Query-Performance Prediction and Cluster Ranking: Two Sides of the Same Coin

Oren Kurland
kurland@ie.technion.ac.il

Fiana Raiber
fiana@tx.technion.ac.il

Anna Shtok
annabel@tx.technion.ac.il

Faculty of Industrial Engineering and Management, Technion
Haifa 32000, Israel

ABSTRACT

We show that two tasks which were independently addressed in the information retrieval literature actually amount to the exact same task. The first is query-performance prediction; i.e., estimating the effectiveness of a search performed in response to a query in the absence of relevance judgments. The second task is cluster ranking, that is, ranking clusters of similar documents by their presumed effectiveness (i.e., relevance) with respect to the query. Furthermore, we show that several state-of-the-art methods that were independently devised for each of the two tasks are based on the same principles. Finally, we empirically demonstrate that using insights gained in work on query-performance prediction can help, in many cases, to improve the performance of a previously proposed cluster ranking method.

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

General Terms: Algorithms, Experimentation

Keywords: query-performance prediction, cluster ranking

1. INTRODUCTION

The observation that the effectiveness of retrieval methods can substantially vary from one query to another gave rise to a large body of work on query-performance prediction (**QPP**) [3]. The task is predicting the effectiveness of a retrieval performed in response to a query when relevance judgments are not available. Post-retrieval prediction methods, which are our focus here, analyze the *result list* of the documents most highly ranked.

Another task that also attracted much research attention is *cluster ranking* (**CR**) [20, 21, 16, 13, 14, 23, 9]; specifically, ranking clusters of similar documents based on their presumed relevance to the information need expressed by the query. Following the *cluster hypothesis* [32], several researchers showed that if the result list of top-retrieved documents is clustered, then among these clusters there are some that contain a very high percentage of relevant documents [12, 30, 14]. Furthermore, positioning the constituent documents of these clusters at the top ranks of the final result list was shown to yield retrieval performance that substantially transcends that of document-based retrieval [12, 30, 14].

The QPP and CR tasks mentioned above might seem at first glance to be quite different. Indeed, all previous reports of methods addressing these tasks have made no connections between the two. However, we show that the two tasks actually amount to the exact same task. That is, estimating the relevance of a *document set* to the information need expresses by a query.

Our second contribution is showing that some (state-of-the-art) methods that were independently devised for QPP and CR rely on the exact same principles; these principles directly touch on the underlying common grounds of the two tasks. Our third contribution is an operational one that emerges following the realization that the QPP and CR tasks are the same. We empirically show that a previously proposed cluster ranking method [21] can be improved in quite a few cases if the induced ranking is, in fact, reversed. The insight for this ranking reversal rises from considering a query-performance prediction method [27] that relies on the same core principle but which uses it in an opposite manner.

2. RELATED WORK

Post-retrieval query performance predictors can be roughly categorized into those that [3] (i) measure the clarity of the result list with respect to the corpus [5, 1, 4, 11], (ii) measure different notions of the robustness of the result list [34, 33, 36, 2, 37], and (iii) analyze retrieval scores in the result list [31, 8, 37, 27, 6, 7]. We show that several state-of-the-art predictors, specifically, some that are based on estimating result list robustness [37] and on analysis of retrieval scores [31, 37, 27, 24, 7] use the exact same principles utilized in some cluster ranking methods [21, 13, 14, 23].

The cluster hypothesis was used in some work to devise query-performance predictors [33, 8]. For example, a result list was assumed to be effective to the extent that similar documents are assigned with similar retrieval scores [8]. The coherence of the result list, as manifested in its clustering tendency, was also used for prediction [33]. Yet, connections to the cluster ranking task and methods devised for this task were not established in contrast to the work we present here.

Query performance predictors were employed upon clusters of similar documents for estimating the quality of relevance models [18] constructed from these clusters [28]. However, the connections between query performance prediction (QPP) and cluster ranking (CR) at the task level and at the method level, which we address here, were not discussed.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

3. ANALYSIS OF TASKS AND METHODS

Let $\mathcal{D}_{res}^{[q;k]}$ denote the result list of the k documents that are the highest ranked by a retrieval method employed in response to query q upon corpus $\mathcal{D}$. The query performance prediction (QPP) task [3] is estimating the effectiveness of $\mathcal{D}_{res}^{[q;k]}$ in lack of relevance judgments. We use $\mathcal{P}_{QPP}(q;\mathcal{D}_{res}^{[q;k]})$ to denote the effectiveness estimate assigned by a prediction method $\mathcal{P}_{QPP}$.

