Sparse Multi-Level Representations for Text Retrieval

by
Charles Lee Isbell, Junior

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
at the
Massachusetts Institute of Technology

Abstract

In text retrieval, our task is to find the subset of a collection of natural language documents that is relevant to a query. The popular Vector Space Model (VSM) treats documents and queries as a vector of word counts, defining relevance as the dot product of two vectors—a measure of the number of common terms. Although both space and time-efficient, VSM ignores a great deal of the natural structure in text. Further, VSM suffers because the presence or absence of a term is not sufficient to determine relevance.

Typical machine learning extensions to VSM use linear dimensionality reduction to extract underlying structure. We show that a popular family of dimensionality reduction techniques induces a distribution over documents of a form that can be very different from the "optimal" distribution induced when clusters of related documents are known beforehand. We characterize the form of the optimal distribution, drawing connections to *independent components,* and *projection pursuit*. We then derive an unsupervised method that approximate characteristics of this distribution. They are *topic-centered* in that they construct a new space whose axes represent clusters of highly-related terms. In all of these techniques, we are able to reject candidate axes as unsuitable by using kurtosis as a measure of "interestingness," improving retrieval performance.

Further, we address the common assumption that queries and documents should be treated as if they live in the same vector space. Rather than accept this wholesale, we construct a multi-level representation of documents using an efficient statistical procedure and simple natural language parsing. Each level of the representation contains different features that embody increasingly more general natural language structure. These features are used to classify queries of varying structural richness. Once classified, a query is matched against a representation of documents that reflects its structure.

Finally, we deal with the computational issues that arise with the extremely large data sets used in real-world retrieval. In particular, we describe Brazil, a research system custom-designed for retrieval on extremely large data sets. This system is based on the Parallel Problems Server, an extremely powerful and novel parallel "linear algebra server" that we have developed. Using this system, we show that our techniques improve retrieval performance and can be realized through algorithms that are space and time efficient.

Thesis Committee:
Paul A. Viola
Assistant Professor of Computer Science and Engineering
Rodney A. Brooks
Professor of Computer Science and Engineering
Lynn Andrea Stein
Associate Professor of Computer Science and Engineering