Estimating Query Representativeness for Query-Performance Prediction

Mor Sondak
mor@tx.technion.ac.il

Anna Shtok
annabel@tx.technion.ac.il

Oren Kurland
kurland@ie.technion.ac.il

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

ABSTRACT

The query-performance prediction (QPP) task is estimating retrieval effectiveness with no relevance judgments. We present a novel probabilistic framework for QPP that gives rise to an important aspect that was not addressed in previous work; namely, the extent to which the query effectively represents the information need for retrieval. Accordingly, we devise a few query-representativeness measures that utilize relevance language models. Experiments show that integrating the most effective measures with state-of-the-art predictors in our framework often yields prediction quality that significantly transcends that of using the predictors alone.

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

General Terms: Algorithms, Experimentation

Keywords: query-performance prediction

1. INTRODUCTION

The task of estimating retrieval effectiveness in the absence of relevance judgments — a.k.a. query-performance prediction (QPP) — has attracted much research attention [2]. Interestingly, an important aspect of search effectiveness has been overlooked, or not explicitly modeled, in previously proposed prediction approaches; namely, the presumed extent to which the query effectively represents the underlying information need for retrieval.

Indeed, an information need can be represented by various queries which in turn might represent various information needs. Some of these queries might be more effective for retrieval over a given corpus than others for the information need at hand. Furthermore, relevance is determined with respect to the information need rather than with respect to the query. These basic observations underlie the development of the novel query-performance prediction framework that we present. A key component of the framework is the use of measures for the query representativeness of the information need. We propose several such measures that are based on using relevance language models [8].

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Empirical evaluation shows that integrating the most effective representativeness measures with state-of-the-art predictors in our framework yields prediction quality that often significantly transcends that of using these predictors alone.

2. RELATED WORK

Our query-performance prediction framework essentially generalizes a recently proposed framework [7], the basis of which was the estimation of the relevance of a result list to a query. Our framework relies on the basic definition of relevance with respect to the information need, and therefore accounts for the connection between the query and the information need. This connection was not (explicitly) addressed in previous work including [7]. For example, pre-retrieval predictors, which use only the query and corpus-based statistics, are mostly based on estimating the discriminative power of the query with respect to the corpus, but do not account for the query-information need connection.

Post-retrieval predictors analyze also the result list of top-retrieved documents [2]. Our framework provides formal grounds to integrating pre-retrieval, post-retrieval, and query-representativeness, which turn out to be three complementary aspects of the prediction task. Furthermore, we demonstrate the merits of integrating post-retrieval predictors with query representativeness measures in the framework.

The query representativeness measures that we devise utilize relevance language models [8]. Relevance models were used for other purposes in various predictors [3, 14, 5, 10]. We demonstrate the merits of integrating in our framework one such state-of-the-art predictor [14].

3. PREDICTION FRAMEWORK

Let q, d and $\mathcal{D}$ denote a query, a document, and a corpus of documents, respectively. The task we pursue is estimating the effectiveness of a retrieval performed over $\mathcal{D}$ in response to q when no relevance judgments are available [2] — i.e., query performance prediction (QPP).

Let I_q be the information need that q represents. Since relevance is determined with respect to I_q rather than with respect to q, the QPP task amounts, in probabilistic terms, to answering the following question:

"What is the probability that the result list $\mathcal{D}_{res}$, of the most highly ranked documents with respect to q, is relevant to I_q?"

Formally, the task is estimating

$$p(r|I_q, \mathcal{D}_{res}) = \frac{p(\mathcal{D}_{res}|I_q, r)p(r|I_q)}{p(\mathcal{D}_{res}|I_q)}, \quad (1)$$

where r is the relevance event and $p(r|I_q, \mathcal{D}_{res})$ is the probability that the result list $\mathcal{D}_{res}$ satisfies I_q.

Estimating $p(\mathcal{D}_{res}|I_q, r)$ is the (implicit) basis of many post-retrieval prediction methods, if q serves for I_q, as recently observed [7]. The denominator, $p(\mathcal{D}_{res}|I_q)$, is the probability that the result list $\mathcal{D}_{res}$ is retrieved using some representation of I_q regardless of relevance. If q is used for I_q, then the probability of retrieving $\mathcal{D}_{res}$ depends on the properties of the retrieval method employed. Accordingly, the denominator in Equation 1 can serve as a normalizer across different retrieval methods [7]. However, standard QPP evaluation [2] is based on estimating the retrieval effectiveness of a *fixed* retrieval method across different queries. Thus, the denominator in Equation 1 need not be computed for such evaluation, if q serves for I_q [7].

