Unit-2

Association Rule Mining :-

Mining Frequent Patterns :-

Frequent patterns :-

The pattern that appear frequently in dataset.

- patterns include data items, data subsequence, data substructure.

Example : Milk and Bread.

Market Basket Analysis (or) Association Rule Mining :-

This is a criteria of finding (or) determining the items (or) patterns which are likely to be purchased by the customer in single transaction.

It is mainly used for sellers.

Strategies used :

1. Placing them together
2. Placing them at two different ends.

- This analysis will help ~~them~~ sellers to plan their shelf space for increased sales.

- Frequent patterns are represented by association rules

Example : computer and Anti-Virus.

Example

Transaction ID	Items.
1	Bread, Milk.
2	Bread, Diaper, Beer, egg
3	Milk, Diaper, Beer, cola
4	Bread, milk, diaper, Beer
5	Bread, milk, Diaper, cola.

Transactions $T = \{t_1, t_2, t_3, t_4, t_5\}$

Items (Total no. of items) $I = \{i_1, i_2, i_3, i_4, i_5\}$

Item set : Set of items.

{ Bread, milk, Beer } - 3 item set.

Support Count : Freqiently an item set occurs.

σ (Bread, Milk, Beer) = 1

σ (Bread, milk) = 3

Support :-

The ratio of Support Count and no. of transacti -ons.

$$\text{Support} = \frac{\text{Support Count}}{\text{Total no. of transactions}}$$

Confidence :-

The ratio of Support of A U B and Support of A.

$$\text{Confidence} = \frac{\text{Support } (A \cup B)}{\text{Support } (A)}$$

Confidence =

Ex :- {Bread, Milk}

$$\text{Support (Bread)} = \frac{4}{5} = 0.8$$

$$\text{Support (Milk)} = \frac{4}{5} = 0.8$$

$$\text{support (Bread, Milk)} = \frac{3}{5} = 0.6$$

$$\text{confidence} = \frac{\text{Support (Bread, Milk)}}{\text{Support (Bread)}}$$

$$= \frac{0.6}{0.8} = \frac{\frac{6}{10}}{\frac{8}{10}} = \frac{6}{8} = 0.75.$$

Association and Correlation :-

Association :-

It is a way to identifying relation between various items.

Ex: It is used to determine sales of items that are frequently purchased together.

Correlation :-

It is Mathematical technique.

It shows how strongly pair of attributes are related together.

Ex: Tall people tend to have more weight.

Mining Methods:-

Mining Methods :-

There are two types of mining methods. They are:

1. Apriori Algorithm
2. FP Growth Algorithm.

(1) Apriori Algorithm :-

- It is given by R. Agarwal and R. Srikant.
- It shows how objects are associated with each other.

objective : To generate an association.

Example :

Minimum Support = 50%

Threshold Confidence = 70%

T.ID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Item set = (1, 2, 3, 4, 5)

Item set	Support	min support
1	2	2/4 = 50%
2	3	3/4 = 75%
3	3	3/4 = 75%
4	1	1/4 = 25% (x)
5	3	3/4 = 75%

∴ our minimum support is 50%

so, in the above table which is less than 50% it is not considered for next process.

New, item set = (1, 2, 3, 5)

- form pairs

(1,2), (1,3), (1,5), (2,3), (2,5), (3,5).

Itemset	Support	minimum support
(1,2)	1	1/4 = 25% (x)
(1,3)	2	2/4 = 50%
(1,5)	1	1/4 = 25% (x)
(2,3)	2	2/4 = 50%
(2,5)	3	3/4 = 75%
(3,5)	2	2/4 = 50%

(1,2) and (1,5) are eliminated because it is less than minimum support.

New, Item set = (1,3), (2,3), (2,5), (3,5)

- form triples.

(1,2,3), (1,2,5), (1,3,5), (2,3,5)

Item set	Support	Min Support
(1,2,3)	1	1/4 = 25% (x)
(1,2,5)	1	1/4 = 25% (x)
(1,3,5)	1	1/4 = 25% (x)
(2,3,5)	2	2/4 = 50%

Item set = (2, 3, 5).

- Now, lets calculate support and confidence.

Confidence = support (A ∪ B) / support of A

Using (2, 3, 5) we can generate association rules.

