Project 2

Frequent Item-set Mining

Date: 13th November ,2019

Algorithm: Optimized FP-Growth

Name	Student ID
Nikunj Arora	40104832
Sarvesh Vora	40081458
Shivam Nautiyal	40090841

Algorithm:

DataSet(baskets) = read Input file
ItemCount = For each distinct item in the *DataSet,* store the number of occurrences.
Pruning (1) = Remove all the items from *ItemCount,* which has count less than Minimum support (Threshold).
Sort the *ItemCount* in the descending order with respect to the count.
Generate Header Table,
    For each item in *ItemCount* create an entry in header table with a pointer, pointing towards null.
Pattern = A descending pattern of the items based on their count.
Rearrange the *DataSet* depending on the *Pattern* and if the item isn't present in the *Pattern*, discard the item from the *DataSet*.
Build FP-Tree,
    Create a root node with its Parent node as null
    For each basket of the *DataSet,*
        previous = root;
        children = children of previous
        for each item in basket
            if item is a child of previous
                increment the count of child
            else
                Create a new node and attach it to the next node for the item in header table
        previous = item
        children = children of item
Generate Frequent Patterns,
    For each Item in the Header Table,
        For each occurrences of the item in the Tree,
            find the path p, from parent of the item to the child of the root node.
            Add all the nodes in path p into the FreqentItemList of item.
            Remove all items from FreqentItemList which has count less than threshold.
            Generate all the combinations of all the nodes in the path with the item and set
            The count as the item count
    Add all generated patterns to the Frequent Patterns.
    Remove all paths from Frequent Patterns which has count less than threshold.
    Store all Patterns in the file.

Test Cases:

FileName	Type	# Baskets	Threshold	Time(Seconds)
proj2sample2.txt	Sparse	200	5	0.137
500_150.txt	Dense	500	150	0.48
Sparse_2500_50.txt	Sparse	2500	50	0.38
Sparse_25000_500.txt	Sparse	25000	500	3.652
Dense_2500_1000.txt	Dense	2500	1000	4.029
Dense_25000_20000.txt	Dense	25000	20000	267.703