Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence

Ying Zhao, Falk Scholer, and Yohannes Tsegay

School of Computer Science and IT, RMIT University, GPO Box 2476v, Melbourne, Australia
{ying.zhao,falk.scholer,yohannes.tsegay}@rmit.edu.au

Abstract. Query performance prediction aims to estimate the quality of answers that a search system will return in response to a particular query. In this paper we propose a new family of pre-retrieval predictors based on information at both the collection and document level. Pre-retrieval predictors are important because they can be calculated from information that is available at indexing time; they are therefore more efficient than predictors that incorporate information obtained from actual search results. Experimental evaluation of our approach shows that the new predictors give more consistent performance than previously proposed pre-retrieval methods across a variety of data types and search tasks.

1 Introduction

As the amount of electronic data continues to grow, the availability of effective information retrieval systems is essential. Despite a continuing increase in the average performance of information retrieval systems, the ability of search systems to find useful answers for individual queries still shows a great deal of variation [14].

An analysis of the chief causes of failure of current information retrieval (IR) systems concluded that, if a search system could identify in advance the problem associated with a particular search request, then the selective application of different retrieval technologies should be able to improve results for the majority of problem searches [8]. The ability to predict the performance of a query in advance would enable search systems to respond more intelligently to user requests. For example, if a user query is predicted to perform poorly, the user could be asked to supply additional information to improve the current search request. Alternatively, a search system could selectively apply different techniques in response to difficult and easy queries, for example the selective application of different retrieval models, or automatic relevance feedback.

Query performance prediction is the problem of trying to identify, without user intervention, whether a search request is likely to return a useful set of answers. The importance of the query difficulty prediction problem has been highlighted in the IR community in recent years; the Text REtrieval Conference (TREC) Robust tracks in 2004 and 2005 included an explicit query difficulty prediction task [14], and prediction has been the subject of specific workshops [4]. Despite this recent growth in attention, the prediction of query difficulty is an open research problem.

In this paper, we propose new pre-retrieval predictors based on two sources of information: the similarity between a query and the underlying collection; and the variability with which query terms occur in documents. Pre-retrieval predictors make use of

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 52–64, 2008.

information that is available at indexing-time, such as collection and term distribution statistics. They can be calculated without needing to first evaluate the query and obtain an answer set, and are therefore more efficient than post-retrieval predictors.

We evaluate the performance of our new predictors by considering the correlation between the predictors and the actual performance of the retrieval system on each query, as measured by mean average precision. We conduct experiments on both newswire and web data, and across informational and navigational search tasks. The results demonstrate that these new predictors show more consistent performance than previously published pre-retrieval predictor baselines across data collections and search tasks.

2 Background

Many different approaches for the prediction of query performance have been proposed. These can be divided into three broad categories: pre-retrieval predictors, post-retrieval predictors, and learning predictors. In this paper we focus on pre-retrieval predictors; the background section therefore concentrates on previous work in this area. We also provide brief descriptions of the other families of predictors for completeness.

Pre-retrieval predictors can be calculated from features of the query or collection, without requiring the search system to evaluate the query itself. The information that these predictors use is available at indexing-time; they are therefore efficient, and impose a minimal overhead on the retrieval system. Pre-retrieval predictors generally make use of evidence based on term distribution statistics such as the inverse document frequency, inverse collection term frequency, or the length of a query.

A range of pre-retrieval predictors were proposed and evaluated by He and Ounis [9]. Their experimental results showed the two best-performing predictors to be the average inverse collection term frequency (AvICTF), and the simplified clarity score (SCS). In their approach, the SCS is obtained by calculating the Kullback-Leibler divergence between a query model and a collection model. We use AvICTF and SCS as baselines in our experiments, and these approaches are explained in detail in Section 4.

Scholer et al. [11] describe results based on using the inverse document frequency (IDF) to predict query performance. They find that using the maximum IDF of any term in a query gives the best correlation on the TREC web data. We present results using the maximum IDF (MaxIDF) as a baseline in our experiments.

Post-retrieval predictors use evidence that is obtained from the actual evaluation of the underlying search query. These predictors can leverage information about the cohesiveness of search results, and can therefore show high levels of effectiveness. However, for the same reason they are less efficient: the search system must first process the query and generate an answer set, and the answer set itself is then usually the subject of further analysis, which may involve fetching and processing individual documents. This can impose a substantial overhead on a retrieval system.

