Using Statistical Decision Theory and Relevance Models for Query-Performance Prediction

Anna Shtok[1] annabel@tx.technion.ac.il
Oren Kurland[1] kurland@ie.technion.ac.il
David Carmel[2] carmel@il.ibm.com

1. Faculty of Industrial Engineering and Management, Technion, Haifa 32000, Israel
2. IBM Research lab, Haifa 31905, Israel

ABSTRACT

We present a novel framework for the *query-performance prediction* task. That is, estimating the effectiveness of a search performed in response to a query in lack of relevance judgments. Our approach is based on using statistical decision theory for estimating the utility that a document ranking provides with respect to an information need expressed by the query. To address the uncertainty in inferring the information need, we estimate utility by the expected similarity between the given ranking and those induced by *relevance models*; the impact of a relevance model is based on its presumed *representativeness* of the information need. Specific query-performance predictors instantiated from the framework substantially outperform state-of-the-art predictors over five TREC corpora.

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

General Terms: Algorithms, Experimentation

Keywords: query-performance prediction, relevance models, statistical decision theory, rank correlation

1. INTRODUCTION

The effectiveness of search engines can significantly vary across queries [27, 9]. Thus, the ability to identify which queries are more difficult than others could be of great benefit. Indeed, there is a large body of work on predicting *query performance*, that is, estimating the effectiveness of a search performed in response to a query in lack of relevance-judgments information. (See Section 2 for a survey.)

We present a novel framework for query-performance prediction that is based on statistical decision theory. Specifically, we consider a ranking induced by a retrieval method in response to a query as a decision taken so as to satisfy the underlying information need [15]. The quality of the ranking — i.e., query-performance — is then estimated based on the utility it provides with respect to the presumed information need. However, there is often uncertainty about the actual information need, especially when examples of relevant documents are not provided and queries are ambiguous.

To address the uncertainty in inferring the information need, we first assume a "true" (latent) model of relevance, e.g., a *relevance language model* [16] that generates terms in the query and in relevant documents. Then, we estimate the utility of the given ranking by its expected similarity with the rankings induced by estimates of this relevance model.

We instantiate various query-performance predictors from the framework by varying the (i) estimates of the relevance model, (ii) measures for the similarity between the given ranking and that induced by a relevance-model estimate, and (iii) measures for the quality of a relevance-model estimate, that is, the extent to which it presumably *represents* the underlying information need.

A key observation, which enables to derive effective predictors from our framework, concerns the representativeness of relevance-model estimates for the underlying information need. We argue, and empirically show, that relevance-model representativeness can be estimated using various query-performance predictors that were originally devised to predict the quality of the document list from which the relevance model is constructed, e.g., the *Clarity* method [6].

Empirical evaluation performed using five TREC corpora shows that the predictors instantiated from our framework substantially and consistently outperform, in terms of prediction quality, four state-of-the-art predictors. A case in point, the average relative prediction improvement over the Clarity method [6] is above 30%.

2. RELATED WORK

Query-performance predictors can be roughly categorized to pre-retrieval and post-retrieval methods [13]. Pre-retrieval methods analyze the query expression before search is performed [6, 13, 21, 19, 11, 31, 10]. Linguistic features and/or statistical properties of the query-terms distribution are often used along with some corpus-based statistics.

Post-retrieval predictors analyze also the *result list* — the list of documents most highly ranked in response to the query. The *Clarity* method [6], for example, estimates the "focus" of the result list with respect to the corpus, as measured by the KL-divergence between their induced (language) models. Variants of clarity were proposed for improving prediction performance [2, 7, 4, 12].

Estimating result-list *robustness* is another effective post-retrieval prediction paradigm. Intuitively, the more robust the result list with respect to different factors, the less "dif-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

ficult" the query is. For example, the cohesion of the result list, as measured by its clustering patterns, indicates query performance [26]. The effects on the result list of document perturbations [26, 33], query perturbations [29], query-model modifications [34], and different retrieval functions [3], have been used to devise performance predictors.

Retrieval scores often reflect document-query similarity. Hence, analyzing retrieval-scores distribution can potentially help to predict query performance. Indeed, the highest retrieval score and the mean of top scores were shown to indicate query performance [25]. Query performance was also shown to be correlated with the extent to which similar documents in the result list are assigned with similar retrieval scores [8]. In addition, high retrieval scores at top ranks of the list with respect to that of the corpus [34], and large variance of retrieval scores in the list [22], were shown to indicate effective performance.

We argue, and empirically show, that some post-retrieval predictors can effectively be used in our framework to estimate the extent to which relevance models [16] presumably *represent* the information need expressed by a query. We use state-of-the-art predictors that represent the different paradigms mentioned above, namely, the *Clarity* method [6], the *query-feedback* (*QF*) approach, which is based on ranking robustness [34], and the *WIG* [34] and *NQC* [22] measures that utilize retrieval scores. (See Section 3.2.2 for a discussion on these predictors.) The effectiveness of the predictors we derive from the framework is substantially better than that of using those four predictors, as originally proposed, to estimate the quality of the result list.

Some work on ranking documents in response to a query [5, 28, 24] addresses the uncertainty in inferring the information need by using multiple relevance models (or more generally, query-expansion-based models) as we propose in our framework. Furthermore, representativeness measures for relevance models have also been used [28, 24], albeit not QF, WIG, and NQC that we use here. More importantly, the task we pursue — predicting the quality of a given document ranking — is different than that of inducing document ranking in response to a query [5, 28, 24].

Our treatment to document ranking as a decision made by the retrieval method in response to a query is inspired by the *risk minimization framework* [15]. However, while the latter aims at inducing document ranking, our framework is intended to estimate the utility of a given ranking.

3. QUERY-PERFORMANCE PREDICTION BASED ON UTILITY ESTIMATION

Let q, d, and $\mathcal{D}$ denote a query, a document and a corpus of documents, respectively. We assume that q is used to express some information need I_q.

