Abstract

Mining frequent itemsets from massive datasets is always being a most important problem of data mining. Apriori is the most popular and simplest algorithm for frequent itemset mining. To enhance the efficiency and scalability of Apriori, a number of algorithms have been proposed addressing the design of efficient data structures, minimizing database scan and parallel and distributed processing. MapReduce is the emerging parallel and distributed technology to process big datasets on Hadoop Cluster. To mine big datasets it is essential to re-design the data mining algorithm on this new paradigm. In this paper, we implement three variations of Apriori algorithm using data structures hash tree, trie and hash table trie i.e. trie with hash technique on MapReduce paradigm. We emphasize and investigate the significance of these three data structures for Apriori algorithm on Hadoop cluster, which has not been given attention yet. Experiments are carried out on both real life and synthetic datasets which shows that hash table trie data structures performs far better than trie and hash tree in terms of execution time. Moreover the performance in case of hash tree becomes worst.
References


Index Terms

Computer Science Algorithms

Keywords

Frequent Itemset Mining; Apriori; Hash Tree; Trie; Hadoop; MapReduce