Rules	Support	Confidence
$(2 \wedge 3) \rightarrow 5$	2	100%
$(3 \wedge 5) \rightarrow 2$	2	100%
$(2 \wedge 5) \rightarrow 3$	2	66% (x)
$2 \rightarrow (3 \wedge 5)$	2	66% (x)
$5 \rightarrow (2 \wedge 3)$	2	66% (x)
$3 \rightarrow (2 \wedge 5)$	2	66% (x)

$(2 \wedge 3) \rightarrow 5$ (A → B) - Confidence $= \dfrac{\text{Support } (A \cup B)}{\text{Support } (A)}$

$= \dfrac{\text{Support } ((2 \wedge 3) \cup 5)}{\text{Support } (2 \wedge 3)}$

$= \dfrac{2}{2} = 1 = 100\%$

$(3 \wedge 5) \rightarrow 2$ - Confidence $= \dfrac{\text{Support } ((3 \wedge 5) \cup 2)}{\text{Support } (3 \wedge 5)}$

$= \dfrac{2}{2} = 100\%$

$(2 \wedge 5) \rightarrow 3$ - Confidence $= \dfrac{\text{Support } ((2 \wedge 5) \cup 3)}{\text{Support } (2 \wedge 5)}$

$= \dfrac{2}{3} = 66\%$

$2 \rightarrow (3 \wedge 5) \rightarrow$ Confidence $= \frac{\text{Support}(2 \cup (3 \wedge 5))}{\text{Support}(2)}$

$= \frac{2}{3} = 66\%$

$5 \rightarrow (2 \wedge 3) \rightarrow$ Confidence $= \frac{\text{Support}(5 \cup (2 \wedge 3))}{\text{Support}(5)}$

$= \frac{2}{3} = 66\%$

$3 \rightarrow (2 \wedge 5) \rightarrow$ Confidence $= \frac{\text{Support}(3 \cup (2 \wedge 5))}{\text{Support}(3)}$

$= \frac{2}{3} = 66\%$

$\therefore (2 \wedge 3) \rightarrow 5$; $(3 \wedge 5) \rightarrow 2$ are association rules.

(2) FP Growth Algorithm :-

- FP Stands for Foequent patteans.
- It is an efficient and Scable method for mining the complete set of FP by using a tree Structure.
- Tree Structure Stores information about FP called FP tree.

Example :

Minimum Support = 30%

Trans. id	Items
1	E, A, D, B
2	D, A, E, C, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

- List out the priorities.

itemset = (A, B, C, D, E)

* more frequency. more priority.
* same frequency FCFS.

itemset	Frequency	priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

Order of priority - (B, D, A, E, C)

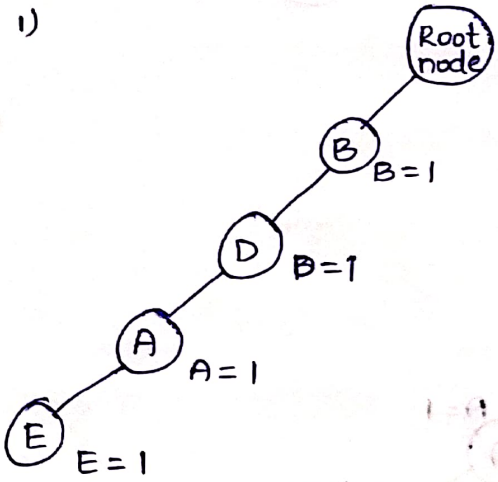
- Order items according to priority.

Trans. Id	Items	Ordered items
1	E, A, D, B	B, D, A, E
2	D, A, E, C, B	B, D, A, E, C
3	C, A, B, E	B, A, E, C
4	B, A, D	B, D, A

