Using Coherence-based Measures to Predict Query Difficulty

Jiyin He, Martha Larson, and Maarten de Rijke

ISLA, University of Amsterdam
{jiyinhe,larson,mdr}@science.uva.nl

Abstract. We investigate the potential of coherence-based scores to predict query difficulty. The coherence of a document set associated with each query word is used to capture the quality of a query topic aspect. A simple query coherence score, QC-1, is proposed that requires the average coherence contribution of individual query terms to be high. Two further query scores, QC-2 and QC-3, are developed by constraining QC-1 in order to capture the semantic similarity among query topic aspects. All three query coherence scores show the correlation with average precision necessary to make them good predictors of query difficulty. Simple and efficient, the measures require no training data and are competitive with language model-based clarity scores.

1 Introduction

Robustness is an important feature of information retrieval (IR) systems [7]. A robust system achieves solid performance across the board and does not display marked sensitivity to difficult queries. IR systems stand to benefit if, prior to performing retrieval, they can be provided with information about problems associated with particular queries [4]. Work devoted to predicting query difficulty [1–3, 5, 8] is pursued with the aim of providing systems with the information necessary to adapt retrieval strategies to problematic queries. We investigate the usefulness of coherence-based scores in predicting query difficulty. The *query coherence scores* we propose are inspired by the *gene expression coherence* score used in the genetics literature [6], which functions as a measure of clustering structures. They are designed to reflect the quality of individual aspects of the query, following the suggestion that "the presence or absence of topic aspects in retrieved documents" is the predominant cause of current system failure [4].

We use document sets associated with individual query terms to assess the quality of query topic aspects (i.e., subtopics), noting that a similar assumption proved fruitful in [8]. We consider that a document set associated with a query term reflects a high-quality query topic aspect when it is: (1) topically constrained or specific and (2) characterized by a clustering structure tighter than that of the background document collection. These two characteristics are captured by coherence and for this reason we chose to investigate the potential of coherence-based scores. Like the clarity score [2, 3], our approach attempts to capture the difference between the language usage associated with the query

and the language usage in the background collection. Our approach promises low run-time computational costs. Additionally, our query coherence scores do not require training data as is the case with the method proposed in [8].

We propose three query coherence scores. The first query coherence score, QC-1, is an average of the coherence contribution of each query word and has only the effect of requiring that all query terms be associated with high-quality topic aspects. This score is simple and efficient. However, it does not require any semantic overlap between the contributions of the query words. A query topic composed of high-quality aspects would receive a QC-1 score even if those aspects were never reflected *together* in a collection document. Hence, we develop two further scores, which impose the requirement that, in addition to being associated with high-quality topic aspects, query words must be topically close. The second query coherence score, QC-2, adds a global constraint to QC-1. It requires the union of the set of documents associated with each query word to be coherent. The third score, QC-3, adds a proximity constraint to QC-1. It requires the document sets associated with individual query words to exhibit a certain closeness. QC-2 and QC-3 require more computational effort than QC-1, but fail to demonstrate an improved ability to predict query difficulty.

The next section further explains our coherence-based scores. After that we describe our experiments and results. We conclude with discussion and outlook.

2 Method

Given a document collection C and query $Q = \{q_i\}_{i=1}^{N}$, where q_i is a query term, $_{q_i}$ is the set of documents associated with that query word, i.e., the set of documents that contain at least one occurrence of the query word. The coherence of R_{q_i} reflects the quality of the aspect of a query topic that is associated with query word q_i. The overall query coherence score of a query is based on a combination of the set coherence contributed by each individual query word. Below, we first discuss set coherence and then present our three query coherence scores.

2.1 The coherence of a set of documents

The coherence of a set of documents is defined as the proportion of "coherent" pairs of documents in the set. A pair of documents is "coherent" if the similarity between them exceeds a given threshold. Formally, given a set of documents $D = \{d_i\}_{i=1}^{M}$ and threshold θ, we have

$$\delta(d_i, d_j) = \begin{cases} 1 & \text{if } \textit{similarity}(d_i, d_j) \geq \theta, \\ 0 & \text{otherwise.} \end{cases} \quad i \neq j \in \{1, \ldots, M\} \tag{1}$$

where the similarity between documents d_i and d_j can be any similarity metric; here we use the cosine similarity as an example. The *coherence* of the document set D is defined as

$$\textit{SetCoherence}(D) = \frac{\sum_{i \neq j \in \{1,\ldots,M\}} \delta(d_i, d_j)}{M(M-1)}. \tag{2}$$

