UNIT 2 DMA Notes

Data Mining And Analytics (SRM Institute of Science and Technology)

UNIT - 2

Frequent patterns :-

The patterns that appear frequently in dataset
(include frequent dataitems, ⇒ eg: buying computer
sequences, ⇒ computer, mouse, keyboard
substructures) ⇒ graph, tree etc.

eg: Eg: milk & bread ⟶ bought together.

Market basket analysis :-

→ Process of analysing customer buying habits by finding the association b/w different items that a customer will place in their baskets.

→ Mainly usefull for sellers. (they can understand what type of products customers chosing)

Strategies Used :-

1. placing them together
2. placing them at ② different ends.

→ This analysis will help sellers to plan their shelf space for increased sales.

– Frequent patterns are represented by association rules.

Eg :- computer + antivirus

Computer ⇒ antivirus software [support – 2% (comp + antivirus are purchased together), confidence – 60%

(denotes among the total transactions of store 2% computers are purchased along with the antivirus)

(denotes 60% of customers who purchased the comp. purchased anti-virus)

Support:

Identifies how frequently a rule is applied to given dataset.

association rules

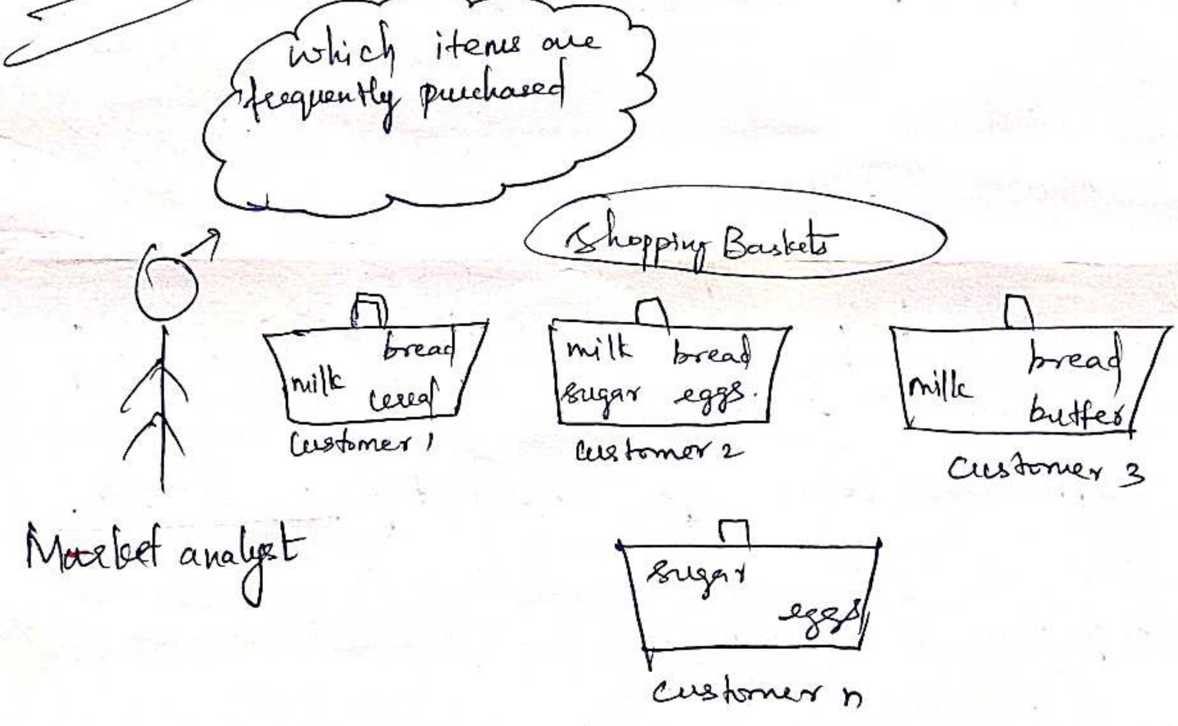
$$S(P \rightarrow Q) = \frac{\sigma(P \cup Q)}{N}$$ N = total transactions

$$\boxed{\Rightarrow P(A \cup B)}$$

Confidence:

Defines frequent occurrence of items of Q in transactions.

$$\boxed{C(P \rightarrow Q) = P(B/A)}$$ ⇒ conditional propability.



[ECLAT Method]

(Equivalence class Transformation algorithm)

* Mining frequent itemset using vertical data format.

Uses intersection based approach

Given: Horizontal data format.

min. supp count = 2.

TID	list of items ID's
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3.

