ABSTRACT

In this electronic age, increasing number of organizations are facing the problem of explosion of data and the size of the databases used in today's enterprises has been growing at exponential rates. With the increase in the number of electronic documents, it is hard to organize, analyze and present these documents efficiently by putting manual effort . These have brought challenges for the effective and efficient organization of text documents automatically. For this, document clustering, an unsupervised machine learning approach is used.

Document clustering, one of the traditional data mining techniques, is an unsupervised learning paradigm where clustering methods try to identify inherent grouping of the text documents.The importance of document clustering emerges from the massive volumes of textual documents created. Also, with more and more development of information technology, data set in many domains is reaching beyond peta-scale; making it difficult to work with the document clustering algorithms in central site and leading to the need of increasing the computational requirements. The concept of distributed computing thus; is explored for document clustering giving rise to distributed document clustering. Here, we propose distributed document clustering usingHadoop and MapReduce. We implemented K- means and tested on single node and then modified the map, reduce functions to run over cluster of three machines. We tested on two datasets consisting of 20000 documents (20-NewsGroups) and 21578 documents (Reuters-21578). The results show that **timing requirement for clustering reduces with addition of nodes in the cluster.**