基于 Hadoop 的大数据频繁模式挖掘算法

李校林 1,2，杜托 1，谢勇 1
（1 重庆邮电大学 通信新技术应用研究中心，重庆 400065；2 重庆信科设计有限公司，重庆 400021）

摘要： 针对传统的频繁模式挖掘算法不能满足大数据环境下的挖掘需要，提出一种高效挖掘大型数据库中频繁模式的并行算法 H_PrePost.首先从压缩数据库、简化数据表示以及采用高效的连接和剪枝策略等方面对 PrePost 算法进行改进，以提高单机模式下的挖掘效率.然后将改进算法迁移到 Hadoop 平台上，利用 MapReduce 模型进行并行计算，同时提出一种负载均衡策略保证集群高效运行.最后使用 kulczynski 度量和不平衡比对所挖掘的频繁模式进行评估，以确保所挖掘模式具有实际应用价值.实验结果表明，H_PrePost 算法可以有效挖掘大数据集中的频繁模式.

关键词： Hadoop；频繁模式；大数据

Algorithm for Mining Frequent Patterns in Big Data Based on Hadoop

LI Xiao-lin1,2，DU Tuo1，XIE Yong1
(1 Chongqing University of Posts and Telecommunications, New Technology Application Research Center, Chongqing 400065, China; 2 Chongqing Information Technology Designing Co. Ltd, Chongqing 400021, China)

Abstract： Aiming at the traditional frequent pattern mining algorithm can not meet the needs of mining in big data environment, a parallel algorithm for efficiently mining frequent patterns in large databases is proposed. Firstly, PrePost algorithm is improved from compressing database, simplifying data representation and using efficient connection and pruning strategy,which improve the efficiency of mining in stand-alone mode. Then, the improved algorithm is migrated to the Hadoop platform and the MapReduce model is used for parallel computing. A load balancing strategy is proposed to ensure the efficient operation of the cluster. Finally, the frequent pattern mining is evaluated using kulczynski metric and unbalance ratio to ensure that the mining pattern has practical value. Experimental results show that this algorithm can effectively mine the frequent patterns in big data sets.

Key words： Hadoop; frequent pattern; big data

作者简介：
李校林男，(1968-)，高级工程师，硕士生导师.研究方向为大数据、移动通信.
杜托（通讯作者）男，(1993-)，硕士研究生.研究方向为大数据、数据挖掘.E-mail:dutuotuo@yeah.net.
谢勇男，(1994-)，硕士研究生.研究方向为分布式计算、数据挖掘.