$\Downarrow$

Vertical data format (1-Itemset)

Itemset	TID Set	Sup. count	
I_1	$\{T_{100}, T_{400}, T_{500}, T_{700}, T_{800}, T_{900}\}$	6	✓
I_2	$\{T_{100}, T_{200}, T_{300}, T_{400}, T_{600}, T_{800}, T_{900}\}$	7	✓
I_3	$\{T_{200}, T_{500}, T_{600}, T_{700}, T_{800}, T_{900}\}$	6	✓
I_4	$\{T_{200}, T_{400}\}$	2	✓
I_5	$\{T_{100}, T_{800}\}$	2	✓

$\therefore I_1, I_2, I_3, I_4, I_5 \Rightarrow$ freq. itemsets ($\because$ sup. >= 2)

2- Itemsets In Vertical Data format:

Itemset	TID Set	Sup. count	
$\{I_1, I_2\}$	$\{T_{100}, T_{400}, T_{800}, T_{900}\}$	4	
$\{I_1, I_3\}$	$\{T_{500}, T_{700}, T_{800}, T_{900}\}$	4	
$\{I_1, I_4\}$	$\{T_{400}\}$	1	×
$\{I_1, I_5\}$	$\{T_{100}, T_{800}\}$	2	
$\{I_2, I_3\}$	$\{T_{300}, T_{600}, T_{800}, T_{900}\}$	4	
$\{I_2, I_4\}$	$\{T_{200}, T_{400}\}$	2	
$\{I_2, I_5\}$	$\{T_{100}, T_{800}\}$	2	
$\{I_3, I_4\}$	$\{-\}$	0	×
$\{I_3, I_5\}$	$\{T_{800}\}$	1	×
$\{I_4, I_5\}$	$\{-\}$	0	×

3- Itemset in Vertical Data format:

Itemset	TID Set	Sup. count	
$\{I_1, I_2, I_3\}$	$\{T_{800}, T_{900}\}$	2	
$\{I_1, I_2, I_5\}$	$\{T_{800}, T_{100}\}$	2	
$\{I_1, I_3, I_5\}$	$\{T_{800}\}$	1	×
$\{I_2, I_3, I_4\}$	$\{-\}$	0	×
$\{I_2, I_3, I_5\}$	$\{T_{800}\}$	1	×
$\{I_2, I_4, I_5\}$	$\{-\}$	0	×

⇒ 3-Itemset format:

Itemset	ID
{I_1, I_2, I_3}	{T800, T900}
{I_1, I_2, I_5}	{T800, T100}.

4- Itemset vertical data format:

Itemset	ID	Sup.count	
{I_1, I_2, I_3, I_5}	{T800}	1	×

∴ If 4-itemset is not frequent we can take 3-itemset as output.

⇒ Mined frequent itemsets are,

{I_1, I_2, I_3}
{I_1, I_2, I_5} } output.

* Improving the efficiency of apriori algorithm :-

Many methods are available for improving the efficiency of apriori algorithm,

① Hash-based technique :-

↳ This method uses a hash-based structure called hash table for generating the k-itemsets and their corresponding count.

↳ It uses a hash function for generating a table.

② Transaction Reduction :-

↳ This method reduces the no. of transactions scanned in iterations.

↳ The transactions which do not contain frequent items are marked or removed.

③ Partitioning :-

↳ This method requires only 2 database scans to mine the frequent itemsets.

↳ It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of database.

④ Sampling :-

↳ This method picks the random sample 'S' from Database 'D' and then searches for frequent itemset in 'S'.

↳ It may be possible to lose a global frequent itemset.

↳ This can be reduced by lowering the min_sup.

• Variation of Apriori which tries to reduce the number of passes made over a transactional db while keeping the no of itemsets counted in a pass relatively low

(5) Dynamic itemset coupling :-

↳ This technique can add new candidate itemsets at any marked start point of the db during the scanning of the database.

* From association rule to correlation analysis :-

↳ Association rule algorithms tends to produce too many rules.

Many of them are uninteresting or redundant.

Redundant if,

{A, B, C} → {D} & {A, B} → {D} have same support & confidence.

Asso. rules consist of support & confidence.

But this support & confidence is insufficient at filtering out uninteresting association rules.

To tackle this weakness, a correlation measures can be used.

