Sem-VI / Comp / 22-12-10 / Datawarehousing & mining.

ws Oct Sacn- 10 56

Con. 6135–10. **(REVISED COURSE)** **GT–7566**

(3 Hours) **[Total Marks : 100**

NOTE: (1) Question No.1 is **Compulsory**.
(2) Solve any **four** questions out of remaining six questions.
(3) All questions carry **equal** marks.
(4) Assume **suitable data** if **required**.
(5) **Figures** to the **right** indicate full marks.

Q.1) a) Explain ETL of data warehousing in detail. (10)

b) Explain data mining as a step in KDD. Give the architecture of typical DM system. (10)

Q.2) a) A dimension table is wide; the fact table is deep. Explain.
What is STAR schema and its advantages (10)

b) What is Clustering? Explain K-means clustering algorithm.
Suppose the data for clustering is {2,4,10,12,3,20,30,11,25} consider K=2,
cluster the given data using above algorithm. (10)

Q.3) a) Consider the transaction database given below. Use Apriori Algorithm with minimum support count 2, Generate the association rules along with its confidence : (10)

TID	List of items
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

b) Explain the characteristics of the data present in the data warehouse. (10)

[TURN OVER

Q.4) a) Explain HITS algorithm. (10)

b) Define Data Warehouse. Explain the architecture of data warehouse with suitable block diagram. (10)

Q.5) a) Distinguish between: (10)
(i) Top-Down and Bottom-Up Approach
(ii) OLAP and OLTP

b) Explain Partitioning Methods for Clustering. (10)

Q.6) a) Explain different OLAP operations. (10)

b) Given the training data for height classification, classify the tuple, t = < Rohit, M, 1.95m > using Bayesian Classification. (10)

Name	Gender	Height	Output
Kiran	F	1.6m	Short
Jatin	M	2m	Tall
Madhuri	F	109m	Medium
Manisha	F	1.88m	Medium
Shilpa	F	1.7m	Short
Bobby	M	1.85m	Medium
Kavita	F	1.6m	Short
Dinesh	M	1.7m	Short
Rahul	M	2.2m	Tall
Shree	M	2.1m	Tall
Divya	F	1.8m	Medium
Tushar	M	1.95m	Medium
Kim	F	1.9m	Medium
Aarti	F	1.8m	Medium
Rajashree	F	1.75m	Medium

Q.7) Write short notes on (Any **Four**): (20)

(a) Web Structure Mining
(b) Decision Tree based Classification approach
(c) Crawlers
(d) Metadata
(e) Web personalization

1. (a) What is dimentional modeling ? Explain in detail. **10**

   (b) What is datamining ? What are techniques and applications of datamining ? Explain the architecture of typical datamining system. **10**

2 (a) Explain how Apriori algorithm is useful in identifying frequent item set ? **10**

   (b) Explain major factors related to performance of DT based datamining techniques. **10**

3. (a) Explain in detail HITS algorithm. **10**

   (b) Explain Architecture of data warehouse. **10**

4. (a) Design Star Schema for autosales analysis of the company. **10**

   (b) Explain the techniques for web structure mining. **10**

5. (a) Explain Snowflake Schema with example. **10**

   (b) Explain in detail metadata and its various types. **10**

6. (a) Explain ID3 algorithum. What are the pros and cones of it ? **10**

   (b) Explain ETL of data warehousing in detail. **10**

7. Write short notes on :–

   (a) Similarity and distance measures in datamining **10**

   (b) Advanced Association rules. **10**

T.E. Sem VI (Rev) 14 Dec -12
Comp Sub:- DWM

ws-Con-2012
Con. 8161-12. KR-9224

(3 Hours) [Total Marks : 100

Note: 1. Question 1 is compulsory
2. Answer any 4 out of the remaining questions.
3. Answers to sub questions must be written together

Q1. (A) Consider the following database for a chain of bookstores.
BOOKS (Booknum, Primary_author, Topic, Total_stock, price)
BOOKSTORE (Storenum, City, State. Zip, Inventory_value)
STOCK (Storenum, Booknum, Qty)
With respect to the above business scenario, answer the following questions. Clearly state any reasonable assumptions you make.
(a) Design an information package diagram. (5)
(b) Design a star schema for the data warehouse clearly identifying the Fact table(s), Dimension table(s), their attributes and measures. (5)

(B) Consider the 5 transactions given below. If minimum support is 30% and minimum confidence is 80%, determine the frequent itemsets and association rules using the a priori algorithm.

