Doctoral Dissertations 2014-current

Indexing Proximity-based Dependencies for Information Retrieval

Samuel Huston, *University of Massachusetts - Amherst*

Follow

Date of Award
2-2014

Document Type
Open Access Dissertation

Degree Name
Doctor of Philosophy (PhD)

Degree Program
Computer Science

First Advisor
W. Bruce Croft

Second Advisor
James Allan,

Third Advisor
Andrew McGregor

Keywords
Efficiency, Evaluation, Information Retrieval, Systems

Subject Categories
Computer Sciences

Abstract
Research into term dependencies for information retrieval has demonstrated that dependency retrieval models are able to consistently improve retrieval effectiveness over bag-of-words models. However, the computation of term dependency statistics is a major efficiency bottleneck in the execution of these retrieval models. This thesis investigates the problem of improving the efficiency of dependency retrieval models without compromising the effectiveness benefits of the term dependency

features.

Despite the large number of published comparisons between dependency models and bag-of-words approaches, there has been a lack of direct comparisons between alternate dependency models. We provide this comparison and investigate different types of proximity features. Several bi-term and many-term dependency models over a range of TREC collections, for both short (title) and long (description) queries, are compared to determine the strongest benchmark models. We observe that the weighted sequential dependence model is the most effective model studied. Additionally, we observe that there is some potential in many-term dependencies, but more selective methods are required to exploit these features.

We then investigate two novel index structures to directly index the proximitybased dependencies used in the sequential dependence model and weighted sequential dependence model. The frequent index and the sketch index data structures can both provide efficient access to collection and document level statistics for all indexed term dependencies, while minimizing space costs, relative to a full inverted index of term dependencies. We test whether these structures can improve retrieval efficiency without incurring large space requirements, or degrading retrieval effectiveness significantly. A secondary requirement is that each data structure must be able to be constructed for an input text collection in a scalable and distributed manner.

Based on the observation that the vast majority of term dependencies extracted from queries are relatively frequent in the collection, the “frequent” index of term dependencies omits data for infrequent term dependencies. The sketch index of term dependencies uses techniques from sketch data structures to store probabilisticallybounded estimates of the required statistics. We present analyses of these data structures that include construction and space costs, retrieval efficiency and investigation of any degradation of retrieval effectiveness.

Finally, we investigate the application of these data structures to the execution of the strongest performing dependency models identified. We compare the retrieval efficiency of each of these structures across two query processing algorithms, and across both short and long queries, using two large web collections. We observe that these newly proposed data structures allow the execution of queries considerably faster than when using positional indexes, and as fast as a full index of term dependencies, but with lowered storage overhead.