Tuesday

Date 29/1/16
Page 1
STUDY BUDDIES

UNIT 4: Introduction to Data Mining

SQL → RDBMS ← simple query
Data Mining → Data Warehouse ← complex query

- Data Mining: access to data
- Data Warehouse: managing the data
- Mining data from data warehouse is an advance version of extracting SQL data from RDBMS.

Problems with today's database management

i) The amount of data in computer files and database, busines areas, banks is growing at phenomenal rate at the same time the users of these data are expecting much sophisticated information from them

ii) The problem is how to store these data, how to manage these data and how to access these data in order to solve this problem, & the solution is data mining, and data warehouse.

When a marketing manager wants to predict what are the most important trends in omer behaviour.

) what is the difference between SQL & data mining?

→

	SQL	Data Mining
Query	· SQL query are simple and well formed	· queries are complex or might not be well-formed or precisely stated.
Data	· The data accessed from the original operational dbs., which are not in consitent format. (RDBMS)	· Data accessed is different version from original operational db. The data have been cleansed and modified to better support the data mining.
Output	The output of SQL is subset of database.	· In data mining, you may vaguely know what you are looking for

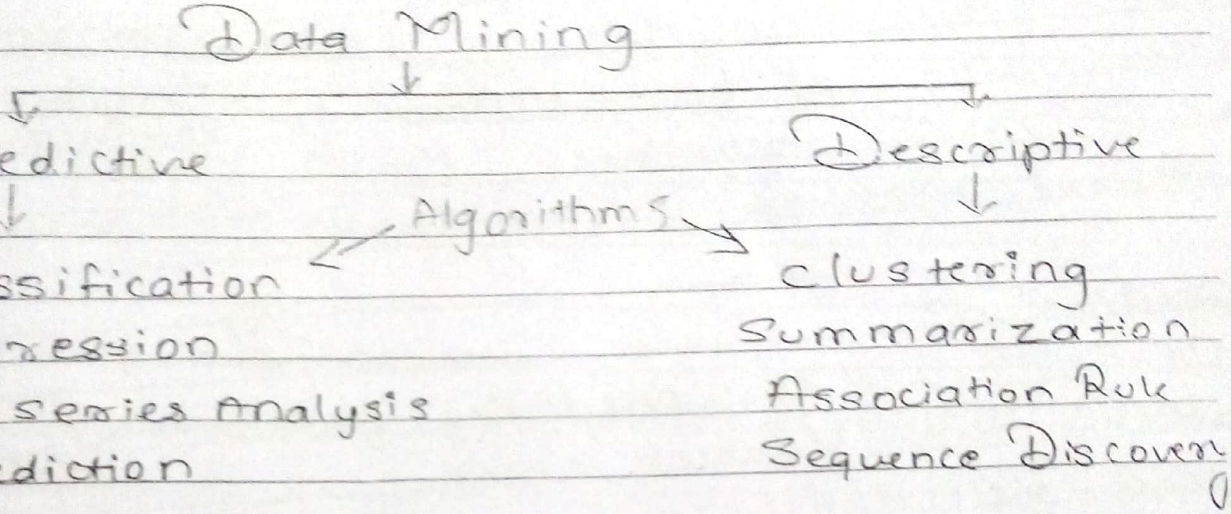
Defination of Data Mining.

Data Mining is defined as finding hidden information in a database. It is also called as exploratory data analysis, data-driv discovery, and deductive learning.

Data mining is used to describe discovering o mining language knowledge from large amt. of

Date / /
Page 5
STUDY BUDDIES

Data mining involves many different algorithm to accomplish task. It has two models: Predictive & Descriptive.



Predictive Model:-

- It makes a prediction about values of data using known results from different data.

Algorithms:

Classification:-

- It maps data into predefined groups or classes.
- Eg: an airport security screening is used to determine if passengers are terrorist or criminals.

Regression:

- It is used to map a data item to a real valued prediction variable.
- If an employee wants to reach a certain level of savings before his/her retirement

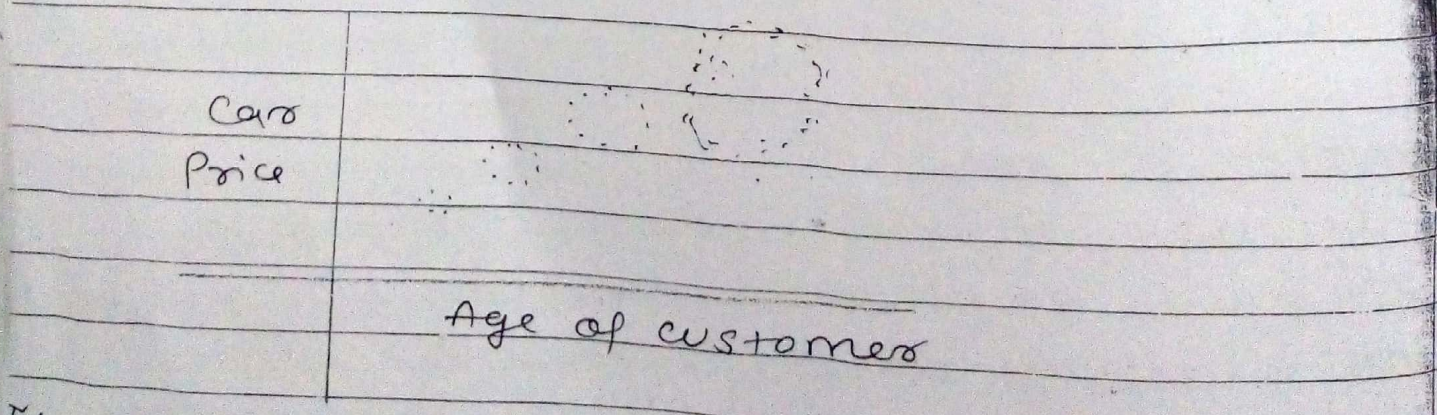
) Saving will be based on current values and past values.

