

File ID 216754
Filename Chapter 5: Using coherence-based score for query difficulty prediction

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation
Title Exploring topic structure: Coherence, diversity and relatedness
Author J. He
Faculty Faculty of Science
Year 2011
Pages xi, 201
ISBN 9789490371814

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/377895>

Copyright

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use.

Chapter 5

Using Coherence-Based Score for Query Difficulty Prediction

Robustness is an important feature of information retrieval (IR) systems [250]. 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 [85]. Work devoted to predicting query difficulty (also called query performance) [7, 33, 51, 52, 90, 93, 100, 265, 272] is pursued with the aim of providing systems with the information necessary to adapt retrieval strategies to problematic queries. For a survey on the work on this research topic, a recent book by Carmel and Yom-Tov [33] covers a wide range of query performance predictors proposed in the literature. Moreover, Hauff [90] has conducted extensive comparative studies on various types of predictors in her thesis, including the predictors we discuss in this chapter.

In this chapter, we investigate the usefulness of the coherence score in predicting query difficulty in a *pre-retrieval* setting. Specifically, we ask the following research questions:

RQ3a. Can we use the coherence score to measure query ambiguity?

RQ3b. Can we use query ambiguity as measured by coherence-based scores to predict query performance in an ad-hoc retrieval setting?

We posit that the performance of a query is correlated with its level of ambiguity. That is, we assume that the user's information need is specific and clearly defined, and therefore a query tends to retrieve non-relevant documents when it is ambiguous. For example, when a user searches for information about "java program," the query "java" may retrieve documents on topics such as "java island" or "java coffee." Here, the retrieval performance of a query is influenced by two factors. The first factor is the query itself. In the above example, if a query is associated with multiple subtopics or interpretations, it is likely that some of the subtopics or interpretations are non-relevant. Second, the performance for a given query also depends on the document collection we

use: an ambiguous query only affects retrieval performance if the collection contains documents associated with non-relevant interpretations of the query.

The *query coherence scores* we propose 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 [85]. 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 [265]. We consider that a document set associated with a query term reflects a high-quality (i.e., non-ambiguous) query topic aspect when it is: (1) topically constrained or specific and (2) characterized by a clustering structure tighter than that of some 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 [51, 52], our approach attempts to capture the difference between the language usage associated with the query and the language usage in the background collection.

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 only has 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 non-zero QC-1 score even if those aspects were never reflected *together* in a document. Hence, we further develop two alternative scores that 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 be close to each other.

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

5.1 Query coherence scores

Given a document collection C and query $Q = \{q_i\}_{i=1}^N$, where q_i is a query term. We define R_{q_i} as the set of documents associated with a 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 scores contributed by each individual query word. Below, we first discuss coherence on a set of documents and then present our three query coherence scores.