We use $\pi_{\mathcal{M}}(q;S)$ to denote the ranking induced over a set of documents S in response to q using retrieval method $\mathcal{M}$. Our goal is to predict the *query-performance* of $\mathcal{M}$ with respect to q. In other words, we would like to quantify the quality (effectiveness) of the ranking of all documents in the corpus, $\pi_{\mathcal{M}}(q;\mathcal{D})$, with respect to the information need I_q in lack of relevance-judgments information.

3.1 Prediction framework

The corpus ranking, $\pi_{\mathcal{M}}(q;\mathcal{D})$, could be viewed as a decision made by the retrieval method $\mathcal{M}$ in response to q so as to satisfy the user's (hidden) information need I_q [15]. The ranking effectiveness reflects the *utility* provided to the user, denoted $U(\pi_{\mathcal{M}}(q;\mathcal{D});I_q)$. In what follows we devise estimates for this utility, i.e., query-performance predictors.

Suppose that there is an oracle that provides us with a *"true"* model of relevance R_{I_q} representing the information need I_q. Suppose also that R_{I_q} can be used by $\mathcal{M}$ for ranking — e.g., that R_{I_q} has a query-model representation to be used by $\mathcal{M}$. A statistical *relevance language model* that generates the terms in relevant documents is an example for such a query model in the language modeling framework [16]. Then, according to the *probability ranking principle* [20], using $\mathcal{M}$ with R_{I_q} yields a ranking $\pi_{\mathcal{M}}(R_{I_q};\mathcal{D})$ of maximal utility (e.g., all relevant documents are positioned at the highest ranks, and all non-relevant documents are positioned below the relevant documents). Thus, we can use the maximal-utility ranking $\pi_{\mathcal{M}}(R_{I_q};\mathcal{D})$ to estimate the utility of the given ranking, $\pi_{\mathcal{M}}(q;\mathcal{D})$, based on their "similarity", measures of which we discuss in Section 3.2.3:

$$U(\pi_{\mathcal{M}}(q;\mathcal{D});I_q) \stackrel{def}{=} Sim(\pi_{\mathcal{M}}(q;\mathcal{D}),\pi_{\mathcal{M}}(R_{I_q};\mathcal{D})). \quad (1)$$

In practice, we have no explicit knowledge of the underlying information need I_q, except for the information in q, nor do we have an oracle to provide us with a model of relevance. Hence, we use estimates $\hat{R}_q$ for R_{I_q} that are based on the information in q and in the corpus. Using statistical decision theory principles, we can approximate Equation 1 by the expected similarity between the given ranking and those induced by the estimates for R_{I_q}:

$$U(\pi_{\mathcal{M}}(q;\mathcal{D});I_q) \approx \int_{\hat{R}_q} Sim(\pi_{\mathcal{M}}(q;\mathcal{D}),\pi_{\mathcal{M}}(\hat{R}_q;\mathcal{D}))p(\hat{R}_q|I_q)\ d\hat{R}_q; \quad (2)$$

$p(\hat{R}_q|I_q)$ is the probability that $\hat{R}_q$ represents I_q — i.e., that $\hat{R}_q$ is the "true" relevance model R_{I_q}.

In practice, users' utility is often determined based on the documents most highly ranked. Indeed, evaluation measures of retrieval effectiveness (e.g., MAP and precision at top k) consider only top-retrieved documents — a.k.a., the *result list*. This is also the case for many *post-retrieval* query-performance prediction methods (e.g., [6, 29, 33, 34, 4]) that analyze the result list. Thus, we confine the utility analysis to the result list $\mathcal{D}^{[k]}_{\mathcal{M};q}$ — the k most highly ranked documents in the corpus $\mathcal{D}$ by $\mathcal{M}$ with respect to q:

$$U(\pi_{\mathcal{M}}(q;\mathcal{D});I_q) \approx \int_{\hat{R}_q} Sim(\pi_{\mathcal{M}}(q;\mathcal{D}^{[k]}_{\mathcal{M};q}),\pi_{\mathcal{M}}(\hat{R}_q;\mathcal{D}^{[k]}_{\mathcal{M};q}))p(\hat{R}_q|I_q)\ d\hat{R}_q; \quad (3)$$

That is, we measure the expected inter-ranking similarity between that of the top-k results of the original ranking and their re-ordering as induced by the relevance model estimates; if we set k to the number of documents in the corpus, then Equation 3 reduces to Equation 2.

Equation 3 can be instantiated in numerous ways to yield specific query-performance predictors. We have to (i) derive estimates $\hat{R}_q$ for the true relevance model (see Section 3.2.1), (ii) estimate the extent to which these estimates represent the hidden information need ($p(\hat{R}_q|I_q)$) — a task which we address in Section 3.2.2, and (iii) select measures of similarity between ranked lists (see Section 3.2.3).

3.2 Framework instantiation

3.2.1 Relevance models

We use relevance language models (RM3) [16, 1] for the estimates $\hat{R}_q$. Relevance-model estimation is performed using a pseudo-feedback-based approach, that is, utilizing a list $\mathcal{D}^{[\nu]}_{QL;q}$ of the ν highest ranked documents by the *query-likelihood* (QL) method [23]. Specifically, if $p(w|x)$ is the probability assigned to term w by a language model induced from text (or text collection) x, then the query-likelihood approach scores x in response to $q = \{q_i\}$ by

$$Score_{QL}(q;x) \stackrel{def}{=} \log p(q|x) \stackrel{def}{=} \log \prod_{q_i} p(q_i|x). \quad (4)$$

A relevance-model estimate, $\hat{R}_{q;S}$, constructed from a set S ($\subset \mathcal{D}^{[\nu]}_{QL;q}$) of highly ranked documents, is a probability distribution over the vocabulary:

$$p(w|\hat{R}_{q;S}) \stackrel{def}{=} \lambda p(w|q) + (1-\lambda) \sum_{d \in S} p(w|d)p(d|q); \quad (5)$$