⇒ Time Series Analysis:
The value of an attribute is examined as it varies overtime.
Share market / stock.

Prediction:
It is used to predicting future state rather than current state.
flood detection.

Descriptive Method - Model:-
This model identifies patterns or relationships between data.

clustering:
It is called as segmentation.
It is similar as classification except that the groups are not predifined
The groups contain similar information.



It abstracts representative info about database.

g: compare students of Mumbai and Pune university. It will use to estimate the intellectual level of student.

ii) Association Rule:

- It describes relationships among data.

g: Bread ↔ Butter
pencil ↔ eraser.

ii) Sequence Discovery:

- It is used to determine sequential pattern in data.

y: <A, B, C, D, E> } The pattern describes 70-1.
<A, B, C, E> } users of page A, follow
<A, B, C> } page C. So we will add link directly from Page A → C.

16 KDD : Knowledge Discovery Process

- Data mining and KDD is often use interchangeably.
- Data mining is the use of algorithm to extract the information and patterns derived by the KDD process.

+ Defination of KDD:-

- It is the process of finding useful informatn and patterns in data.
Input to this process is data and o/p is useful information derived by the user.

The KDD process consists of 5 steps:-

i) Selection:-
The data needed for data mining process obtained from many different and heterogeneous data sources.

ii) Preprocessing:-
Erroneous data may be corrected or removed.

i) Bank Branch at Thane

CID	Name	Add
1	ABC	Thane
2	XYZ	Bhandup

Bank Branch at Kurla

CID	Name	Address
1	ABC	Kurla
2	XYZ	Bhandup.

customer 1 has got different address in different databases, which can either removed or corrected.

Missing data must be supplied or predicted or removed.

Bank Branch at Thane

CID	NAME	Address
1	ABC	Thane
2	XYZ	Bhandup

Bank branch at Kurla

CID	Name	Address
1	ABC	-
2	XYZ	Bhandup

It shud be supplied predicted or removed.

Date 1/1
Page 1
STUDY BUDDIES

4) Remove false information :-

CID	Name	Add	Magazine Purchase Date	Magazine
1	ABC	Thane	1-1-1901	car
2	XYZ	Bhandup	1-1-1968	cosmetics

→ Here, on the date of 1901, the company is not existing that means false information is present in the database, which should be corrected or removed.

5) Remove Redundancy :-

→ The preprocessing is used to correct missing values, de-duplication, disambiguation before transferring data from operational system to data warehouse.

ii) Transformation :-

Data from different sources must be converted into a common format for processing.

It is used to reduce the no. of possible data values.

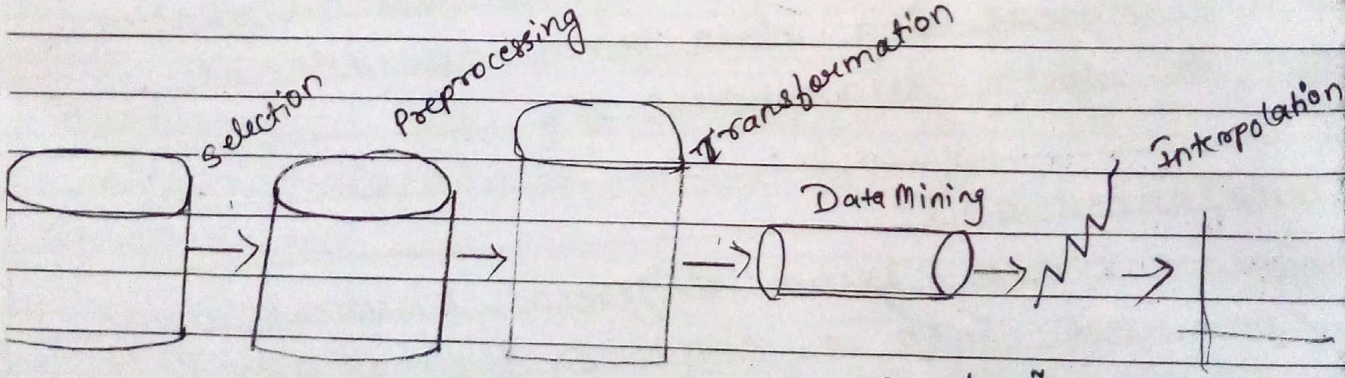
g: Either keep age or dob.
Either keep no. of years of experience or date of joining

iv) Data Mining :-

→ It is used to apply many algorithms like

clustering, classification, association rule etc. on the data present in data warehouse.

Interpolation / Evaluation (Visualizatn Technique)
How data mining results are presented to the users, is extremely important because usefulness of the result is dependent on it. For this visualization and GUI's Strategy are used.
Visualization Technique include:
graphical
Icon based
pixel based
Hierarchical
Hybrid.



Data Mining Issues:-
Human Interaction
Visualization of results
Multimedia data
Missing data e.) Ease of use.

* Data Mining Data *

- Categorical (divide data in specific groups)
  - Nominal — eg: marital status, gender
  - Ordinal — height, credit score
- Numerical (represent numerical value)
  - interval — temperature
  - Ratio — mass, energy.

- Data refers to collection of facts usually obtained as result of experiences and observations.

Data & Data mining refers as follows:-

* Categorical Data:- It represents labels of multiple classes used to divide a variable into specific groups.

i.) Nominal Data:- It contains obj. which are not measurements. Eg: marital status can be categorized as single / married / divorced.

ii.) Ordinal Data:- It contains code or object dat represent the rank order among them.
g: credit score can be categorised as low/medium/high

* Numerical Data:- It represents numeric value of specific variables. i.) Interval Data:-
g: age, income, temperature. These are variables measured on interval scale. eg. temp.

ii) Ratio: It includes variables, commonly found in the physical sciences and engineering
Eg: mass, energy, time length.