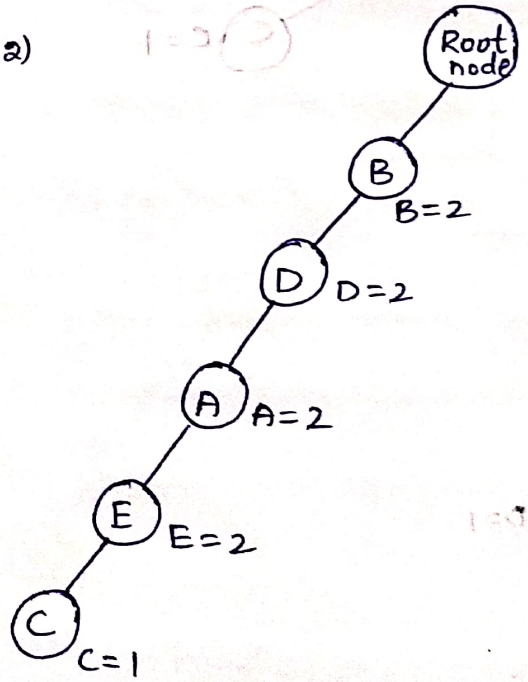
5	D	D
6	D, B	B, D
7	A, D, E	D, A, E
8	B, C	B, C

- Construct FP tree.

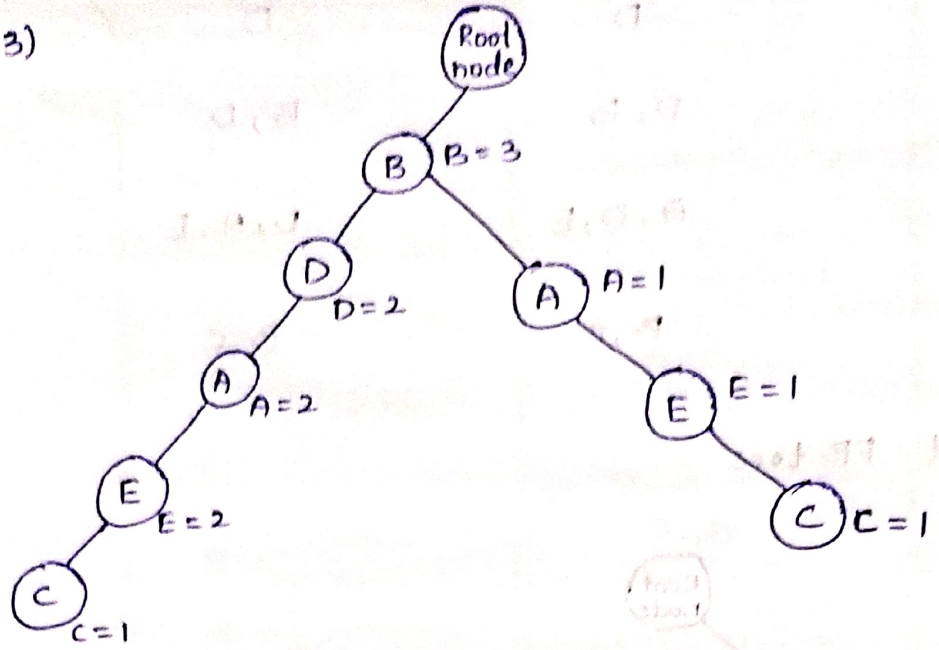
1)



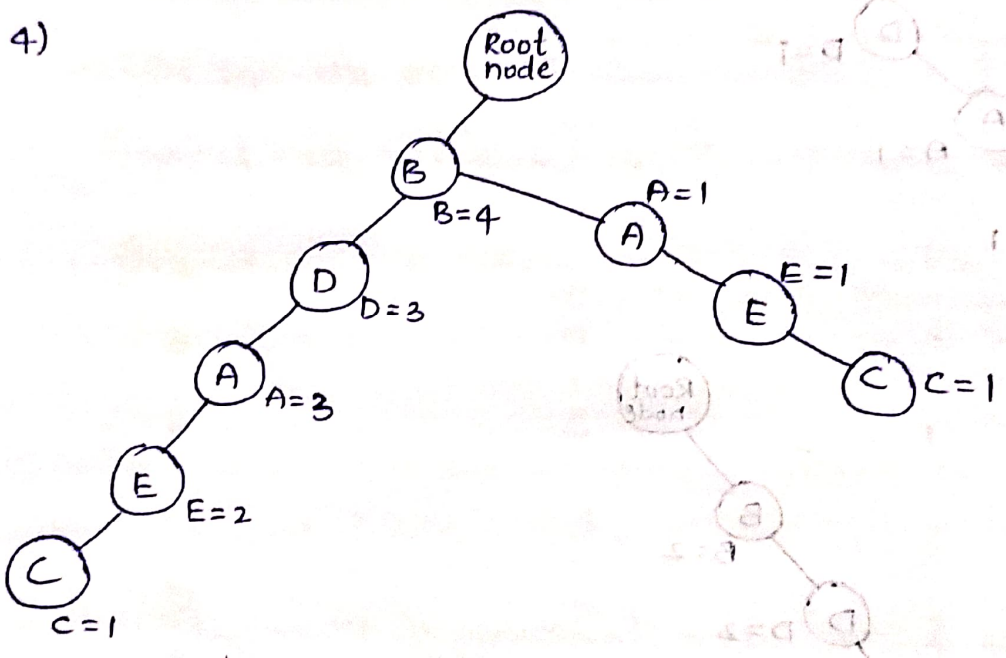
2)