Transaction	Items
T1	Bread, Jelly, Butter
T2	Bread, Butter
T3	Bread, Milk, Butter
T4	Coke, Bread
T5	Coke, Milk

(10)

Q2. Define the following terms by giving examples
(a) Factless Fact tables
(b) Snowflake Schema
(c) Web structure Mining
(d) Classification (5 X 4 =20)

Q3. (a) Explain the ETL cycle for a data warehouse in detail. (10)
(b) Give five examples of applications that can use Clustering. Describe any one clustering algorithm with the help of an example. (10)

Q 4. (a) Consider a data warehouse storing sales details of various goods sold, and the time of the sale. Using this example describe the following OLAP operations
(1) Slice (2) Dice (3) Rollup (4) Drill down (10)
(b) With a neat diagram describe the KDD process (10)

Q5. (a) What do you mean by web mining? Explain any one web mining algorithm. (10)
(b) Describe the different features of a web enabled data warehouse. Give two example applications where such a system would be used. (10)

Q6. (a) Explain spatial and temporal data mining (10)
(b) What is the role of Meta data in a data warehouse? Illustrate with examples (10)

Q7. Describe through a short note each of the following topics:
(a) DMQL
(b) Visualization techniques for Data warehousing and mining (10 X 2 = 20)

Con. 9243–13. **LJ–11650**

(3 Hours) **[Total Marks : 100**

N.B.: (1) Question No. **1** is **compulsory**.
(2) Answer any **four** questions out of the remaining **six** questions.
(3) Assume data if **required**, and state **clearly**.

Q1. (A) What is a Data ware house? Explain the three tier architecture of a Data Ware house with a block diagram. 10

Q1. (B) Explain Data mining as a step in KDD. Explain the architecture of a typical DM system. 10

Q2. (A) What is meant by market- basket analysis? Explain with an example. State and explain with formula the meaning of following terms 10

(I) Support

(ii) Confidence

(iii) Iceberg Queries

Hence explain how to mine multilevel Association rules from transaction databases, with examples.

Q2. (B) What is meant by Web Mining? Explain any one Web mining Algorithm. 10

Q3. (A) All Electronic company have sales department sales, consider three dimensions namely 10

(i) Time (ii) Product (iii) Store

The schema contains central fact table sales with two measures

(I) Dollars-cost and (ii) Units-Sold

Using the above example, describe the following OLAP operations

(I) Dice (ii) Slice (iii) Roll-Up (IV) Drill-Down

Q3. (B) Explain ETL (Extract Transform Load) cycle in a Data Warehouse in detail 10

Q4. (A) Compare between OLAP and OLTP 10

Q4. (B) Explain in detail the HITS Algorithm 10

Q5. (A) What is meant by Information Package Diagram, For recording the information requirements for "Hotel Occupancy" having dimensions like time, hotel etc., give the information package diagram for the same, also draw the star schema and snowflake schema 10

Q5. (B) Consider the following transactions: - 10

TID	Items
01	1,3,4,6
02	2,3,5,7
03	1,2,3,5,8
04	2,5,9,10
05	1,4

Apply the Apriori Algorithm with minimum support of 30 % and minimum confidence of 75 and find the large item set L

Q6. (A) Give five examples of application that can use clustering. Describe any one clustering algorithm with an example. 10

Q6. (B) What is meant by meta data? Explain with example. Explain the different types of meta data stored in a data ware house. Illustrate with examples. 10

Q7. Write short Notes on (Any Two) 20

(a) Web Personalization
(b) Decision Tree based classification Approach
(c) Trends in Data Ware Housing
(d) Attribute Oriented Induction

Sem VI / Data ware housing & mining COMP

11-12-14

QP Code :15195

(3Hours) **Total Marks : 100**

N. B. : (1) Question **No.1** is **compulsory**.
(2) Answer any **four** out of the remaining questions.
(3) Answer to **sub** questions must be **written** together.

