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"
Index Terms

Computer Science  Information Sciences

Keywords

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