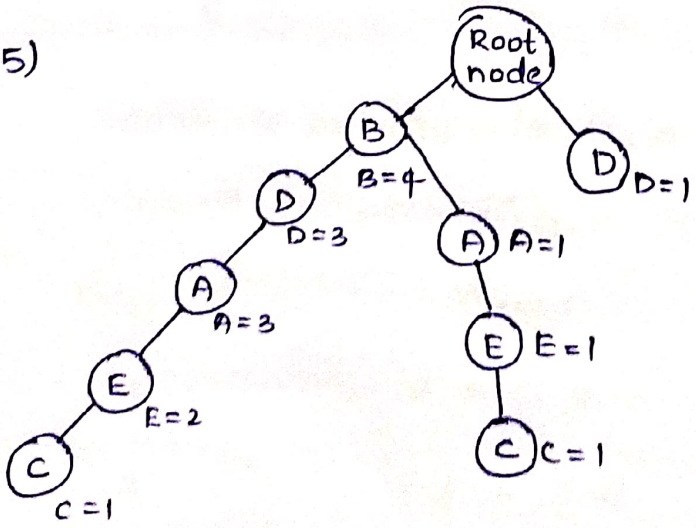
3)


Root node
B
B=3
D
D=2
A
A=1
A
A=2
E
E=1
E
E=2
C
C=1
C
C=1

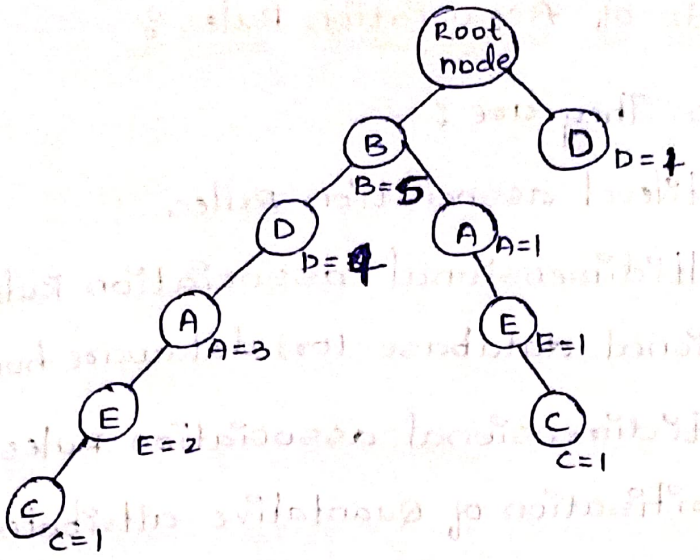
4)