1. (a) What are the different characteristics of a Data Warehouse? **5**
(b) For a Supermarket Chain consider the following dimensions, namely Product, store , time, promotion. The schema contains a central fact table, sales facts with three measures unit_sales, dollars_sales and dollar_cost. Design star schema for this application. **5**
(c) Explain Web usage mining. **5**
(d) Illustrate how the supermarket can use clustering methods to improve sales. **5**

2. Define the following terms :– **20**
(a) Dimension Tables
(b) Snowflake Schema
(c) Web Structure Mining
(d) Supervised learning

3. (a) Explain Hierarchical Clustering methods. **10**
(b) Explain the Page Rank algorithm **10**

4. (a) Describe the following OLAP operations using an example: **10**
(1) Slice
(2) Dice
(3) Rollup
(4) Drill Down
(5) Pivot
(b) Consider the following transaction database: **10**

TID	Items
01	A,B,C,D
02	A,B,C,D,E,G
03	A.C.G.H.K
04	B,C,D,E,K
05	D,E,F,H,L
06	A,B,C,D,L
07	B,I,E,K,L
08	A,B,D,E,K
09	A,E,F,H,L
10	B,C,D,F

[TURN OVER

LM-Con.:11103-14.

Apply the Apriori algorithm with minimum support of 30% and minimum confidence of 70% and find all the association rule in the data set. 10

5. (a) Explain Classification Algorithms 10
   (b) Explain the ETL (Exptract, Trausform Load) cycle. 10

6. (a) Define multidimensional and multilevel association mining. 10
   (b) Explain the role of Meta data in a data warehouse. 10

7. (a) Write detailed notes on 20
   (a) Data Warehouse Architecture
   (b) K-Means Clustering

BE/Sem VIII/ Comp/ CBGS / DWM

Datawarehousing & Mining / 02·12·16

Q.P. Code : 733601

(3 Hours) **[Total Marks : 80**

Note: 1. Question No.1 is compulsory
2. Attempt any **Three** questions out of remaining questions
3. Assume suitable data wherever necessary and state them clearly

Q1 a) Consider following dimensions for a Hypermarket chain: Product, Store, Time and Promotion. With respect to this business scenario, answer the following questions. Clearly state any reasonable assumptions you make. Design a star schema. Whether the star schema can be converted to snowflake schema? Justify your answer and draw snowflake schema for the data warehouse (clearly mention the Fact table(s), Dimension table(s), their attributes and measures). [10]

b) Define linear, non-linear and multiple regressions. Plan a regression model for Disease development with respect to change in weather parameters. [10]

Q2 a) What is meant by metadata in the context of a Data warehouse? Explain the different types of meta data stored in a data warehouse. Illustrate with a suitable example. [10]

b) Describe the various functionalities of Data mining as a step in the process of knowledge Discovery. [10]

Q3 a) In what way ETL cycle can be used in typical data ware house, explain with suitable instance. [10]

b) What is Clustering Technique? Discuss the Agglomerative algorithm with the following data and plot a Dendrogram using single link approach. The table below comprises sample data items indicting the distance between the elements. [10]

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

TURN OVER

Q4 a) Discuss how computations can be performed efficiently on data cubes. [10]

b) A database has five transactions. Let min-support=60% and min-confidence = 80%. Find all Find frequent item sets by using Apriori Algorithm. T_ID is the transaction ID. [10]

T_ID	Items bought
T-1000	M, O, N , K, E, Y
T-1001	D, O, N, K, E, Y
T-1002	M, A, K, E
T-1003	M, U, C, K, Y
T-1004	C, O, O, K, E

Q5 a) Differentiate [10]

i. OLTP Vs. OLAP

ii. Data Warehouse Vs. Data Mart

b) Why naive Bayesian classification is called "naive"? Briefly outline the major ideas of naive Bayesian classification. [10]

Q6 Write short notes on any **four** of the following: [20]

i. Application of Data Mining to Financial Analysis

ii. Fact less Fact Table

iii. Indexing OLAP data

iv. Data Quality

v. Decision Tree based Classification Approach

Q. P. Code: 25962

Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory

2. Answer any three out of remaining questions.

Q1 A) A manufacturing company has a huge sales network. To control the sales, it is divided into regions. Each region has multiple zones. Each zone has different cities. Each sales person is allocated different cities. The objective is to track sales figure at different granularity levels of region and to count no. of products sold. Design a star schema by considering granularity levels for region, sales person and time. Convert the star schema to snowflake schema. [10]

B) Discuss: [10]

i) Architecture of a typical data mining system.

ii) Application and major issues in Data Mining

Q2 A) Consider a data warehouse for a hospital where there are three dimension [10]
a) Doctor b) Patient c) Time
Consider two measures i) Count ii) Charge where charge is the fee that the doctor charges a patient for a visit. For the above example create a cube and illustrate the following OLAP operations.
1) Rollup 2) Drill down 3) Slice 4) Dice 5) Pivot.