$p(d|q) \stackrel{def}{=} \frac{p(q|d)}{\sum_{d' \in S} p(q|d')}$ is d's normalized query likelihood; λ is a free parameter. To rank documents using $\hat{R}_{q;S}$, e.g., in Equation 3, the negative cross entropy (CE) is used:

$$Score_{CE}(\hat{R}_{q;S};d) \stackrel{def}{=} \sum_w p(w|\hat{R}_{q;S}) \log p(w|d). \quad (6)$$

Following some work on addressing the performance robustness issues of pseudo-feedback-based methods [5, 17, 24], we create relevance-model estimates by sampling sets S of documents from $\mathcal{D}^{[\nu]}_{QL;q}$. (See Section 4 for specific details.) Then, we construct from each set S a relevance-model estimate $\hat{R}_{q;S}$ using Equation 5, that is used to rank $\mathcal{D}^{[k]}_{\mathcal{M};q}$ according to Equation 6. Finally, we approximate the utility in Equation 3 by:

$$U(\pi_{\mathcal{M}}(q;\mathcal{D});I_q) \approx \quad (7)$$
$$\sum_{\hat{R}_{q;S}} Sim(\pi_{\mathcal{M}}(q;\mathcal{D}^{[k]}_{\mathcal{M};q}), \pi_{\mathcal{M}}(\hat{R}_{q;S};\mathcal{D}^{[k]}_{\mathcal{M};q}))p(\hat{R}_{q;S}|I_q).$$

3.2.2 Representativeness of relevance models

Quantifying the extent to which a relevance-model estimate, $\hat{R}_{q;S}$, *represents* the information need I_q (i.e., $p(\hat{R}_{q;S}|I_q)$) is a prediction challenge at its own right [7, 28, 24]. It is important to point out that the framework proposed above is not committed to any specific paradigm of quantifying representativeness. Specifically, there is no coupling in the framework between the way relevance model estimates are devised and the methods used for quantifying their representativeness. Here, we adapt four effective query-performance measures that were originally proposed for predicting the quality of a result list as surrogates for estimates $\hat{p}(\hat{R}_{q;S}|I_q)$ of relevance model representativeness. These four measures represent the different post-retrieval query-performance prediction paradigms surveyed in Section 2.

Clarity. A natural measure for the representativeness of a relevance-model estimate $\hat{R}_{q;S}$ is its *clarity* [6]; that is, the "distance" between $\hat{R}_{q;S}$ and the corpus model, which can be measured by the KL divergence:

$$\hat{p}_{Clarity}(\hat{R}_{q;S}|I_q) \propto \sum_w p(w|\hat{R}_{q;S}) \log \frac{p(w|\hat{R}_{q;S})}{p(w|\mathcal{D})}.$$

The larger the KL divergence is, the more distant $\hat{R}_{q;S}$ is from the corpus model, and hence, is considered more coherent ("clear"). Since $\hat{R}_{q;S}$ is constructed from documents highly ranked in response to q, we assume, as in some recent work on utilizing multiple relevance models for retrieval [28, 24], that higher clarity indicates better representativeness of the information need. We note that originally [6], the clarity of *some* relevance model was used as an estimate for the quality of the result list from which the model was constructed, rather than as an estimate for the representativeness of the relevance model.

WIG. The Clarity measure estimates the representativeness of $\hat{R}_{q;S}$ by directly measuring its "quality" as a language model. An alternative paradigm is based on estimating the presumed percentage of relevant documents in the set S from which $\hat{R}_{q;S}$ is constructed. The higher this presumed percentage, the better representative $\hat{R}_{q;S}$ is assumed to be. Recall that S is a set of documents highly ranked by the query-likelihood method. Hence, estimating the relevant-document percentage in S is a form of a query-performance prediction task. To that end, we use the WIG measure [34][1]:

$$\hat{p}_{WIG}(\hat{R}_{q;S}|I_q) \propto$$
$$\frac{1}{\sqrt{|q|}} \frac{1}{|S|} \sum_{d \in S} (Score_{QL}(q;d) - Score_{QL}(q;\mathcal{D}));$$

normalization with respect to the query length, which affects query-likelihood scores (see Equation 4), is performed for inter-query compatibility.

Note that WIG relies on the premise that high retrieval scores with respect to that of the corpus imply to relevance. Thus, WIG was originally [34] computed based on the retrieval scores of the most highly ranked documents in the result list so as to predict the quality of the result list itself. Here, we use WIG to measure the presumed representativeness of a relevance model constructed from S ($\subset \mathcal{D}^{[\nu]}_{QL;q}$), as we assume that WIG is correlated with the percentage of relevant documents in S.[2]

NQC. A recently proposed query-performance predictor [22], NQC, is based on the hypothesis that the standard deviation of retrieval scores in the result list is negatively correlated with the potential amount of *query drift* [18] — i.e., non-query-related information manifested in the list. Specifically, the mean (QL) retrieval score in the list was shown to be the retrieval score of a pseudo non-relevant document, namely, a centroid of the result list. Thus, result lists with retrieval scores much higher/lower than the mean were argued, and shown, to be of high quality.

Hence, NQC can potentially help to estimate the quality of $\mathcal{D}^{[\nu]}_{QL;q}$, and thereby, the extent to which a relevance-model

[1] WIG was originally proposed in the Markov Random Field framework [34]. If no term-dependencies are considered, WIG measures the query-performance of the query-likelihood approach, and is effective to this end [32, 22].

