Abstract

This paper proposes a new and efficient methodology for clustering of html documents. The topic wise categorization of documents into different clusters makes searching easier and efficient. This technique can be utilized by search engines to provide relevant results to the user according to query and also utilized by online journal domains that are maintaining large set of documents. This paper suggests a good word matching and naming of automatic generated clusters, so, the time consume for finding the appropriate cluster for a document will be reduced. This paper shows the use of an efficient technique for finding the similarity between the documents and assigns them a proper cluster. The proper clustering of documents will be further utilized by multidocument summarization system, which produces a summary for the documents related to each other.
Automatic Naming of Domain Specific Clusters for Efficient Searching

- M. Sahami, S. Dumais, D. Heckerman, E. Horvitz. A Bayesian approach to filtering
  junk e-mail. AAAI Workshop on Learning for Text Categorization. Tech. Rep. WS-98-05,
- A. Y. Ng, M. I. Jordan. On discriminative vs. generative classifiers: a comparison of
- Fabrizio Silvestri, Raffaele Perego and Salvatore Orlando. “Assigning Document
  Identifiers to Enhance Compressibility of Web Search Engines Indexes”; In the
  demonstration”; In the proceedings of SIGIR, 1998.
  Artificial Intelligence, 2004
  Information and Cybernetics.

Index Terms

Computer Science

Information Science

Keywords

Keywords are clustering similarity clusters etc.