B) Consider the data given below. Create adjacency matrix. Apply single link algorithm to cluster the given data set and draw the dendogram [10]

Object	Attribute 1 (X):	Attribute 2 (Y):
A	2	2
B	3	2
C	1	1
D	3	1
E	1.5	0.5

Q3 A) Define Metadata. Discuss the types of Metadata stored in a data warehouse. Illustrate with an example. [10]

B) Discuss different steps involved in Data Pre-processing [10]

Q4 A) Discuss various OLAP Models and their architecture [10]

B) Define Classification. Discuss the issues in Classification. A simple example from the stock market involving only discrete ranges has profit as categorical attribute, with values { Up, Down} and the training data is: [10]

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Apply decision tree algorithm and show the generated rules.

Q5 A) Differentiate top-down and bottom-up approaches for building data warehouse. [10]
Discuss the merits and limitations of each approach.

B) i) Discuss Association Rule Mining and Apriori Algorithm. [10]

ii) A database has four transactions. Let minimum support = 50% and minimum confidence = 50%

TID	Items-bought
T100	A,B,C
T200	A,C
T300	A,D
T400	B,E,F

Find all frequent item sets using apriori algorithm. List strong association rules.

Q6 Write short note on the following (Answer any **FOUR**) [20]

a) Fact Constellation

b) Data visualization

c) FP Tree

d) DBSCAN

e) ETL Process

****

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

Department of Computer Engineering
Academic Year 2021-2022
Term Test – I

Course Name: Data Mining and Warehousing **Course Code: DJ19CEC501**
Class: TE (A & B) **Sem:** V
Date: 22/10/2021 **Time**: 11:10 am – 12:10 pm
Maximum Marks: 25

Instructions:

1. Please solve questions in order with clear and dark ink pens
2. Draw figures wherever required
3. Write SAPID on each page top right corner and Sign with Name at the end of each page

Q. No	Questions	Marks
1.	Comment on all the data pre-processing techniques that uses binning. Explain in detail how binning is used in all these techniques?	04
2.	Consider a fashion brand has two stores that sell their products. The brand recorded the number of sales made each month at each store. In the past 12 months, the following was the sales data. Store 1: 350, 460, 20, 160, 580, 250, 210, 120, 200, 510, 290, 380 Store 2: 520, 180, 260, 380, 80, 500, 630, 420, 210, 70, 440, 140 a. Draw box plots for each sales store. b. Give your analysis for the sales of each store.	 02 01
3 a.	Differentiate between bagging and boosting.	02
3 b.	Explain with an example the process of learning a rule in rule-based classification.	02
3.	The following table consists of training data from an employee database. The data have been generalized. For example, "31 ... 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. Let status be the class label attribute. Construct a decision tree using ID3 technique. Draw the final tree.	06

department	*status*	*age*	*salary*
sales	senior	31...35	46K...50K
sales	junior	26...30	26K...30K
sales	junior	31...35	31K...35K
systems	junior	21...25	46K...50K
systems	senior	31...35	66K...70K
systems	junior	26...30	46K...50K
systems	senior	41...45	66K...70K
marketing	senior	36...40	46K...50K
marketing	junior	31...35	41K...45K
secretary	senior	46...50	36K...40K
secretary	junior	26...30	26K...30K

4 a.	Discuss the advantages and disadvantages of k-means clustering method.	02
4 b.	Apply k-mediods for the following distance matrix for 2 clusters and find the clusters. Justify your choice of final cluster	06

Item	A	B	C	D
A	0	1	4	5
B	1	0	2	6
C	4	2	0	3
D	5	6	3	0

T.E. Com / Sem VI / Rev.
Data warehousing & Mining

Con. 3881-10. **(REVISED COURSE)** **AN-4472**

(3 Hours) **[Total Marks : 100**

N.B. (1) Question No. **1** is **compulsory**.
(2) Attempt any **four** questions out of remaining **six** questions.

