Predicting Query Performance by Query-Drift Estimation

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

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

Abstract. Predicting *query performance*, that is, the effectiveness of a search performed in response to a query, is a highly important and challenging problem. Our novel approach to addressing this challenge is based on estimating the potential amount of *query drift* in the result list, i.e., the presence (and dominance) of aspects or topics not related to the query in top-retrieved documents. We argue that query-drift can potentially be estimated by measuring the *diversity* (e.g., standard deviation) of the retrieval scores of these documents. Empirical evaluation demonstrates the prediction effectiveness of our approach for several retrieval models. Specifically, the prediction success is better, over most tested TREC corpora, than that of state-of-the-art prediction methods.

Keywords: query-performance prediction, query drift, score distribution.

1 Introduction

Many information retrieval (IR) systems suffer from a radical variance in performance when responding to users' queries. Even for systems that succeed very well on average, the quality of results returned for some of the queries is poor [1]. Thus, it is desirable that IR systems will be able to identify "difficult" queries in order to handle them properly.

We present a novel approach to query-performance prediction that is based on estimating the potential amount of *query drift* in the *result list* — the documents most highly ranked in response to the query. That is, the presence and dominance of non-query-related aspects or topics manifested in documents in the list.

As it turns out, we potentially do not need to directly identify query-drift, rather we can use a proxy for its estimation. Specifically, using insights from work on pseudo-feedback-based query expansion [2] we argue that high standard deviation of retrieval scores in the result list correlates with reduced query-drift, and consequently, with improved effectiveness. Empirical evaluation demonstrates the prediction-effectiveness of our predictor for several retrieval methods, specifically, with respect to that of state-of-the-art predictors.

L. Azzopardi et al. (Eds.): ICTIR 2009, LNCS 5766, pp. 305–312, 2009.

2 Related Work

Pre-retrieval query-performance prediction methods [3] analyze the query expression. However, the (short) query alone is often not expressive enough for reliable prediction [3]. The most effective prediction approaches employ post-retrieval analysis of the *result list* — the documents most highly ranked in response to the query. In what follows we discuss three such prominent paradigms.

The *clarity* prediction paradigm [4] is based on measuring the "focus" (clarity) of the result-list *with respect* to the corpus by computing different forms of their "distance" [5,6,7]. In Sect. 4 we show that our predictor is more effective than the clarity measure [4] over most tested collections.

Different notions of the *robustness* (e.g., with respect to document and query perturbations), and cohesion, of the result list [8,9,10,11,12] were shown to indicate query performance. Our proposed predictor, which measures the diversity of retrieval scores in the result list, can be thought of as a surrogate for estimating robustness with respect to document perturbations [9,10] — small *random* document perturbations are unlikely to result in major changes to documents' retrieval scores, and hence, are unlikely to significantly change the result list if retrieval scores are quite spread.

Work on analyzing *retrieval-scores distributions* to predict query performance showed that (i) the highest retrieval score [13], (ii) the difference between retrieval-scores produced in a query-independent and a query-dependent manner [14], and (iii) the extent to which similar documents receive similar retrieval scores [15] can indicate query performance. These techniques are complementary to ours. A state-of-the-art predictor, *Weighted Information Gain* (WIG) [12], measures the divergence between the mean retrieval score of top-ranked documents and that of the entire corpus. In contrast, our predictor essentially computes the divergence between the retrieval scores of top-ranked documents and that of a *pseudo non-relevant document* that exhibits a relatively high query-similarity. We demonstrate the merits of our predictor with respect to WIG in Sect. 4.

3 Prediction Framework

Let q, d, $\mathcal{D}$ and $\mathcal{M}$ be a query, document, corpus, and retrieval method, respectively. We use $Score(d)$ to denote the retrieval score assigned to d in response to q by $\mathcal{M}$. Our goal is to devise an estimate (predictor) for the *effectiveness* of the ranking induced by $\mathcal{M}$ over $\mathcal{D}$ in the *absence* of *relevance judgment* information. The estimated effectiveness is the *query performance* we attribute to $\mathcal{M}$ with respect to q. The methods we present utilize the result list $\mathcal{D}_q^{[k]}$ of the k documents that are the most highly ranked; k is a free parameter that is fixed to some value prior to retrieval (and prediction) time. As in many retrieval paradigms, we assume that $\mathcal{D}_q^{[k]}$ is composed of the documents that exhibit the highest (non-zero) surface-level similarity to q.

