Declaration and statement of authorship

I, bearing Registration Number 106118036, agree and acknowledge that:

1. The assessment was answered by me as per the instructions applicable to each assessment, and that I have not resorted to any unfair means to deliberately improve my performance.

2. I have neither impersonated anyone, nor have I been impersonated by any person for the purpose of assessments.

Signature of the Student : 

Full Name : Harshit Agrawal

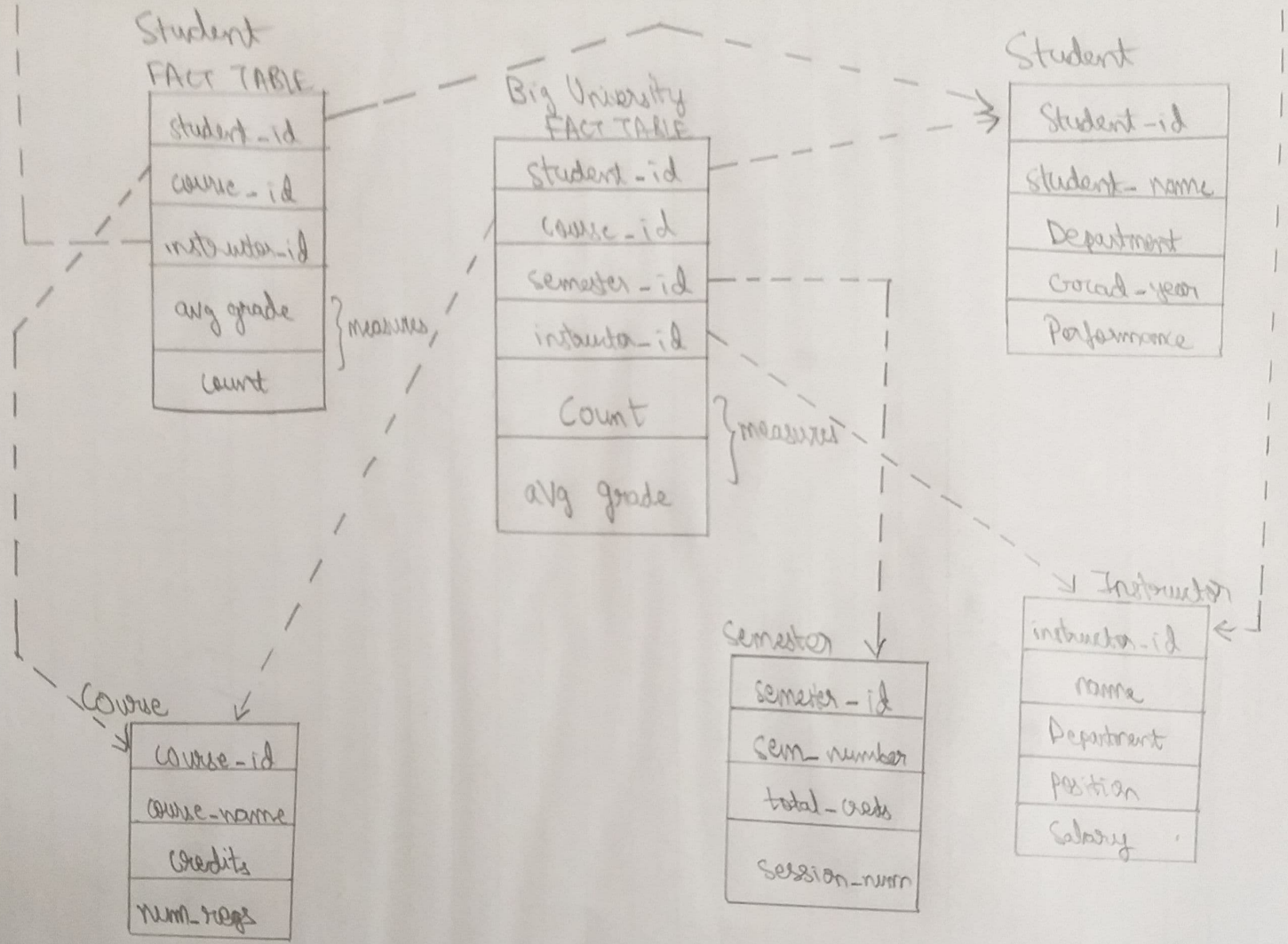
Roll No. : 106118036

Sub Code : CSPE 14

Mobile No. : 7016004637

1.) a) Galaxy Schema

106118036 DWDM ENDSEM



Expt. No. _____ Date _____

Page No. _____

b) DMQL :-

- define CUBE BIG-UNIVERSITY [student-id, course-id, semester-id, instructor-id] :
  count = count (*)
  avg-grade = avg (avg-grade)
- define CUBE STUDENT [student-id, course-id, instructor-id] :
  avg-grade = avg (avg-grade)
  count = count (*)
- define dimension Student as (student-id, student-name, Department, grad-year, performance)
- define dimension Course as (course-id, course-name, credits, num-reqs)
- define dimension Semester as (semester-id, sem_number, total-creds, session-num)
- define dimension Instructor as (instructor-id, name, Department, position, salary)