Root node
B
B=4
A
A=1
D
D=3
E
E=1
A
A=3
C
C=1
E
E=2
C
C=1

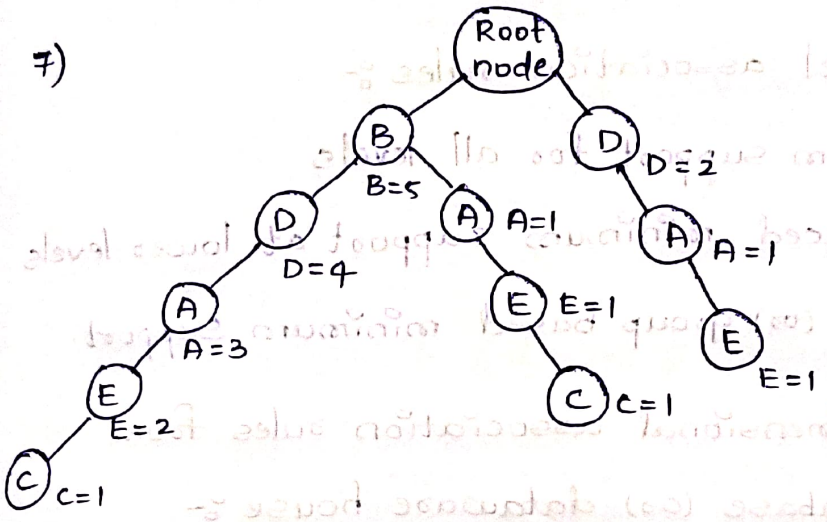
5)


Root node
B
B=4
D
D=1
D
D=3
A
A=1
A
A=3
E
E=1
E
E=2
C
C=1
C
C=1

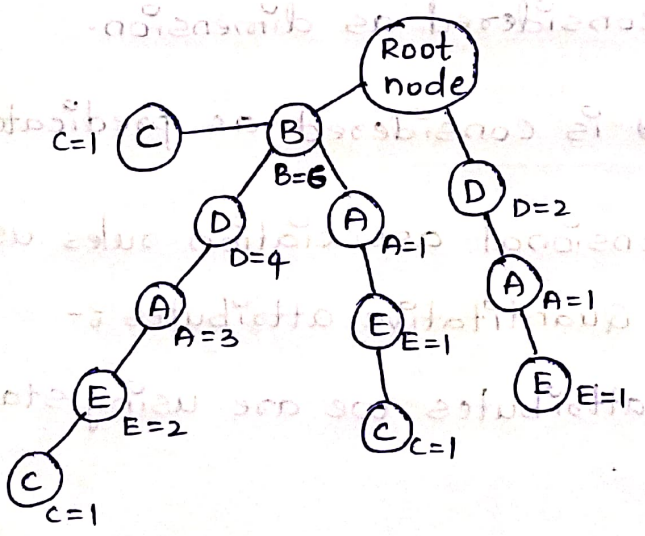
6)



7)



8)



B = 6 times

D = 6 times

A = 5 times

E = 4 times

C = 3 times.

Mining Various kinds of Association Rules :-

There are 4 types. They are :

1. Mining Multilevel association Rules.
2. Mining Multidimensional association Rules. from relational database (or) dataware house.
3. Mining Multidimensional association Rules using static discritisation of Quantative attributes.
4. Mining Quantitative Association rules.

1. Mining Multilevel association rules :-
   - Using Uniform support for all levels
   - Using reduced minimum support at lower levels
   - Using item (or) group based minimum support.

2. Mining Multidimensional association rules from relational database (or) dataware house :-
   - Each group is considered as dimension.
   - Each attributed is considered as predicate.

3. Mining Multidimensional association rules using staticdiscritisation of Quantitative attributes :-
   - For Numerical attributes we are using static discritisation.
   - Static discritisation means dividing into intervals.

4. Mining Quantitative Association rules :-
   - For Numerical attributes we are using dynamic discritisation
   - It is more efficient.

Correlation Analysis :

- Correlation Analysis is used to measure the relationship between two variables.
- One variable increases, then other variable also increases is called positive correlation.
- One variable ~~decreases~~ increases, then other variable decreases is called negative correlation.

$$\gamma_{A,B} = \frac{\Sigma(A-A')(B-B')}{(n-1)\sigma A \sigma B}$$

Where,

$\gamma_{A,B}$ = Karle pearson correlation coefficient

A', B' = mean of A and B

$\sigma A, \sigma B$ = Standard deviation of A and B.

n = no. of tuples in d.B.

γ has 3 values $(0, -1, +1)$

$\gamma \to +1 \Rightarrow$ perfect positive correlation

$\gamma \to -1 \Rightarrow$ perfect negative correlation

$\gamma \to 0 \Rightarrow$ No correlation (no dependence).