This leads to the correlation rule of the form;

$$\boxed{A \Rightarrow B\,[\text{support, confidence, correlation}]}$$

i.e a corelation rule is measured by not only sup & confidence but also the correlation b/w the itemsets A & B.

correlation measures :-

There are many different correlation measures,

Lift :-

Lift is a simple correlation measure that is given as follows,
The occurence of itemset A is independent of the occurrence of itemset B if P(AUB)

$\Rightarrow$ $\boxed{P(A \cup B) = P(A) P(B)}$;

itemsets A & B are dependant & correlated.

if there are more than 2 itemset,

$$\boxed{lift(A, B) = P(A \cup B) / P(A) P(B)}$$

↳ if value of lift is less than 1 $\Rightarrow$ the occurence of 'A' is negatively correlated with the occurence of 'B'. i.e, the occurence of one likely leads to the absence of the other one.

↳ lift value greater than 1 $\Rightarrow$ A & B are positively correlated, meaning that the occurence of one implies the occurence of other.

↳ value equal to 1 $\Rightarrow$ A & B are independant & there is no correlation b/w them.

Eg:-

If the probability of purchasing a computer game is P({game}) = 0.60,
the prob. of purchasing a video is P({video}) = 0.75,
" " both is P({game, video}) = 0.40.

∴ Lift value is,

$$\boxed{P(\{game, video\}) / (P\{game\}) \times P(\{video\}))}$$

$$= 0.40 / (0.60 \times 0.75) = 0.89.$$

∴ the value is less than 1,

∴ there is a negative correlation b/w the occurence of {game} & {video}.

the second correlation measure is,

$$\boxed{\chi^2 = \Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}}}$$

eg: contingency table,

	game	$\overline{\text{game}}$	Σ_{row}
video	4000(4500)	3500(3000)	7500
$\overline{\text{video}}$	2000(1500)	500(1000)	2500
Σ_{col}	6000	4000	10,000

$$\Rightarrow \chi^2 = \Sigma \frac{(ob - exp)^2}{\text{expected}}$$

$$= \frac{(4000-4500)^2}{4500} + \frac{(3500-3000)^2}{3000} + \frac{(2000-1500)^2}{1500}$$

$$+ \frac{(500-1000)^2}{1000} = 555.6$$

⇒ χ^2 value is greater than 1, then observed value of {game, vedio} = 4000 which is less than expected value 4500,

∴ buying game & buying video are negatively correlated.

Apriori Algorithm

$\frac{50}{100} \times 4 = 2$

TID	items	
100	1 3 4	min support = 50%
200	2 3 5	Threshold Confidence = 70%
300	1 2 3 5	
400	2 5	

Iteration 1: (step 1:)

Itemset	Support	minsupport
1	2	2/4 = 50%
2	3	3/4 = 75%
3	3	3/4 = 75%
4	1	1/4 = 25% ✗
5	3	3/4 = 75%

Itemset (1, 2, 3, 5)

Iteration 2 (step 2:)

Form pairs (1,2) (1,3) (1,5) (2,3) (2,5) (3,5)

itemset	support	min supp
(1,2)	1	1/4 = 25% ✗

(1,3)	2	2/4 = 50%
(1,5)	1	1/4 = 25% ×
(2,3)	2	2/4 = 50%
(2,5)	3	3/4 = 75%
(3,5)	2	2/4 = 50%

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

Iteration 3 : Form Triplets (Step 3)

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

Itemset	Support	min support
(1,2,3)	1	1/4 = 25% ×
(1,2,5)	1	1/4 = 25% ×
(1,3,5)	1	1/4 = 25% ×
(2,3,5)	2	2/4 = 50%

Itemset = (2,3,5)

Now Calc. Support & Confidence

freq count of an item

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

Generate Association Rules using (2,3,5)

Rules	Support	Confidence
(2,3) → 5	2	2/2 = 100% ✓
(3,5) → 2	2	2/2 = 100% ✓
(2,5) → 3	2	2/3 = 66% ✗
2 → (3,5)	2	2/3 = 66% ✗
5 → (2,3)	2	2/3 = 66% ✗
3 → (2,5)	2	2/3 = 66% ✗

$(2,3) \to 5$ → Confidence = $\frac{\text{Support}(A \cup B)}{\text{Supp}(A)}$

A = (2,3), B = 5

$$\frac{S((2,3) \cup 5)}{S(2,3)} = 2/2 = 100\%$$

A $\quad$ B

$$2 \rightarrow (3,5) = \frac{S(2 \cup (3,5))}{S(2)} = 2/3 = 66\%$$

$\therefore$ $(2,3) \rightarrow 5$, $(3,5) \rightarrow 2$ are

association rules

Frequent pattern growth (FP growth) Algorithm

ms = 2

TID	List of items ID's
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

Sol)

