 2: The effect of the number of feedback documents (k) on the prediction quality of $\widehat{AP}$ and $\mathcal{P}_{Res:RF}$. Note: for $k = 0$ the prediction quality of zero-feedback-based prediction ($\mathcal{P}_{Res}$) is presented.

We see in Table 3 that for a single feedback ($k = 1$), $\mathcal{P}_{Res:RF:\widehat{AP}}$ posts consistent improvements over $\widehat{AP}$, but is often outperformed by $\mathcal{P}_{Res:RF}$. This finding could be explained by the following observations. First, $\mathcal{P}_{Res:RF}$ is consistently better than $\widehat{AP}$. Now, if the judged document is non-relevant, $\mathcal{P}_{Res:RF:\widehat{AP}}$ assigns a heavily discounted zero-feedback-based prediction value ($\epsilon(1-\lambda)\mathcal{P}_{Res}(q)$), which is much smaller than $(1-\lambda)\mathcal{P}_{Res}(q)$, the value assigned by $\mathcal{P}_{Res:RF}$. Thus, $\mathcal{P}_{Res:RF:\widehat{AP}}$ "leans" towards the 0 prediction value assigned by $\widehat{AP}$. If the document is relevant, $\mathcal{P}_{Res:RF:\widehat{AP}}$ essentially uses $\mathcal{P}_{Res:RF}$; the latter integrates prediction based on this document with zero-feedback-based prediction. Thus, while for the relevant document case $\mathcal{P}_{Res:RF:\widehat{AP}}$ leverages the relative strength of $\mathcal{P}_{Res:RF}$ over $\widehat{AP}$, for the non-relevant document case it does not.

When using 10 feedback documents, we see in Table 3 that $\mathcal{P}_{Res:RF:\widehat{AP}}$ often improves over $\widehat{AP}$ and $\mathcal{P}_{Res:RF}$. This is not a surprise as $\widehat{AP}$ is based on quite sufficient statistics that is complementary to the information used by $\mathcal{P}_{Res:RF}$. Furthermore, we see that with both single and ten feedback documents, the standard deviation of the prediction quality of $\mathcal{P}_{Res:RF:\widehat{AP}}$, $\mathcal{P}_{Res:RF}$ and $\widehat{AP}$ is quite similar.

4.2.5 *Fixing the number of relevant documents*

We demonstrated the effectiveness of applying zero-feedback-based predictors to only relevant documents — the $\mathcal{P}_{RF}$ predictors. Now, the number of relevant documents (r) among the (fixed) k judged varies across queries. Thus, to study the effect of r on $\mathcal{P}_{RF}$, we fix r and vary k, as follows. Scanning the ranking top to bottom we gather r relevant documents, and then apply $\mathcal{P}_{RF}$ on these. In Figure 3 we present the prediction quality for WT10G, ROBUST, GOV2, and TREC123. The patterns for TREC4 and TREC5 are similar. These are omitted due to space considerations.

Figure 3 shows that, in general, increasing the number of relevant documents results in increased prediction quality that might level off. For WIG and QF using a single relevant document ($r = 1$) yields in almost all cases prediction quality that transcends that of zero-feedback-based prediction. These predictors were originally devised to operate with zero feedback. Increasing the number of relevant documents used by WIG and QF results in very high prediction quality.

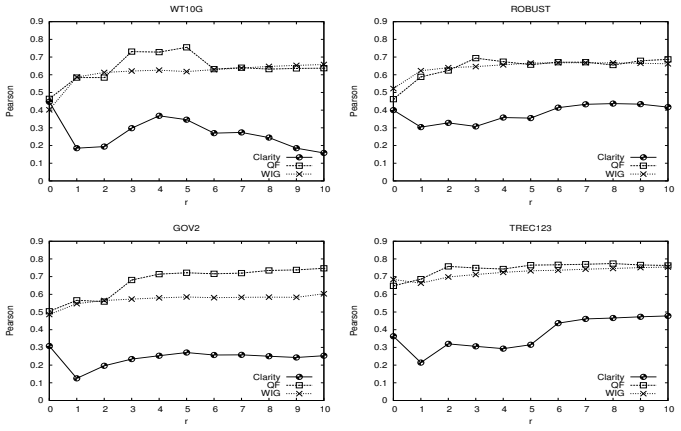

Figure 3: Prediction quality of applying Clarity, QF and WIG on the top-r relevant documents (i.e., the $\mathcal{P}_{RF}$ predictors). For $r = 0$ the presented quality is for the zero-feedback-based predictors ($\mathcal{P}_{Res}$).

We also see in Figure 3 that QF is somewhat more effective than WIG; both substantially outperform Clarity, which is quite not effective when employed over only relevant documents. As discussed in Section 3.2.1, when using relevant documents QF estimates the result list effectiveness more directly than WIG and Clarity which measure characteristics of relevant documents that somewhat indirectly imply to retrieval effectiveness. WIG's superiority over Clarity shows that the focus of relevant documents with respect to the corpus (Clarity) is less indicative of query performance than the potential vocabulary mismatch between relevant documents and the query (WIG) when using document-query surface-level similarities for ranking.

