Abstract

In the last two decades association rules has found its application in various domains. Due to the wide uses of association rules, study of association rule mining has drawn the attention of many researchers. As a result many algorithms have been proposed to efficiently mine association rules, and many of those algorithms greatly rely on available physical memory. Number of database scan required and limited physical memory are the two main challenges faced by these algorithms. FP-Growth is one of them which is found to be efficient from database scan point of view but the limited physical memory is a serious bottle neck for it. So, it is necessary to develop new method that does not fully rely on physical memory; new methods that utilize the secondary storage in the mining process are to be found. In this report we propose a new approach, (Secondary storage Based Frequent Pattern) SBFP-Growth, which is a modification to the FP-Growth algorithm. SBFP-Growth uses secondary storage to store the tree, and so it overcomes the main memory bottleneck at FP-Growth algorithm. This way, we are able to mine for frequent itemsets from databases of arbitrary sizes without being restricted by the available physical memory.

Moreover SBFP-Growth construct complete tree, i.e. a tree with minimum support count equal to one. Complete tree has the advantage that it provides the freedom of mining for lower minimum support values without the need to reconstruct the tree for different minimum supports. We can store the SBFP-Tree in secondary storage so that it can be mined latter in any period of time and with any minimum support value; this is useful as the time to rebuild the SBFP-tree is saved. It is observed from the results that SBFP-Tree file size is less than market basket dataset and hence we can say that SBFP-Tree provides compression to market basket datasets.

Keywords: Association rule mining,FP-Growth, SBFP-Growth, Complete tree.