基于 N-list 的并行频繁项集挖掘算法

陈　奇，张曦煌

（江南大学 物联网工程学院，江苏 无锡 214122）

摘　要：N-list 是近几年提出来的一种新的数据结构，它在频繁项集挖掘中有很高的效率.本文基于N-list提出了一种新型的并行频繁项集挖掘算法PPF算法.该算法通过扫描数据库创建一颗 PPC-tree 树，利用 PPC-Tree 树生成一系列 N-list，将 N-list 数据表项分配到不同的节点进行深度挖掘,最后汇总所有节点的结果挖掘出所有的频繁项集.本文在四种不同的数据集上对 PPF 算法就行了测试和分析,实验结果表明在任何数据集上 PPF 算法的运行速度都是最优的.

关键词：数据挖掘；频繁项集挖掘；并行；N-list

An N-list based Parallel Algorithm for Mining Frequent Itemsets

CHEN Qi，ZHANG Xi-huang

(School of Internet Things Engineering Jiangnan University, Wuxi 214122, China)

Abstract：N-list is a novel data structure proposed in recent years. It has been proven to be very efficient for mining frequent itemsets. In this paper, we present PPF, a new parallel algorithm for mining frequent itemsets. The algorithm directly scans dataset to construct a PPC-Tree. Then, the algorithm uses PPC-Tree to generate a series of N-lists which will be assigned to different nodes to mining frequent itemsets. We have conducted extensive experiments to evaluate PPF against PrePost algorithm on four various real datasets. The experimental results show that PPF algorithm is always the fastest one on all datasets.

Key words：data mining; frequent itemset mining; parallel; N-list

作者简介：

陈　奇　　男，(1991-)，硕士研究生.研究方向为数据挖掘.

E-mail：876447590@qq.com.

张曦煌　　男，(1962-)，博士，教授.研究方向为数据挖掘、嵌入式系统.