Teacher's Signature : _____

2) Component of a Data Mining System:-

a) Data Source:-
→ This is the actual source of the database
→ It consists of data warehouses, world wide web databases.
→ This data is cleaned and can be used.

b) Data Preprocessing:-
→ Before passing data to the databases or data warehouses, it has to be cleaned, integrated and filtered.
→ Data comes in various formats and cannot be used directly due to compatibility issues.
→ Therefore data has to be cleaned and modified.

c) Data Base or data Warehouse Server:-
→ This is responsible for fetching the relevant data, based on user's data mining request.

d) Knowledge Base:-
→ This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
→ Such knowledge can include concept hierarchy and is used to organize attributes or attribute values into different levels of abstraction.

e) Data Mining Engine:-

→ Contains several modules for operating association rule mining, classification, clustering, predicting, time-series analysis etc.

→ This is the core of our data mining architecture.

→ It has software, instruments to derive knowledge, insights etc.

f) Pattern Evaluation:-

→ This component typically employs interestingness measures and interacts with a data mining module to focus the search towards interesting patterns.

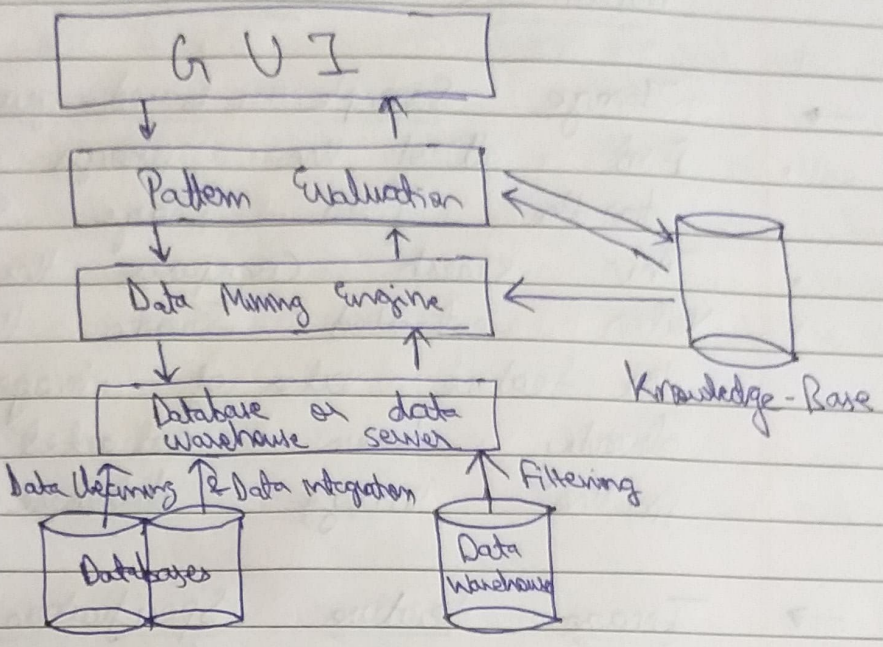
→ It is used to confine the search only to the interesting patterns.

g) Graphical User Interface:-

→ This module communicates between the user and the data mining system.

→ It allows the user to interact with the system by specifying a data mining query or provide information to focus the search providing intermediate data mining results.

Architecture of DM System



→ Most banks and financial institutions offer a wide variety of banking services, credit and investment services.

→ Multi-dimensional data analysis can be used to analyze properties like viewing debt, and revenue & other spatial features.

→ Loan payment prediction & customer credit policy analysis. Data mining methods such as feature selection and attribute relevance ranking may help identify important factors and eliminate irrelevant ones.

→ Classification and clustering of customers for targeted marketing. These methods can be used for customer group identification and targeted marketing.

106118036

DWDM Endsem

3.) Naive Bayes:-

We have to predict the class label for
Outlook (Ol.) = Rain
Temperature (T) = Mild
Humidity (H) = Normal
Wind (W) = Strong.

→ So, we create look up tables for all the 4 attributes:-

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

Wind	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

$P(Yes) = 9/14$

$P(No) = 5/14$.

