Javier-EchavarrenSuarez-HW3

1.1 Tan, Ch. 5 (Association Analysis)

Questions 15

15. Answer the following questions using the data sets shown in Figure 5.34. Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the *Apriori* algorithm to extract frequent itemsets with *minsup* = 10% (i.e., itemsets must be contained in at least 1000 transactions).

(a) Which data set(s) will produce the most number of frequent itemsets?

To solve this problem, we start knowing that an itemset with k items can generate $2^k - 1$ itemsets, and the number of frequent itemsets in a dataset where m number of items are simultaneously present in n number of rows, (provided m/n ≥ minsup) is $2^l - 1$.

That's why we are looking for the dataset(s) that have the most number of items present in the same share of the data.

The dataset that clearly has the most number of items present in the same share of the data is the (e) of the figure 4.34, as most of the items from 0 to 800 are present in the 2000-4000 transaction range (20% of the data), having approximately $2^{800} - 1$ frequent itemsets.

(b) Which data set(s) will produce the fewest number of frequent itemsets?

The dataset with the fewest number of frequent itemsets is the (d), because it doesn't have items present in 10% of the samples (minsup), and that's why it has no itemsets apart from {null}.

(c) Which data set(s) will produce the longest frequent itemset?

The longest frequent itemset will be produced in the (e) dataset, because in the transactions 2000-4000 there are approximately 800 items present in the same transactions, and these transactions are approximately 20% of the dataset. This doesn't happen in the other datasets.

(d) Which data set(s) will produce frequent itemsets with highest maximum support?

The (b) dataset will produce frequent itemsets with highest maximum support, as it has the longest vertical line of all charts.

(e) Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%)?

The (e) dataset will produce frequent itemsets containing items with wide-varying support levels, as it contains items with a lot of support (the ones where there are two vertical lines or a big vertical line) and also has other items that are present in less transactions, such as the ones of one little vertical line.

1.2 Zaki, Chapter 8 (Frequent Pattern Mining)

Questions 1(a), 4

1(a) Using minsup = 3/8, show how the Apriori algorithm enumerates all frequent patterns from this dataset.

tid	itemset
t_1	$ABCD$
t_2	$ACDF$
t_3	$ACDEG$
t_4	$ABDF$
t_5	BCG
t_6	DFG
t_7	ABG
t_8	$CDFG$

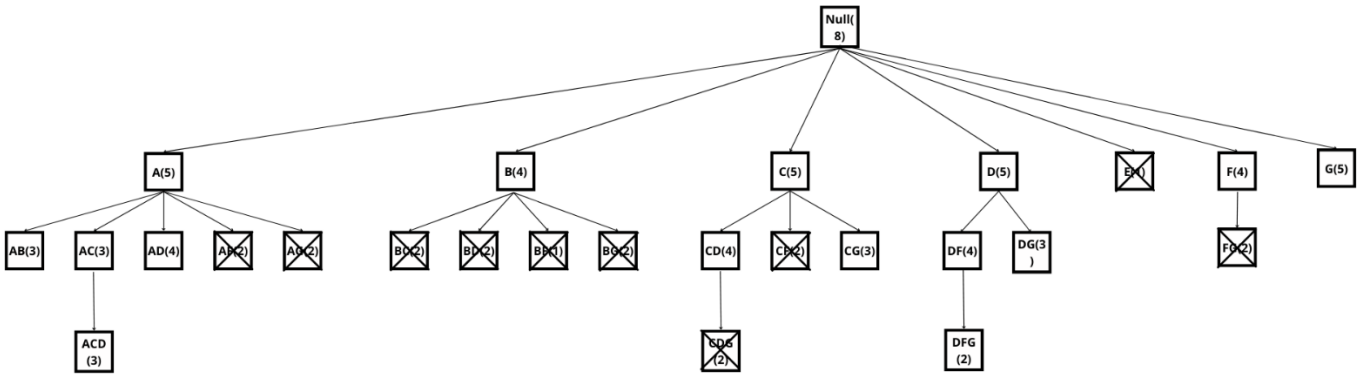

Figure 1: Apriori Tree

The frequent patterns of this dataset (minsup=3/8) are:

1 Null
2 A
3 B
4 C
5 D
6 F
7 G
8 AB
9 AC
10 AD
11 CD
12 CG
13 DF
14 DG
15 ACD
16 DFG

4. Given the database in Table 8.4. Show all rules that one can generate from the set ABE

Table 8.4. Dataset for Q4

tid	itemset
t_1	ACD
t_2	BCE
t_3	$ABCE$
t_4	BDE
t_5	$ABCE$
t_6	$ABCD$

The rules that one can generate from the database, the support of its subsets and rules confidences are shown in the following table.

Itemset	Support	Rule	Confidence
		A->B	0,75
AB	3	B->A	0,6
		A->E	0,5
AE	2	E->A	0,5
		B->E	0,8
BE	4	E->B	0,8
		AB->E	0,666
		AE->B	1
		A->BE	0,5
		B->AE	0,4
ABE	2	E->AB	0,5