Let $Cl(\mathcal{D}_{res}^{[q;k]})$ denote the set of clusters of similar documents that are created from the result list $\mathcal{D}_{res}^{[q;k]}$ using *some* clustering algorithm. The cluster ranking (CR) task [20, 21, 16, 13, 14, 23, 9] is estimating the effectiveness (i.e., relevance) of each cluster c ($\in Cl(\mathcal{D}_{res}^{[q;k]})$) with respect to the information need expressed by q; $\mathcal{P}_{CR}(q;c)$ denotes the estimate. The estimate is used for ranking the clusters in $Cl(\mathcal{D}_{res}^{[q;k]})$. The cluster ranking can then be transformed into document ranking using various approaches [15, 20].

Thus, we get that while $\mathcal{P}_{QPP}(q;\mathcal{D}_{res}^{[q;k]})$ is an estimate for the effectiveness of the entire result list, $\mathcal{P}_{CR}(q;c)$ is an estimate for the effectiveness of a subset of documents, c, in $\mathcal{D}_{res}^{[q;k]}$. Hence, both the QPP and the CR tasks actually amount to the same task. Formally, let S be a set of documents. The task is estimating, in the absence of relevance judgments, the probability $p(S|q)$ that S contains information pertaining to the information need expressed by q. Indeed, $\mathcal{P}_{QPP}$ and $\mathcal{P}_{CR}$ are two such estimation methods. In Section 3.1 we show that several state-of-the-art query-performance prediction methods are based on the exact same principles that underlie some cluster ranking methods.

On some differences between QPP and CR methods. The estimates $\{\mathcal{P}_{CR}(q;c_i)\}$ assigned to the set of clusters c_i ($\in Cl(\mathcal{D}_{res}^{[q;k]})$), which were created for the same query q, are used for ranking the clusters relatively to each other. On the other hand, $\mathcal{P}_{QPP}$ is used to assign a prediction value to a single result list retrieved for a query. This difference means, for example, that the approach of using inter-cluster similarities for devising $\mathcal{P}_{CR}$ [13] is not applicable in the QPP case, since using inter-list similarities cannot convey much information for lists retrieved for different queries.

The prediction quality of $\mathcal{P}_{QPP}$ is measured by the correlation between the estimates it assigns to queries and the ground truth effectiveness for the queries [3]. Thus, the estimates are sometimes normalized for inter-query compatibility [37, 27], unless they are transformed to direct measures of retrieval effectiveness [10]. This normalization is not needed when using $\mathcal{P}_{CR}$ to rank clusters for the same query.

3.1 Analogous QPP and CR methods

In what follows we discuss the analogies between principles used for devising several QPP and CR methods. These shared principles rise, not surprisingly, due to the equivalence of the tasks that was discussed above.

3.1.1 The query-drift-based principle

The state-of-the-art query feedback (QF) QPP method [37] is based on the following principle. A query model is induced from the result list $\mathcal{D}_{res}^{[q;k]}$ and used for retrieval over the corpus. The number of highest ranked documents that are also among the highest ranked in $\mathcal{D}_{res}^{[q;k]}$ is the prediction value. The hypothesis is that for a result list $\mathcal{D}_{res}^{[q;k]}$ that does not manifest much "query drift" and hence is effective, the ranking induced over the corpus using a model of $\mathcal{D}_{res}^{[q;k]}$ will not drift much from the ranking used to create $\mathcal{D}_{res}^{[q;k]}$.

As it turns out, the same approach employed by QF is also implemented by an effective cluster ranking method ("self faithfulness") [14]. Specifically, a query model is induced from cluster c and used for retrieval over the corpus. The higher c's constituent documents are ranked by this retrieval, the less query drift c is presumed to manifest; accordingly, the higher is the estimate for the effectiveness of c.

3.1.2 Methods based on analysis of retrieval scores

We next discuss three QPP methods that are based on analysis of retrieval scores. These methods, as it turns out, were also employed by several CR approaches.

Maximal retrieval score. The maximal retrieval score in $\mathcal{D}_{res}^{[q;k]}$ was shown to be an effective QPP estimate [31]. Clusters were also ranked, in some work [19, 26, 23], based on the highest retrieval score of their constituent documents.