The (novel) task we focus on is estimating the probability $p(r|I_q)$ from Equation 1 that a relevance event happens for I_q. Obviously, the ability to satisfy I_q depends on the corpus $\mathcal{D}$; e.g., if there are no documents in $\mathcal{D}$ that pertain to I_q then the estimate should be zero. Furthermore, the satisfaction of I_q also depends on the query q used to represent it. Thus, the estimate for $p(r|I_q)$ can be approximated by:

$$\hat{p}(r|I_q) \approx \hat{p}(r|I_q, q, \mathcal{D}) = \frac{\hat{p}(q|I_q, \mathcal{D}, r)\hat{p}(r|I_q, \mathcal{D})}{\hat{p}(q|I_q, \mathcal{D})}, \quad (2)$$

where $\hat{p}(\cdot)$ is an estimate for $p(\cdot)$.

The estimate $\hat{p}(q|I_q, \mathcal{D})$ for the probability that q is chosen to represent I_q for retrieval over $\mathcal{D}$ can be used to account, for example, for personalization aspects. We leave this task for future work, and assume here a fixed user model, and accordingly, a fixed (across queries) $\hat{p}(q|I_q, \mathcal{D})$.

If we use q for I_q in the estimate $\hat{p}(r|I_q, \mathcal{D})$, we get the probabilistic basis for pre-retrieval prediction methods [6, 4]. These predictors implicitly estimate the probability for a relevance event using information induced from the query and the corpus, but not from the result list ($\mathcal{D}_{res}$).

The task left for completing the instantiation of Equation 2, and as a result that of Equation 1, is devising $\hat{p}(q|I_q, \mathcal{D}, r)$ — the estimate for the probability that q is the most likely query to effectively represent I_q for retrieval over $\mathcal{D}$.

3.1 Estimating query representativeness

The only signal about the information need I_q is the (short) query q. To induce a "richer" representation for I_q, we use the generative theory for relevance [8]. Specifically, we construct a (unigram) relevance language model R from documents in the corpus $\mathcal{D}$. (Details are provided in Section 4.1.) Then, estimating q's representativeness amounts to estimating the probability $p(q|R, \mathcal{D}, r)$ of generating q by R. Henceforth, we refer to such estimates as measures of q's "representativeness", denoted $X(q; R)$.

We assume, as in the original relevance model's formulation [8], that q's terms ($\{q_i\}$) are generated independently by R: $\hat{p}(q|R, \mathcal{D}, r) \stackrel{def}{=} \prod_{q_i} p(q_i|R)$; $p(q_i|R)$ is the probability assigned to q_i by R. To prevent the query-length bias, we use the geometric mean of the generation probabilities which results in the **GEO** measure:

$$GEO(q; R) \stackrel{def}{=} \sqrt[|q|]{\prod_{q_i \in q} p(q_i|R)};$$

$|q|$ is the number of terms in q.

We also consider the arithmetic mean of the generation probabilities, **ARITH**, as a representativeness measure:

$$ARITH(q; R) \stackrel{def}{=} \frac{1}{|q|} \sum_{q_i \in q} \hat{p}(q_i|R).$$

For comparison purposes, we study the min and max aggregators of the generation probabilities:

$$MIN(q; R) \stackrel{def}{=} \min_{q_i \in q} p(q_i|R);$$

$$MAX(q; R) \stackrel{def}{=} \max_{q_i \in q} p(q_i|R).$$

Another measure that we consider is the *weighted* entropy of R, where q's terms are assigned with a unit weight and all other terms in the vocabulary are assigned a zero weight:

$$ENT(q; R) \stackrel{def}{=} -\sum_{q_i \in q} \hat{p}(q_i|R) \log \hat{p}(q_i|R).$$

The underlying assumption is that high entropy, which implies to a relatively uniform importance assigned to q's terms by R, is indicative of effective representation by q. Indeed, too little emphasis on some query aspects was identified as a major cause for retrieval failures [1].

