AUTOMATIC TEXT STRUCTURING AND RETRIEVAL IN LARGE NATURAL-LANGUAGE TEXT FILES

Gerard Salton

Cornell University
Computer Science Department
Ithaca, Nueva York

Abstract

Very large text files are now available for automatic processing applicable for a wide variety of tasks, and covering many different subject areas. Methods must be provided that allow easy access to the stored data and make it possible to locate particular items on demand. The conventional text analysis methods based on preconstructed knowledge-bases or other vocabulary-control tools are difficult to apply when the subject coverage is unrestricted.

An alternative approach, applicable to text collections in any subject area, is introduced which uses the document collections themselves as a basis for the text analysis, together with sophisticated text matching operations carried out at several levels of detail. . Methods are described for relating semantically similar pieces of text, and for using the resulting hypertext structures for collection browsing and information retrieval.