[2] There are similar measures used to this end in work on cluster-based retrieval [14].

estimate constructed from $\mathcal{D}^{[\nu]}_{QL;q}$ is a good representative of the information need. However, NQC does not have a natural implementation for a subset S of $\mathcal{D}^{[\nu]}_{QL;q}$. Therefore, we use (as a rough approximation) for each such S the NQC computed for $\mathcal{D}^{[\nu]}_{QL;q}$:

$$\hat{p}_{NQC}(\hat{R}_{q;S}|I_q) \propto \frac{\sqrt{\frac{1}{\nu}\sum_{d\in\mathcal{D}^{[\nu]}_{QL;q}}(Score_{QL}(q;d)-\mu)^2}}{|Score_{QL}(q;\mathcal{D})|};$$

μ is the mean retrieval score in $\mathcal{D}^{[\nu]}_{QL;q}$; normalization with the corpus score is for inter-query compatibility [22].

QF. The query-feedback (QF) performance predictor measures ranking robustness [34]. Specifically, the quality of the list $\mathcal{D}^{[\nu]}_{QL;q}$ is presumed to be correlated with the overlap between the top-n_{QF} ranked documents in $\mathcal{D}^{[\nu]}_{QL;q}$ and the top-n_{QF} ranked documents in the corpus by a search performed using $\hat{R}_{q;\mathcal{D}^{[\nu]}_{QL;q}}$ — a relevance model constructed from $\mathcal{D}^{[\nu]}_{QL;q}$.[3] The overlap is simply the number of shared documents. The idea is that a relevance model constructed from a high quality list would not yield a ranking that *drifts* much from the original ranking. Thus, here we utilize QF as a measure for the representativeness of $\hat{R}_{q;\mathcal{D}^{[\nu]}_{QL;q}}$ — the less drift the ranking it induces manifests, the more likely it is to represent the information need [7, 28, 24].

In our experiments we use QF only with $\hat{R}_{q;\mathcal{D}^{[\nu]}_{QL;q}}$ and not with relevance-model estimates that are constructed from subsets of $\mathcal{D}^{[\nu]}_{QL;q}$. (See Section 4 for details.)

3.2.3 Similarity between ranked lists

The remaining task for instantiating Equation 3 is the estimation of the similarity $Sim(\pi_{\mathcal{M}}(q;\mathcal{D}^{[k]}_{\mathcal{M};q}), \pi_{\mathcal{M}}(\hat{R}_{q;S};\mathcal{D}^{[k]}_{\mathcal{M};q}))$ between two rankings of the given result list $\mathcal{D}^{[k]}_{\mathcal{M};q}$. We use three popular measures for similarity between rankings: Pearson's coefficient, which measures the linear correlation between the retrieval scores used to induce the two rankings, and Spearman's-ρ and Kendall's-τ that rely only on ranks. All three correlation measures assign values in $[-1, +1]$.

4. EVALUATION

In what follows we evaluate the effectiveness of predictors instantiated from our proposed framework, denoted **UEF** for *utility estimation framework*. As noted above, to derive a specific predictor from Equation 7 we have to (i) devise a sampling technique for document sets from which relevance-model estimates are created, (ii) select a representativeness measure for relevance-model estimates, and (iii) select a measure of similarity between ranked lists.

As noted above, we use Clarity, WIG, NQC, and QF as measures for the representativeness of relevance model estimates; and, Pearson's coefficient, Spearman's-ρ, and Kendall's-τ as inter-ranking similarity measures. We use two strategies for sampling sets S of documents from $\mathcal{D}^{[\nu]}_{QL;q}$ — the documents most highly ranked by the query likelihood approach — so as to define relevance-model estimates. The first, denoted **Single**, is using $\mathcal{D}^{[\nu]}_{QL;q}$ as a single sampled set. As this is the standard, highly effective, approach for selecting documents for relevance-model estimation [16, 1], the resultant predictor instantiated from Equation 7 could be regarded as the posterior-mode estimate for the integral in Equation 3 [15]. Note that the predictor in this case is the similarity between the ranking induced over $\mathcal{D}^{[k]}_{\mathcal{M};q}$ by the single relevance model and the original ranking of $\mathcal{D}^{[k]}_{\mathcal{M};q}$, scaled by the representativeness estimate of the relevance model.

The second document sampling strategy, **Multi**, is based on using multiple *clusters* of similar documents from $\mathcal{D}^{[\nu]}_{QL;q}$ [17, 24]. Specifically, we employ a simple nearest-neighbors-based clustering approach wherein each document d ($\in \mathcal{D}^{[\nu]}_{QL;q}$) and its $\delta-1$ nearest neighbors from $\mathcal{D}^{[\nu]}_{QL;q}$ serve as a cluster [14]; we use the KL divergence between document language models for a similarity measure [14]. Thus, we sample ν (overlapping) clusters of δ documents.

4.1 Experimental setup

We evaluate the prediction quality of a query-performance predictor by measuring Pearson's correlation between the actual average precision (AP at cutoff 1000) for a set of queries — as measured by using relevance judgments — and the values assigned to the queries by the predictor [4, 34]. All correlation numbers that we report are statistically significant at a 95% confidence level.

We conducted experiments on several TREC collections that were used in previous query-performance-prediction studies [29, 33, 32, 8]. Table 1 provides the details of the collections and topics used.

Collection	Data	Num Docs	Topics	Rels/topic
TREC4	Disks 2&3	567,529	201-250	130.06
TREC5	Disks 2&4	524,929	251-300	110.48
WT10G	WT10g	1,692,096	451-550	61.14
ROBUST	Disk 4&5-CR	528,155	301-450, 601-700	69.92
GOV2	GOV2	25,205,179	701-850	181.79

Table 1: Test collections and topics. The last column reports the average number of relevant documents per topic.

We use titles of TREC topics for queries, except for TREC4 for which no titles are provided, and hence, topic descriptions are used. We applied tokenization, Porter stemming, and stopword removal (using the INQUERY list) to all data via the Lemur/Indri toolkit (www.lemurproject.org), which was also used for experiments.