Data Mining Applications:

1) Customer Relationship mngmt. (CRM): The goal of CRM is to create one to one relationship with customers by developing & understand their needs & wants.

2) Banking:- it identify base to maximize customer value by selling them products and services like atm machine, credit card, banking branches etc.

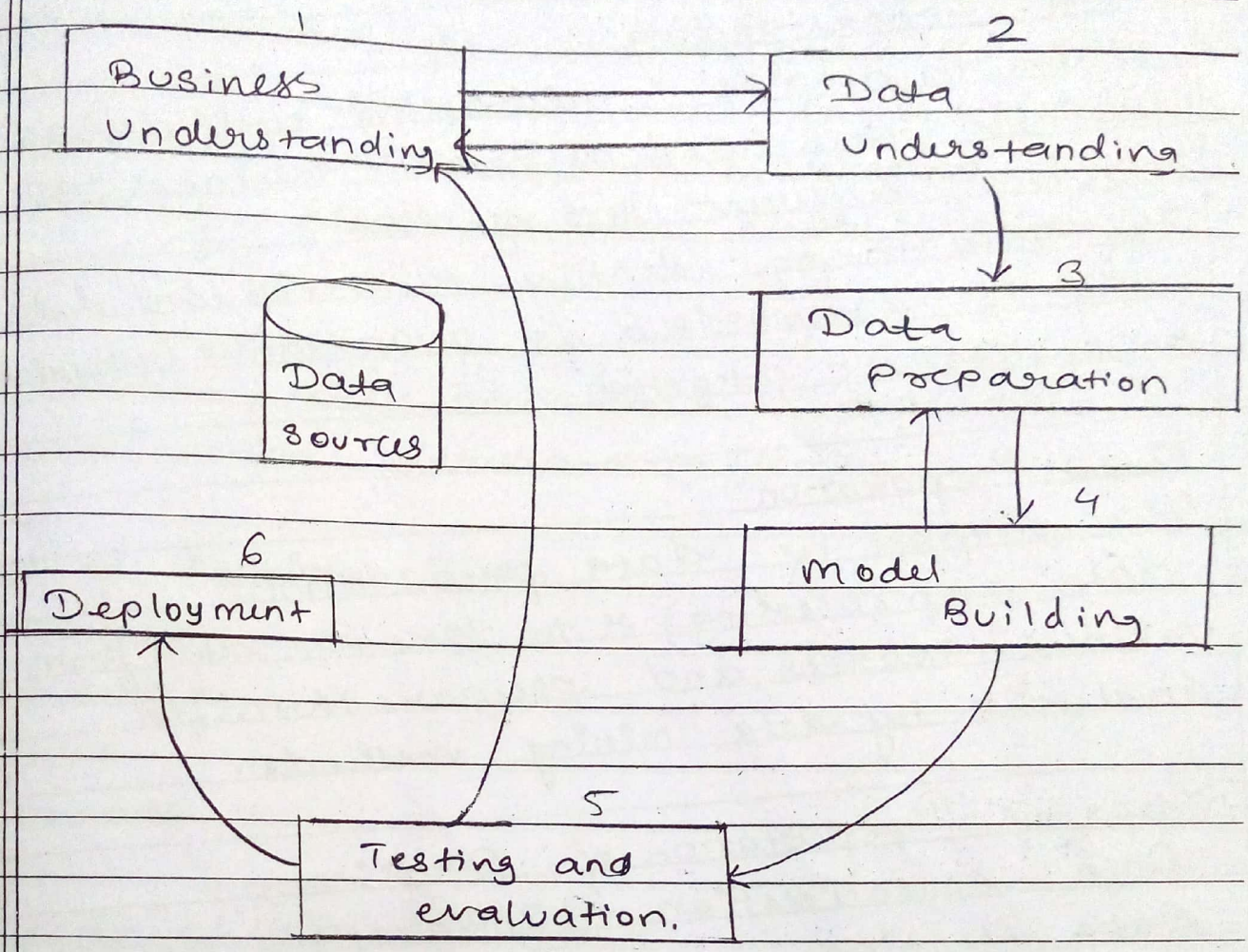
3) Insurance: It uses data mining technique to predict which customer are more likely to buy new policy

4) Computer H/W & S/W: It uses data mining to detect & prevent comp. n/w security bridge and identify potentially unsecur S/W products.

5) Entertainment Industry: Data Mining is used to analyze viewer's data to decide what programs to show during prime time & how to maximize returns by knowing where to insert advertisements.

6) Medicines:- Data Mining is used 2 discover the relationships between symptoms & illness. It predicts success rates of organ transplantation patients to develope better donor organ matching policies

* Data Mining Process *



- Data Mining researches have proposed several researches to maximize the chance of success in conducting data mining projects.

1) Business Understanding:-

→ It is the thorough understanding of the managerial need for new knowledge, an explicit specification of the business objective regarding the study to be conduc

2.) Data Understanding:-
→ Data mining is addressing well defined business task and different business task require different sets of data
→ The Data mining identify the relavant data which is categorised as quantitative (numerical & qualitative (categorial)

3.) Data Preparation:
→ The purpose of data preparation (data preprocessing) is to take the data from various sources and prepare them for analysis by data mining methods.

Steps of Preparation of Data:-
1.) Data consolidation (Integration/selection)
2.) Data cleaning
3.) Data Transformation
4.) Data Reduction

Model Building:-
On prepared data, we will apply one of the data mining model to get the output.

Testing and Evaluation:-
This step will test the output and evaluate it.

Deployment:-
Once testing has been successful, we will deploy the output.

Association Rule

Date 1/1
Page 13
STUDY BUDDIES

The purchasing of one product, when another product is purchased, we present an association rule. It is used by retail stores, to assist in marketing advertising, floor placement and inventory control.

Association rules are used to show the relationships between data items.

