FiDoop: Parallel Mining of Frequent Item sets Using Map Reduce

Existing parallel mining algorithms for frequent item sets lack a mechanism that enables automatic parallelization, load balancing, data distribution, and fault tolerance on large clusters. As a solution to this problem, we design a parallel frequent item sets mining algorithm called FiDoop using the Map Reduce programming model. To achieve compressed storage and avoid building conditional pattern bases, FiDoop incorporates the frequent items ultra metric tree, rather than conventional FP trees. In FiDoop, three Map Reduce jobs are implemented to complete the mining task. In the crucial third Map Reduce job, the mappers independently decompose item sets; the reducers perform combination operations by constructing small ultra metric trees, and the actual mining of these trees separately. We implement FiDoop on our in-house Hadoop cluster. We show that FiDoop on the cluster is sensitive to data distribution and dimensions, because item sets with different lengths have different decomposition and construction costs. To improve FiDoop's performance, we develop a workload balance metric to measure load balance across the cluster's computing nodes. We develop FiDoop-HD, an extension of FiDoop, to speed up the mining performance for high-dimensional data analysis. Extensive experiments using real-world celestial spectral data demonstrate that our proposed solution is efficient and scalable.