Roya Feizi

Project report

For this project, I've implemented a version of Apriori which usese Dynamic Itemset Counting. This method is mentioned in : S. Brin R. Motwani, J. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. SIGMOD'97

This method is used to reduce number of DB scans. Apriori scans the DB equal to length of the frequent itemset with maximum length. In DIC we use a chunk of data of size M , when we reach to M'th transaction we start counting the itemsets of length n+1 which are immediate supersets of frequent itemsets of size n so far. So it is important to choose an optimal M for any dataset that we are working on. Choosing smaller M will lead to counting more candidates in each step. However we might not choose the M to very low number , because after reading a low number of transactions there are no frequent candidates of length n, to be used for producing the candidates of size n+1.

I checked different values for M in adult data set and the result is as follows for a fixed minimum support equal to 20% :

Minimum Support=20%

M	DB Scans for DIC	DB scans for Apriori
Less than 25000	6	8
30000 and more	8	9

For the adult dataset the optimal value for M is 2500 . if the dataset was more intense , itemsets became frequent after a couple of transactions and M will get a smaller value in more intense datasets.

Another important factor is the choose of minimum support. I checked different minimum supports for the data set with fixed value for M. the result is that if minimum support is large, for instance 50% or more, maximum length of frequent itemsets becomes low. So DB scans for Apriori and DIC are not much different. But when we reduce the minimum support to say, 10% or 20% maximum length for frequent itemsets increases and we can see that DB scans are different.(recall that Apriori scans the DB equal to maximum length of frequent itemsets).

At the table below I've listed the result with different minimum supports. For fixed value of M .

M=25000

Minimum Support	DB Scans for DIC	DB scans for Apriori
10%	7	9
30%	5	6
50%	4	5
70%	3	3

As well as the M and minimum support , size of the dataset is also important DIC increases the efficiency when dataset is larger . (adult data set is not that much large) .

For running the code, just run the Apriori.java . adult.data should be placed in the same place. You can choose different values for M and minimum support in the main method.