* Uses of Association Rule:-
a) Placement
b) Advertising
c) Sales
d) Coupons

* Objective of Association Rule:
→ Increased sales & Reduced costs.

* Defination of Association Rule:-
→ Given a set of Items $I = \{ I_1, I_2, \cdots, I_m \}$ and a database of transaction $D = \{ t_1, t_2, \cdots t_n \}$ where $t_i = \{ I_{i1}, I_{i2}, \cdots I_{ik} \}$ and $I_{ij} \in I$, an association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \in I$ are sets of items called itemset and $X \cap Y = \phi$

Eg:

Transaction	Items
t_1	Bread, Jelly, Peanut Butter
t_2	Bread, Peanut Butter
t_3	Bread, milk, Peanut Butter
t_4	Beer, Bread
t_5	Beer, milk

I_{00} I_{01} I_{10} I_{11}

5/1/16
Monday

* Defination of Support (S)

Association Rule

→ The support (S) for an A.R $X \Rightarrow Y$ is the percentage of transactions in the database that contain

$$\left[\frac{X \cup Y}{\text{Total no. of items}} \times 100\right]$$

* Defination of confidence or Strength (α)

→ The confidence or Strength for an association $X \Rightarrow Y$ is the ratio of no. of transactions that contain $X \cup Y$ to the no. of transactions that contain X

$$\left[\frac{X \cup Y}{X} \times 100\right]$$

Importance is measured by two features
Support & Confidence

The selection of Association rules is based on support and confidence.

Confidence measures the strength of the rule.
Support measures how often it should occur in the database.

X => Y	Support (s)	Confidence (α)
X Y		
Beer => Bread	$\frac{1}{5} \times \frac{20}{100} = 20\%$	$\frac{1}{2} \times \frac{50}{100} = 50\%$
X Y		
Peanut Butter => Jelly	$\frac{1}{5} \times 100 = 20\%$	$\frac{1}{3} \times 100 = 33.3\%$
X Y		
Jelly => Peanut Butter	$\frac{1}{5} \times \frac{20}{100} = 20\%$	$\frac{1}{1} \times 100 = 100\%$
X Y		
Jelly => Milk	$\frac{0}{5} \times 100 = 0\%$	$\frac{0}{1} \times 100 = 0\%$
X Y		
Bread => Peanut Butter	$\frac{3}{5} \times \frac{20}{100} = 60\%$	$\frac{3}{4} \times \frac{25}{50} \times 100 = 75\%$
X Y		
Peanut Butter => Bread	$\frac{3}{5} \times 100 = 60\%$	$\frac{3}{3} \times 100 = 100\%$

* **Lift:**

The defination of lift is defined as

$$\text{lift}(X \Rightarrow Y) = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$

or the ratio of the observed support to that of expected that if X and Y were independent.

* **Conviction:**

→ The conviction of a.r. is defined as

$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{Supp}(X)}{1 - \text{conf}(X \Rightarrow Y)}$$

: Calculate lift of Bread ⇒ PeanutButter

→ $= \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$

$= \frac{\text{Supp}(\text{Bread} \Rightarrow \text{PeanutButter})}{\text{Supp}(\text{Bread}) \times \text{Supp}(\text{PeanutBu}$

$= \frac{60\%}{80\% \times 60\%} = \frac{1}{80\%} = \frac{8/1}{8} \times 0.01$

$= \frac{100}{80} = 1.2$

ii) Conviction of Bread => PeanutButter:

→ $= \frac{1 - \text{supp}(X)}{1 - \text{conf}(X \Rightarrow Y)}$

$= \frac{1 - 80\%}{1 - 75\%}$

$= \frac{1 - 0.8}{1 - 0.75} = \frac{0.2}{0.25} = 0.8 = 1.0$

2. Calculate lift & Conviction for jelly & PeanutButter

→

$\text{lift} = \frac{\text{supp}(\text{jelly} \cup \text{PeanutButter})}{\text{supp}(\text{jelly}) * \text{sup}(\text{PeanutButter})}$

$= \frac{20\%}{20\% \times 60\%} = \frac{1}{60} = \frac{0.1}{6} = 1.6$

ii Conviction: $\frac{1 - \text{supp}(\text{jelly})}{1 - \text{conf}(\text{jelly} \Rightarrow \text{PeanutButter})}$

$= \frac{1 - 20}{1 - 100} = \frac{1 - 0.2}{1 - 1} = \infty.$

* Frequent Item Set (Large Item Set)

→ A large frequent item set is an item set whose no. of occurances is above a threshold value (S, α) we use the notation to indicate the complete set of large item set & l to indicate a specific large item set.

eg: Suppose that the input support & confidence are S = 30 % and confidence = 50 %. respectively

L = { { Beer }, { Bread }, { Milk }, { Peanut Butter }, { Bread }, { Peanut Butter } }

l = { Bread, Peanut Butter }

$l \in L$

There are 2 non-empty subsets of l

{ Bread } & { Peanut Butter }

(i) frequent itemset (large itemset with 1st one is { Bread }

→ $\frac{\text{Support (Bread, Peanut Butter)}}{\text{support (Bread)}}$

$= \frac{60\%}{80\%} = \frac{0.6}{0.8} = \frac{3}{4} = 0.75 = 75\%$.

(ii) with 2nd one l = { Peanut Butter }

$\frac{\text{Support (Bread, PB)}}{\text{Sup (PB)}} = \frac{60\%}{60\%} = \frac{1}{1} \times 100 = 100\%$

$X \Rightarrow (I - X)$

Bread ⇒ PB | PB ⇒ Bread.

i) The a.r. rule is given by $X \Rightarrow (I - X)$ which is {Bread ⇒ PeanutButter}

confidence of a.r. {Bread ⇒ PeanutButter} is 75% which is above threshold value., hence it is a valid association rule & is added to R ← (Set of valid A.R) = {{Bread ⇒ PeanutButter}}

ii) The a.r. is given by $X \Rightarrow (I - X)$ which is {PeanutButter ⇒ Bread}

Confidence of a.r. {PeanutButter ⇒ Bread} is 100%, which is above threshold value hence, it is a valid a.r. & is added to R = {{PeanutButter ⇒ Bread}}

Apriori Algorithm:-

This algorithm is most well known a.r. & is used in most commercial product.