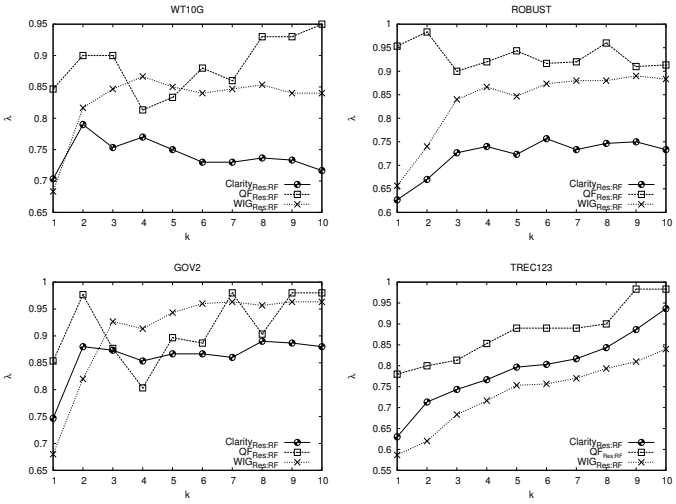

Figure 4: The value of λ that is learned for the $\mathcal{P}_{Res:RF}$ predictors (averaged over 30 train sets) as a function of the number of feedback documents (k). Note: figures are not to the same scale.

4.2.6 *Balancing between $\mathcal{P}_{Res}$ and $\mathcal{P}_{RF}$*

Balancing zero-feedback-based prediction and prediction based on relevant documents in the $\mathcal{P}_{Res:RF}$ predictors is controlled by the relative number of relevant documents among those judged ($\frac{r}{k}$) and the regularization parameter λ. (See Equation 6.) Higher values of λ result in more reliance on the relevant documents. Insofar, λ's value, as those of the other free parameters, was set using a train query set.

Figure 4 presents the learned value of λ (averaged over 30 train sets) as a function of the number of feedback documents (k). (The curves for TREC4 and TREC5 exhibit the same patterns as those for WT10G and ROBUST and are omitted due to space considerations.) In several cases, λ's learned value rises when increasing k. Indeed, increased k results, *on average*, in increased number of relevant documents; thus, prediction can be improved by relying more on these. As both λ and $\frac{r}{k}$ control the reliance, and the values of the other free parameters are simultaneously learned, there are cases of no monotonic increase of λ's learned value.

We also see in Figure 4 that the learned value of λ is, in general, higher for $QF_{Res:RF}$ than for $Clarity_{Res:RF}$ and $WIG_{Res:RF}$. As mentioned above, QF is more effective than Clarity and WIG when employed over relevant documents. Thus, the reliance on relevant documents is higher for QF than for Clarity and WIG. The fluctuations in the learned-λ curves for $QF_{Res:RF}$ can be attributed to its two additional free parameters (on top of λ) in contrast to the single additional one incorporated by $Clarity_{Res:RF}$ and $WIG_{Res:RF}$.

Finally, we see that even for a relatively large k, λ's learned value, which optimizes prediction over the train set, can be (quite) smaller than 1. This attests to the importance of integrating zero-feedback-based prediction even when a relatively descent amount of feedback is available.

5. CONCLUSIONS

We addressed the challenge of predicting query performance using relevance feedback for very few top ranked documents. We showed that applying some zero-feedback-based post-retrieval predictors over only a few relevant documents yields high prediction quality. We also demonstrated the merits of integrating zero-feedback-based prediction with prediction based on relevant documents. The prediction quality often substantially transcends that of state-of-the-art direct estimates of retrieval effectiveness.

6. ACKNOWLEDGMENTS

We thank the reviewers for their comments. Part of the work reported here was done while David Carmel was at IBM and Olga Butman was at the Technion. This paper is based on work supported in part by the Israel Science Foundation under grant no. 433/12, by a Google faculty research award, and by an IBM Ph.D. fellowship.

7. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMASS at TREC 2004 — novelty and hard. In *Proc. of TREC-13*, 2004.

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

[3] C. Buckley. Why current IR engines fail. *Information Retrieval*, 12(6):652–665, 2009.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of SIGIR*, pages 25–32, 2004.

[5] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2010.

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

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Precision prediction based on ranked list coherence. *Information Retrieval*, 9(6):723–755, 2006.

[8] O. Kurland, A. Shtok, D. Carmel, and S. Hummel. A unified framework for post-retrieval query-performance prediction. In *Proc. of ICTIR*, pages 15–26, 2011.

[9] V. Lavrenko and W. B. Croft. Relevance models in information retrieval. In *Language Modeling for Information Retrieval*, pages 11–56. Kluwer, 2003.

[10] M. Lease. Incorporating relevance and pseudo-relevance feedback in the markov random field model. In *Proc. of TREC 2008*, 2008.

[11] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *Proc. of CIKM*, pages 255–264, 2009.

[12] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

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

[14] F. Song and W. B. Croft. A general language model for information retrieval (poster abstract). In *Proc. of SIGIR*, pages 279–280, 1999.

[15] E. M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. In *Proc. of TREC-13*, 2004.

[16] R. W. White, J. M. Jose, and I. Ruthven. An implicit feedback approach for interactive information retrieval. *Inf. Process. Manage.*, 42:166–190, January 2006.

[17] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. of CIKM*, pages 102–111, 2006.

[18] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of SIGIR*, pages 334–342, 2001.

[19] L. Zhao, C. Liang, and J. Callan. Extending relevance model for relevance feedback. In *Proc. of TREC 2008*, 2008.

[20] Y. Zhou. *Retrieval Performance Prediction and Document Quality*. PhD thesis, University of Massachusetts, 2007.

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