Abstract

Document clustering has emerged as a widely used technique with the increase in large number of documents that is getting accumulated day by day in various fields like news groups, government organizations, Internet and digital libraries. Document clustering is the process of grouping similar documents into clusters. A good document clustering algorithm should have high intra-cluster similarity and less inter-cluster similarity. i.e. the documents with the clusters should be more relevant compared to the documents of other clusters. In this paper, the implementation of document clustering in distributed environment based on peer to peer network architecture is reviewed. The documents in local site are clustered using K-means algorithm. Hierarchical clustering is obtained when clusters in each peer combine to form the
next level of cluster. This process repeats until a global cluster is formed and is made available in all the peers. These clustered documents find its application in search engines.

References

- Jiawei Han and Micheline Kamber, "Data Mining Concepts and techniques," Second Edition
- Hinrich Schiitze, Craig Silverstein "Projections For Efficient Document Clustering," Xerox Palo Alto Research Centre
- Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey "Scatter/Gather: A Cluster based Approach to Browsing Large Document Collections."
- Debnaz Deb, M. Muztaba Faud and Rafal A. Angryk, "Distributed Hierarchical Document Clustering," Montana State University, Bozeman, MT 59717, USA.

Index Terms

Computer Science
Distributed Computing

Keywords
Clustering
Distributed Knowledge Discovery
K-means Algorithm
Supernodes
Intercluster
Intracluster