Tidset-based Parallel FP-tree Algorithm for the Frequent Pattern Mining Problem on PC Clusters

周嘉奕,游坤明
Computer Science & Information Engineering
Computer Science and Informatics
yu@chu.edu.tw

Abstract

Mining association rules from a transaction-oriented database is a problem in data mining. Frequent patterns are essential for generating association rules, time series analysis, classification, etc. There are two categories of algorithms for data mining, the generate-and-test approach (Apriori-like) and the pattern growth approach (FP-tree). Recently, many methods have been proposed for solving this problem based on an FP-tree as a replacement for Apriori-like algorithms, because these need to scan the database many times. However, even for the pattern growth method, the
execution time takes long when the database is large or the given support is low. Parallel- distributed computing is good strategy for solving this problem.
Some parallel algorithms have been proposed, however, the execution time increases rapidly when the database increases or when the given minimum threshold is small. In this study, an efficient parallel- distributed mining algorithm based on an FP-tree structure - the Tidset-based Parallel FP-tree (TPFP-tree) - is proposed. In order to exchange transactions efficiently, transaction identification set (Tidset) was used to directly choose transactions without scanning databases. The algorithm was verified on a Linux cluster with 16 computing nodes. It was also compared with a PFP-tree algorithm. The dataset generated by IBM's Quest Synthetic Data Generator to verify the
performance of algorithms was used. The experimental results showed that this algorithm can reduce the execution time when the database grows. Moreover, it
was also observed that this algorithm had better scalability than the PFP-tree.

Keyword : frequent pattern mining, parallel processing, association rule, data
mining, tidset