

1 Activities and Findings

1.1 Project Activities and Findings

1.1.1 Research and education activities

In the past year, we have made progress on a number of fronts with regard to representations and retrieval models for searching with long queries. In one research direction, we have continued to enhance the theory of the Markov Random Field-based retrieval model to include a more graph-based representation for longer queries and algorithms for weighting the concepts in the graph (Bendersky, Metzler and Croft, 2011). The majority of current information retrieval models weight the query concepts (e.g., terms or phrases) in an unsupervised manner, based solely on the collection statistics. In this work, we go beyond the unsupervised estimation of concept weights, and propose a parameterized concept weighting model. In our model, the weight of each query concept is determined using a parameterized combination of diverse importance features. Unlike the existing supervised ranking methods, our model learns importance weights not only for the explicit query concepts, but also for the latent concepts that are associated with the query through pseudo-relevance feedback.

In a related work, we have shown how the graph-based representation of queries can be generated using limited training data (Bendersky, Croft and Smith, 2010 and 2011). Marking up queries with a set of structural annotations such as part-of-speech tags, capitalization, segmentation, and stopword indicators, is an important part of many approaches to query processing and understanding. Due to their brevity and idiosyncratic grammatical structure, search queries pose a challenge to existing annotation tools that are commonly trained on full-length documents. To address this challenge, we view the query as an explicit representation of a latent user information need. This view allows us to use pseudo-relevance feedback, and to leverage additional information from the document corpus on which the retrieval is performed, in order to improve the accuracy of the query annotation. In addition, we propose a probabilistic framework for performing joint query annotation that exploits the dependency between the different unsupervised annotations to further improve the accuracy over the entire set of annotations, even with a very limited amount of available training data. Unlike previous work on query annotation, our methods do not require access to search logs, user click-data, or large amounts of training data and can be used to perform several annotation tasks in tandem.

In a further development of our retrieval model framework, we incorporated features based on document quality (Bendersky, Croft, and Diao, 2011). Many existing retrieval approaches do not take into account the content quality of the retrieved documents, although link-based measures such as PageRank are commonly used as a form of document prior. In this work, we developed a quality-biased ranking method that promotes documents containing high-quality content, and penalizes low-quality documents. The quality of the document content can be determined by its readability, layout and ease-of-navigation, among other factors. Accordingly, instead of using a single estimate for document quality, we consider multiple content-based features that are directly integrated into a state-of-the-art retrieval method. These content-based features are easy to compute, store and retrieve, even for large web collections. This technique is particularly important to longer, “tail” queries that are more susceptible to spam.

Query reformulation techniques modify and expand the original query with the aim of better matching the vocabulary of the relevant documents, and consequently improving ranking effectiveness. Previous models typically generate words and phrases related to the original query, but do not consider how these words and phrases would fit together in new queries. To address this issue, we have developed a novel framework that models reformulation as a distribution of queries, where each query is a variation of the original query (Xue and Croft, 2011). This approach considers a query as a basic unit and can capture important dependencies between words and phrases in the query. Previous reformulation models are special cases of the proposed framework by making certain assumptions. An implementation of this framework consists of a query generation step that analyzes the passages containing query words to generate reformulated queries and a probability estimation step that learns a distribution for reformulated queries by optimizing the retrieval performance.

Finally, one feature of long queries is that they contain significantly more syntactic structure than short queries. On the other hand, incorporating syntactic features in a retrieval model has had very limited success in the past, with the exception of term dependencies. We have developed a new term dependency modeling approach based on a dependency parsing technique used for both queries and documents (Park, Croft and Smith, 2011). Our model is inspired by a quasi-synchronous stochastic process for machine translation. It describes four different types of syntactic relationships between dependent terms and allows inexact matching between documents and queries to deal with possible syntactic transformations. We have also proposed a machine learning technique for predicting optimal parameter settings for a retrieval model incorporating the syntactic relationships.

1.1.2 Findings

The experimental results with the Markov Random field model on both newswire and web TREC corpora show that our approach consistently and significantly outperforms a wide range of state-of-the-art retrieval models. In addition, our model significantly reduces the number of latent concepts used for query expansion compared to the non-parameterized pseudo-relevance feedback based models.

The joint annotation approach to query representation was evaluated using a range of queries taken from a web search log. Experimental results verify the effectiveness of this approach across all query types, and especially for longer natural language queries.

The quality-biased retrieval features was evaluated using several query sets and web collections. In each case, our method consistently improves by a large margin the retrieval performance of text-based and link-based retrieval methods that do not take into account the quality of the document content.

Experiments on a range of TREC collections have shown that our proposed query distribution model can significantly outperform previous reformulation models.

The results on two TREC collections have shown that the quasi-synchronous dependence model can improve retrieval performance and outperform a strong state-of-art baseline when we use predicted optimal parameters. This is one of the first retrieval experiments showing effectiveness benefits for syntactic features in these collections.