Set coherence is a measure for the relative tightness of the clustering of a specific set of data with respect to the background collection. In a random subset drawn

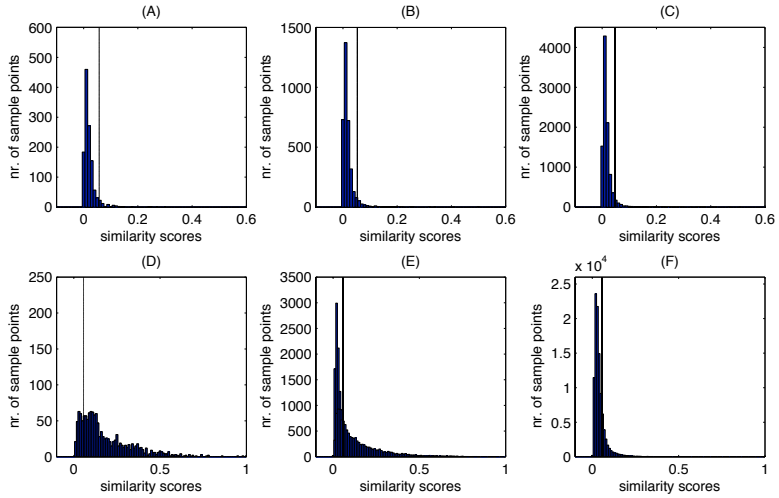

Fig. 1. Distribution of document similarities from subsets of TREC AP89+88. (A)–(C) Randomly sampled 50, 100, and 500 documents, respectively; (D) R_Q determined by query21, $SetCoherence(R_{Q21}) = 0.8483$; AP(Q21)=0.1328; (E) R_Q determined by query57, $SetCoherence(R_{Q57}) = 0.7216$; AP(Q57)=0.0472; (F) R determined by query75, $SetCoherence(R_{Q75}) = 0.2504$; AP(Q75)=0.0027.

from a document collection, few pairs of documents have high similarities. Plots A, B, and C in Figure 1 show that pairs having similarity scores higher than the threshold θ (the vertical line) are proportionally rare cases in a random sample, independently of sample size. Plots D, E and F show the distribution of document similarities for a collection subset associated with a one-word query, which we use to illustrate the properties of the R_{q_i}, the collection subset associated with a single query word q_i. Plots D, E, and F are ordered by decreasing coherence score, which can be seen to correspond to an increasing proportion of dissimilar document pairs. Plot F approaches the distribution of the random samples from the background collection. Initial support for the legitimacy of our approach derives from the fact that across these three queries decreasing set coherence of R_{q_i} corresponds to decreasing average precision.

2.2 Scoring queries based on coherence

For a given query $Q = \{q_i\}_{i=1}^{N}$, we propose three types of query coherence scores. The first requires that each query word have a high contribution to the coherence of the query. This score reflects the overall quality of the aspects of a topic.

QC-1 Average query term coherence:

$$QC\text{-}1(Q) = \frac{1}{N} \sum_{i=1}^{N} SetCoherence(R_{q_i}), \tag{3}$$

where $SetCoherence(R_{q_i})$ is the coherence score of the set R_{q_i} determined by the query word q_i. This score is simple, but leaves open the question of whether query aspects must also be semantically related. Therefore, we investigate whether

QC-1 can be improved by adding limitations that would force the R_{q_i}'s to be semantically constrained. The second query coherence score adds a constraint on global coherence, multiplying QC-1 by the coherence of $R_Q = \bigcup_{i=1}^{N} R_{q_i}$.

QC-2 Average query term coherence with global constraint:

$$QC\text{-}2(Q) = SetCoherence(R_Q) \frac{1}{N} \sum_{i=1}^{N} SetCoherence(R_{q_i}). \quad (4)$$

The third query coherence score adds a constraint on the proximity of the R_{q_i}'s, multiplying QC-1 by the average of the closeness of the centers of the R_{q_i}'s.

