CA2 Assignment

ANKITA DUTTA, 11500221040, CSE B, Data Warehousing & Data Mining (PECIT602B)

Q1. Transactional dataset:

Transactional ID	Itemset
T_1	a, b, c
T_2	a, b, d
T_3	b, c
T_4	a, c, e
T_5	a, b, c, d
T_6	a, c, d
T_7	b, c, e
T_8	a, b, e
T_9	b, c, d

a) Apply ECLAT algorithm to identify frequent pattern. Minimum support 25%

Ans. Total transactions = 9

Min. support count = 25% of 9 = $\frac{25}{100} \times 9 = \lceil 2.25 \rceil \approx 3$

We can create a table of items and the transactions they are appearing in. Here, 'eligible' implies eligibility for next level clustering. (i.e., is no. of transactions ≥ support count?)

Items	Transactions	Eligible?
a	$T_1, T_2, T_4, T_5, T_6, T_8$	yes
b	$T_1, T_2, T_3, T_5, T_7, T_8, T_9$	yes
c	$T_1, T_3, T_4,$ ~~T~~ T_5, T_6, T_7, T_9	yes
d	T_2, T_5, T_6, T_9	yes
e	T_4, T_7, T_8	yes.

Item set 2:

Items	Transactions	Eligible?
ab	T_1, T_2, T_5, T_8	yes
ac	T_1, T_4, T_5, T_6	yes
ad	T_2, T_5, T_6	yes
bc	T_1, T_3, T_5, T_7, T_9	yes
bd	T_2, T_5, T_9	yes
cd	T_5, T_6, T_9	yes
ae	T_4, T_8	no
be	T_7, T_8	no
ce	T_4, T_7	no
de	Null.	no

Itemset 3:

Items	Transactions	Eligible?
abc	T_1, T_5	no
abd	T_2, T_5	no.
acd	T_5, T_6	no.
bcd	T_5, T_9	no.

Since no more supersets can be built as the current item set doesn't have support count ≥ min support count. ∴ The frequent pattern is –

Items brought	Recommended Products
a	b
a	c
a	d
b	c
b	d
c	d

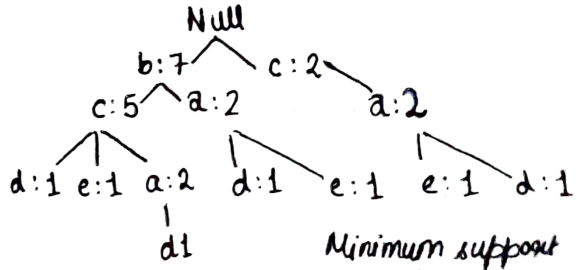
b) Build FP tree to identify frequent pattern. Minimum support count 3.

Ans. Let T_n be no. of transactions where items are part of. Items are in list in decreasing order, and arranging alphabetically if they have same no. of occurance

Items	T_n
b	7
c	7
a	6
d	4
e	3

Transaction ID	Items
T_1	b, c, d
T_2	b, a, d
T_3	b, c
T_4	a, c, e
T_5	b, c, a, d
T_6	c, a, d
T_7	b, c, e
T_8	b, a, e
T_9	b, c, d

From this let us construct the FP tree



Minimum support count 3.

Items	Conditional Pattern Base	Conditional Freq. Pattern Tree	Frequent Pattern
e	{b,a}:1, {b,c}:2, {c,a}:1	x	x
d	{b,c}:1, {b,a}:1, {b,c,a}:1, {c,a}:1	{b}:3	{b,d}:3
a	{b}:2, {c}:2, {b,c}:2	{b}:4	{b,a}:4
c	{b}:5	{b}:5	{b,c}:5
b	—	—	—

:. Frequent Pattern using FP tree : (b,d), (b,a), (b,c)

c) For conditional FP tree, identify the association rules with confidence 70%.

