Abstract

In day today life huge amount of electronic data is generated from various resources. Such data is literally large and not easy to work with for storage and retrieval. This type of data can be treated with various efficient techniques for cleaning, compression and sorting of data. Preprocessing can be used to remove basic English stop-words from data making it compact and easy for further processing; later dimensionality reduction techniques make data more efficient and specific. This data later can be clustered for better information retrieval. This paper elaborates the various dimensionality reduction and clustering techniques applied on sample dataset C50test of 2500 documents giving promising results, their comparison and better approach for relevant information retrieval.

References

1. V. Srividhya, R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization
Efficient Information Retrieval through Comparison of Dimensionality Reduction Techniques with Clustering Approach

2. Nguyen Hung Son, "Data Cleaning and Data Preprocessing".
3. Lei Yu Binghamton University, Jieping Ye, Huan Liu, Arizona State University, “Dimensionality Reduction for data mining-Techniques, Applications and Trends”.
6. Rui Tang, Simon Fong, Xin-She Yang, Suash Deb, "Integrating nature-inspired optimization algorithms to k-means clustering", 978-1-4673-2430-4/12/$31.00 ©2012 IEEE.
7. Carlos Cobos, Henry Muñoz-Collazos, Richard Urbano-Muñoz, Martha Mendoza, Elizabeth León, Enrique Herrera-Viedma "Clustering Of Web Search Results Based On The Cuckoo Search Algorithm And Balanced Bayesian Information Criterion " ELSEVIER Publication, 2014 Elsevier Inc. All rights reserved, 21 May 2014.
11. https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

Index Terms

Computer Science
Information Sciences

Keywords

High Dimensional Datasets, Dimensionality reduction, SVD, PCA, Clustering, K-means.