1. (a) Define Data Warehouse. Explain the architecture of data warehouse with suitable block diagram. **10**
   (b) Explain data mining as a step in KDD. Give the architecture of typical DM system. **10**

2. (a) How are top-down and bottom-up approaches for building data warehouse differ ? Discuss the merits and limitation of each approach. **10**
   (b) What is K-means clustering ? Confer the K-means algorithm with the following data for two clusters. Data set { 10, 4, 2, 12, 3, 20, 30, 11, 25, 31 } **10**

3. (a) Give information package for recording information requirement for "Hotel Occupancy" considering dimensions like time, Hotel etc. Design star schema from the information package. **10**
   (b) Explain HITS algorithm. **10**

4. (a) What is Classification ? What are the issues in classification ? Apply statistical based algorithm to obtain the actual probabilities of each event to classify the new tuple as a tall. Use the following data – **10**

Person ID	Name	Gender	Height	Class
1	Kristina	Female	1·6 m	Short
2	Jim	Male	2 m	Tall
3	Maggi	Female	1·9 m	Medium
4	Marya	Female	2·1 m	Tall
5	Stephanie	Female	1·7 m	Short
6	Bob	Male	1·85 m	Medium
7	Catherine	Female	1·6 m	Short
8	Dave	Male	1·7 m	Short
9	Wilson	Male	2·2 m	Tall

   (b) Define Metadata. What are the different types of metadata stored in a data warehouse ? Illustrate with a simple customer sales data warehouse. **10**

[TURN OVER

5. (a) What is Clustering Techniques ? Discuss the Agglomerative algorithm using following data and plot a Dendrogram using single link approach. The following figure contains sample data items indicting the distance between the elements :– **10**

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

(b) All electronics company have sales deptartment Sales consider three dimensions namely **10**

(i) Time (ii) Product (iii) Store.

The schema contain a central fact table sales with two measures.

(i) dollars–cost and (ii) units-sold

Using the above example describe the following OLAP operations :–

(i) Dice (ii) Slice (iii) Roll-up (iv) Drill-down

6. (a) Explain ETL of data warehousing in detail. **10**

(b) Consider the following transactions :– **10**

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3, 5, 8
04	2, 5, 9, 10
05	1, 4

Apply the Apriori Algorithm with minimum support of 30% and minimum confidence of 75% and find the large item set L.

7. Write short notes on any **four** :– **20**

(a) Trends in data warehousing

(b) Decision tree based classification approach

(c) Key restructuring

(d) Crawlers

(e) Web personalization.

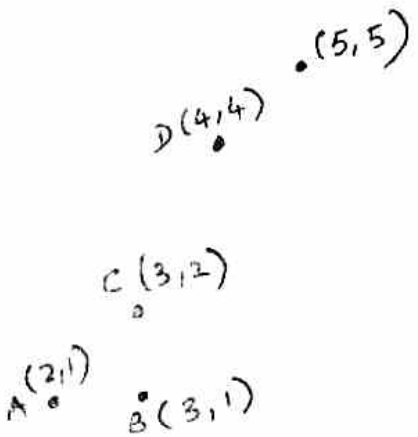
1. (a) Describe the steps in the KDD process with a suitable block diagram. 5
   (b) Compare between OLTP and OLAP. 5
   (c) What will be the effect of performing attribute oriented Induction (AOI) on the initial working relation **student** with attributes such as name, gender, birth-date, birth place, address, phone-no, and gpa. 10

2. (a) Using the table given below, create a classification model using decision true technique. Indicate how to utilize the model to estimate the risk category of the customer with (**Credit-History** – bad, **Debt** – high, **Collateral** – none, **Income** – (15-35k)). 10

Sr. No.	Debt	Collateral	Income	Credit - History	Risk
1	high	none	0-15 k	bad	high risk
2	high	none	15-35 k	unknown	high risk
3	low	none	15-35 k	unknown	Moderate risk
4	low	none	0-15 k	unknown	high risk
5	low	none	over 35 k	unknown	low risk
6	low	adequate	over 35 k	unknown	low risk
7	low	none	0-15 k	bad	high risk
8	low	adequate	over 35 k	bad	Moderate risk
9	low	none	over 35 k	good	low risk
10	high	adequate	over 35 k	good	low risk
11	high	none	0-15 k	good	high risk
12	high	none	15-35 k	good	Moderate risk.