4. EVALUATION

We next present an evaluation of our query-performance prediction (QPP) framework. We begin by describing the experimental setup in Section 4.1. In Section 4.2.1 we focus on using the query-representativeness measures. To that end, we use an oracle-based experiment where the relevance model is constructed only from relevant documents. In Section 4.2.2 we study the integration of the representativeness measures with post-retrieval predictors in our framework.

Collection	Data	# of Docs	Topics	Avg. query length
TREC12	Disks 1,2	741,854	51-200	3.52
TREC5	Disks 2,4	524,929	251-300	3.08
ROBUST	Disks 4,5-CR	528,155	301-450, 601-700	2.64
WT10G	WT10g	1,692,096	451-550	2.66

Table 1: TREC datasets used for experiments.

4.1 Experimental setup

Table 1 presents the TREC datasets used for experiments. TREC12, TREC5 and ROBUST are composed (mostly) of newswire documents, while WT10G is a noisy Web collection. Titles of TREC topics serve for queries. Documents and queries were stemmed with the Krovetz stemmer and stopwords (on the INQUERY list) were removed. The Indri toolkit (www.lemurproject.org) was used for experiments.

Following common practice [2], prediction quality is measured by the Pearson correlation between the true average precision (AP@1000) for the queries, as determined using the relevance judgments in the qrels files, and the values assigned to these queries by a predictor.

The query likelihood method [11] serves for the retrieval method, the effectiveness of which we predict. Document d's retrieval score is the log query likelihood: $\log \prod_{q_i \in q} p(q_i|d)$; $p(q_i|d)$ is the probability assigned to q_i by a Dirichlet

smoothed unigram language model induced from d with the smoothing parameter set to 1000 [13].

We use relevance model #1 (RM1) [8] in the query representativeness measures: $p(w|R) \stackrel{def}{=} \sum_{d \in S} p(w|d)p(d|q)$; S is a set of documents; $p(w|d)$ is the maximum likelihood estimate of term w with respect to d; $p(d|q)$ is (i) $\frac{1}{|S|}$ when S is a set of relevant documents as is the case in Section 4.2.1; and, (ii) d's normalized query likelihood: $\frac{p(q|d)}{\sum_{d' \in S} p(q|d')}$, when S is the set of all documents in the corpus that contain at least one query term as is the case in Section 4.2.2. No term clipping was employed for RM1.

4.2 Experimental results

4.2.1 The query-representativeness measures

The query-representativeness measures play an important role in our QPP framework, and are novel to this study. Thus, we first perform a *controlled* experiment to explore the potential extent to which these measures can attest to query performance. To that end, we let the measures use a relevance model of a (very) high quality. Specifically, RM1 is constructed from *all* relevant documents in the qrels files as described in Section 4.1. Table 2 presents the prediction quality of using the representativeness measures *by themselves* as query-performance predictors. As can be seen, the prediction quality numbers are in many cases quite high. All these numbers — which are Pearson correlations — are different than zero to a statistically significant degree according to the two-tailed t-test with a 95% confidence level.

We can also see in Table 2 that GEO is the most effective measure except for TREC5. ARITH and MIN are also quite effective, although often less than GEO. ENT is highly effective for TREC5 and WT10G but much less effective for TREC12 and ROBUST. The MAX measure is evidently less effective than the others, except for TREC5. All in all, we see that different statistics of the generation probabilities assigned by the relevance model to the query terms can serve as effective query representativeness measures for query-performance prediction.

	TREC12	TREC5	ROBUST	WT10G
GEO	**0.588**	0.295	**0.376**	**0.414**
ARITH	0.457^{g}	0.398	0.274	0.356
MIN	0.523^{g}	0.334	0.328	0.373
MAX	$0.216^{g,a}_{n}$	0.351	$0.153^{g,a}_{n}$	$0.24^{g,a}$
ENT	$0.251^{g,a}_{n}$	**0.526**$_{x}$	0.222^{g}_{x}	0.375_{x}

Table 2: Using the representativeness measures by themselves as query-performance predictors with RM1 constructed from relevant documents. Boldface: the best result in a column. 'g', 'a', 'n', 'x' and 'e' mark statistically significant differences in correlation [12] with GEO, ARITH, MIN, MAX, and ENT, respectively.

