基于 MapReduce 和矩阵的频繁项集挖掘算法

周国军，龚榆桐

（玉林师范学院 数学与信息科学学院，广西 玉林 537000）

摘 要： 为了能高效地从大数据集中挖掘所有频繁项集，提出了一种基于 MapReduce 和矩阵的频繁项集挖掘算法.该算法将事务数据库转化为矩阵，按照垂直划分方法将矩阵分解成多个规模相同的子矩阵，根据 MapReduce 模型将子矩阵分配给计算机集群的节点，各节点并行对子矩阵计算候选项集的支持度.该算法在执行过程中产生的通信量较少，实现了节点计算任务的负载平衡.在 Hadoop 平台上测试了算法的性能，实验结果表明该算法具有较好的加速比和可扩展性，适合对大数据集挖掘频繁项集.

关键词： MapReduce；Hadoop 平台；矩阵；频繁项集；关联规则

Frequent Itemsets Mining Algorithm Based on MapReduce and Matrix

ZHOU Guo-jun，GONG Yu-tong

(School of Mathematics and Information Science, Yulin Normal University, Yulin 537000,China)

Abstract： To efficiently find all frequent itemsets from large data sets, a frequent itemsets mining algorithm based on MapReduce and matrix is proposed. The algorithm includes the following points: Transaction database is transformed into matrix, then the matrix is divided into multiple submatrices by vertical partitioning method. According to MapReduce model, submatrices are assigned to the nodes of cluster, and the support of candidate itemsets is parallelly computed by the nodes. The amount of communication generated during the execution of the algorithm is small. Load balance among nodes is achieved . The performance of the algorithm is tested on Hadoop platform. Experimental results show that the algorithm has good speedup and scalability, which is suitable for mining frequent itemsets from large data sets.

Key words： MapReduce; Hadoo platform; matrix; frequent itemsets; association rules

作者简介：

周国军 男，（1975-），硕士，讲师. 研究方向为数据挖掘、云计算. E-mail：ylsyzgj@126.com.

龚榆桐 男，（1982-），硕士，副教授. 研究方向为计算机应用、电子商务.