Cronen-Townsend et al. [6] proposed a post-retrieval predictor based on language models: they calculate the divergence between a statistical model of the language used in the overall collection and a model of the language used in the query, to obtain an estimate of the ambiguity of the query. Unlike the simplified clarity score pre-retrieval predictor discussed previously, this approach estimates the query language model from

the documents that are returned in the answer set of a retrieval system. The approach was demonstrated to be highly effective on newswire data. Post-retrieval predictors for web data were developed by Zhou and Croft [19], who use a weighted information gain approach that shows a high correlation with system performance for both navigational and informational web search tasks. Other post-retrieval predictors have considered factors such as the variability of similarity scores; for example, Kwok et al. divide a search results list into groups of adjacent documents and compare the similarity among these [10]. Zhou and Croft [18] introduced ranking robustness scores to predict query performance, by proposing noise channel from information theory. This approach has shown higher effectiveness than the clarity score [6].

Learning predictors incorporate a variety of statistical regression [10] or machine learning algorithms [16], such as neural networks or support vector machines, to train on labeled examples of easy and difficult queries. The learned estimator is then used to predict the difficulty of previously unseen queries. While some learning predictors have shown high levels of correlation with system performance, this family of predictors requires suitable training data to be available; a corresponding overhead is therefore incurred during the training stage.

3 Pre-retrieval Prediction of Query Performance

In this section, we present several predictors of query performance. The predictors are concerned with *pre-retrieval* prediction. The information required by such prediction is obtained from various collection, document and term occurrence statistics. These are all obtained at indexing time, and can be efficiently fetched from inverted index structures that are widely used in information retrieval [20]. The computation of these predictors can therefore be carried out prior to query evaluation. This has significant advantages in terms of simplicity and efficiency, factors whose importance increases as the size of collections continues to grow. We propose two broad classes of pre-retrieval predictors: first, predictors that are based on the similarity between queries and the collection; and second, predictors that are based on the variability of how query terms are distributed in the collection, by exploring the in-document statistics for the input queries.

3.1 The Similarity between a Query and a Collection

While many retrieval models have been proposed in the IR literature, at their core these rely to a greater or lesser extent on various collection, document and term distribution statistics, which are used as sources of evidence for relevance [20]. In particular, two of the most commonly-used sources of evidence are the term frequency (TF), and the inverse document frequency (IDF). The former represents the intuitive concept that the higher the frequency with which a query term occurs in a document, the more likely it is that the document is about that term. The latter is used to discriminate between query terms that are highly selective (they occur in fewer documents, and therefore have a high IDF), and those that are less selective (occurring in many documents, and therefore having a lower IDF).

An intuitive geometric interpretation of the similarity between documents and queries is provided by the vector space model [15]. Here, queries and documents are represented

as vectors in n-dimensional space, where n is the number of unique terms that occur in a collection. The estimated similarity between a document and a given query is defined as the closeness of a document vector and a query vector, where the closeness is measured by degree of the angle between these two vectors. Documents whose vectors are closely aligned with a query vector are considered to have a high similarity, and are likely to be on similar subjects or topics to the query. In a similar manner, the collection itself can also be represented as an n-dimensional vector, and the similarity of a query to the collection as a whole can be calculated. Those query vectors which are more similar to the collection vector are considered to be easier to evaluate—the evidence suggests that the collection contains documents that are similar to the query. Such queries and therefore more likely to have higher performance.

Predictor 1 (SCQ): Given a query $Q(t_1, \ldots, t_n)$, the similarity score between the query and collection can be defined as:

$$SCQ = \sum_{t \in Q} (1 + \ln(f_{c,t})) \times \ln\left(1 + \frac{N}{f_t}\right) \tag{1}$$

where N is the total number of documents in the collection C, $f_{c,t}$ is the frequency of term t in the collection, and f_t is the number of documents that contain term t. In this version of the metric we simply sum up the contributions of the collection term frequencies and inverse document frequencies of all query terms. Such a process will be biased towards longer queries. We therefore calculate a normalised metric:

Predictor 2 (NSCQ): We define the NSCQ score as the SCQ score, divided by the query length, where only terms in the collection vocabulary are considered:

$$NSCQ = \frac{SCQ}{|Q|_{t \in \mathcal{V}}} \tag{2}$$

where $\mathcal{V}$ is the vocabulary (all unique terms in the collection).