   (b) Define a data warehouse. Explain the architecture of data warehouse with suitable block diagram. 10

3. (a) Consider the data set given. Create the adjacency matrix. Use single link agglomerative technique to cluster the given data. Draw the dendogram. **4+4+2**



   (b) What are the different ways of finding the distance between two clusters ? 5
   (c) Define Factless Fact tables with a suitable example. 5

[TURN OVER

comps - sem 6 - data warehouse & mining

4. (a) What is Association Rule Mining ? Give the Apriori algorithm. Apply AR Mining to find all frequent itemsets from the following table :— **2+4+4**

Transcation – ID	Items
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	1, 3
700	1, 3, 2, 5
800	1, 3
900	1, 2, 3

Minimum Support Count = 2
Minimum – Confidence – 70%.

(b) Explain the major steps in the ETL process with a suitable diagram and an example. **10**

5. (a) The college wants to record the grades for the courses completed by students. There are four dimensions :—
   (i) Course (iii) Student
   (ii) Professor (iv) Period.
   The only fact that is to be recorded in the table is course-grade :–
   (i) Design star schema. **5**
   (ii) Write DMQL for the above star schema. **5**

   (b) Using the above example describe the following OLAP operations :— **10**
   Slice, Dice, Roll-up, Drill-down, Pivot.

6. (a) What are crawlers ? How do periodic crawlers differ from incremental crawlers ? Give the architecture of focussed crawlers and explain how it is used. **2+2+6**

   (b) Explain how HITS algorithm finds hubs and authoritative pages. **10**

7. Write short notes on (any **four**) :— **20**
   (a) Qutlier Mining
   (b) Applications of Web Usage Mining
   (c) Snowflake Schema
   (d) Generalized Association Rules
   (e) Top-down and Bottom-up approaches in data warehousing.

T.E Sem VI (comp) Rev
DWM
5/6/12

AGJ 1st half 1

Con. 4721-12. **GN–9218**

(3 Hours) **[Total Marks : 100**

N.B. : (1) Question No. **1** is **compulsory**.
(2) Solve any **four** out of the **remaining**.
(3) Draw suitable diagrams wherever **necessary**.
(4) Assume **suitable** data (if required)

1. (a) Define a data warehouse. Explain what is the need for developing a data warehouse and hence explain its architecture. **10**
   (b) Compare OLTP and OLAP systems. Explain the steps in KDD with a suitable block diagram. **10**

2. (a) What is meant by ETL ? Explain the ETL process in detail. **10**
   (b) State and explain the various schemas used in data warehousing with examples for each of them. **10**

3. (a) Differentiate between top down and bottom-up approaches for building a data warehouse. Explain the advantages and disadvantages of each of them. **10**
   (b) Define what is meant by information package diagram. For recording the information requirements for "hotel occupancy" having dimensions like time, hotel etc, give the information package diagram for the same, also draw the star schema and snow flake schema. **10**

4. (a) What is meant by meta data ? Explain with an example. Explain the different types of meta data stored in a data warehouse. **10**
   (b) Explain what is meant by association rule mining. For the table given below perform opriori algorithm. Also – **10**
      (i) Determine the k-item sets (frequent) obtained.
      (ii) Justify the strong association rule that has been determined i.e. specify which is the strongest rule obtained.

      The table is as follows -

TID	Items
01	1, 3, 4, 6
02	2, 3, 5, 7
03	1, 2, 3, 5, 8
04	2, 5, 9, 10
05	1, 4

      Assume Minimum support of 30% and
      Minimum confidence of 75%.

[TURN OVER

5. (a) Explain dimension modelling in detail. 10
(b) Explain what is meant by clustering. State and explain the various types with suitable example for each. 10

6. (a) What is meant by classification ? Justify why clustering is said to be supervised learning. How is the classifier accuracy determined and also explain its various types. 10
(b) What is meant by market-basket analysis ? Explain with an example. State and explain with formula the meaning of the terms :– 10
(i) Support
(ii) Confidence
(iii) Iceberg queries.
Hence explain how to mine multi level association rules from transaction databases, with example for each.