Average retrieval score. The state-of-the-art WIG QPP method [37] measures the difference between the average retrieval score in $\mathcal{D}_{res}^{[q;k]}$ and the retrieval score assigned to the corpus. The basic premise is that high retrieval scores in $\mathcal{D}_{res}^{[q;k]}$ attest to its effectiveness. The use of the corpus retrieval score is to ensure inter-query compatibility of prediction values as the retrieval scores themselves are not compatible across queries for various retrieval methods [37]. Indeed, some recent work [29] shows that WIG is highly effective if retrieval scores are normalized to begin with and the corpus retrieval score is not used; in this case, WIG amounts to using only the average retrieval score as the prediction value [29]. This prediction principle was also employed for the CR task. Clusters were ranked based on the mean retrieval score of their constituent documents [23, 13, 25].

Using the average retrieval score of a set of documents either for QPP or for CR, as discussed above, results in the following interesting interpretation when retrieval scores are induced using the language modeling framework. Specifically, the retrieval score assigned to document d in response to q is $\log \hat{p}(q|d)$, where $\hat{p}(q|d)$ is an estimate for the probability that q can be generated from a language model induced from d. Then, the average retrieval score in $\mathcal{D}_{res}^{[q;k]}$, which is used for QPP, is $\log \sqrt[k]{\prod_{d\in\mathcal{D}_{res}^{[q;k]}} \hat{p}(q|d)}$. This average score is equivalent to the retrieval score assigned to a geometric mean representation of $\mathcal{D}_{res}^{[q;k]}$ as was recently shown [27]. Interestingly, work on CR in the language modeling framework [23, 25] demonstrated the empirical merits of ranking a cluster using the exact same approach just mentioned for estimating QPP for $\mathcal{D}_{res}^{[q;k]}$; that is, by using the retrieval score assigned to the cluster's geometric-mean-based representation. Arguments based on information geometry were used to advocate a geometric-mean-based representation for sets of documents (e.g., in the context of CR) [25].

Variance of retrieval scores. The state-of-the art NQC QPP method [27], as well as other QPP approaches [24, 7], measure the standard deviation of retrieval scores in $\mathcal{D}_{res}^{[q;k]}$. The hypothesis is that *high* deviation corresponds to reduced query drift, and thereby, indicates improved retrieval effec-

corpus	# of docs	queries	disk(s)
AP	242,918	51-150	1-3
WSJ	173,252	151-200	1-2
SJMN	90,257	51-150	3
TREC8	528,155	401-450	4-5
WT10G	1,692,096	451-550	WT10g
GOV2	25,205,179	701-850	GOV2

Table 1: Data used for experiments.

tiveness [27]. The (formal) support for the hypothesis was based on the fact that the mean retrieval score in $\mathcal{D}_{res}^{[q;k]}$ is, for several retrieval methods, the retrieval score of a centroid-based representation of $\mathcal{D}_{res}^{[q;k]}$ that manifests query drift.

As it turns out, there is work on CR that ranks a cluster by the deviation of the retrieval scores of its constituent documents from the retrieval score of the cluster [21]. The cluster's retrieval score is based on the similarity between the query and a specific form of a centroid-based representation of the cluster, namely, the big document that results from concatenating its constituent documents[1]. However, in contrast to the case for QPP, clusters were ranked in *ascending* order of deviation [21], based on the premise that effective clusters are those for which the deviation is small.

Thus, while the exact same principle of using the variance of retrieval scores of documents in a set was employed for both the QPP and the CR tasks, it was used in a reversed manner. In Section 4 we show that ranking clusters in descending order of variance, a principle adapted from the work on QPP mentioned above [27, 24, 7], can outperform in many cases the use of an ascending order, which was originally proposed for ranking clusters [21].

4. EMPIRICAL EXPLORATION

Following the discussion in Section 3.1.2, the goal of the exploration to follow is contrasting two opposite hypotheses with regard to the connection between the variance of retrieval scores in a cluster and the cluster's (presumed) relevance to the query; that is, whether decreased variance as originally proposed [21], or increased variance as implied by work on QPP [27, 24, 7], attests to increased relevance.

4.1 Experimental setup

The TREC corpora and queries used for experiments are specified in Table 1. Titles of TREC topics served for queries. We applied Krovetz stemming to queries and documents; stopwords (on the INQUERY list) were removed only from queries. The Lemur/Indri toolkit was used for experiments.

