Abstract:

Bottleneck issues handled in the field of information retrieval are Query analysis and Data storage management. Hadoop is a large scale environment that is supported with larger storage and faster processing. Even though, it suffers from these challenging issues while the number of information requesters is higher. This research work addresses these two bottleneck issues in Hadoop by retrieving the information with the design of Query Analysis and Ontology based Clustering (QAOC) architecture. The research work Novel Query Analysis and Ontology based Clustering for data management in Hadoop consists of four stages: Query Management; User Management; Data Management and Data Retrieval. In the Query Management, the query are preprocessed and consolidated using Enhanced Krovetz (EK) Stemmer and Yet Another Suffix Stripper (YASS) stemmer which eliminates unwanted words from the query and refines the query. Then in User Management, Neuro-Fuzzy Logic based query scheduler is used to handle the multiple queries from multiple users. In Data Management, data which are stored in HDFS are organized in cluster like structure using Ontology which facilitate the query result accuracy. The query execution time is considerable reduced by the use of Binary tree like structure used in Data Retrieval process and lastly the relevant results are ranked using Okapi BM25 and delivered to user. This QAOC architecture is conduct experiment on Hadoop 2.7 and the results are compared in terms of execution time, processing speed and memory consumption.