The query likelihood (QL) model [23] from Equation 4 serves as the retrieval model $\mathcal{M}$; k, the size of the result-list ($\mathcal{D}^{[k]}_{\mathcal{M};q}$) considered for predicting performance, is set to 150. (Experiments with $k = 100$ yielded slightly worse performance.) Recall that documents highly ranked by QL (i.e., in the list $\mathcal{D}^{[\nu]}_{QL;q}$) are those used for relevance-model estimation. To downplay the effect of parameter tuning, we set $\nu = k$ in all experiments to follow; hence, $\mathcal{D}^{[\nu]}_{QL;q} \equiv \mathcal{D}^{[k]}_{\mathcal{M};q}$. Thus, the document list for which we predict query performance is also the one utilized for devising estimates of relevance models. We come back to this point in Section 4.2.2. For the *Multi* sampling strategy, we set the cluster size, δ, to 20; smaller clusters yield less effective prediction.

The representativeness measures of relevance model estimates that we use — Clarity, WIG, NQC, and QF — are

[3] To maintain consistency with all other predictors, we use RM3 as defined in Equation 5, which is somewhat different than the query model used originally [34].

state-of-the-art query-performance predictors at their own right. Hence, we use optimized versions of these as reference comparisons to our predictors. The Clarity method uses a relevance model constructed from all documents in $\mathcal{D}_{QL;q}^{[\nu]}$; $\nu = 150$ indeed yields optimal Clarity performance. $\nu = 150$ also yields highly effective prediction performance for the NQC measure as previously reported [22]. The QF measure, as Clarity, uses a relevance model constructed from $\mathcal{D}_{QL;q}^{[\nu]}$; optimal QF prediction performance is attained when setting the number of top documents it depends on, n_{QF}, to 50. The WIG measure is highly effective when using the 5 top-ranked documents in $\mathcal{D}_{QL;q}^{[\nu]}$ as previously reported [34].

To facilitate comparison with the optimized reference comparisons just described, we use those as relevance-model representativeness measures in our framework; except, for WIG and Clarity with *Multi* sampling that use the information within a cluster for devising a representative measure for the relevance model constructed from it.

Language models. We use Dirichlet smoothed unigram document language models with the smoothing parameter, μ, set to 1000 [30]. To *construct* a relevance model from a document set using Equation 5, we set $\mu = 0$ for language models of documents in the set, and $\lambda = 0$ (i.e., we use RM1); all relevance models use 100 terms [1]. These parameter values yield very good performance both for our predictors and for Clarity and QF that utilize relevance models.

4.2 Experimental results

In Section 4.2.5 we study the effect of varying the inter-ranking-similarity measure, the document-sets sampling strategy, and the representativeness measure on the instantiated predictors' performance. We show that the best performing predictors are those that use Pearson correlation as inter-ranking similarity measure, and a single relevance model estimate (*Single*). Hence, in Sections 4.2.1-4.2.4 we present an in-depth analysis of the performance of these predictors.

We note that the computational overhead posted by these predictors on top of computing the representativeness estimates they incorporate, which is performed by current predictors adapted to this end, is quite small. That is, a relevance model is constructed from documents in the (short) result list and is used to rank this list; then, Pearson correlation between the original list ranking and its relevance-model-based ranking is computed.

4.2.1 Main results

In what follows we fix the inter-ranking similarity measure to Pearson's coefficient, and use a single relevance model estimate constructed from the entire initial list, $\mathcal{D}_{QL;q}^{[\nu]}$. The prediction quality of the UEF-based predictors, when using the four representativeness measures for the relevance model estimate, is presented in Table 2.

Evidently, our predictors consistently and substantially improve over using the representativeness measures as predictors at their own right — i.e., to directly predict search effectiveness; recall that these are state-of-the-art predictors[4]. A case in point, using UEF with Clarity improves prediction quality by more than 30% on average, over the 5 TREC benchmarks, with respect to direct use of Clarity for performance prediction. Furthermore, the improvements over all four representative measures for the WT10G benchmark are quite striking as WT10G is known to post a hard challenge for performance prediction [12]. The relative improvements for GOV2, on the other hand, are in general smaller than those for the other collections.

4.2.2 Quality of representativeness measures

As noted above, the representativeness measures that we use were *originally* shown to be highly effective predictors for the quality of the initial QL-based ranking from which $\mathcal{D}_{QL;q}^{[\nu]}$ was created — the task that we pursue here as well — rather than for the representativeness of a relevance model constructed from $\mathcal{D}_{QL;q}^{[\nu]}$. Thus, when using a single relevance model estimate (*Single*), our UEF-based predictors could be viewed in this specific setting as combining two predictors for the ranking quality of $\mathcal{D}_{QL;q}^{[\nu]}$ *itself*. The first is the representativeness measure and the second is the similarity between the original ranking and that induced by the relevance model estimate[5]. We hasten to point out, however, that this specific operational consequence does not contradict the fundamentals of our framework. On the contrary, highly effective predictors for the ranking quality of $\mathcal{D}_{QL;q}^{[\nu]}$ should serve as effective measures for the representativeness of the single relevance model constructed from $\mathcal{D}_{QL;q}^{[\nu]}$, as was argued in Section 3.2.2. That is, the more relevant documents there are in $\mathcal{D}_{QL;q}^{[\nu]}$, and the higher they are ranked, the higher the quality of the relevance model estimate is (refer to Equation 5).

We thus turn to empirically examine the premise just stated that effective predictors of the quality of the ranking using which $\mathcal{D}_{QL;q}^{[\nu]}$ is created are indeed effective measures of the representativeness of the relevance-model constructed from $\mathcal{D}_{QL;q}^{[\nu]}$. To that end, we study the performance of the representativeness measures (Clarity, WIG, NQC, QF) when predicting the quality of the ranking induced by the *relevance model* over the *entire* corpus[6]. Prediction performance is measured, as usual, by the Pearson correlation between the true AP of the relevance-model-based corpus ranking (at cutoff 1000) and that which corresponds to the predicted values. For reference comparison, we report the performance of using the measures to *directly* predict the quality of the initial QL-based ranking, as originally proposed.