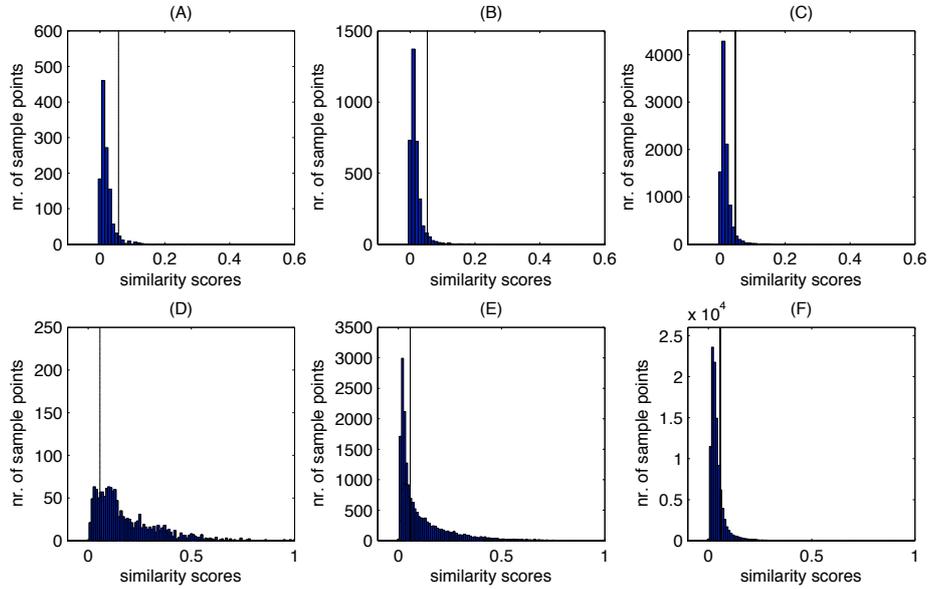


Figure 5.1: Distribution of document similarities from subsets of TREC AP89+88 (as introduced in Section 3.3.1 on page 40). (A)–(C) Randomly sampled 50, 100, and 500 documents, respectively; (D) R_Q determined by query21, $Co(R_{Q21}) = 0.8483$; $AP(Q21) = 0.1328$; (E) R_Q determined by query57, $Co(R_{Q57}) = 0.7216$; $AP(Q57) = 0.0472$; (F) R determined by query75, $Co(R_{Q75}) = 0.2504$; $AP(Q75) = 0.0027$.

5.1.1 The coherence of a set of documents

As defined in Chapter 3, the coherence score 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 from a document collection, few pairs of documents have high similarities. In Figure 5.1 we illustrate coherence of documents collected in different ways. The coherence of each document set is calculated as defined in Eq. 3.2 on page 36. Plots A, B, and C in Figure 5.1 show that pairs having similarity scores higher than the threshold τ (the vertical line) are proportionally rare cases in a random sample, independent 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 corresponds 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 is derived from the fact that across these three queries decreasing set coherence of R_{q_i} corresponds to decreasing AP (as introduced in Section 2.4.2 on page 27).

5.1.2 Scoring queries based on coherence

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

QC-1 Average query term coherence:

$$QC-1(Q) = \frac{1}{N} \sum_{i=1}^N Co(R_{q_i}), \quad (5.1)$$

where $Co(R_{q_i})$ 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 constraints that would force the R_{q_i} 's to be semantically related. 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:

$$QC-2(Q) = Co(R_Q) \frac{1}{N} \sum_{i=1}^N Co(R_{q_i}). \quad (5.2)$$

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:

$$QC-3(Q) = \frac{S}{N} \sum_{i=1}^N Co(R_{q_i}) \quad (5.3)$$

$$S = \frac{\sum_{l \neq k}^N \text{Similarity}(c(q_k), c(q_l))}{N(N-1)}, \quad (5.4)$$

where S is the mean similarity score of each pair of cluster centers $c(q_i)$ of the R_{q_i} 's. Here, $c(q_i)$ is calculated as

$$c(q_i) = \frac{1}{M} \sum_{d \in R_{q_i}} \vec{d}, \quad (5.5)$$

where M is the total number of documents contained in R_{q_i} and \vec{d} is a document in R_{q_i} represented using a vector space model.

Below, we compare the performance of the three query coherence scores.

5.2 Evaluation

5.2.1 Experimental setup

We run experiments to analyze the correlation between the proposed query coherence scores and the retrieval performance. Following [51], TREC datasets AP88+89 (as

introduced in Section 3.3.1) are selected as our document collection. We use TREC topics 1–200 with the “title” field. We experiment with a number of retrieval models, including BM25 [202], TFIDF [203, 234], and the DFR model with the PL2 and the DH13 weighting schemes [6]. We use the Terrier [186] implementation of these models with default parameter settings.

We calculate the coherence score for a document set associated with a query term as defined in Eq. 3.2 on page 36. The threshold τ is determined as described in Section 3.1.1 on page 36, and cosine similarity is used as the measure of similarity between documents. For large sets R (e.g., $> 10,000$ documents), we approximate the coherence score by using the “collection” score (the threshold τ); we posit that a set R with many documents has a coherence score similar to the collection.

5.2.2 Evaluation measure

We use Spearman’s ρ to measure the rank correlation between the coherence score and the Average Precision. 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 of our observations across models.