It uses following property any subset of large Itemset must be large.

Working of apriori algorithm:-

It is used to generate candidate item set of a particular size and then scan the database to count to see if they are large.

→ During scan i (index 0...n) candidate of size i (C_i) are counted only those candidate that are large are use to generate candidates for the next pass i.e. L_i are use to generate C_{i+1}

→ An item set is consider as a candidate only in all its subset are large to generate candidate of size i+1 joins are made of large item set found in the previous pass.

Eg: S = 30%. & 50%.

Pass	Candidates	Large Itemset.
	40 80 20	
1	{Beer}, {Bread}, {Jelly}, {Milk}, {PB}	{Beer} {Bread} {Milk} {PB}
	40 60	
	20 20	
2	{Beer, Bread} {Beer, Milk} {Beer, PeanutButter} {Bread Milk} {Bread, PeanutButter} {Milk, PB}	{Bread, PB}
3	~~{Bread, PB}~~ NULL	~~{Bread, PB}~~ ϕ

Define support and confidence to measure the strength of association rule calculate support and confidence for the association rule.

TID	Item
1	{Bread, Milk}
2	{Bread, Diaper, Beer, Eggs}
3	{Milk, Diaper, Beer, Cola}
4	{Bread, Milk, Diaper, Beer}
5	{Bread, Milk, Diaper, Cola}

X Y

i) Support ({Milk, Diapers} → {Beer}) =

$$\frac{X \cup Y}{\text{total no. of items}} \times 100$$

$$= \frac{2}{5} \times \frac{2}{100} = 2 \times 20 = 40\%$$

X Y

ii) Confidence ({Milk, Diapers} → {Beers}) =)

$$\frac{X \cup Y}{X} \times 100$$

$$= \frac{2}{3} \times 100^{33.3}$$

$$= 2 \times 3.33 = 66.6\%$$

2.) Eg: What do you mean by frequent item set. use apriori algo to generate frequent item set for the following by taking S = 60%.

ass	Candidates	Large Itemset
1	{Bread} {Milk} {Beer} 80 80 60 {Diapers} {Eggs} {Cold} 80 20 40	{Bread} {Milk} {Diaper} {Beer}
2.	{Bread, Milk} {Bread, Diaper} 60 60 {Milk, Diaper} {Bread, Beer} 40 {Milk, Beer} {Diaper, Beer} 60	{Bread, Milk} {Bread, Diaper} {Milk, Diaper} {Diaper, Beer} {Bread, Milk, Beer}
3.	{Bread, Milk, Diaper} {Bread, Milk, Diaper, Beer} 20 {Bread, Diaper, Beer} 40 {Milk, Diaper, Beer} 40	ϕ
4.	ϕ	ϕ

* Issues with Apriori Algorithm:

-) Apriori algo seems simple only for small data sets.

-) In much larger data set especially those with huge amount of items present in low quantities and small amount of items present in big quantities.

→ The search and calculations becomes a computationally intensive process.

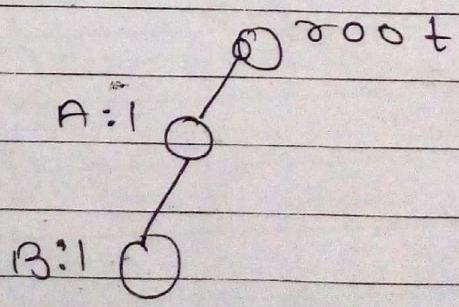
* FP Tree [Frequent Pattern]

→ FP tree is a compress representation of the database once an FP tree has been constructed it uses a recursive divide & conquer approach to mine frequent item set.

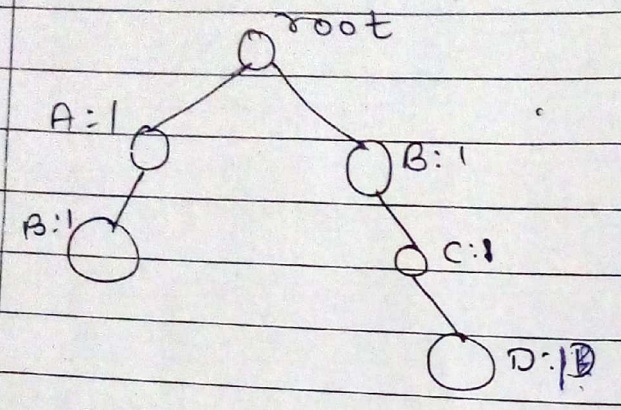
Develop FP tree for following database :

TID	Items
1	{A, B}
2	{B, C, D}
3	{A, C, D, E}
4	{A, D, E}
5	{A, B, C}
6	{A, B, C, D}
7	{B, C}
8	{A, B, C}
9	{A, B, D}
10	{B, C, E}