4.2.2 Integrating query-representativeness measures with post-retrieval predictors

Query-representativeness measures are one component of our QPP framework. Other important components are post-retrieval and pre-retrieval prediction as described in Section 3. Since (i) the query representativeness measures constitute a novel contribution of this paper, (ii) the merits of the integration of post-retrieval and pre-retrieval prediction were already demonstrated in previous work [7], and, (iii) post-retrieval predictors often yield prediction quality that is substantially better than that of pre-retrieval predictors [2], we focus on the integration of the representativeness measures with the post-retrieval predictors in our framework. The integration is performed using Equations 1 and 2. In contrast to the case in Section 4.2.1, we use the standard practical QPP setting; that is, no relevance judgments are available. The relevance model used by the query-representativeness measures is constructed as described in Section 4.1 from all the documents in the corpus that contain at least one query term. Using only top-retrieved documents for constructing the relevance model resulted in inferior prediction quality.

Three state-of-the-art post-retrieval predictors, NQC [9], WIG [14] and QF [14], are used. As these predictors incorporate free parameters, we apply a train-test approach to set the values of the parameters. Since Pearson correlation is the evaluation metric for prediction quality, there should be as many queries as possible in both the train and test sets. Thus, each query set is randomly spit into two folds (train and test) of equal size. We use 40 such splits and report the average prediction quality over the test folds. For each split, we set the free-parameter values of each predictor by maximizing prediction quality over the train fold.

NQC and WIG analyze the retrieval scores of top-retrieved documents, the number of which is set to values in $\{5, 10, 50, 100, 500, 1000\}$. QF incorporates three parameters. The number of top-retrieved documents used to construct the relevance model (RM1) utilized by QF is selected from $\{5, 10, 25, 50, 75, 100, 200, 500, 700, 1000\}$ and the number of terms used by this RM1 is set to 100 following previous recommendations [10]. The cuttoff used by the overlap-based similarity measure in QF is set to values in $\{5, 10, 50, 100, 500, 1000\}$.

In Table 3 we present the average (over the 40 test folds) prediction quality of using the query-representativeness measures alone; using the post-retrieval predictors alone; and, integrating the representativeness measures with the post-retrieval predictors in our framework. Although the query-representativeness measures do not incorporate free parameters, we report their prediction quality when used alone using the same test splits. When the measures are integrated with the post-retrieval predictors, the free-parameters of the integration are those of the post-retrieval predictors. In this case, the parameters are tuned by optimizing the prediction quality of the integration over the train folds, as is the case when using the post-retrieval predictors alone. Differences of prediction quality (i.e., Pearson correlations) are tested for statistical significance using the two tailed paired t-test computed over the 40 splits with a 95% confidence level.[1]

We first see in Table 3 — specifically, by referring to the underlined numbers — that the best prediction quality for the majority of the corpora is attained by integrating a representativeness measure with a post-retrieval predictor.

Further exploration of Table 3 reveals the following. The GEO and ARITH measures are effective — specifically, in comparison to the other representativeness measures which is reminiscent of the case in Table 2 — both as stand-alone

[1]Note that the numbers in Table 2 are not comparable to those in Table 3. This is because the latter presents averages over the train-test splits while the former is based on using the all queries for the test set. Furthermore, as noted above, the relevance models used for the representativeness measures are constructed using different sets of documents.

