Document Clustering Through Non-Negative Matrix Factorization: A Case Study of Hadoop for Computational Time Reduction of Large Scale Documents.

Bishnu Prasad Gautam and Dipesh Shrestha

● **Abstract**

In this paper we discuss a new model for document clustering which has been adapted using non-negative matrix factorization during our research. The key idea is to cluster the documents after measuring the proximity of the documents with the extracted features. The extracted features are considered as the final cluster labels and clustering is done using cosine similarity which is equivalent to k-means with a single turn. An application was developed using apache lucene for indexing documents and mapreduce framework of apache hadoop project was used for parallel implementation of kmeans algorithm from apache mahout project. Since experiments were carried only in one cluster of Hadoop, the significant reduction in time was obtained by mapreduce implementation when clusters size exceeded 9 i.e. 40 documents averaging 1.5 kilobytes. Thus it's concluded that the feature extracted using NMF can be used to cluster documents considering them to be final cluster labels as in kmeans, and for large scale documents the parallel implementation using mapreduce can lead to reduction of computational time.

● **Key words**

Document Clustering
Non-negative Matrix
Map Reduce