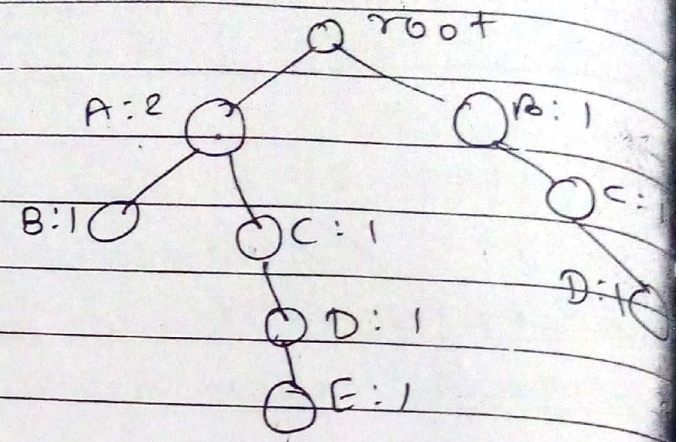
1.) After reading TID = 1



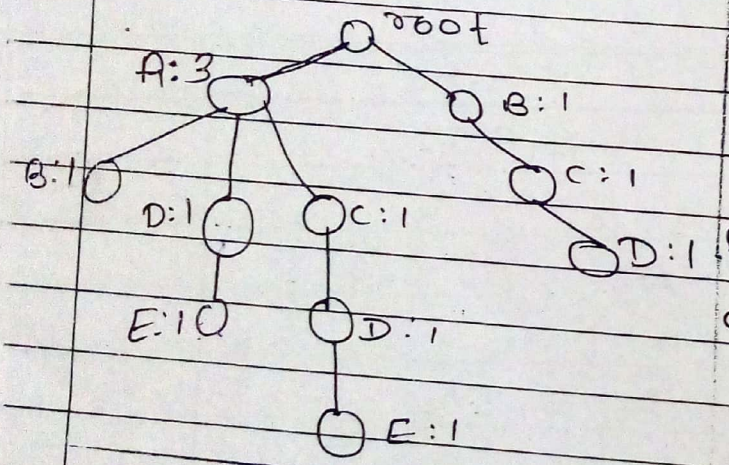
2. After reading TID = 2



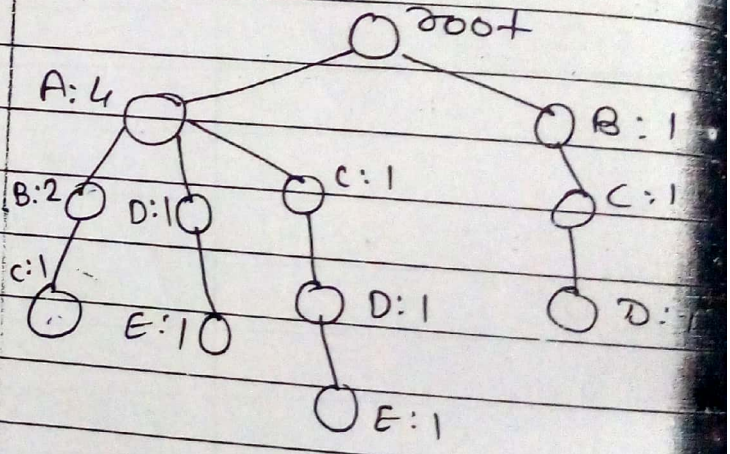
3. After reading TID = 3



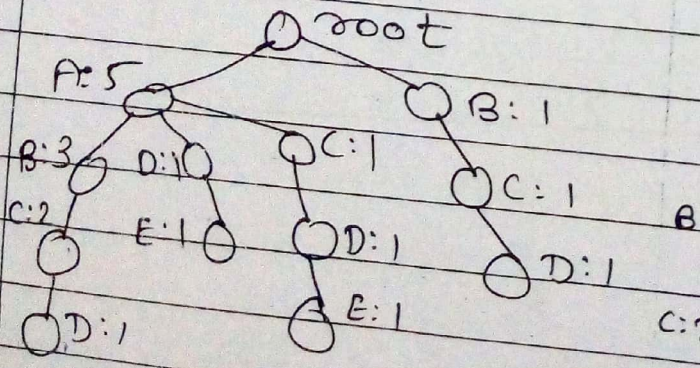
4. After reading TID = 4



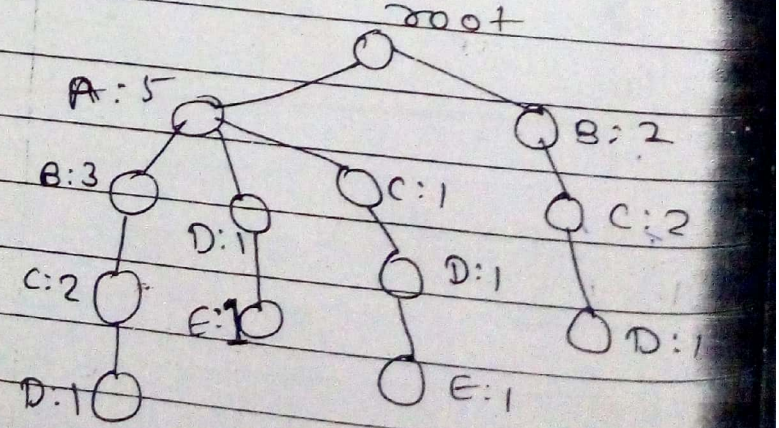
5. After reading TID = 5



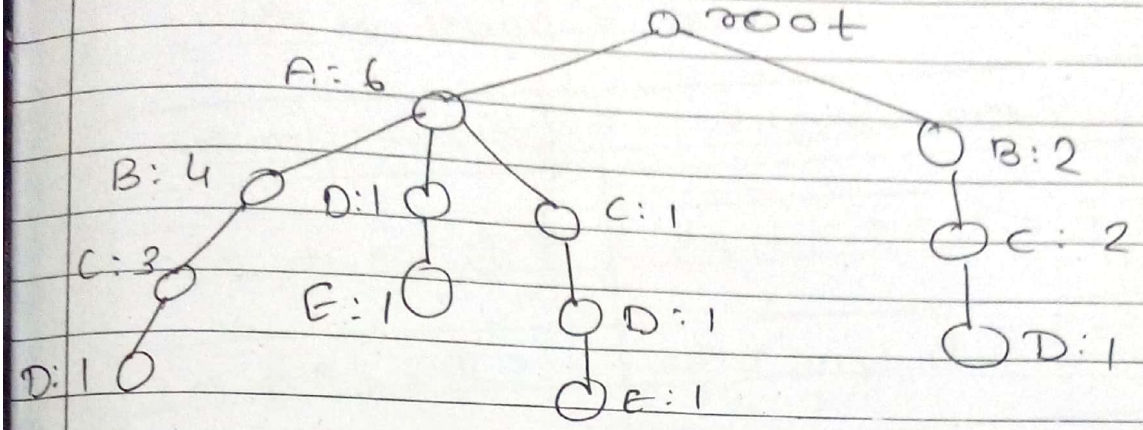