Instead of normalising by query length, an alternative approach is to choose the maximum SCQ score of any query term. The intuition behind this approach is that, since web search queries tend to be short, if at least one of the terms has a high score then the query as a whole can be expected to perform well:

Predictor 3 (MaxSCQ): MaxSCQ considers that the performance of a query is determined by the "best" term in the query—the term that has the highest SCQ score:

$$MaxSCQ = max\left[\forall_{t \in Q} (1 + \ln(f_{c,t})) \times \ln\left(1 + \frac{N}{f_t}\right)\right] \tag{3}$$

We note that it is not rare to encounter a query term t that is missing in $\mathcal{V}$. For simplicity, we assign such terms with 0 scores in a query.

3.2 Variability Score

The previous group of predictors explored the surface features of a collection, such as the frequency with which terms occur in the collection as a whole. In this section, we

propose alternative predictors which are concerned with the distribution of query terms over the collection, taking into account the variability of term-occurrences within individual documents. The standard deviation is a statistical measure of dispersion, reflecting how widely spread the values in a data set are around the mean: if the data points are close to the mean, then the standard deviation is small, while a wide dispersion of data points leads to a high standard deviation.

We incorporate such a mechanism in the prediction task, estimating the standard deviation of term occurrence weights across the collection. We hypothesise that if the standard deviation of term weights is high, then the term should be easier to evaluate. This is because the retrieval system will be able to differentiate with higher confidence between answer documents. Conversely, a low standard deviation would indicate that the system does not have much evidence on which to choose the best answer documents; such a query would therefore be predicted to perform less well.

In general, each query term t can be assigned with a weight value $w_{d,t}$ if it occurs in document d. From all the documents that contain term t in a collection, the distribution of t can then be estimated. We use a simple $TF.IDF$ approach to compute the term weight, $w_{d,t}$, within a document [20]:

$$w_{d,t} = 1 + \ln(f_{d,t}) \times \ln\left(1 + \frac{N}{f_t}\right)$$

Again, for query terms that are missing in $\mathcal{V}$, we assign $w_{d,t} = 0$.

Predictor 4 (σ_1): Given a query Q $(t_1, \ldots, t_n)$, the basic variability score is defined as the sum of the deviations:

$$\sigma_1 = \sum_{t \in Q} \sqrt{\frac{1}{f_t} \sum_{d \in \mathcal{D}_t} (w_{d,t} - \overline{w}_t)^2} \qquad \text{where} \tag{4}$$

$$\overline{w}_t = \frac{\sum_{d \in \mathcal{D}_t} w_{d,t}}{|D_t|}$$

where $f_{d,t}$ is the frequency of term t in document d, and $\mathcal{D}_t$ is the set of documents that contain query term t. This predictor sums the deviations across query terms, and thus reflects the variability of the query as a whole. An alternative is to use a metric normalised for query length:

Predictor 5 (σ_2): Normalising the σ_1 score by the number of valid query terms, we obtain the σ_2 score for a given query Q:

$$\sigma_2 = \frac{\sigma_1}{|Q|_{t \in \mathcal{V}}} \tag{5}$$

As for the SCQ score, a further intuitive alternative to simply normalising by query length is to select the largest variability score of any query term:

Predictor 6 (σ_3): σ_3 estimates the performance of a query based on the maximum deviation from the mean that is observed for any one query term:

$$\sigma_3 = max\left[\forall_{t \in Q} \sqrt{\frac{1}{f_t} \sum_{d \in \mathcal{D}_t} (w_{d,t} - \overline{w}_t)^2}\right] \tag{6}$$

where $\overline{w}_{d,t}$ is defined as in **Predictor 4**.

The proposed SCQ score and variability score predictors are based on different sources of evidence—the former considers evidence at the high collection level, while the latter is based on the distribution of terms across documents. Combining both sources of information could therefore lead to additional prediction accuracy:

Predictor 7 ($joint$): For each query term t in query Q, this predictor combines both the MaxSCQ and σ_1 scores. We use a simple linear interpolation approach to combine the two scores; the computation is defined as:

$$joint = \alpha \cdot MaxSCQ + (1 - \alpha) \cdot \sigma_1 \tag{7}$$

where α is a parameter that determines the weight given to the SCQ and variability score components. The parameter is set using separate training data (for example, for the WT10g collection below, the parameter is trained using topics 501–550 for experiments on topics 451–500, and vice-versa). We find little variation in performance for a region of settings between 0.7 and 0.85.

4 Experimental Methodology

Query performance prediction aims to identify whether a set of search results is likely to contains useful answers. The established information retrieval methodology for this type of investigation involves testing the performance of a predictor across a set of queries that are run on a collection of documents. The performance of the predictor is measured by calculating the correlation between the predicted performance levels with actual system performance levels.

