Home > ETDS > Doctoral Dissertations 2014-current > 41

< <u>Previous</u> <u>Next</u> >

SHARE

Doctoral Dissertations 2014-current

Off-campus UMass Amherst users: To download campus access dissertations, please use the following link to log into our proxy server with your UMass Amherst user name and password.

Non-UMass Amherst users: Please talk to your librarian about requesting this dissertation through interlibrary loan.

Dissertations that have an embargo placed on them will not be available to anyone until the embargo expires.

Indexing Proximity-based Dependencies for Information Retrieval

Samuel Huston, University of Massachusetts - Amherst

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

	Home	<u>About</u>	FAQ	My Account
ext >	Enter se	earch terr	ns:	Search
ollowing	Advanced Search Notify me via email or RSS			
ugh	Browse			
mbargo	Collections Disciplines Authors			
	Author Corner			
Download	<u>Author</u> Submit	FAQ Disserta	tion	

Follow

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.

Recommended Citation

Huston, Samuel, "Indexing Proximity-based Dependencies for Information Retrieval" (2014). *Doctoral Dissertations 2014-current*. Paper 41. http://scholarworks.umass.edu/dissertations_2/41

> This page is sponsored by the <u>University Libraries.</u> © 2009 University of Massachusetts Amherst • Site Policies