6. After reading TID = 6



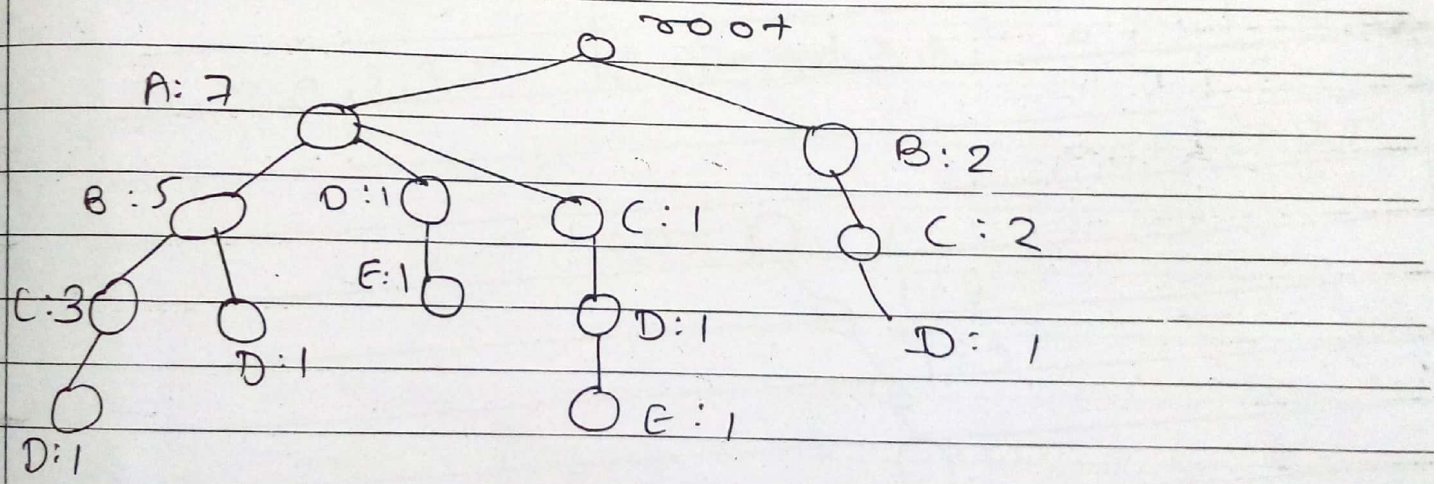
7. After reading TID = 7



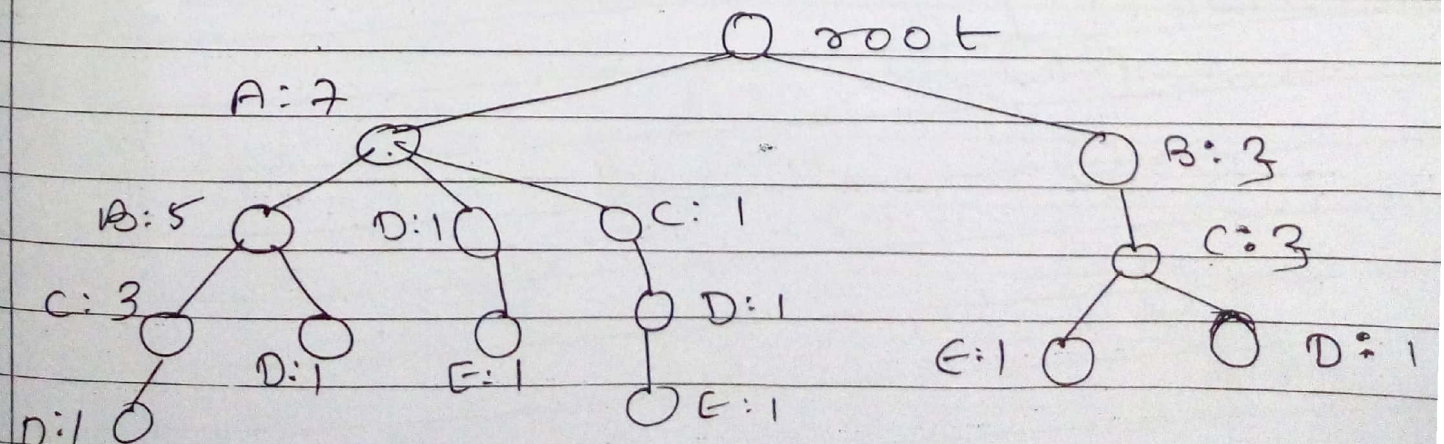
8. After reading TID = 8



9. After reading TID = 9



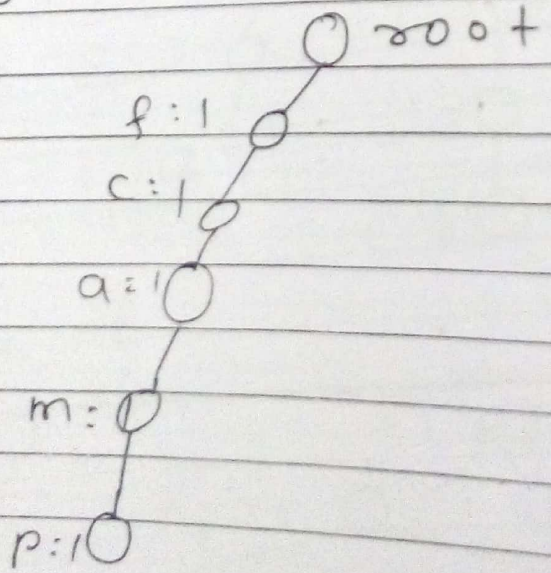
10. After reading TID: 10



3. Construct FPT for following transaction database with minimum support as true.

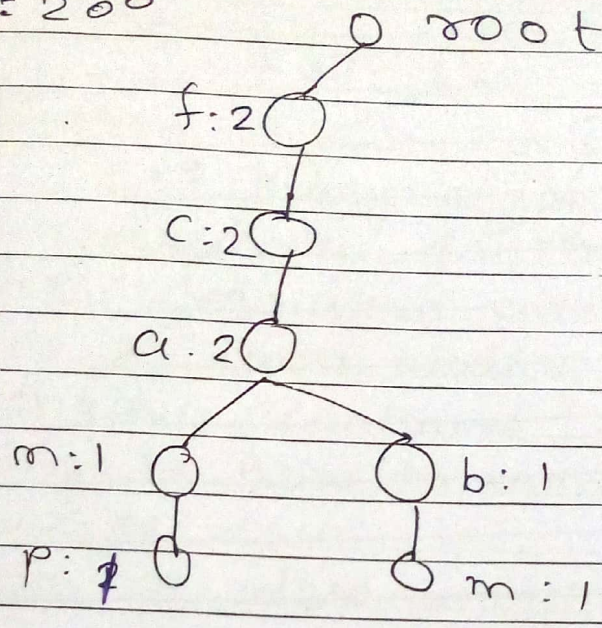