3.1 Estimating Query Drift

We refer to non-relevant documents in $\mathcal{D}_q^{[k]}$ as *misleaders* because they "mislead" the retrieval method into "believing" that they are relevant as they exhibit relatively high query-similarity. Misleaders are usually dominated by non query-related aspects (topics) that "drift away" from those represented by q [2].

As it turns out, we can potentially identify (at least) one (pseudo) misleader. Work on pseudo-feedback-based query expansion often uses a *centroid* representation, $Cent(\mathcal{D}_q^{[k]})$, of the list $\mathcal{D}_q^{[k]}$ as an expanded "query model" [16,17]. While using *only* the centroid yields poor retrieval performance [16,18,19], anchoring it to the query q via interpolation [18,19] yields improved performance, leading to the conclusion that the centroid manifests query drift [2]. Thus, $Cent(\mathcal{D}_q^{[k]})$ could be viewed as a prototypical misleader as it exhibits (some) similarity to the query by virtue of the way it is constructed (from documents in $\mathcal{D}_q^{[k]}$), but this similarity is dominated by non-query-related aspects that lead to query drift.

The degree of relevance of $Cent(\mathcal{D}_q^{[k]})$ to q is presumed by the retrieval method $\mathcal{M}$ to be correlated with its retrieval score, $\mu \stackrel{def}{=} Score(Cent(\mathcal{D}_q^{[k]}))$. In fact, we need not directly compute μ, because the mean retrieval score of documents in $\mathcal{D}_q^{[k]}$, $\hat{\mu} \stackrel{def}{=} \frac{1}{k}\sum_{d \in \mathcal{D}_q^{[k]}} Score(d)$, corresponds in several retrieval methods to the retrieval score, μ, of some centroid-based representation of $\mathcal{D}_q^{[k]}$. (We show that in Sect. 3.2). Thus, $\hat{\mu}$ represents the retrieval score of a prototypical misleader.

Estimates of Retrieval Effectiveness. Documents with retrieval scores (much) higher than $\hat{\mu}$, the score of a prototypical misleader, are potentially less probable to manifest query drift, and hence, be misleaders. Such documents could be considered as exhibiting positive ("+") *query-commitment* (QC). We therefore hypothesize that high divergence from $\hat{\mu}$ of the retrieval scores of these documents correlates with improved retrieval effectiveness. Since retrieval scores are query dependent, we normalize the divergence with respect to the retrieval score of a *general* prototypical non-relevant document, namely, the corpus. (We assume that the corpus can be represented as a single "pseudo" document, e.g., by using a centroid representation.) The resultant positive ("+") normalized-query-commitment (NQC) estimate is:

$$NQC_{+}(q, \mathcal{M}) \stackrel{def}{=} \frac{1}{Score(\mathcal{D})} \sqrt{\frac{1}{k} \sum_{d \in \mathcal{D}_q^{[k]} : Score(d) \geq \hat{\mu}} (Score(d) - \hat{\mu})^2} \ .$$

If we assume that there are only a few *relevant* documents in the corpus that yield "reasonable" query similarity, then a small overall number of documents exhibiting "reasonable" query-similarity can potentially indicate a small number of misleaders. The lower the retrieval score of a document is with respect to $\hat{\mu}$, the less we consider it to exhibit "reasonable" query-similarity (i.e., query-commitment). Hence, we hypothesize that the overall number of misleaders decreases (and hence, retrieval effectiveness increases) with increased (normalized) negative ("-") query-commitment measured by:

$$NQC_{-}(q,\mathcal{M}) \overset{def}{=} \frac{1}{Score(\mathcal{D})}\sqrt{\frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}:Score(d)<\hat{\mu}}(Score(d)-\hat{\mu})^2} \enspace .$$

We integrate the NQC_{+} and NQC_{-} measures to yield our main query-performance predictor, NQC, the (normalized) *standard deviation* of the retrieval scores in $\mathcal{D}_q^{[k]}$:

$$NQC(q,\mathcal{M}) \overset{def}{=} \sqrt{NQC_{+}(q,\mathcal{M})^2 + NQC_{-}(q,\mathcal{M})^2} = \frac{\sqrt{\frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}}(Score(d)-\hat{\mu})^2}}{Score(\mathcal{D})};$$

this measure has an appealing geometric interpretation exemplified in Fig. 1.

3.2 Use Case: Language Modeling Framework

The proposed performance-prediction measures can be employed with retrieval methods that estimate relevance based on surface-level document-query similarities. Here, we focus on the language modeling framework [21].