The results in Figure 1 support our premise. That is, the measures are indeed high-quality representativeness estimates for the relevance model, as the high correlation numbers attest. While in general the measures are more effective in directly predicting the quality of the QL-based initial ranking rather than serving as representativeness estimates, the reverse holds for the QF measure over most collections.

[4]Somewhat similar relative prediction performance patterns are observed when using Kendall's-τ rather than Pearson's coefficient to measure *prediction quality*. Specifically, the average improvements over Clarity, WIG, NQC and QF are 26%, 24%, 27%, and 21%, respectively. Specific prediction results are omitted due to space considerations.

[5]This ranking-similarity-based predictor yields prediction performance of .607, .579, .46, .578 and .402 for TREC4, TREC5, WT10G, ROBUST, and GOV2, respectively. Hence, while it is an effective predictor at its own right, its integration with the representativeness measure yields, in general, much better prediction performance.

[6]Similar prediction performance patterns — actual numbers are omitted to avoid cluttering the presentation — are observed with respect to the ranking induced by the relevance model over the inital list $\mathcal{D}_{QL;q}^{[\nu]}$.

Predictor	TREC4	TREC5	WT10G	ROBUST	GOV2	avg. improv.
Clarity	0.453	0.42	0.348	0.512	0.433	
UEF(Clarity)	0.623 (+37.5%)	**0.629** (+49.8%)	0.483 (+38.8%)	0.635 (+24%)	0.462 (+6.7%)	(+31.4%)
WIG	0.544	0.297	0.376	0.543	0.479	
UEF(WIG)	0.638 (+17.3%)	0.555 (+86.9%)	0.453 (+20.5%)	**0.644** (+18.6%)	0.458 (−4.4%)	(+27.8%)
NQC	0.588	0.354	0.488	0.566	0.36	
UEF(NQC)	0.641 (+9%)	0.545 (+53.9%)	0.522 (+6.9%)	0.619 (+9.4%)	0.393 (+9.2%)	(+17.7%)
QF	0.627	0.414	0.426	0.285	0.476	
UEF(QF)	**0.666** (+6.2%)	0.538 (+29.9%)	**0.526** (+23.5%)	0.459 (+61%)	**0.491** (+3.1%)	(+24.7%)

Table 2: Prediction quality of UEF when fixing the sampling technique to *Single*, fixing the inter-ranking similarity measure to *Pearson*, and using the four representativeness estimates. The prediction quality of each representativeness estimate when used to directly predict search effectiveness is presented for reference in the first row of a block. Best result in a column is boldfaced.

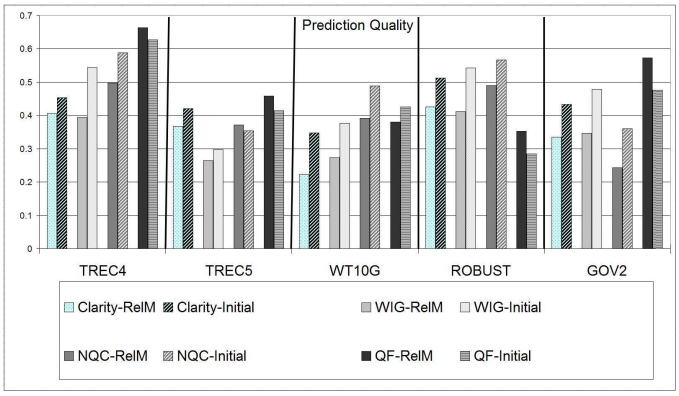

Figure 1: Effectiveness of the performance predictors for estimating the representativeness of a relevance model (*RelM*) as measured by the prediction quality of the ranking it induces over the corpus. The prediction quality of the *initial* query-likelihood-based ranking is presented for reference.

The finding regarding QF sheds some light on its high effectiveness when used in our framework. (See Table 2.) Recall that UEF(QF) operates as follows: a relevance model is constructed from $\mathcal{D}^{[\nu]}_{QL;q}$ and is used to rank the corpus. The overlap between top-ranked documents and those at top-ranks of $\mathcal{D}^{[\nu]}_{QL;q}$ serves for the relevance-model representativeness estimate; small overlap presumably attests to query-drift manifested by the relevance model. Then, the estimate is multiplied by the similarity (Pearson's coefficient) between $\mathcal{D}^{[\nu]}_{QL;q}$'s original ranking and that induced over it by the relevance model.

4.2.3 Prediction for rankings produced by other retrieval methods

All the predictors that we have studied insofar operate in the language modeling framework. More specifically, we have focused on predicting the effectiveness of the ranking used to create the list $\mathcal{D}^{[\nu]}_{QL;q}$ using a relevance language model constructed from this list. We now turn to study the effectiveness of our framework in predicting the quality of a ranking produced by another retrieval method. This challenge fits, for example, the scenario of predicting the performance of a retrieval method that may not be known to the predictor, but rather only the induced ranking and/or retrieval scores.

We predict performance for vector-space-based (VS) retrieval with the cosine measure, and for Okapi's BM25 retrieval model. In both cases Lemur's implementation with default parameter settings is used.

Note that the effectiveness of a ranking of the corpus by $\mathcal{M}$ (VS or BM25) is estimated by our predictors as follows. We measure the similarity (Pearson's coefficient) between the original ranking of the result list $\mathcal{D}^{[k]}_{\mathcal{M};q}$ of top-k retrieved documents by $\mathcal{M}$, and the ranking of $\mathcal{D}^{[k]}_{\mathcal{M};q}$ by a single relevance model constructed from $\mathcal{D}^{[\nu]}_{QL;q}$ — the QL-based result list; this similarity is then scaled by the estimated representativeness of the relevance model. Thus, the result list for which we predict performance, $\mathcal{D}^{[k]}_{\mathcal{M};q}$, is different than the list $\mathcal{D}^{[\nu]}_{QL;q}$ used to construct a relevance model.