Example :

A	B
20	8
12	34
9	4

$$\gamma_{A,B} = \frac{\Sigma(A-A')(B-B')}{(n-1)\sigma A \sigma B}$$

$n = 3$

A' = mean of A

$$A' = \frac{20 + 12 + 9}{3} = \frac{41}{3} = 13.66$$

B' = Mean of B

$$B' = \frac{8 + 34 + 4}{3} = \frac{46}{3} = 15.33$$

$$\sigma A = \sqrt{\frac{\Sigma (A - A')^2}{(n-1)}}$$

$$= \sqrt{\frac{(20-13.66)^2 + (12-13.66)^2 + (9-13.66)^2}{(3-1)}}$$

$$= \sqrt{\frac{(6.34)^2 + (-1.66)^2 + (-4.66)^2}{2}}$$

$$= \sqrt{\frac{40.19 + 2.75 + 21.71}{2}}$$

$$= \sqrt{\frac{64.65}{2}}$$

$$= \sqrt{32.32}$$

$$= 5.68$$

$$\sigma B = \sqrt{\frac{\Sigma (B - B')^2}{n-1}}$$

$$\sigma B = \sqrt{\frac{(8-15.33)^2+(34-15.33)^2+(4-15.33)^2}{(3-1)}}$$

$$= \sqrt{\frac{(-7.33)^2+(18.67)^2+(-11.33)^2}{2}}$$

$$= \sqrt{\frac{53.72+348.56+128.36}{2}}$$

$$= \sqrt{\frac{530.64}{2}} = \sqrt{265.32} = 16.28$$

$$\gamma_{A,B} = \frac{\sum[(A-A')(B-B')]}{(n-1)\sigma A \sigma B}$$

$$= \frac{(20-13.66)(8-15.33)+(12-13.66)(34-15.33)+(9-13.66)(4-15.33)}{(3-1)\times 5.68 \times 16.28}$$

$$= \frac{6.34\times(-7.33)+(-1.66)\times 18.67+(-4.66)\times(-11.33)}{2\times 5.68\times 16.28}$$

$$= \frac{-46.47+(-30.99)+52.79}{184.94}$$

$$= \frac{-24.67}{184.94} = -1.\_\_$$

$\cong -1$

i.e Negative Correlation.

Constraint Based Association Mining :-

- Constraint means Condition.
- association rules are generated based on conditions.

Types of Constraints :-

1. Knowledge type :

   Specifies the knowledge you want to do mining.

   - association, correlation, Regression, etc.

2. Data Constraints :

   specifies the data on which you generate association rules.

   - only task relevant data.

3. dimension/level Constraints :

   Specifies the dimension (or) level.

   - Concept hierarchy.

4. Interestingness Constraints :

   Support, Confidence are used to identify.

5. Rule Constraints :

   Specifies the form of rules to be mined.

   1. Metarules Guided mining
   2. Constraint pushing.

Graph Pattern Mining :-

- Set of tools and techniques are used to mine freguents subgraphs.
- It is used to analyze the properties of real world graph.

- It is used to analyse the how structure of graph will effect the rules.
- There are two methods. They are
  - Apriori based Approaches
  - Pattern growth Approaches.

Algorithm used :
- Gspan
- Closed Graph.

Application :
- In xml structures
- Anomaly detection
- Network analysis
- Control flow Analysis
- Biological structures, etc.

Sequential Pattern Mining :-

- SPM stands for Sequential pattern mining.
- Sequence = ~~order~~ set of Ordered events.

  Ex: $S = \{e_1, e_2, e_3, e_4, e_5\}$.
- SPM → Process of finding frequent subsequences from a set of sequence.
- Sequences are represented by < >

Normal Transactional data

CID	TID	Transactions
1	100	a,b,e,d
3	111	a,f,d,e
1	122	d,e,p
3	133	b,f,s,a

Sequential data.

CID	sequences
1	<(abed),(d,ep)>
3	<(afde),(bfsa)>

Challenges in SPM:

- Finding all subsequences:

min. sup = 2

<(ab)c>

Sid	Sequence
10	✓ <a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ac)>
30	✓ <(ef)(ab)(dfcb)>
40	<eg(af)cbc>

<(ab)c> is present in 1st sequence

<(ab)c> is present in 3rd sequence.

∴ min sup = 2.