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.

References

- S. Chakrabarti, B. Dom. R. Agrawal, P. Raghavan. Using taxonomy, discriminants
- Fabrizio Silvestri, Raffaele Perego and Salvatore Orlando. &quot;Assigning Document Identifiers to Enhance Compressibility of Web Search Engines Indexes&quot; In the proceedings of SAC, 2004.
- Van Rijsbergen C. J. &quot;Information Retrieval&quot; Butterworth 1979
- Oren Zamir and Oren Etzioni. &quot;Web Document Clustering: A feasibility demonstration&quot; In the proceedings of SIGIR, 1998.
- Jain and R. Dubes. &quot;Algorithms for Clustering Data. &quot; Prentice Hall, 1988
- Sanjiv K. Bhatia. &quot;Adaptive K Means Clustering&quot; American Association for Artificial Intelligence, 2004

Index Terms

Computer Science Information Science

Keywords

Keywords are clustering similarity clusters etc.