Now,

$X = \{Ot. = Rain, T = Mild, H = Normal, W = Strong\}$

So, ~~for~~ Play = Yes, we have the following:

$P(Rain \mid Yes) = 3/9$

$P(Mild \mid Yes) = 4/9$

$P(Normal \mid Yes) = 6/9$

$P(Strong \mid Yes) = 3/9$

$P(Yes) = 9/14$

$\therefore$ ~~$P(\text{Play} \mid X)$~~ $= P(Yes \mid X) = \dfrac{P(X \mid Yes) \cdot P(Yes)}{P(X)}$

$\therefore P(Yes \mid X) \propto P(X \mid Yes) \cdot P(Yes)$

$= \frac{3}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14}$

$= \boxed{0.021164} \rightarrow (i)$

~~for~~ Play = No,

$P(No \mid X) \propto P(X \mid No) \cdot P(No)$

$= \frac{2}{5} * \frac{2}{5} * \frac{1}{5} * \frac{3}{5} * \frac{5}{14}$

$= \boxed{0.006857} \rightarrow (ii)$

So, we can see

$$P(yes \mid X) > P(No \mid X)$$

∴ YES, the person has higher chance to play tennis in the given whether.

4)

Min Support = 30%.

Total transactions = 10

∴ Min Support count = 30% of 10 = 3

Now, we find the frequent item sets:-

C1:

Items	Support count
{Laptop}	5
{Printer}	7
{Tablet}	9
{Headset}	6
{Monitor}	5

L1:

Items	Support count
{Laptop}	5
{Printer}	7
{Tablet}	9
{Headset}	6
{Monitor}	5

C2: Items	Support count
{Laptop, Printer}	3
{Laptop, Tablet}	4
{Laptop, Headset}	4
{Laptop, Monitor}	(2)
{Printer, Tablet}	6
{Printer, Headset}	4
{Printer, Monitor}	3
{Tablet, Headset}	6
{Tablet, Monitor}	4
{Headset, Monitor}	(2)

L2: Items	Support count
{Laptop, Printer}	3
{Laptop, Tablet}	4
{Laptop, Headset}	4
{Printer, Tablet}	6
{Printer, Headset}	4
{Printer, Monitor}	3
{Tablet, Headset}	6
{Tablet, Monitor}	4

C3: Items	Support count
{Laptop, Printer, Tablet}	(2)
{Laptop, Printer, Headset}	(2)
{Laptop, Tablet, Headset}	4
{Printer, Monitor, Tablet}	(2)
{Tablet, Printer, Headset}	4

L_3:

Items	Support count
{Laptop, Tablet, Headset}	4
{Tablet, Printer, Headset}	4

We stop here as one of its subset of C_4 is not in L_3.

So, the frequent item sets are :-

Items
{Laptop}
{Printer}
{Tablet}
{Headset}
{Monitor}
{Laptop, Printer}
{Laptop, Tablet}
{Laptop, Headset}
{Printer, Tablet}
{Printer, Headset}
{Printer, Monitor}
{Tablet, Headset}
{Tablet, Monitor}
{Laptop, Tablet, Headset}
{Tablet, Printer, Headset}

5) ⊛ Similarity based retrieval and search in multimedia databases :-

→ Image Sample based queries :-

- Find all of the images that are similar to the given image sample.
- This search compares the feature vectors extracted from the sample with the feature vectors of images that have already been extracted and indexed in the image database.

→ Image feature Specification queries :-

- Specify or sketch image features like color, texture or shape which are translated into a feature vector to be matched with the feature vectors of the images in the database.

→ Techniques for similarity based retrieval :-

- Color histogram-based signature
- Multifeature composed signature.
- Wavelet based signature.
- Wavelet-based signature with region-based granularity.

Mining association rules in multimedia databases:

→ Associations b/w image content and non-image content features:

- A rule like "if atleast 50% of the upper part of the picture is RED, it is likely to represent fire" belongs to this category since it links the image content to the keyword sky.

→ Associations among image contents that are not related to spatial relationships :-

- A rule like "a picture contains two GREEN squares, it is likely to contain one BLUE circle as well" belongs to this category since the associations are all regarding image contents.

→ Associations among image contents related to spatial relationships :-

- A rule like "A red triangle is between two yellow squares, it is likely there is a big oval-shaped object underneath" belongs to this category since it associates objects in the image with spatial relationships.

→ Data preprocessing is important when mining image data and can include data cleaning, data transformation and feature extraction.

→ Standard methods used in pattern recognition, such as edge detection can be explored.

→ Since the image data are often in large volumes and may require substantial processing power, parallel and distributed processing is useful.

→ To mine association among multimedia objects, we can treat each image as a transaction and find frequently occurring patterns among different images.