Technology Profile Fact Sheet

Title: Automatically Describing and Categorizing the Topic of Text

Aliases: None

Technical Challenge: To provide a means of document retrieval by topic that will serve as the basis of a maintenance scheme for a semantic interface to a knowledge base. Without automatic maintenance, a large-scale knowledge base would degrade in a dynamically changing environment and be inoperable. Such an automatic mechanism does not exist now.

Description: This invention is a solution to the basic document index problem for document retrieval in its most primitive form. It applies particularly to the automatic generation of a topic description and categorization for text, and the searching and sorting of text by topic. The text may be of any length, in any language, and may be derived from any source, to include machine translated speech or optical character reader.

Common keyword-based searches may lead to problems of over-specification or over-generalization. In contrast, this method relies upon scored n-tuples of singular form nouns in order to determine topic descriptions. It has the advantage of not needing stop lists or extensive stemming, and owes its accuracy to the use of the n-tuples to limit the search space concisely, together with a probability based measure of distance to categorize each topic. Several alternative embodiments of the concept are described, in order to optimize its use in any given environment.

Demonstration Capability: The code can be demonstrated on any document database.

Potential Commercial Application(s): Document retrieval by example rather than SQL queries would be the basis of collaborative filtering approach to document database retrieval.

Patent Status: A patent application has been filed with the USPTO.

Reference Number: 1397