Let x and y be a query, a document, or a cluster of documents. A cluster is represented by the concatenation of its constituent documents as is standard in work on cluster-based retrieval [20, 16, 22, 21, 13, 14]; the order of concatenation has no effect since we use unigram language models that assume term independence. Specifically, $p_x^{[\mu]}(\cdot)$ denotes the Dirichlet-smoothed unigram language model induced from x with the smoothing parameter μ. We use $Sim(x,y) \stackrel{def}{=} \exp\left(-CE\left(p_x^{[0]}(\cdot) \,\middle|\middle|\, p_y^{[\mu]}(\cdot)\right)\right)$ to denote the language-model-based similarity between x and y [15]; $\mu = 1000$ following previous recommendations [35].

We rank all documents d in the corpus in response to query q by $Sim(q,d)$. This initial ranking, henceforth referred to as **init. rank.**, amounts to the standard KL retrieval approach [17]. The $k = 50$ highest ranked documents, which form the result list $\mathcal{D}_{res}^{[q;k]}$, are then clustered using a simple nearest-neighbors-based clustering method [16, 13, 14, 23]. Specifically, for each d ($\in \mathcal{D}_{res}^{[q;k]}$) we create a cluster that contains d and the 4 documents d' ($\in \mathcal{D}_{res}^{[q;k]}$; $d' \neq d$) that yield the highest $Sim(d,d')$. Using such small overlapping nearest-neighbors-based clusters was shown to be highly effective for cluster-based retrieval [15, 16, 21, 13, 14, 23].

Let $\sum_{d\in c}(Sim(q,d) - Sim(q,c))^2$ be the sum of squared deviations of retrieval scores of documents in c from c's retrieval score. (All clusters contain 5 documents as noted above.) In what follows we use **DevAsc** to refer to a ranking of the clusters in ascending order of the deviation, which corresponds to previous work on cluster ranking [23]; **DevDesc** refers to a ranking in descending order of the deviation, which corresponds to the principle employed in QPP methods [27, 24, 7]. As reference comparisons we use **ArithMean** [13, 23] and **GeomMean** [23, 25] which refer to a ranking of the clusters based on the arithmetic mean, and geometric mean, respectively, of the retrieval scores ($Sim(q,d)$) of their constituent documents. To transform a ranking of clusters to a ranking of documents in $\mathcal{D}_{res}^{[q;k]}$ — i.e., to re-rank $\mathcal{D}_{res}^{[q;k]}$ — we follow previous work [20, 21, 23] and order the clusters top to bottom and then replace them by their constituent documents omitting repeats; documents within a cluster are ranked by their retrieval scores.

We report the MAP@50 (as the number of documents in $\mathcal{D}_{res}^{[q;k]}$ is 50) and precision at 5 (p@5) performance numbers. Note that p@5 corresponds to the percentage of relevant documents in the highest ranked cluster, as all clusters contain 5 documents. Statistically significant differences of performance are determined using the two-tailed paired t-test at the 95% confidence level.

4.2 Experimental results

Table 2 presents the experimental results. Except for GOV2[2], it is always the case that DevDesc outperforms DevAsc; in several cases, the performance differences are quite substantial and statistically significant. Hence, we see that the hypothesis postulated in work on QPP [27], about increased variance of retrieval scores within a set of documents attesting to improved effectiveness, can translate to a more effective cluster ranking method than that originally proposed (i.e., that based on the opposite hypothesis) [21]. Yet, both DevDesc and DevAsc are substantially outperformed by ArithMean and GeomMean and the initial ranking. Integrating methods based on the average and variance of retrieval scores is a future venue to explore.

5. CONCLUSIONS AND FUTURE WORK

We showed that the query-performance prediction (QPP) task and the cluster ranking (CR) task essentially amount to the exact same task. It is not surprising, therefore, that

[1]Using these clusters' retrieval scores, which reflect surface-level query similarities, for ranking clusters is known to yield relatively poor retrieval performance [16, 13, 14]. This finding potentially implies that this cluster representation also manifests query drift.

[2]Initial experiments with several settings for the ClueWeb collection resulted in findings similar to those observed for GOV2.