TID	Item bought	Frequent item set
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, n, i, o}	{f, b}
400	{b, e, k, s, p}	{f, b, p}
500	{a, f, c, e, i, p, m, n}	{f, c, a, m, p}

TID: 100

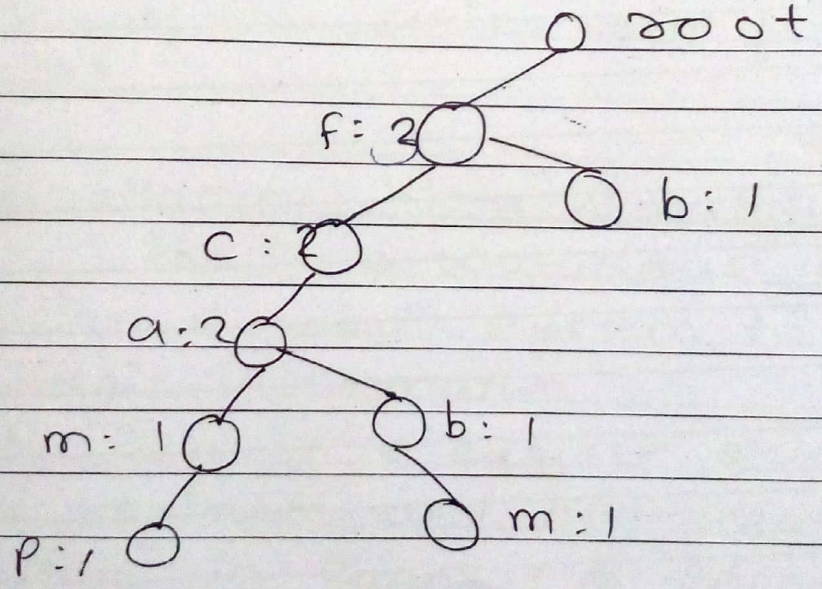


Date 1/1
Page 29
STUDY BUDDIES

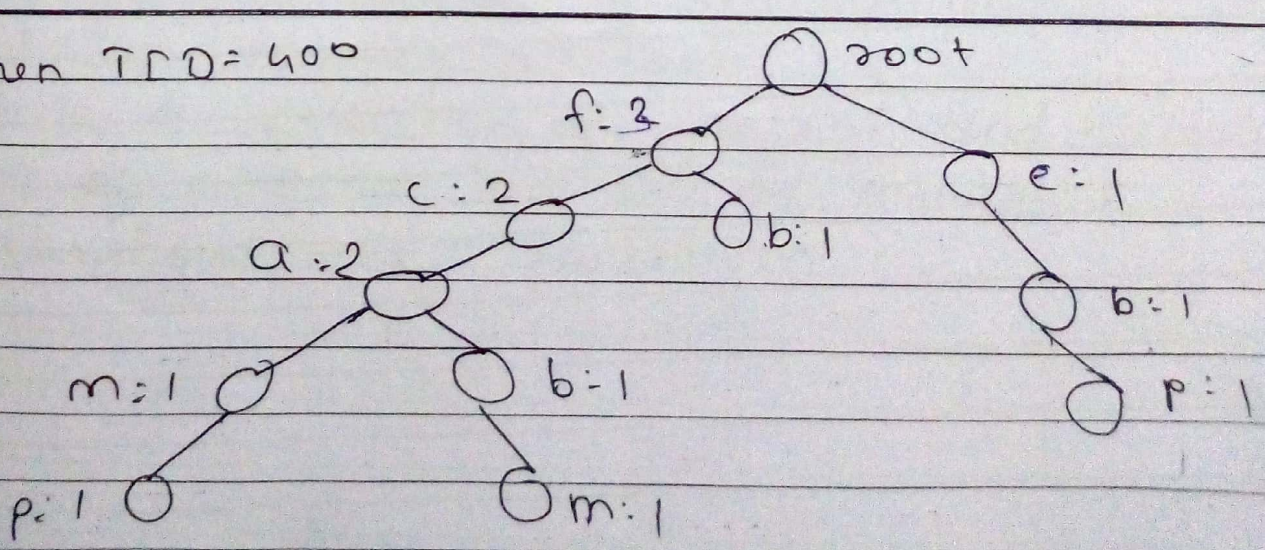
i) TID: 200



ii) TID: 300



) when TID: 400



v) when TID= 500