Let $p(w|d)$ be the probability assigned to term w by a (smoothed) unigram language model induced from document d. The commonly-used *query likelihood* (QL) retrieval method [20] scores document d in response to query $q = \{q_i\}$ by

$$Score_{QL}(d) = \sum_{q_i} \log p(q_i|d) \enspace . \tag{1}$$

To compute the corpus retrieval score $Score_{QL}(\mathcal{D})$, we treat $\mathcal{D}$ as the document that results from concatenating all documents in $\mathcal{D}$; the order of concatenation has no effect, since we use unigram language models.

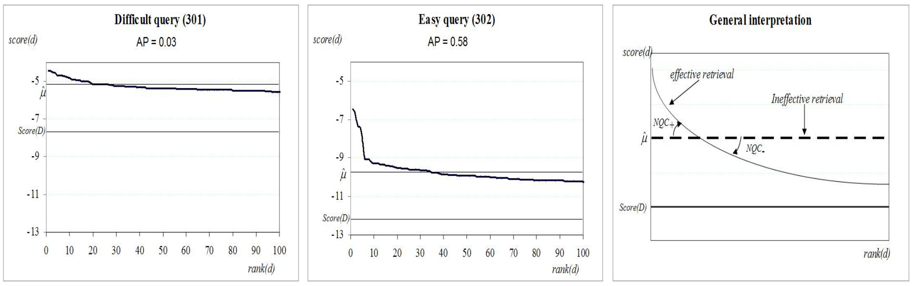

Fig. 1. Geometric interpretation of NQC. The two leftmost graphs present retrieval-scores curves for "difficult" and "easy" queries chosen by average-precision (AP) performance (query-likelihood model [20], ROBUST benchmark). Right: the shift between these two scenarios amounts to clockwise rotation of the retrieval-scores line.

The Centroid We stated in Sect. 3.1 that the mean retrieval score ($\hat{\mu}$) of documents in $\mathcal{D}_q^{[k]}$ corresponds to the retrieval score of a centroid-based representation of $\mathcal{D}_q^{[k]}$. We now demonstrate this correspondence for the query likelihood model.

Proposition 1. *The mean of the QL-retrieval-scores of documents in $\mathcal{D}_q^{[k]}$ is the QL score of a geometric-centroid language-model-based representation of $\mathcal{D}_q^{[k]}$.*

Proof. Let $\hat{\mu} = \frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}} Score_{QL}(d)$. By definition, $\hat{\mu} = \frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}}\sum_{q_i}\log p(q_i|d)$. We can re-arrange the summation and write $\hat{\mu} = \sum_{q_i}\frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}}\log p(q_i|d)$ $= \sum_{q_i}\log\sqrt[k]{\prod_{d\in\mathcal{D}_q^{[k]}} p(q_i|d)}$. We define $p(w|Cent(\mathcal{D}_q^{[k]})) \overset{def}{=} \sqrt[k]{\prod_{d\in\mathcal{D}_q^{[k]}} p(w|d)}$ — a language model (modulo normalization details) that corresponds to the geometric-centroid of language models of documents in $\mathcal{D}_q^{[k]}$; similar centroid was used in recent work on cluster-based retrieval [22]. By Eq. 1, $Score_{QL}(Cent(\mathcal{D}_q^{[k]})) = \hat{\mu}$.

The connection between the mean retrieval score of documents in $\mathcal{D}_q^{[k]}$ and the retrieval score of a centroid of $\mathcal{D}_q^{[k]}$ holds for other retrieval functions that are linear in features. For example, let $\boldsymbol{x}$ be the vector-space representation of text x. Now, if $Cent(\mathcal{D}_q^{[k]}) \overset{def}{=} \frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}} \boldsymbol{d}$ is the algebraic-centroid of $\mathcal{D}_q^{[k]}$, and the inner product is used as a retrieval function, then $\hat{\mu} \overset{def}{=} \frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}} <\boldsymbol{q},\boldsymbol{d}>=<\boldsymbol{q},\frac{1}{k}\sum_{d\in\mathcal{D}_q^{[k]}}\boldsymbol{d}>=<\boldsymbol{q}, Cent(\mathcal{D}_q^{[k]})>$.

4 Evaluation

We evaluate prediction quality by measuring Pearson's [7] and Kendall's$-\tau$ [1] correlation between the actual performance (average precision at cutoff 1000), and accordingly, induced ordering, of queries in a given set (as determined by using relevance judgments), and the values (and accordingly, induced ordering) assigned to these queries by a performance predictor. For both measures, higher correlation values indicate increased prediction quality. All correlation numbers that we report are statistically significant at a 95% confidence level.

