A load-balanced distributed parallel mining algorithm

游坤明,Jiayi Zhou,Tzung-Pei Hong,Jia-Ling Zhou
Computer Science & Information Engineering
Computer Science and Informatics
yu@chu.edu.tw

Abstract

Due to the exponential growth in worldwide information, companies have to deal with an ever growing amount of digital information. One of the most important challenges for data mining is quickly and correctly finding the relationship among data. The Apriori algorithm has been the most popular technique in finding frequent patterns. However, when applying this method, a database has to be scanned many times to calculate the counts of a huge number of candidate itemsets. Parallel and distribute computing is an effective strategy for accelerating the mining process. In this paper, the Distributed Parallel Apriori algorithm (DPA) is proposed as a solution to this problem. In the proposed method, metadata are stored in the form of Transaction Identifiers (TIDs), such that only a single scan to the database is needed. The approach also takes the factor of itemset counts into consideration, thus generating a balanced workload among processors and reducing processor idle time. Experiments on a PC cluster with 16 computing nodes are also made to show the performance of the proposed approach and compare it with some other parallel mining algorithms. The experimental results show that the proposed approach outperforms the others, especially while the minimum supports are low.

Keyword：Parallel and distributed processing, Cluster computing, Frequent patterns, Association rules, Data mining