	WSJ		AP		SJMN		TREC8		WT10G		GOV2	
	MAP	p@5	MAP	p@5	MAP	p@5	MAP	p@5	MAP	p@5	MAP	p@5
init. rank.	**19.8**	51.2	9.2	46.1	14.3	**35.5**	16.9	**46.8**	13.8	33.4	11.7	56.2
ArithMean	19.5	**53.2**	**9.8**	48.1	**15.1**^{i}	35.3	16.6	44.8	14.0	**35.9**	**12.4**^{i}	**61.8**^{i}
GeomMean	19.3	51.6	**9.8**^{i}	**49.5**	15.0^{i}	34.9	**17.0**	**46.8**	**14.1**	35.5	12.3^{i}	60.4^{i}
DevAsc	12.6^{i}	29.6^{i}	7.6	35.2^{i}	10.1^{i}	24.0^{i}	8.4^{i}	22.8^{i}	6.5^{i}	21.2^{i}	10.3^{i}	47.2^{i}
DevDesc	13.6^{i}	30.0^{i}	8.2	37.6^{i}	13.3$_{a}$	27.7^{i}	14.7$_{a}$	38.8$_{a}$	10.8$^{i}_{a}$	22.3^{i}	10.1^{i}	41.8^{i}

Table 2: Performance of the cluster ranking methods. Boldface: best result in a column; 'i': a statistically significant difference with init. rank.; 'a': a statistically significant difference between DevDesc and DevAsc.

although the two tasks were independently addressed in previous work, several of the methods used to tackle them use the same principles. We discussed these shared principles for a variety of QPP and CR methods.

In addition, we empirically showed that using insights gained in some work on query-performance prediction can help to improve the retrieval performance of a cluster ranking method. For future work we intend to explore whether additional cluster ranking methods can be devised based on principles employed by query-performance prediction methods and vice versa.

Acknowledgments We thank the reviewers for their comments. This paper is based upon work supported in part by the Israel Science Foundation under grant no. 557/09, by IBM's Ph.D. fellowship and SUR award, by Google's and Yahoo!'s faculty research awards, and by Miriam and Aaron Gutwirth Memorial Fellowship. Any opinions, findings and conclusions or recommendations expressed here are the authors' and do not necessarily reflect those of the sponsors.

6. REFERENCES

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

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

[3] D. Carmel and E. Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool, 2010.

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

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

[6] R. Cummins. Predicting query performance directly from score distributions. In *Proc. of AIRS*, pages 315–326, 2011.

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

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

[9] N. Fuhr, M. Lechtenfeld, B. Stein, and T. Gollub. The optimum clustering framework: implementing the cluster hypothesis. *Informaltion Retrieval*, 15(2):93–115, 2012.

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

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

[12] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proc. of SIGIR*, pages 76–84, 1996.

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

[14] O. Kurland and C. Domshlak. A rank-aggregation approach to searching for optimal query-specific clusters. In *Proc. of SIGIR*, pages 547–554, 2008.

[15] O. Kurland and L. Lee. Corpus structure, language models, and ad hoc information retrieval. In *Proc. of SIGIR*, pages 194–201, 2004.

[16] O. Kurland and L. Lee. Respect my authority! HITS without hyperlinks utilizing cluster-based language models. In *Proc. of SIGIR*, pages 83–90, 2006.

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

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

[19] A. Leuski. Evaluating document clustering for interactive information retrieval. In *Proc. of CIKM*, pages 33–40, 2001.

[20] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proc. of SIGIR*, pages 186–193, 2004.

[21] X. Liu and W. B. Croft. Experiments on retrieval of optimal clusters. Technical Report IR-478, Center for Intelligent Information Retrieval (CIIR), University of Massachusetts, 2006.

[22] X. Liu and W. B. Croft. Representing clusters for retrieval. In *Proc. of SIGIR*, pages 671–672, 2006. Poster.

[23] X. Liu and W. B. Croft. Evaluating text representations for retrieval of the best group of documents. In *Proc. of ECIR*, pages 454–462, 2008.

[24] J. Pérez-Iglesias and L. Araujo. Standard deviation as a query hardness estimator. In *Proc. of SPIRE*, pages 207–212, 2010.

[25] J. Seo and W. B. Croft. Geometric representations for multiple documents. In *Proc. of SIGIR*, pages 251–258, 2010.

[26] J. G. Shanahan, J. Bennett, D. A. Evans, D. A. Hull, and J. Montgomery. Clairvoyance Corporation experiments in the TREC 2003. High accuracy retrieval from documents (HARD) track. In *Proc. of TREC-12*, pages 152–160, 2003.

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

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

[29] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems*, 30(2), 2012.

[30] A. Tombros, R. Villa, and C. van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management*, 38(4):559–582, 2002.

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

[32] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, second edition, 1979.

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

[34] 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 *Proc. of SIGIR*, pages 512–519, 2005.

[35] 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.

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

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