5.2.3 Results

Table 5.1 shows that all three coherence scores have a significant correlation with AP. In general, QC-2 and QC-3 show a higher positive correlation with the AP than QC-1. However, their predictive ability is not substantially stronger than QC-1, judging from the difference between the correlation coefficients of QC-1 and that of QC-2 and QC-3, though they do take the semantic relation between query words into account. Since the coherence score is the proportion of “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 of the “coherent pairs.” 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 that need to be processed at run-time. Therefore, taking into account its computational efficiency and the limited improvements seen in the alternative QC’s, QC-1 is the preferred score. Moreover, even though it is a pre-retrieval predictor, QC-1 has a competitive prediction ability compared to other post-retrieval methods such as the clarity score [51]; see Table 5.2.

5.2.4 Hauff’s experiments

In addition to our preliminary experiments, Hauff [90] has conducted further experiments in analyzing the performance of QC-1 and QC-2. In Hauff’s experiments, the coherence score is implemented as described in [100] with the same experimental settings as described here. Additional TREC test collections (TREC 4+5 [248],

Model	QC-1		QC-2		QC-3	
	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 5.1: The Spearman’s rank correlation of query coherence scores with average precision. Queries: TREC topics 1–200; document collection: AP89+88.

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

Table 5.2: The Spearman’s rank 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; the clarity scores of column 1 are taken from [51].

WT10g [232] and Gov2 [39]) are used. TREC4+5 is similar to AP89+88: it is relatively small compared to the other two collections, consisting of news articles. WT10g and Gov2 are large collections consisting of Web crawls. TFIDF, BM25 and Language Model (LM) are included as retrieval models.

The conclusion of Hauff’s experiments can be summarized as follows.

- First, the performance of QC-1 and QC-2 varies across collections and the better performance is achieved on smaller collections. Best performance is achieved on TREC 4+5, where both predictors show relatively stable positive correlations across all three retrieval models and significant results are achieved. The performance on WT10g and Gov2 are not as stable, in many cases only insignificant correlations are found between our predictors (i.e., QC-1 and QC-2) and the AP.
- Second, when LM is used as retrieval model, the rank correlation between the query coherence scores and AP increases with an increasing amount of smoothing.

Combining the observations made from our experiments and those of Hauff’s experiments, one important conclusion here is that the query coherence scores are more effective on small collections (particularly, on a specific domain such as news) than on large and Web based collections. One possible explanation is that, in smaller collections, especially in a single domain such as news articles where the language usage is often more confined compared to that on the Web, the query term ambiguity is captured well by the topical coherence of documents associated with it. In a Web collection, however, *every* query term may be associated with more diverse documents, including spam, which may reduce the distinction between non-ambiguous terms and ambiguous

terms. Particularly, when using the heuristic approximation for large document sets associated with a query term (as described in 5.2.1 on page 72), in large collections, it is very likely that this approximation is used for most of the queries, as most of the queries may be associated with a document set with more than 10,000 documents.

5.3 Discussion and conclusions

With respect to our two research questions RQ3a and RQ3b as stated on page 69, we have the following answers. We introduced coherence-based measures for query difficulty prediction. The coherence score of the set of documents associated with a single query term is used as a measure of the quality (i.e., level of non-ambiguity) of the query term. We then experimented with three ways to combine the coherence scores of each query term into a single score as performance predictors for a query. Our initial experiments on short queries show that the coherence score has a strong positive correlation with average precision, which reflects the predictive ability of the proposed score.

Hauff's experiments, on the other hand, have raised further open issues for these predictors. For example, what makes our predictors less effective on large collections? how do we measure query ambiguity on large collections such as the Web? Further, with respect to Web retrieval, it is an open question whether query ambiguity is an important factor responsible for query performance. For example, strategies such as result diversification are often used to deal with ambiguous or multi-faceted queries. That is, without knowing the actual user's information need, the retrieval system presents a list of documents covering as many as possible subtopics associated with the query. Within this specific task scenario, the importance of query ambiguity with respect to the query performance may need to be reconsidered, which we leave as future work.