4.1 Experimental Setup

We conducted experiments on TREC collections used in previous query-performance-prediction studies [8,10,23,15]: (i) WT10G (topics 451-550), (ii) ROBUST (disks 4&5-CR, topics 301-450, 601-700), (iii) TREC123 (disks 1&2, topics 51-200), (iv) TREC4 (disks 2&3, topics 201-250), and (v) TREC5 (disks 2&4, topics 251-300).

We use the titles of TREC topics for queries, except for the TREC4 case, where 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 toolkit (www.lemurproject.org), which was also used for retrieval. The query likelihood model [20] described in Sect. 3.2 served as the

Table 1. Comparison of NQC with state-of-the-art predictors. The best result per collection and evaluation measure is boldfaced.

		Pearson			Kendall's$-\tau$		
Corpus	#topics	Clarity	WIG	NQC	Clarity	WIG	NQC
WT10G	100	0.331	0.376	**0.527**	0.285	0.3	**0.303**
ROBUST	249	0.513	0.543	**0.563**	0.411	0.386	**0.419**
TREC123	150	0.462	**0.624**	0.376	0.351	**0.437**	0.273
TREC4	50	0.478	0.543	**0.556**	0.389	**0.489**	0.414
TREC5	50	**0.441**	0.297	0.431	**0.312**	0.253	0.3

retrieval model. (We used Dirichlet smoothing with the smoothing parameter set to 1000 following previous recommendations [24].)

We compare the prediction quality of NQC with that of two state-of-the-art predictors: Clarity [4] and WIG [12]. Clarity measures the KL divergence between a relevance language model (RM1) [17] constructed from the result-list $\mathcal{D}_q^{[k]}$ and the corpus model. We use Lemur's Clarity implementation.[1]

WIG was originally proposed in the MRF framework [25]. If term-dependencies are not used, MRF reduces to the query likelihood model with unigram language models. (It was noted that WIG is very effective with such implementation [23].) In this case, $WIG(q, QL) \overset{def}{=} \frac{1}{k}\sum_{d_i \in \mathcal{D}_q^{[k]}} \frac{1}{\sqrt{|q|}}(Score_{QL}(d_i) - Score_{QL}(\mathcal{D}))$.

Following experiments (results omitted due to space considerations) with different values of k, the number of documents in the result-list $\mathcal{D}_q^{[k]}$, we set its value to 100 for both our NQC measure and the Clarity predictor, and to 5 for WIG (which is in accordance with previous recommendations [23]).[2]

4.2 Experimental Results

The results in Table 1 show that NQC predicts query-performance very well over most collections. Specifically, NQC outperforms each of the baselines, WIG and Clarity, over three out of the five collections with respect to both evaluation measures. We attribute the relatively low prediction quality of NQC for TREC123 to the fact that TREC123 has extremely high average number of relevant documents per topic with respect to the other collections. Indeed, if NQC is employed for TREC123 over a much larger result-list, then prediction success can improve up to a Pearson correlation of 0.7; the same holds for WIG.

Table 2 shows that both NQC$_+$ and NQC$_-$ that are integrated by NQC are effective performance predictors. (Note the relatively high correlation numbers.) We also see that NQC is more effective than NQC$_+$ and NQC$_-$ over three collections with respect to both evaluation measures. These findings support the importance of considering both NQC$_+$ and NQC$_-$ as described in Sect. 3.

[1] We found that *clipping* RM1 so as to use 100 terms yields much better prediction-quality than using all terms as previously suggested [6].

[2] The prediction quality of (i) the Clarity measure is highly stable with respect to k, (ii) the WIG measure is in general optimal for low values of k (specifically, $k = 5$), and (iii) our NQC measure is in general quite stable for $k \in [80 - 500]$.

Table 2. Prediction quality of NQC sub-components: NQC$_+$ and NQC$_-$. Best result per collection and evaluation measure is boldfaced.

