

doi:10.1631/FITEE.1800467

题目：基于条件模式的一种无分组并行频繁模式挖掘算法

概要：频繁项集挖掘是关联规则挖掘的主要方法。由于计算空间和性能限制，特别是当数据集剧增时，挖掘频繁项的关联需要大量时间和资源。在大数据环境下的关联数据挖掘过程中，通常采用MapReduce模型进行任务划分及并行处理，从而提高算法执行效率。为确保关联规则在任务划分和并行处理期间不被破坏，需要将内部关系数据存储在计算机空间中。与原始数据集相比，存储冗余的内部关系数据将显著增加空间的使用。研究发现，频繁模式挖掘算法的形成主要依赖于条件模式基。基于并行频繁模式（PFP）算法理论，本文提出一种无分组的PFP（NG-PFP）挖掘算法。该算法取消了分组模式，减少了子任务之间的数据冗余。实验结果表明，无分组模型在计算效率和空间利用率方面都有显著提高。

关键词：频繁模式挖掘；并行算法；条件模式基；MapReduce；大数据