By: Joseph Jourekian & Joseph El-Ghaname

Mining Frequent Itemsets

Project I | COMP 4250 | Dr. Samet

Abstract

Contents of this Report

Summary

This report consists of graphical results that provide a basis for comparison of CPU time used when using A-Priori and PCY algorithms for mining frequent itemset. The scalability study was applied on this dataset using different sample sizes including: 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. Also, within each sample size, the support threshold variable was set to the following numbers: 1%, 5%, and 10%. The graphs on the following pages visually display this data and the correlation with the efficiency of each algorithm according to their CPU times.

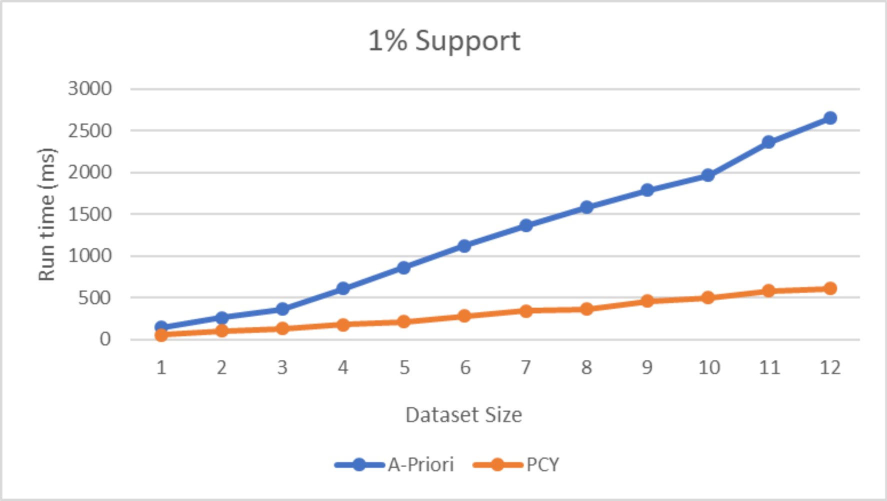
Method 1: A-Priori Algorithm

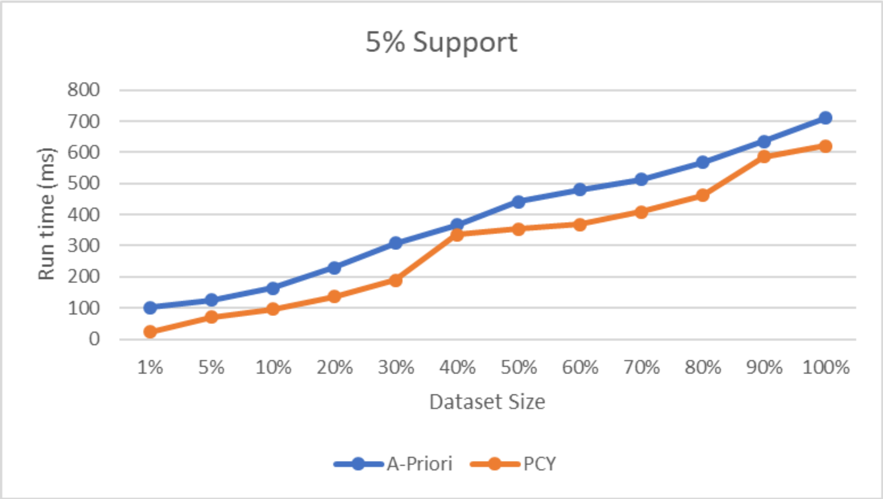
In the A-Priori algorithm for searching for frequent pairs, the method first starts off with *Pass 1*, in which baskets are read and counted into main memory, including the occurrences of each individual item. Then, in *Pass 2*, the baskets are read again into main memory, except this time, the algorithm searches for pairs within the same basket, that are both frequent – using information about frequency from *Pass 1*. The frequent pairs are then accumulated according to this process in A-Priori.

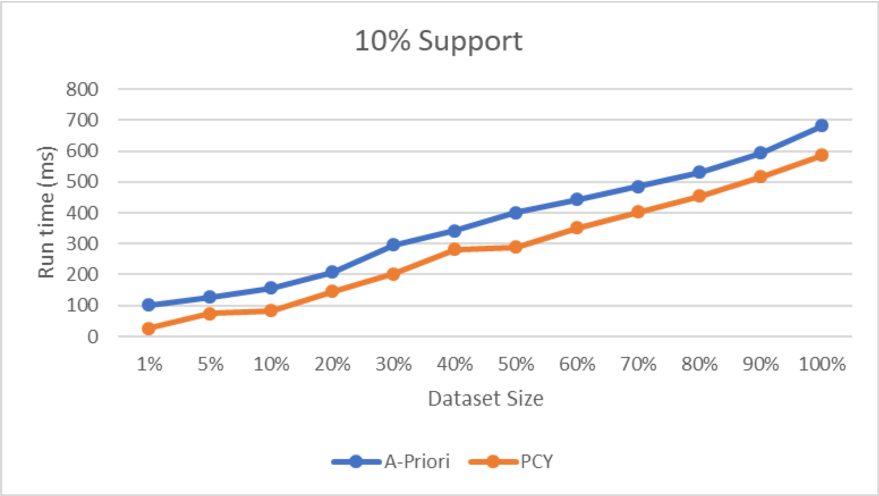
Method 2: PCY Algorithm

In the PCY algorithm, in *Pass 1*, each basket is read, and within that loop, each item in the basket is also read, while the item's frequency is incremented. Then within the basket itself, each pair of items are hashed into a bucket using a hash function. Each time a pair is hashed to the same bucket, that hash function frequency is incremented. The hash buckets are then converted into bit-vectors (or bitmaps), in which each bucket is either frequent (1) or non-frequent, according to a support threshold. Then, in *Pass 2*, all pairs {i, j} are gathered, and if they both are frequent and hash to a frequent bucket, then they are considered a frequent pair.

Results







These results were calculated using the hash function: hashValue = (itemI + itemJ) mod 20000

Specifications of the Laptops Used

Joseph Jourekian’s Asus X555UA

Operating system: Windows 10 Education

CPU: 2.5GHz Intel core i7-6500U

RAM: 12GB

Joseph El-Ghaname’s MacBook

Operating system: macOS Mojave

CPU: 2.3GHz Intel core i5

RAM: 8GB