As in Table 2, we want to compare our predictor's performance with that of applying the representativeness measure as a predictor at its own right directly to $\mathcal{D}^{[k]}_{\mathcal{M};q}$. Among the four measures used, NQC is the only predictor that has a non language-model-based implementation; specifically, it was shown to be effective with VS and BM25 [22]. Hence, in Table 3 we compare the performance of our framework using NQC as a representativeness measure, UEF(NQC), with that of using NQC directly to predict performance[7]. We set ν=k=150 as at the above.

The results in Table 3 clearly attest to the general effectiveness of our framework. Indeed, UEF(NQC) yields better prediction performance for vector-space and Okapi-BM25 retrieval than that of NQC over most collections.

	TREC4	TREC5	WT10G	ROBUST
VS-NQC	0.660	0.440	**0.407**	0.535
VS-UEF(NQC)	**0.678**	**0.503**	0.393	**0.625**
BM25-NQC	0.578	0.423	0.31	**0.6**
BM25-UEF(NQC)	**0.668**	**0.499**	**0.508**	0.592

Table 3: Using our framework, UEF(NQC), to predict the performance of vector space (VS) and Okapi-BM25 retrieval in comparison to using NQC to directly predict their performance. Boldface marks the best result in a block.

4.2.4 Integrating predictors

Integrating predictors using linear interpolation was shown to be of merit [8, 34]. As such integration yields a predictor for the quality of the initial list from which we construct a relevance model, we can use it in our framework as a repre-

[7]The original reports for NQC have not used GOV2 [22].

sentative measure. Specifically, when integrating WIG and QF [34] the resultant predictor is denoted UEF(WIG+QF), where WIG+QF is the interpolation-based predictor. We also study the performance of UEF(WIG)+UEF(QF) that interpolates UEF(WIG) and UEF(QF). Interpolation with equal weights is performed in all cases upon the min-max normalized values assigned by predictors. The prediction performance numbers are presented in Table 4.

In accordance with previous findings [34], we see in Table 4 that integrating WIG and QF (WIG+QF) results in performance superior to that of each over most corpora. Using the integrated predictor in our framework (UEF(WIG+QF)) yields further improvements for 3 out of the 5 corpora. Furthermore, for all corpora, except for GOV2, it is better to integrate our predictors that are based on WIG and QF — i.e., UEF(WIG) and UEF(QF) — than to integrate WIG and QF directly. (Compare UEF(WIG)+UEF(QF) and WIG+QF.)

	TREC4	TREC5	WT10G	ROBUST	GOV2
WIG	0.544	0.297	0.376	0.543	0.479
UEF(WIG)	0.638	0.555	0.453	**0.644**	0.458
QF	0.627	0.414	0.426	0.285	0.476
UEF(QF)	0.666	0.538	**0.526**	0.459	0.491
WIG+QF	0.676	0.446	0.472	0.503	**0.555**
UEF(WIG+QF)	0.663	0.562	0.513	0.586	0.501
UEF(WIG)+ UEF(QF)	**0.679**	**0.591**	0.521	0.591	0.490

Table 4: Integrating predictors using linear interpolation (+). Best result in a column is boldfaced.

4.2.5 Deeper inside UEF

Heretofore, we have focused on instantiating our framework by utilizing the *Single* sampling strategy (i.e., a single relevance model estimate), and using Pearson's coefficient to measure inter-ranking similarities. We now turn to examine the effect of varying these factors, along with the measures used for estimating representativeness. To study the effectiveness of the latter when using *Multi* sampling (i.e., constructing relevance models from clusters), we use as a reference comparison the *Uniform* measure that assigns all relevance models the same representativeness value. Table 5 presents the performance of all predictors.[8]

We can see in Table 5 that all instantiated predictors are effective as the relatively high performance numbers indicate. Specifically, almost all of these predictors yield positive average improvements (refer to the last column) over the representativeness measures that they incorporate when the latter are used to directly predict performance. A notable exception is the ROBUST benchmark with the *Multi* strategy. (See the below for further discussion.) All in all, as the representativeness measures are state-of-the-art predictors at their own right, we find these results gratifying.

Among the inter-ranking similarity measures, we see that Pearson's coefficient in general performs best, attesting to the importance of considering the retrieval scores used to induce the rankings.

We can also see in Table 5 that using a single relevance model estimate (*Single*) yields superior performance to that of utilizing multiple relevance models (*Multi*) constructed from clusters. This finding is not surprising as most representativeness measures that we use are not well suited for estimating representativeness of a relevance model constructed from a small cluster, which is composed of *some* top-retrieved documents. This is of utmost importance as potentially very few of the clusters contain a high percentage of relevant documents, and identifying these is a hard challenge [14]. Indeed, WIG, which is based on retrieval scores within the clusters, yields performance that is inferior under *Multi* sampling to that of using uniform values for representativeness. On the other hand, Clarity, which measures the "quality" of the relevance model constructed from the cluster with respect to the corpus, does improve consistently over uniform representativeness scores. This finding attests to the potential of cluster-based sampling.

[8]Results for QF with *Multi* are not presented as those require running tens of relevance models per query over the corpus. This is computationally demanding, and accordingly, does not constitute a realistic prediction scenario.

5. CONCLUSION AND FUTURE WORK

We presented a novel framework, which is based on statistical decision theory, for predicting query performance. The quality of a given document ranking is predicted based on its expected similarity with those induced by estimates of relevance models; the presumed representativeness of a relevance-model estimate of the underlying information need determines its impact. Relevance-model representativeness is measured using state-of-the-art query-performance predictors that were originally designed to estimate the quality of the initial search. Empirical evaluation shows that predictors instantiated from our framework are substantially more effective than current state-of-the-art predictors.

Improving the sampling technique used for relevance-model construction, and devising and adapting [24] better measures of representativeness for relevance models constructed from clusters, are future directions we intend to explore.

Acknowledgments We thank the reviewers for their comments. This paper is based upon work supported in part by Israel's Science Foundation under grant no. 890015. and by G. S. Elkin research fund at the Technion. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsoring institutions.

