Modeling and Solving Term Mismatch for Full-Text Retrieval

Le Zhao
Language Technologies Institute
School of Computer Science
Carnegie Mellon University
lezhao@cs.cmu.edu

October 22, 2012

Even though modern retrieval systems typically use a multitude of features to rank documents, the backbone for search ranking is usually the standard tf.idf retrieval models.

This thesis addresses a limitation of the fundamental retrieval models, the term mismatch problem, which happens when query terms fail to appear in the documents that are relevant to the query. The term mismatch problem is a long standing problem in information retrieval. However, it was not well understood how often term mismatch happens in retrieval, how important it is for retrieval, or how it affects retrieval performance. This thesis answers the above questions, and proposes principled solutions to address this limitation. The new understandings of the retrieval models will benefit its users, as well as inform the development of software applications built on top of them.

This new direction of research is enabled by the formal definition of the probability of term mismatch, and quantitative data analyses around it. In this thesis, term mismatch is defined as the probability of a term not appearing in a document that is relevant to the query. The complement of term mismatch is the term recall, the probability of a term appearing in relevant documents. Even though the term recall probability is known to be a fundamental quantity in the theory of probabilistic information retrieval, prior research in ad hoc retrieval provided few clues about how to estimate term recall reliably.

This dissertation research designs two term mismatch prediction methods. With exploratory data analyses, this research first identifies common reasons that user-specified query terms fail to appear in documents relevant to the query, develops features correlated with each reason, and integrates them into a predictive model that can be trained from data. This prediction model uses training queries with relevance judgments to predict term mismatch for test queries without known relevance, and can be viewed as a type of transfer learning where training queries represent related ranking tasks that are used by the learning algorithm to facilitate the ranking for new test tasks. Further data analyses focus on the variation of the term mismatch probability for the same term across different queries, and demonstrate that query dependent features are needed for effective term mismatch prediction. At the same time, because the cross-query variation of term mismatch

is small for most of the repeating term occurrences, a second mismatch prediction method is designed to use historic occurrences of the same term to predict the mismatch probability for its test occurrences. This provides an alternative and more efficient procedure to predict term mismatch.

Effective term mismatch predictions can be used in several different ways to improve retrieval. The probabilistic retrieval theory suggests to use the term recall probabilities as term weights in the retrieval models. Experiments on 6 different TREC Ad hoc track and Web track datasets show that this automatic intervention improves both retrieval recall and precision substantially for long queries. Even though term weighting does not substantially improve retrieval accuracy for short queries which typically have a higher baseline performance, much larger gains are possible by solving mismatch using user expanded Conjunctive Normal Form queries. These queries try to fix the mismatch problem by expanding every query term individually. Our method uses the automatic term mismatch predictions as a diagnostic tool to guide interactive interventions, so that the users can expand the query terms that need expansion most. Simulated expansion interactions based on real user-expanded queries on TREC Ad hoc and Legal track datasets show that expanding the terms that have the highest predicted mismatch probabilities effectively improves retrieval performance. The resulting Boolean Conjunctive Normal Form expansion queries are both compact and effective, substantially outperforming the short keyword queries as well as the traditional bag of word expansion that may use the same set of high quality manual expansion terms.

Promising problems for future research are identified, together with research areas where the term mismatch research may make an impact.

Available online at `http://www.cs.cmu.edu/~lezhao`