IPFP: An Improved Parallel FP-Growth Algorithm for Frequent Itemsets Mining

Dawen. Xia*
Southwest University, Chongqing, China
Guizhou Minzu University, Guiyang, China gzmdxdw@swu.edu.cn

Zili. Zhang*
Southwest University, Chongqing, China
Deakin University, Victoria, Australia zhangzl@swu.edu.cn

Yanhui. Zhou
Southwest University, Chongqing, China xiaohui@swu.edu.cn

Zhuobo. Rong
Southwest University, Chongqing, China zhuobo@swu.edu.cn

As an important part of discovering association rules, frequent itemsets mining plays a key role in mining associations, correlations, causality and other important data mining tasks. Since some traditional frequent itemsets mining algorithms are unable to handle massive small files datasets effectively, such as high memory cost, high I/O overhead, and low computing performance, we propose an improved Parallel FP-Growth (IPFP) algorithm and discuss its applications in this paper. In particular, we introduce a small files processing strategy for massive small files datasets to compensate defects of low read/write speed and low processing efficiency in Hadoop. Moreover, we use MapReduce to implement the parallelization of FP-Growth algorithm, thereby improving the overall performance of frequent itemsets mining. The experimental results show that the IPFP algorithm is feasible and valid with a good speedup and a higher mining efficiency, and can meet the rapidly growing needs of frequent itemsets mining for massive small files datasets.

Key Words: Frequent itemsets mining, Hadoop MapReduce, Parallel FP-Growth, Small files problem