Ans. For FP tree in previous question we get frequent patterns : (b,d), (b,a), (b,c)

- for (b,d) - $\text{confidence}(b \to d) = \frac{\text{support}(b \cup d)}{\text{support}(d)} = \frac{\text{no. of transaction } b,d \text{ together}}{\text{no. of transaction with } b \text{ only}} = \frac{3}{7} \times 100 = 42.85\%$

  $\text{confidencece}(d \to b) = \frac{\text{support}(b \cup d)}{\text{support}(d)} = \frac{3}{4} \times 100\% = 75\%$

- for (b,a) - $\text{confidence}(b \to a) = \frac{\text{support}(b \cup a)}{\text{support}(b)} = \frac{4}{7} \times 100\% = 57.14\%$

  $\text{confidence}(a \to b) = \frac{\text{support}(b \cup a)}{\text{support}(a)} = \frac{4}{6} \times 100\% = 66.67\%$

- for (b,c) - $\text{confidence}(b \to c) = \frac{\text{support}(b \cup c)}{\text{support}(b)} = \frac{5}{7} \times 100\% = 71.42\%$

  $\text{confidence}(c \to b) = \frac{\text{support}(b \cup c)}{\text{support}(c)} = \frac{5}{7} \times 100\% = 71.42\%$

so the associates with confidence 70% : (d→b), (b→c), (c→b)

d) Compute interest measure for the association rules

Ans. Calculating only lift as association measure of interest for association rules.

Frequent patterns : (b,d), (b,a), (b,c)

$$\text{Lift}(b \to d) = \frac{\text{confidence}(b \to d)}{\text{support}(d)} = \frac{42.8}{4/9} = 0.964125$$

$$\text{Lift}(d \to b) = \frac{\text{confidence}(d \to b)}{\text{support}(b)} = \frac{75}{7/9} = 0.964285$$

$$\text{Lift}(a \to b) = \frac{\text{confidence}(a \to b)}{\text{support}(b)} = \frac{66.67}{7/9} = 0.857185$$

$$\text{Lift}(b \to a) = \frac{\text{confidence}(b \to a)}{\text{support}(a)} = \frac{57.14}{6/9} = 0.8571$$

$$\text{Lift}(b \to c) = \frac{\text{confidence}(b \to c)}{\text{support}(c)} = \frac{71.42}{7/9} = 0.918$$

$$\text{Lift}(c \to b) = \frac{\text{confidence}(c \to b)}{\text{support}(b)} = \frac{71.42}{7/9} = 0.918$$

Lift measures how much more frequently the left hand item is found with the right hand item is found without the right

$$\text{confidence } (x \rightarrow y) = \frac{\text{support } (x \cup y)}{\text{support } (x)}$$

$$\text{support } (x) = \frac{\text{no. of transactions with } x \text{ present}}{\text{total no. of transactions}}$$

$$\text{lift } (x \rightarrow y) = \frac{\text{confidence } (x \rightarrow y)}{\text{support } (y)} = \frac{\text{support } (x \cup y)}{\text{support } (x) \cdot \text{support } (y)}$$

---

d) What are the advantage of fp tree algorithm over market basket analysis?

Ans i. Efficiency & Time Complexity : Can be more efficient than apriori algorithm as it reduces the number of passes to 2. First to build the fp tree and 2nd pass to detect frequent pattern from a tree. But also the apriori algo./market basket analysis involves multiple passes, generate candidate item set at each iteration, which takes no computation.

ii. Scalability : due to if its efficiency in memory usage and time complexity, FP growth algo. is more scalable of the 2, so more suitable for large scale datasets

iii. Handles sparse data efficiently : This is because it doesnt generate huge numb of item sets as market basket analysis.

iv. Memory usage efficiency : FP growth algo uses FP tree to compress transactional data into compact structure to reduce memory requirement, whereas in market basket analysis, super dataset is created on each iteration until support count is less than minimum support count, which increases memory usage.
