

SEMANTICS BASED DISTRIBUTED DOCUMENT CLUSTERING: PROPOSAL

NEEPA SHAH¹ & SUNITA MAHAJAN²

¹Assistant Professor, Department of IT, DJSCOE, Vile Parle (West), Mumbai, Maharashtra, India

²Principal, Institute of Computer Science, MET, Bandra (West), Mumbai, Maharashtra, India

ABSTRACT

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 the high performance document clustering using Hadoop and MapReduce. Hadoop is open source software framework to support data intensive distributed applications and MapReduce is parallel programming technique. In this proposed model, we also introduce high level of semantics for better accuracy and quality of document clustering.

KEYWORDS: Document Clustering, Distributed Document Clustering, Parallel Document Clustering, Semantic Document Clustering, Hadoop, MapReduce