Abstract

Importance of document clustering is now widely acknowledged by researchers for better management, smart navigation, efficient filtering, and concise summarization of large collection of documents like World Wide Web (WWW). The next challenge lies in semantically performing clustering based on the semantic contents of the document. The problem of document clustering has two main components: (1) to represent the document in such a form that inherently captures semantics of the text. This may also help to reduce dimensionality of the document, and (2) to define a similarity measure based on the semantic representation such that it assigns higher numerical values to document pairs which have higher semantic relationship. Feature space of the documents can be very challenging for document clustering. A document may contain multiple topics, it may contain a large set of class-independent general-words, and a handful class-specific core-words. With these features in mind, traditional
agglomerative clustering algorithms, which are based on either Document Vector model (DVM) or Suffix Tree model (STC), are less efficient in producing results with high cluster quality. This paper introduces a new approach for document clustering based on the Topic Map representation of the documents. The document is being transformed into a compact form. A similarity measure is proposed based upon the inferred information through topic maps data and structures. The suggested method is implemented using agglomerative hierarchal clustering and tested on standard Information retrieval (IR) datasets. The comparative experiment reveals that the proposed approach is effective in improving the cluster quality.

Reference

Index Terms

Computer Science

Information Retrieval

Key words

Text Document

Document Clustering

Algorithm

Performance measure