		Pearson			Kendall's$-\tau$		
Corpus	#topics	NQC$_+$	NQC$_-$	NQC	NQC$_+$	NQC$_-$	NQC
WT10G	100	**0.531**	0.479	0.527	**0.326**	0.274	0.303
ROBUST	249	0.560	0.519	**0.563**	0.416	0.397	**0.419**
TREC123	150	0.307	**0.48**	0.376	0.236	**0.336**	0.273
TREC4	50	0.526	**0.614**	0.556	0.388	**0.471**	0.414
TREC5	50	**0.491**	0.287	0.431	**0.333**	0.297	0.300

Table 3. Prediction quality (Pearson correlation) of NQC for the vector space model (with the cosine measure), Okapi, and the language model (LM) approach used so far

	Vector space	Okapi	LM
WT10G	0.407	0.311	0.527
ROBUST	0.535	0.603	0.563
TREC123	0.609	0.369	0.376
TREC4	0.664	0.578	0.556
TREC5	0.448	0.423	0.431

Table 3 presents the Pearson correlation for using NQC with the cosine measure in the vector space and with the Okapi BM25 method[3]. The (relatively high) correlation for both methods, which sometimes transcends that for the language-model approach used insofar, attests to the general effectiveness of NQC as a query-performance predictor.

5 Summary

We presented a novel approach to predicting query performance that is based on estimating the potential amount of *query drift* in the list of top-retrieved documents using the standard deviation of their retrieval scores. Empirical evaluation demonstrates the effectiveness of our predictor with several retrieval methods.

Acknowledgments. This paper is based upon work supported in part by Google's and IBM's faculty awards. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of the sponsors.

References

1. Voorhees, E.M.: Overview of the TREC 2004 Robust Retrieval Track. In: Proceedings of TREC-13 (2004)
2. Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: Proceedings of SIGIR, pp. 206–214 (1998)
3. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proceedings of CIKM, pp. 1419–1420 (2008)

[3] Since cosine scores are embedded in the unit sphere, normalization with the corpus retrieval-score is redundant (and, degrades prediction quality); it is therefore not employed. To avoid underflow issues caused by the document-length-normalization in Okapi, we use the centroid of all documents in the corpus to represent it. Lemur's implementation of both methods is used with default parameter settings.

4. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of SIGIR, pp. 299–306 (2002)
5. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness and selective application of query expansion. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004)
6. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Precision prediction based on ranked list coherence. Information Retrieval 9(6), 723–755 (2006)
7. Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: Proceedings of SIGIR, pp. 390–397 (2006)
8. Yom-Tov, E., Fine, S., Carmel, D., Darlow, A.: Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In: Proceedings of SIGIR, pp. 512–519 (2005)
9. Vinay, V., Cox, I.J., Milic-Frayling, N., Wood, K.R.: On ranking the effectiveness of searches. In: Proceedings of SIGIR, pp. 398–404 (2006)
10. Zhou, Y., Croft, W.B.: Ranking robustness: a novel framework to predict query performance. In: Proceedings of CIKM, pp. 567–574 (2006)
11. Aslam, J.A., Pavlu, V.: Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 198–209. Springer, Heidelberg (2007)
12. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of SIGIR, pp. 543–550 (2007)
13. Tomlinson, S.: Robust, Web and Terabyte Retrieval with Hummingbird Search Server at TREC 2004. In: Proceedings of TREC-13 (2004)
14. Bernstein, Y., Billerbeck, B., Garcia, S., Lester, N., Scholer, F., Zobel, J.: RMIT university at TREC 2005: Terabyte and robust track. In: Proceedings of TREC-14 (2005)
15. Diaz, F.: Performance prediction using spatial autocorrelation. In: Proceedings of SIGIR, pp. 583–590 (2007)
16. Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall, Englewood Cliffs (1971)
17. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of SIGIR, pp. 120–127 (2001)
18. Zhai, C., Lafferty, J.D.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of CIKM, pp. 403–410 (2001)
19. Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: UMASS at TREC 2004 — novelty and hard. In: Proceedings of TREC-13 (2004)
20. Song, F., Croft, W.B.: A general language model for information retrieval (poster abstract). In: Proceedings of SIGIR, pp. 279–280 (1999)
21. Croft, W.B., Lafferty, J. (eds.): Language Modeling for Information Retrieval. Information Retrieval Book Series, vol. 13. Kluwer, Dordrecht (2003)
22. Liu, X., Croft, W.B.: Evaluating text representations for retrieval of the best group of documents. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 454–462. Springer, Heidelberg (2008)
23. Zhou, Y.: Retrieval Performance Prediction and Document Quality. PhD thesis, University of Massachusetts (September 2007)
24. Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of SIGIR, pp. 334–342 (2001)
25. Metzler, D., Croft, W.B.: A Markov random field model for term dependencies. In: Proceedings of SIGIR, pp. 472–479 (2005)