Correlation Coefficients. In the query performance prediction literature, three different correlation coefficients are widely used (although individual papers often report only one or two of the three available variants): the Pearson product-moment correlation; Spearman's rank order correlation; and, Kendall's tau.

Although they make different assumptions about the data, each of the coefficients varies in the range $[+1, -1]$; the closer the absolute value of the coefficient is to 1, the stronger the correlation, with a value of zero indicating that there is no relationship between the variables. Moreover, each of the correlation coefficients can be used to conduct a hypothesis test to determine whether there is a significant relationship between the two variables, up to a specified level of confidence [12]. In this paper we report significance at the 0.05, 0.01 and 0.001 levels.

The *Pearson* correlation determines the degree to which two variables have a linear relationship, and takes the actual value of observations into account. *Spearman's* correlation coefficient is calculated based on the rank positions of observations; it therefore measures the degree to which a monotonic relationship exists between the variables (that is, a relationship where a change in the value of one variable is accompanied by a corresponding increase (decrease) in the value of the other variable). *Kendall's* tau is also calculated from rank information, but in contrast to Spearman's coefficient is based on the relative ordering of of all possible pairs of observations. For a comprehensive treatment of the properties of the different correlation coefficients the reader is

referred to Sheskin[12]. We note that there is currently no consensus on which measure of correlation is the most appropriate for information retrieval experiments, with different measures being reported in different studies. However, the Pearson correlation assumes a linear relationship between variables [7]; there is no reason to assume that this assumption holds between various query performance predictors and retrieval system performance.

Retrieval Performance Metrics. Information retrieval experimentation has a strong underlying experimental methodology as used for example in the ongoing series of Text REtrieval Conferences (TREC): a set of queries is run on a static collection of documents, with each query returning a list of answer resources. Humans assess the relevance of each document-query combination, and from this a variety of system performance metrics can be calculated. *Precision* is defined as the proportion of relevant and retrieved documents out of the total number of documents that have been retrieved. The *average precision* (AP) of a single query is then the mean of the precision of each relevant item in a search results list. Mean average precision (MAP) is the mean of AP over a set of queries. MAP gives a single overall measure of system performance, and has been shown to be a stable metric [3]. For the purpose of evaluating predictors of query performance, we calculate the correlation between the predicted ordering of a run of queries, and the AP score for that run of queries.

For navigational search tasks (see below), where it is assumed that the user is looking for a single specific answer resource, the reciprocal rank (RR) at which the item is found in the result list is used to measure system performance [3].

Collections, Queries and Relevance Judgements. We evaluate our predictors using several different collections and query sets; the aim is to examine the performance of the predictors on different types of data and search tasks. For *web* data, we use two collections: the first is the TREC GOV2 collection, a 425Gb crawl of the `gov` domain in 2004, which contains HTML documents and text extracted from PDF, PS and Word files [5]. We also test our predictors on the WT10g collection, a 10Gb crawl of the web in 1998 [1]. For *newswire* data, we use the collection of the 2005 Robust track, consisting of around 528,000 news articles from sources such as the Financial Times and LA Times (TREC disks 4 and 5, minus the congressional record data).

Each of these collections has associated queries and relevance judgements; full details are provided in Table 1. In our experiments we use only the title fields of the TREC topics, which consists of a few key words that are representative of the information need. Being short, the title field are the most representative of typical queries that might be submitted as part of web search.

Users of web search engines engage in different types of search tasks depending on their information need. In an *informational* search task, the user is interested in learning about a particular topic, while in a *navigational* task the user is looking for a specific named resource [2]. A key difference between these two tasks is that for navigational tasks, it is generally assumed that the user is interested in a single named page. For an informational task, on the other hand it is assumed that there may be several resources that are relevant to the information need. We test our predictors on both types of task; in this paper, navigational queries are identified with the prefix NP.

