Indexing by Latent Semantic Analysis

Scott Deerwester
Center for Information and Language Studies, University of Chicago, Chicago, IL 60637

Susan T. Dumais*, George W. Furnas, and Thomas K. Landauer
Bell Communications Research, 445 South St., Morristown, NJ 07960

Richard Harshman
University of Western Ontario, London, Ontario Canada

A new method for automatic indexing and retrieval is described. The approach is to take advantage of implicit higher-order structure in the association of terms with documents ("semantic structure") in order to improve the detection of relevant documents on the basis of terms found in queries. The particular technique used is singular-value decomposition, in which a large term by document matrix is decomposed into a set of ca. 100 orthogonal factors from which the original matrix can be approximated by linear combination. Documents are represented by ca. 100 item vectors of factor weights. Queries are represented as pseudo-document vectors formed from weighted combinations of terms, and documents with supra-threshold cosine values are returned. Initial tests find this completely automatic method for retrieval to be promising.

Introduction

We describe here a new approach to automatic indexing and retrieval. It is designed to overcome a fundamental problem that plagues existing retrieval techniques that try to match words of queries with words of documents. The problem is that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document. There are usually many ways to express a given concept, so the literal terms in a user's query may not match those of a relevant document. In addition, most words have multiple meanings, so terms in a user's query will literally match terms in documents that are not of interest to the user.

The proposed approach tries to overcome the deficiencies of term-matching retrieval by treating the unreliability of observed term-document association data as a statistical problem. We assume there is some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval. We use statistical techniques to estimate this latent structure, and get rid of the obscuring "noise." A description of terms and documents based on the latent semantic structure is used for indexing and retrieval.[1]

The particular "latent semantic indexing" (LSI) analysis that we have tried uses singular-value decomposition. We take a large matrix of term-document association data and construct a "semantic" space wherein terms and documents that are closely associated are placed near one another. Singular-value decomposition allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences. As a result, terms that did not actually appear in a document may still end up close to the document, if that is consistent with the major patterns of association in the data. Position in the space then serves as the new kind of semantic indexing. Retrieval proceeds by using the terms in a query to identify a point in the space, and documents in its neighborhood are returned to the user.

Deficiencies of Current Automatic Indexing and Retrieval Methods

A fundamental deficiency of current information retrieval methods is that the words searchers use often are not the same as those by which the information they seek has been indexed. There are actually two sides to the issue; we will call them broadly *synonymy* and *polysemy*. We use *synonymy* in a very general sense to describe the fact that

*To whom all correspondence should be addressed.

Received August 26, 1987; revised April 4, 1988; accepted April 5, 1988.

[1] By "semantic structure" we mean here only the correlation structure in the way in which individual words appear in documents; "semantic" implies only the fact that terms in a document may be taken as referents to the document itself or to its topic.