Abstract

Clustering the documents based on similarity of words and searching the text is major search procedure and widely used for large set of documents. Documents can be clustered using many clustering algorithms such as Nearest Neighbor, K-Means, Hierarchical, Graph Theoretic etc [4] [5] [7]. The performance measurement in terms of space complexity and execution time and searched output in terms of accuracy and redundancy of these algorithms is a needful study [3]. This paper mainly focuses on performance measurement of Nearest Neighbor, K-Means and Hierarchical agglomerative clustering algorithms on text documents as well as compares them in terms of space complexity, execution time, accuracy and redundancy. In particular, preprocess the input text document and convert it into the document graph represented in the form of matrix. Then convert that document graph into relation matrix which gives relation (similarity score) among all the nodes from 0 to 1 [2]. Implementation and the results of applied clustering algorithms (Nearest Neighbor, K-Means and Hierarchical agglomerative) on documents are discussed and implemented here.

References

- Sholom Weiss, Brian White and Chidanand Apte, "A Lightweight Document Clustering", IBM T. J. Watson Research Centre NY10598, USA.
Performance based Analysis and Comparison of Multi-Algorithmic Clustering Techniques

- A. K. Jain, Michigan State University, M. N. Murthy, Indian Institute of Science and P. J. Flynn, The Ohio State University, "Data Clustering: A Review".

Index Terms

Computer Science

Artificial Intelligence

Keywords
Analysis And Comparison Of K-means Nearest Neighbor Agglomerative Hierarchical Document Graph. Clustering Algorithm