QC-3 Average query term coherence with proximity constraint:

$$QC\text{-}3(Q) = \frac{S}{N} \sum_{i=1}^{N} SetCoherence(R_{q_i}) \quad (5)$$

$$S = \frac{\sum_{l \neq k}^{N} Similarity(c(q_k), c(q_l))}{N(N-1)} \quad (6)$$

where S is the mean similarity score of each pair of cluster centers of the R_{q_i}'s. Below, we compare the performance of these three query coherence scores.

3 Evaluation

We run experiments to analyze the correlation between the proposed query coherence scores and the retrieval performance. Following [2], TREC datasets AP88 and AP89 are selected as our document collection. We use TREC topics 1–200 with the "title" field. The threshold θ is determined heuristically: we randomly sample different numbers of documents from the collection, and take the mean of the similarity scores at the top 5% of each sampled document set as the value of θ. For large sets R (e.g., $> 10,000$ documents), we approximate the *SetCoherence* by using the "collection" score (the threshold θ); a set R with many documents has a *SetCoherence* similar to the collection.

We use Spearman's ρ to measure the rank correlation between the coherence score and the average precision (AP). The higher this correlation, the more effective the scoring method is in terms of predicting query difficulty. Different retrieval models are applied so as to show stability across models.

Table 1 shows that all three coherence scores have significant correlation with AP. However, QC-2 and QC-3 do not have a substantially stronger predictive ability than QC-1, though they take the semantic relation between query words into account. Since the coherence score is the proportion of the "coherent pairs" among all the pairs of data points, and the similarity score can be pre-calculated without seeing any queries, the run-time operation for QC-1 is a simple counting. The same holds for QC-2, but with more effort for the extra term R_Q. Both are much easier to compute than QC-3, which requires the calculation of the centers of the R_{q_i}'s. Therefore, taking into account its computational efficiency, QC-1 is the preferred score. QC-1 is also more efficient at run time than other methods such as the clarity score [2] and has competitive prediction ability; see Table 2.

4 Conclusions

We introduced coherence-based measures for query difficulty prediction. Our initial experiments on short queries, reported here, show that the coherence score has a strong positive correlation with average precision, which reflects the

Table 1. The Spearman's correlation of query coherence scores with average precision. The queries are TREC topic 1–200, and the document collection is AP89+88.

	QC-1		QC-2		QC-3	
Model	ρ	p-value	ρ	p-value	ρ	p-value
BM25	0.3295	1.8897e-06	0.3389	0.0920e-05	0.3813	2.5509e-08
DLH13	0.2949	2.2462e-05	0.3096	0.8180e-05	0.3531	2.9097e-07
PL2	0.3024	1.3501e-05	0.3135	0.6167e-05	0.3608	1.5317e-07
TFIDF	0.2594	2.0842e-04	0.3301	0.1805e-05	0.3749	4.5006e-08

Table 2. The Spearman's correlation of clarity score (CS) and query coherence score (QC) with AP: the correlation coefficient ρ and its corresponding p-value. The queries are TREC topics 101–200, using title only. AP values obtained by running BM25; scores of column 1 taken from [2].

Score	CS	QC-1	QC-2	QC-3
ρ	0.368	0.3443	0.3625	0.3222
p-value	1.2e-04	4.5171e-04	2.1075e-04	0.0011

predictive ability of the proposed score. As similarity scores can be computed offline or at indexing time, this method promises run-time efficiency. Moreover, as the only parameter, θ, is obtained from the background collection, the method requires no training data. We plan to evaluate our coherence scores on more and larger data sets, e.g., the collection used in the TREC Robust track, as well as to investigate their behaviors on long queries. We will also use our approach in applications such as resource selection, and selective query expansion.

Acknowledgments

This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104 and by the Netherlands Organization for Scientific Research (NWO) by a grant under project numbers 220-80-001, 640.001.501, 640.002.501.

References

[1] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness and selective application of query expansion. In *ECIR'04*, pages 127–137, 2004.

[2] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *HLT'02*, pages 94–98, 2002.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR'02*, pages 299–306, 2002.

[4] D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *SIGIR'04*, pages 528–529, 2004.

[5] B. He and I. Ounis. Query performance prediction. *Inf. Syst.*, 31(7):585–594, 2006.

[6] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatiorial analysis of promoter elements. *Nat. Genet.*, 29:153–159, 2001.

[7] E. M. Voorhees. The TREC robust retrieval track. *SIGIR Forum*, 39:11–20, 2005.

[8] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty. In *SIGIR'05*, pages 512–519, 2005.