Table 1. Experimental setup summary: collections and corresponding query sets

Task	Collection	Query Set
TB04	GOV2	701–750
TB05	GOV2	751–800
TB06	GOV2	801–850
TB05-NP	GOV2	NP601–NP872
Robust04	TREC 4+5 (minus CR)	301–450; 601–700
TREC-9	WT10G	451–500
TREC-2001	WT10G	501–550

Retrieval Models. We experiment with two retrieval models: a probabilistic model, and a language model. For the probabilistic model, we use the Okapi BM25 similarity function [13], with the recommended parameter settings of $k_1 = 1.2$, $k_3 = 7$ and $b = 0.75$. For language models, we use the Dirichlet smoothing approach which has been shown to be successful across a wide range of collections and queries [17], with the smoothing parameter set to a value of $\mu = 1000$ [6]. In our experiments, we use version 4.5 of the Lemur Toolkit, an information retrieval toolkit developed jointly by the University of Massachusetts and Carnegie Mellon University[1].

Baselines. We compare our proposed prediction approaches to the two best-performing pre-retrieval predictors evaluated by He and Ounis [9]: the average inverse collection term frequency (AvICTF), and the Simplified Clarity Score (SCS). In their approach, the SCS is obtained by calculating the Kullback-Leibler divergence between a query model (θ_q) and a collection model (θ_c): $SCS = \sum_{t \in Q} \theta_q \cdot \log_2 \frac{\theta_q}{\theta_c}$, where the query model is given by the number of occurrences of a query term in the query ($f_{q,t}$), normalised by query length ($|Q|$), $\theta_q = \frac{f_{q,t}}{|Q|}$. The collection model is given by the number of occurrences of a query term in the entire collection ($f_{c,t}$), normalised by the number of tokens in the collection ($|C|$): $\theta_c = \frac{f_{c,t}}{|C|}$. For a third baseline, we use the maximum inverse document frequency (MaxIDF), which was found to be an effective pre-retrieval predictor for web data by Scholer et al. [11].

5 Results and Discussion

In this section we present the results of our experiments, comparing the performance of our predictors across a range of data types and search tasks.

Web Data, Informational Topics. The correlations between our pre-retrieval predictors with topic-finding queries on the *GOV2* collection are shown in Table 2. We show results separately by TREC data sets (701–750[2], 751–800 and 801–850), as well as the performance over all 149 topics taken together.

Overall, it is apparent that the performance of all predictors varies depending on the query set. The data shows that the similarity between a query and the collection (SCQ) can provide useful information for the prediction of how well a query will perform; the most effective of the three proposed variants here is $MaxSCQ$, which considers

[1] `http://www.lemurproject.org`

[2] Topic 703 is excluded because there is no relevance judgement for this query.

Table 2. Pearson (Cor), Kendall (Tau), and Spearman (Rho) correlation test between average precision (AP) and predictors on the *GOV2* collection. Asterisk, dagger and double dagger indicate significance at the 0.05, 0.01 and 0.001 levels, respectively.