7. Write short notes on (any **two**) :– 20
(a) OLAP operations
(b) Data warehouse deployment and maintenance
(c) Attribute oriented induction
(d) Web mining.

T.E (CMPN) sem VI (Rev)
Data Warehouse &
Mining

501 : Con. No.-JP

Con. 9998–13. **GS–1369**

(3 Hours) **[Total Marks : 100**

Note: 1. Question 1 is compulsory
2. Answer any 4 out of the remaining questions.
3. Answers to sub questions must be written together

Q.1 (a) What are differences between Data Warehouse and Data Mart ? **(05)**

(b) For a Supermarket Chain consider the following dimensions, namely Product, store, time, promotion. The schema contains a central fact table, sales facts with three measures unit_sales, dollars_sales and dollar_cost. Design star schema for this application. **(05)**

(c) Calculate the maximum number of base fact table records for warehouse with the following values given below : **(05)**

- Time period: 5 years
- Store: 300 stores reporting daily sales
- Product: 40,000 products in each store (about 4000 sell in each store daily)

(d) Illustrate how the supermarket can use clustering methods to improve sales. **(05)**

Q2. Define the following terms by giving examples
(a) Factless fact tables
(b) Snowflake Schema
(c) Web Structure Mining
(d) Concept Hierarchy **(5 X 4 =20)**

Q.3 (a) Apply Agglomerative Hierarchical Clustering and draw single link and average link dendrogram for the following distance matrix. **(10)**

	A	B	C	D	E
A	0	2	6	10	9
B	2	0	3	9	8
C	6	3	0	7	5
D	10	9	7	0	4
E	9	8	5	4	0

(b) Explain the Page Rank technique with algorithm. **(10)**

Q 4.(a) Consider a data warehouse for a hospital, where there are three dimensions:
(1) Doctor (2) Patient (3) Time; and two measures: (1) Count & (2) Fees;
For this example create a OLAP cube and describe the following OLAP operations:
(1) Slice (2) Dice (3) Rollup (4) Drill Down (5) Pivot **(10)**

[TURN OVER

(b) Consider the following transaction database:

TID	Items
01	A, B, C, D
02	A, B, C, D, E, G
03	A, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Apply the **Apriori** algorithm with minimum support of 30% and minimum confidence of 70%, and find all the association rules in the data set. **(10)**

Q 5.(a) A simple example from the stock market involving only discrete ranges has Profit as categorical attribute, with values {up, down}. and the training data is:

AGE	COMPETITION	TYPE	PROFIT
Old	Yes	software	Down
Old	No	software	Down
Old	No	hardware	Down
Mid	Yes	software	Down
Mid	Yes	hardware	Down
Mid	No	hardware	Up
Mid	No	software	Up
New	Yes	software	Up
New	No	hardware	Up
New	No	software	Up

Apply the decision tree algorithm and show the generated rules. **(10)**

(b) Describe the steps of the ETL (Extract - Transform - Load) cycle. **(10)**

Q6. (a) Define multidimensional and multilevel association mining. **(10)**

(b) Explain the role of Meta data in a data warehouse. **(10)**

Q7. Write detailed notes on:

(a) Data Warehouse Architecture

(b) K-Means Clustering **(10 X 2 = 20)**

T.E. (comp) - Data warehousing & Min

Sem - VI Date :→ 05/06/2014

QP Code : MV-18251

TIME - 3 Hrs **Marks – 100**

Note: 1. Question 1 is compulsory
2. Answer any 4 out of the remaining questions.
3. Answers to sub questions must be written together

Q1. A bank wants to develop a data warehouse for effective decision-making about their loan schemes. The bank provides loans to customers for various purposes like House Building Loan, Car Loan, Educational Loan, Personal Loan, etc. The whole country is categorized into a number of regions, namely, North, South, East and West. Each region consists of a set of states. Loan is disbursed to customers at interest rates that change from time to time. Also, at any given point of time, the different types of loans have different rates. The data warehouse should record an entry for each disbursement of loan to customer.

a) Design an information package diagram for the application. **(05)**
b) Design a star schema for the data warehouse clearly identifying the fact table(s), Dimensional table(s), their attributes and measures. **(05)**
c) Describe an algorithm the bank can use to cluster its potential customers, based on their attributes. **(05)**
d) Describe how data warehousing and mining help the bank increase its productivity. **(05)**

