Name: Phu Pham

CSC 240

Project 1: Frequent Item Set Mining

Report

*6.7 (Implementation project) Using a programming language that you are familiar with, such as C++ or Java, implement three frequent itemset mining algorithms introduced in this chapter: (1) Apriori (2) FP-growth (3) One Improvement of choice for Apriori.*

*Compare the performance of each algorithm with various kinds of large data sets. Write a report to analyze the situations (e.g., data size, data distribution, minimal support threshold setting, and pattern density) where one algorithm may perform better than the others, and state why.*

* Apriori algorithm: Run-time 120.147

This is the slowest algorithm because it does not store temporary information and has to scan the database for every iteration.

* Improved Apriori: Run-time 0.604

For an improved version of Apriori algorithm, a hashtable was used to map the item ID with its count so the algorithm only need to check the hashtable instead of consistently check the entire database. This is much quicker than Apriori.

* FP-Growth: Run-time 0.383

This is the fastest algorithm since it scans the database only twice and is able to transverse the tree in the process of generating frequent patterns.

References:

Yutong He, Frequent Pattern Mining, open-source GitHub repository, <https://github.com/KellyYutongHe/Frequent-Pattern-Mining>.