		LM			Okapi		
Query	Predictors	Cor	Tau	Rho	Cor	Tau	Rho
701–750	MaxIdf	0.343∗	0.241∗	0.328∗	0.433†	0.311†	0.420†
	SCS	0.154	0.112	0.144	0.202	0.145	0.190
	AvICTF	0.345∗	0.241∗	0.331∗	0.446†	0.321‡	0.439†
	SCQ	0.244	0.175	0.236	0.231	0.156	0.212
	NSCQ	0.388†	0.264†	0.352∗	0.467‡	0.310†	0.431†
	MaxSCQ	0.412†	0.275†	0.399†	0.485‡	0.351‡	0.485‡
	σ_1	0.441†	0.310†	0.426†	0.477‡	0.294†	0.430†
	σ_2	0.401†	0.291†	0.442†	0.466‡	0.320†	0.476‡
	σ_3	0.418†	0.258†	0.394†	0.475‡	0.287†	0.448†
	joint	0.457‡	0.284†	0.399†	0.513‡	0.314†	0.447†
751–800	MaxIdf	0.267	0.238∗	0.334∗	0.308∗	0.247∗	0.354∗
	SCS	0.082	0.068	0.131	0.094	0.094	0.155
	AvICTF	0.276	0.210∗	0.302∗	0.249	0.180	0.271
	SCQ	0.257	0.167	0.267	0.275	0.203∗	0.313∗
	NSCQ	0.395†	0.267†	0.399†	0.359	0.247	0.368†
	MaxSCQ	0.396†	0.251∗	0.379†	0.448‡	0.287†	0.417†
	σ_1	0.379†	0.309†	0.449†	0.397†	0.332‡	0.470‡
	σ_2	0.424†	0.321‡	0.470‡	0.420†	0.324‡	0.450†
	σ_3	0.373†	0.251∗	0.391†	0.415†	0.290†	0.420†
	joint	0.423†	0.309†	0.461‡	0.466‡	0.336‡	0.486‡
801–850	MaxIdf	0.277	0.190	0.285∗	0.293∗	0.191	0.290∗
	SCS	−0.128	−0.048	−0.091	−0.111	−0.057	−0.094
	AvICTF	0.217	0.166	0.241	0.236	0.175	0.263
	SCQ	0.296∗	0.241∗	0.332∗	0.280∗	0.238∗	0.327∗
	NSCQ	0.094	0.113	0.167	0.090	0.108	0.161
	MaxSCQ	0.280∗	0.172	0.257	0.278	0.180	0.265
	σ_1	0.367†	0.319†	0.414†	0.361†	0.317†	0.397†
	σ_2	0.230	0.234∗	0.319∗	0.227	0.219∗	0.298∗
	σ_3	0.304∗	0.227∗	0.316∗	0.311∗	0.221∗	0.311∗
	joint	0.369†	0.283†	0.383†	0.365†	0.270†	0.367†
701–850	MaxIdf	0.297‡	0.219‡	0.326‡	0.343‡	0.247‡	0.367‡
	SCS	0.041	0.053	0.076	0.067	0.064	0.094
	AvICTF	0.269‡	0.187‡	0.282‡	0.295‡	0.205‡	0.307‡
	SCQ	0.260†	0.191‡	0.277‡	0.254†	0.189‡	0.273‡
	NSCQ	0.278‡	0.206‡	0.305‡	0.289‡	0.214‡	0.314‡
	MaxSCQ	0.357‡	0.231‡	0.347‡	0.395‡	0.266‡	0.388‡
	σ_1	0.392‡	0.285‡	0.407‡	0.401‡	0.290‡	0.411‡
	σ_2	0.384‡	0.291‡	0.411‡	0.396‡	0.293‡	0.412‡
	σ_3	0.359‡	0.247‡	0.369‡	0.392‡	0.266‡	0.390‡
	joint	0.410‡	0.287‡	0.415‡	0.436‡	0.297‡	0.430‡

the alignment between the most similar query term and the collection overall. The variability score predictors are extremely effective for the GOV2 data; in particular, the correlation of the $\sigma_1$predictor is statistically significant ($p < 0.01$) for all topics sets. The $joint$ predictor similarly shows consistent significant performance.

Considering performance over 149 topics, the $joint$ predictor, which uses information from both the collection and the document level, consistently outperforms all baseline predictors, and shows highly significant correlation ($p < 0.01$) across correlation types and retrieval models.

Correlation results for the *WT10g* collection are shown in Table 3. Collection-level information is again useful for prediction; the most effective variant is $MaxSCQ$, which considers the alignment between the most similar query term and the collection overall; the $MaxSCQ$ predictor is statistically significant for all correlation coefficients,

Table 3. Pearson (Cor), Kendall (Tau), and Spearman (Rho) correlation test between average precision (AP) and predictors on the *WT10g* collection. Asterisk, dagger and double dagger indicate significance at the 0.05, 0.01 and 0.001 levels, respectively.

		Okapi			LM		
Query	Predictors	Cor	Tau	Rho	Cor	Tau	Rho
451-500	MaxIdf	0.086	0.221∗	0.291∗	0.090	0.227∗	0.302∗
	SCS	0.194	0.226∗	0.333∗	0.197	0.227∗	0.332∗
	AvICTF	−0.056	0.012	0.016	−0.057	0.020	0.028
	SCQ	0.124	0.148	0.204	0.130	0.151	0.211
	NSCQ	0.402†	0.347‡	0.516‡	0.403†	0.348‡	0.520‡
	MaxSCQ	0.443†	0.453‡	0.620‡	0.447†	0.456‡	0.624‡
	σ_1	0.252	0.272†	0.405†	0.259	0.275†	0.410†
	σ_2	0.253	0.281†	0.436†	0.257	0.286†	0.438†
	σ_3	0.347∗	0.350‡	0.523‡	0.354∗	0.358‡	0.529‡
	joint	0.337‡	0.352‡	0.510‡	0.344‡	0.358‡	0.514‡
501-550	MaxIdf	0.491‡	0.284†	0.396†	0.507‡	0.291†	0.408†
	SCS	0.155	0.084	0.128	0.162	0.089	0.132
	AvICTF	−0.007	0.008	0.002	0.007	0.007	0.010
	SCQ	0.260	0.136	0.224	0.235	0.135	0.215
	NSCQ	0.099	0.097	0.140	0.114	0.106	0.157
	MaxSCQ	0.399†	0.267†	0.364*	0.418†	0.274†	0.378†
	σ_1	0.654‡	0.425‡	0.612‡	0.652‡	0.435‡	0.619‡
	σ_2	0.282	0.269†	0.389†	0.304*	0.278†	0.407†
	σ_3	0.518‡	0.358‡	0.486‡	0.538‡	0.372‡	0.502‡
	joint	0.640‡	0.449‡	0.622‡	0.644‡	0.466‡	0.634‡