	TREC12	TREC5	ROBUST	WT10G
GEO	0.642	0.380	0.407	0.317
ARITH	0.635	0.435	0.419	0.287
MIN	0.583	0.272	0.352	0.305
MAX	0.465	0.396	0.366	0.210
ENT	0.277	0.381	0.309	0.256
NQC	0.666	0.289	0.506	**0.422**
NQC∧GEO	0.705^{p}_{q}	0.303_{q}	0.520^{p}_{q}	0.411_{q}
NQC∧ARITH	$\mathbf{0.713}^{p}_{q}$	$\mathbf{0.323}_{q}$	$\mathbf{0.534}^{p}_{q}$	0.375^{p}_{q}
NQC∧MIN	0.663_{q}	0.272	0.456^{p}_{q}	0.405_{q}
NQC∧MAX	0.672_{q}	0.303_{q}	0.508_{q}	0.309^{p}_{q}
NQC∧ENT	0.598^{p}_{q}	0.299_{q}	0.491^{p}_{q}	0.421_{q}
WIG	0.665	0.250	**0.514**	**0.393**
WIG∧GEO	0.688^{p}_{q}	0.371^{p}	0.467^{p}_{q}	0.316^{p}
WIG∧ARITH	$\mathbf{0.689}^{p}_{q}$	$\mathbf{0.373}^{p}_{q}$	0.480^{p}_{q}	0.285^{p}
WIG∧MIN	0.645^{p}_{q}	0.319^{p}_{q}	0.416^{p}_{q}	0.313^{p}
WIG∧MAX	0.604^{p}_{q}	0.338^{p}_{q}	0.447^{p}_{q}	0.240^{p}_{q}
WIG∧ENT	0.462^{p}_{q}	0.334^{p}_{q}	0.409^{p}_{q}	0.333^{p}_{q}
QF	0.673	0.313	0.500	0.267
QF∧GEO	$\mathbf{0.723}^{p}_{q}$	0.378^{p}	0.518^{p}_{q}	$\mathbf{0.372}^{p}_{q}$
QF∧ARITH	0.711^{p}_{q}	0.429^{p}	$\mathbf{0.528}^{p}_{q}$	0.353^{p}_{q}
QF∧MIN	0.692_{q}	0.314_{q}	0.471^{p}_{q}	0.361^{p}_{q}
QF∧MAX	0.608^{p}_{q}	$\mathbf{0.438}^{p}_{q}$	0.504_{q}	0.272_{q}
QF∧ENT	0.498^{p}_{q}	0.393^{p}	0.485^{p}_{q}	0.307^{p}_{q}

Table 3: Average prediction quality over the test folds of the query-representativeness measures, post-retrieval predictors, and their integration (marked with ∧). Boldface: the best result per corpus and a post-retrieval block; underline: the best result in a column. 'q' and 'p' mark statistically significant differences with using the query-representativeness measure alone and the post-retrieval predictor alone, respectively.

predictors and when integrated with the post-retrieval predictors. Indeed, integrating each of GEO and ARITH with a post-retrieval predictor yields prediction quality that transcends that of using the post-retrieval predictor alone in 9 out of the 12 relevant comparisons (three post-retrieval predictors and four corpora); many of these improvements are substantial and statistically significant.

These findings, as those presented above, attest to the merits of our QPP framework that integrates two different, and evidently complementary, aspects of prediction; namely, post-retrieval analysis of the result list and query-representativeness estimation.[2]

In comparing the prediction quality numbers in Table 3 for the three post-retrieval predictors we make the following observation. For QF and WIG the integration with the query-representativeness measures yields the highest and lowest number, respectively, of cases of improvement over using the post-retrieval predictor alone.

[2]It is not a surprise, therefore, that the post-retrieval predictors when used alone outperform in most cases the representativeness measures when used alone. This is because the post-retrieval predictors analyze the result list, while the representativeness measures do not. For TREC5, however, the reverse holds. Presumably, this is because there are only 50 queries for TREC5, while for all other corpora there are at least 100 queries. A relatively small query set makes it difficult to learn the free-parameter values of the post-retrieval predictors, while representativeness measures do not incorporate free parameters.

5. CONCLUSIONS AND FUTURE WORK

We presented a novel probabilistic framework for the query-performance prediction task. The framework gives rise to an important aspect that was not addressed in previous work: the extent to which the query effectively represents the underlying information need for retrieval. We devised query-representativeness measures using relevance language models. Empirical evaluation showed that integrating the most effective measures with state-of-the-art post-retrieval predictors in our framework often yields prediction quality that significantly transcends that of using the predictors alone.

Devising additional query-representativeness measures, and integrating pre-retrieval predictors with post-retrieval predictors and query-representativeness measures in our framework, are future venues to explore.

6. ACKNOWLEDGMENTS

We thank the reviewers for their comments. This work has been supported in part by the Israel Science Foundation under grant no. 433/12 and by a Google faculty research award. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

7. REFERENCES

[1] C. Buckley. Why current IR engines fail. In *Proceedings of SIGIR*, pages 584–585, 2004. Poster.

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

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

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

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

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

[7] O. Kurland, A. Shtok, S. Hummel, F. Raiber, D. Carmel, and O. Rom. Back to the roots: A probabilistic framework for query-performance prediction. In *Proceedings of CIKM*, pages 823–832, 2012.

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

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

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

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

[12] J. H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251, 1980.

[13] 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.

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