FiDoop-DP: Data Partitioning in Frequent Itemset Mining on Hadoop Clusters

ABSTRACT:

Traditional parallel algorithms for mining frequent itemsets aim to balance load by equally partitioning data among a group of computing nodes. We start this study by discovering a serious performance problem of the existing parallel Frequent Itemset Mining algorithms. Given a large dataset, data partitioning strategies in the existing solutions suffer high communication and mining overhead induced by redundant transactions transmitted among computing nodes. We address this problem by developing a data partitioning approach called FiDoop-DP using the MapReduce programming model. The overarching goal of FiDoop-DP is to boost the performance of parallel Frequent Itemset Mining on Hadoop clusters. At the heart of FiDoop-DP is the Voronoi diagram-based data partitioning technique, which exploits correlations among transactions. Incorporating the similarity metric and the Locality-Sensitive Hashing technique, FiDoop-DP places highly similar transactions into a data partition to improve locality without creating an excessive number of redundant transactions.

SHIELD TECHNOLOGIES,
2232, 3RD FLOOR, 16TH B CROSS, YELAHANKA NEW TOWN, BANGALORE-64
Mail us: shieldtechnoblr@gmail.com / manager@shieldtechno.com
Contact: 9972364704 / 8073744810