6. 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 *Proceedings of TREC-13*, pages 715–725, 2004.

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

[3] J. A. Aslam and V. Pavlu. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceeding of ECIR*, pages 198–209, 2007.

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

[5] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of SIGIR*, pages 303–310, 2007.

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

[7] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A language modeling framework for selective query expansion. Technical Report IR-338, Center for Intelligent Information Retrieval, University of Massachusetts, 2004.

Sampling	Similarity	Rep.	TREC4	TREC5	WT10G	ROBUST	GOV2	avg. improv.
		Clarity	0.453	0.42	0.348	0.512	0.433	
		WIG	0.544	0.297	0.376	0.543	0.479	
		NQC	0.588	0.354	0.488	0.566	0.36	
		QF	0.627	0.414	0.426	0.285	0.476	
Single	*Pearson*	Clarity	0.623(+37%)	0.629(+50%)	0.483(+39%)	0.635(+24%)	0.462(+7%)	+31%
		WIG	0.638(+17%)	0.555(+87%)	0.453(+20%)	0.644(+19%)	0.458(−4%)	+28%
		NQC	0.641(+9%)	0.545(+54%)	0.522(+7%)	0.619(+9%)	0.393(+9%)	+18%
		QF	0.666(+6%)	0.538(+30%)	0.526(+23%)	0.459(+61%)	0.491(+3%)	+25%
	Spearman's-ρ	Clarity	0.687(+52%)	0.419(−0.2%)	0.387(+11%)	0.489(−4%)	0.458(+6%)	+13%
		WIG	0.644(+18%)	0.369(+24%)	0.351(−7%)	0.521(−4%)	0.459(−4%)	+5%
		NQC	0.633(+8%)	0.369(+4%)	0.436(−11%)	0.526(−7%)	0.39(+8%)	+0.4%
		QF	0.642(+2%)	0.464(+12%)	0.433(+2%)	0.369(+29%)	0.503(+6%)	+10%
	Kendall's-τ	Clarity	0.685(+51%)	0.428(+2%)	0.369(+6%)	0.51(−0.4%)	0.461(+7%)	+13%
		WIG	0.633(+16%)	0.377(+27%)	0.322(−14%)	0.537(−1%)	0.457(−4%)	+5%
		NQC	0.617(+5%)	0.374(+6%)	0.411(−16%)	0.535(−5%)	0.376(+5%)	−1%
		QF	0.632(+0.8%)	0.470(+13%)	0.425(−0.3%)	0.384(+35%)	0.501(+5%)	+11%
Multi	*Pearson*	Uniform	0.617	0.441	0.443	0.391	0.435	
		Clarity	0.641(+41%)	0.479(+14%)	0.454(+30%)	0.457(−11%)	0.436(+1%)	+15%
		WIG	0.603(+11%)	0.433(+46%)	0.425(+13%)	0.443(−19%)	0.471(−2%)	+10%
		NQC	0.611(+4%)	0.416(+17%)	0.531(+9%)	0.539(−5%)	0.418(+16%)	+8%
	Spearman's-ρ	Uniform	0.636	0.478	0.371	0.3	0.447	
		Clarity	0.653(+44%)	0.489(+16%)	0.379(+9%)	0.356(−30%)	0.441(+2%)	+8%
		WIG	0.612(+13%)	0.441(+48%)	0.349(−7%)	0.358(−34%)	0.474(−1%)	+4%
		NQC	0.599(+2%)	0.418(+18%)	0.454(−7%)	0.454(−20%)	0.42(+17%)	+2%
	Kendall's-τ	Uniform	0.644	0.493	0.377	0.33	0.445	
		Clarity	0.656(+45%)	0.5(+19%)	0.382(+10%)	0.381(−26%)	0.439(+1%)	+10%
		WIG	0.61(+12%)	0.451(+52%)	0.347(−8%)	0.379(−30%)	0.467(−2%)	+5%
		NQC	0.591(+0.5%)	0.426(+20%)	0.45(−8%)	0.47(−17%)	0.403(+12%)	+1%

Table 5: The effectiveness of predictors instantiated from our framework. We vary the document-sets sampling approach, the inter-rankings similarity measure, and the representativeness (rep.) measure of the relevance model estimate; "Uniform" stands for using the same constant representativeness value for all relevance models. The first four rows present the performance of the representativeness measures when used to directly predict performance; percentages of improvements are with respect to these values.

[8] F. Diaz. Performance prediction using spatial autocorrelation. In *Proceedings of SIGIR*, pages 583–590, 2007.

[9] D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *Proceedings of SIGIR*, pages 528–529, 2004.

[10] C. Hauff, L. Azzopardi, and D. Hiemstra. The combination and evaluation of query performance prediction methods. In *Proceedings of ECIR*, pages 301–312, 2009.

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

[12] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proceedings of CIKM*, pages 439–448, 2008.

[13] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *Proceedings of SPIRE*, pages 43–54, 2004.

[14] O. Kurland. The opposite of smoothing: A language model approach to ranking query-specific document clusters. In *Proceedings of SIGIR*, pages 171–178, 2008.

[15] J. D. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR*, pages 111–119, 2001.

[16] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of SIGIR*, pages 120–127, 2001.

[17] K.-S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of SIGIR*, pages 235–242, 2008.

[18] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of SIGIR*, pages 206–214, 1998.

[19] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications*, 2005.

[20] S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, pages 294–304, 1977.

[21] F. Scholer, H. E. Williams, and A. Turpin. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(7):637–650, 2004.

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

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

[24] N. Soskin, O. Kurland, and C. Domshlak. Navigating in the dark: Modeling uncertainty in ad hoc retrieval using multiple relevance models. In *Proceedings of ICTIR*, pages 79–91, 2009.

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

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

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

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

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

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

[31] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *ECIR*, pages 52–64, 2008.

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

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

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