Table 4. Pearson (Cor), Kendall (Tau), and Spearman (Rho) correlation test between reciprocal rank (RR) and predictors on the *GOV2* collection, for navigational topics. Asterisk, dagger and double dagger indicate significance at the 0.05, 0.01 and 0.001 levels, respectively.

		Okapi			LM		
Query	Predictors	Cor	Tau	Rho	Cor	Tau	Rho
NP601-NP872	MaxIdf	0.531‡	0.466‡	0.625‡	0.562‡	0.510‡	0.678‡
	SCS	0.257‡	0.200‡	0.282‡	0.281‡	0.236‡	0.331‡
	AvICTF	0.411‡	0.355‡	0.488‡	0.446‡	0.380‡	0.523‡
	SCQ	0.361‡	0.279‡	0.389‡	0.383‡	0.313‡	0.436‡
	NSCQ	0.385‡	0.343‡	0.478‡	0.423‡	0.374‡	0.521‡
	MaxSCQ	0.426‡	0.415‡	0.571‡	0.440‡	0.453‡	0.623‡
	σ_1	0.457‡	0.423‡	0.582‡	0.516‡	0.470‡	0.643‡
	σ_2	0.318‡	0.405‡	0.555‡	0.380‡	0.448‡	0.608‡
	σ_3	0.409‡	0.430‡	0.591‡	0.453‡	0.479‡	0.649‡
	joint	0.478‡	0.445‡	0.608‡	0.522‡	0.490‡	0.666‡

across all queries ($p < 0.05$). The variability score predictors are extremely effective for the TREC-2001 topics ($p < 0.01$), but show less consistent performance for TREC-9, where the linear Spearman correlation is not significant. The $joint$ predictor using both sources of information consistently performs well over both sets of topics, showing highly significant correlation ($p < 0.001$), and outperforming all three baselines.

We note that the $AvICTF$ baseline performs particularly poorly for this collection; the reason appears to be that presence of queries that contain terms that do not occur in the collection. The $MaxIDF$ predictor is highly effective on the TREC-2001 topics, but performs relatively poorly on the TREC-9 data. Basing prediction on just a single characteristic of queries therefore does not appear provide sufficient information to give consistent performance across informational searches on web data.

Table 5. Pearson (Cor), Kendall (Tau), and Spearman (Rho) correlation test between average precision (AP) and predictors on the *Robust* collection. Asterisk, dagger and double dagger indicate significance at the 0.05, 0.01 and 0.001 levels, respectively.

		LM			Okapi		
Query	Predictors	Cor	Tau	Rho	Cor	Tau	Rho
301-450;	MaxIdf	0.505‡	0.359‡	0.496‡	0.466‡	0.326‡	0.456‡
601-700	SCS	0.376‡	0.204‡	0.293‡	0.343‡	0.183‡	0.266‡
	AvICTF	0.386‡	0.234‡	0.336‡	0.355‡	0.203‡	0.294‡
	SCQ	0.062	0.102∗	0.149∗	0.058	0.090	0.132∗
	NSCQ	0.338‡	0.258‡	0.375‡	0.304‡	0.227‡	0.333‡
	MaxSCQ	0.371‡	0.348‡	0.493‡	0.335‡	0.316‡	0.450‡
	σ_1	0.329‡	0.323‡	0.458‡	0.310‡	0.302‡	0.434‡
	σ_2	0.237†	0.368‡	0.514‡	0.223‡	0.353‡	0.495‡
	σ_3	0.478‡	0.382‡	0.528‡	0.444‡	0.356‡	0.496‡
	joint	0.379‡	0.370‡	0.514‡	0.284‡	0.363‡	0.505‡

Web Data, Navigational Topics. Table 4 shows the results for navigational topics 601–872 on the *GOV2* collection. The performance of the $MaxIDF$ baseline is very strong for named page finding topics; this predictor gives the highest performance for the task across correlation coefficients and retrieval models. While less strong than the $MaxIDF$ correlation, the performance of the $joint$ predictor is consistently the second-highest, and is competitive for this task.