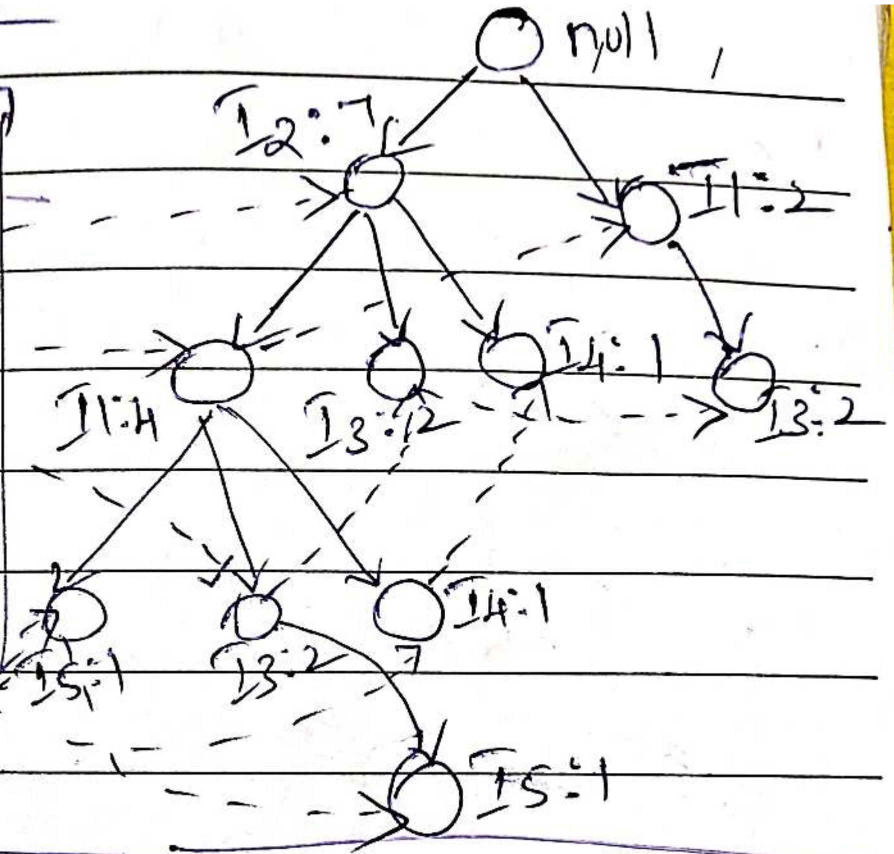
Itemset	Sup Count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

$\rightarrow$

items	sup count
I_2	7
I_1	6
I_3	6
I_4	2
I_5	2

is called FP tree

Itemid	sc	node link
I2	7	
I1	6	
I3	6	
I4	2	
I5	2	



item	Conditional pattern Base	Conditional FP tree	FP's generated
I5	{{I2, I1:1}, {I2, I1, I3:1}}	{I2:2, I1:2}	{I2, I5:2}, {I1, I5:2}, {I2, I1, I5:2}
I4	{{I2, I1:1}, {I2:1}}	{I2:2}	{I2, I4:2}
I3	{{I2, I1:2}, {I2:2}, {I1:2}}	{I2:4, I1:2}, {I1:2}	{I2, I3:4}, {I1, I3:4}, {I2, I1, I3:2}
I1	{{I2:4}}	{I2:4}	{I2, I1:4}

1) Mining Frequent patterns/itemsets / Market Basket Analysis

Synopsis

* What is itemset?
* what is Freql- itemsset | pattern? (FP) with example
* Application of FP (Market basket Analysis) (MBA)
  → What is MBA? with diagram
* Purpose of MBA
* How MBA works?
  → (works on Association rule mining)
→ What is Association rule mining with example
* Strategies used on MBA with example (Computer & antivirus)
* Support
* Confidence
* Mining methods
  a) Apriori Algorithm (defl-)
  b) FP growth Algorithm (def)

Apriori Algorithm

Synopsis

* Definition of Apriori Algorithm
* What is support? with formula
* What is Confidence? with formula
* what is association rule? ~~with~~
* Any 2 techniques of improving efficiency of Apriori Algorithm.

Frequent itemset & Closed itemset

Synopsis

* What is itemset?
* " " Freq-itemset?
* Support with formula
* Confidence. " "
* What is closed itemset?
* Example of Closed itemset?

4) pattern growth approach | Frequent pattern (FP) growth Algorithm.

Synopsis

* Definition of Frequent pattern growth.
* Own example with FP tree.

5) Mining Frequent itemsets Using Vertical data format

Synopsis

* What is itemset ?
* what is Freq-item set ?
* What is mining freqt itemset ?
* own Example.

6. Association analysis to correlation Analysis

Refer the pdf notes. for answer.