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, 2, 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

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 689–694, 2008.

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, R_{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

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

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

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

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