Newswire Data, Informational Topics. Experimental results for predictors on newswire data, using topics from the 2004 TREC *Robust* track, are shown in Table 5. The relative performance of the schemes shows more variation here than for other collections. In general, the two most effective predictors are $MaxIDF$, σ_3 and $joint$. The actual ordering varies depending on the correlation coefficient: $MaxIDF$ shows the highest correlation using the Pearson coefficient. Using the non-linear correlation coefficients leads to different conclusions, with σ_3 showing the highest correlation with the performance of the language model retrieval system, and $joint$ giving the highest correlation with the Okapi model.

We note that our SCS and AvICTF baseline results are slightly lower than those reported by He and Ounis [9]. We believe that this is due to differences in retrieval systems (Terrier and Lemur) used to calculate the system MAP scores.

6 Conclusions

We have introduced two new families of pre-retrieval predictors, based on the similarity between a query and the overall document collection, and the variability in how query terms are distributed across documents. Our experimental results show that these sources of information are both important for different collections, and are significantly correlated with the mean average precision of two different retrieval models. The best performance is obtained when combining both sources of information in the $joint$ predictor; this strongly outperforms three pre-retrieval baseline predictors for informational search tasks on web data, while giving comparable performance with the best baseline

on newswire data, and for navigational search tasks. The new predictors offer a significant advantage over previously proposed pre-retrieval predictors, because the performance of the latter varies drastically between search tasks and data types.

In our results, it can be seen that different correlation coefficients may in some cases lead to different conclusions about the performance of predictors. In future work we intend to explore the methodology of query performance prediction, to investigate which are the more appropriate measures for this task. We also plan to consider a variety of post-retrieval predictors to complement the pre-retrieval approaches.

References

1. Bailey, P., Craswell, N., Hawking, D.: Engineering a multi-purpose test collection for web retrieval experiments. Information Processing and Management 39(6), 853–871 (2003)
2. Broder, A.: A taxonomy of web search. SIGIR Forum 36(2), 3–10 (2002)
3. Buckley, C., Voorhees, E.M.: Retrieval system evaluation. In: Voorhees, E.M., Harman, D.K. (eds.) TREC: experiment and evaluation in information retrieval, MIT Press, Cambridge (2005)
4. Carmel, D., Yom-Tov, E., Soboroff, I.: SIGIR workshop report: predicting query difficulty - methods and applications. SIGIR Forum 39(2), 25–28 (2005)
5. Clarke, C., Craswell, N., Soboroff, I.: Overview of the TREC, terabyte track. In: The Thirteenth Text REtrieval Conference (TREC 2004), Gaithersburg, MD, 2005. National Institute of Standards and Technology Special Publication 500-261 (2004)
6. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 299–306 (2005)
7. Freund, J.E.: Modern Elementary Statistics, 10th edn. (2001)
8. Harman, D., Buckley, C.: The NRRC reliable information access (RIA) workshop. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, pp. 528–529 (2004)
9. He, B., Ounis, I.: Query performance prediction. Information System 31(7), 585–594 (2006)
10. Kwok, K.L.: An attempt to identify weakest and strongest queries. In: Predicting Query Difficulty, SIGIR 2005 Workshop (2005)
11. Scholer, F., Williams, H.E., Turpin, A.: Query association surrogates for web search. Journal of the American Society for Information Science and Technology 55(7), 637–650 (2004)
12. Sheskin, D.: Handbook of parametric and nonparametric statistical proceedures. CRC Press, Boca Raton (1997)
13. Sparck Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Part 1. *Information Processing and Management* 36(6), 779–808 (2000)
14. Voorhees, E.M.: Overview of the TREC, robust retrieval track. In: The Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, MD, 2006. National Institute of Standards and Technology Special Publication 500-266 (2005)
15. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd edn. Morgan Kaufmann, San Francisco (1999)
16. 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 the ACM SIGIR International Conference on Research and Development in Information Retrieval, Salvador, Brazil, pp. 512–519 (2005)

17. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions On Information Systems 22(2), 179–214 (2004)
18. Zhou, Y., Croft, W.B.: Ranking robustness: a novel framework to predict query performance. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, Arlington, Virginia, pp. 567–574 (2006)
19. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, pp. 543–550 (2007)
20. Zobel, J., Moffat, A.: Inverted files for text search engines. ACM Computing Surveys 38(2) (2006)