Q2. Define the following terms by giving examples
(a) Fact Constellation
(b) Snowflake Schema
(c) Aggregate Fact tables
(d) Snapshot and Transaction Tables **(5 X 4 =20)**

Q 3.(a) Consider an online travel agency that helps customers to plan and schedule their holidays. The agency maintains all past history in a data warehouse. Describe the different classes of users who could access this data warehouse and design the information delivery framework for this data warehouse. **(10)**

(b) Describe the working of the K-Means clustering algorithm with the help of the a sample dataset. **(10)**

Q 4.(a) Consider a data warehouse for a hospital, where there are three dimensions: (1) Doctor (2) Patient (3) Time; and two measures: (1) Count & (2) Fees;
For this example create a OLAP cube and describe the following OLAP operations:
(1) Slice (2) Dice (3) Rollup (4) Drill Down (5) Pivot **(10)**

(b) Consider the following transaction database:

TID	Items
01	A, B, C, D
02	A, B, C, D, E, G
03	A, C, G, H, K
04	B, C, D, E, K
05	D, E, F, H, L
06	A, B, C, D, L
07	B, I, E, K, L
08	A, B, D, E, K
09	A, E, F, H, L
10	B, C, D, F

Apply the **Apriori** algorithm with minimum support of 30% and minimum confidence of 70%, and find all the association rules in the data set. **(10)**

Q 5.(a)

Transaction	Income	Credit	Decision
1	Very High	Excellent	AUTHORIZE
2	High	Good	AUTHORIZE
3	Medium	Excellent	AUTHORIZE
4	High	Good	AUTHORIZE
5	Very High	Good	AUTHORIZE
6	Medium	Excellent	AUTHORIZE
7	High	Bad	REQUEST ID
8	Medium	Bad	REQUEST ID
9	High	Bad	REJECT
10	Low	Bad	CALL POLICE

Using the above table illustrate any one classification technique. Further indicate how we can classify a new transaction, with **(Income = Medium and Credit=Good) .** **(10)**

(b) (a) Describe clearly the different steps of the ETL (Extract - Transform - Load) cycle in Data Warehousing **(10)**

Q6.(a) Give a brief description of web mining **(10)**

(b) Explain clearly the role of Meta data in a data warehouse. **(10)**

Q7. Write detailed notes on:-

(a) Data Warehouse Architecture

(b) Hierarchical Clustering methods. **(10 X 2 = 20)**

18/5/16 Data warehousing & Mining BE comp Sem VIII

Q.P. Code : 733600 CBGS

(3 Hours) **[Total Marks : 80**

Note: 1. Question No.1 is compulsory
2. Attempt any **Three** questions out of remaining questions
3. Assume suitable data wherever necessary and state them clearly

Q1 a) For a Super market chain, consider the following dimensions namely product, store, time and promotion. The schema contains a central fact table for sales. [10]
i. Design star schema for the above application.
ii. Calculate the maximum number of base fact table records for warehouse with the following values given below:
- Time period – 5 years
- Store – 300 stores reporting daily sales
- Product – 40,000 products in each store (about 4000 sell in each store daily)

b) Discuss: [10]
i. The steps in KDD process
ii. The architecture of a typical DM system

Q2 a) We would like to view sales data of a company with respect to three dimensions namely Location, Item and Time. Represent the sales data in the form of a 3-D data cube for the above and Perform Roll up, Drill down, Slice and Dice OLAP operations on the above data cube and Illustrate. [10]

b) A simple example from the stock market involving only discrete ranges has profit as categorical attribute, with values {Up, Down} and the training data set is given below. [10]

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Apply decision tree algorithm and show the generated rules.

Q3 a) Illustrate the architecture of a typical DW system. Differentiate DW and Data Mart. [10]

b) Discuss different steps involved in Data Preprocessing. [10]

Q4 a) Discuss various OLAP Models. [10]

b) Explain K-Means clustering algorithm? Apply K-Means algorithms for the following data set with two clusters. Data Set = {1, 2, 6, 7, 8, 10, 15, 17, 20} [10]

TURN OVER

Q5 a) Describe the steps of ETL process. [10]

b) Discuss Association Rule Mining and Apriori Algorithm. Apply AR Mining to find all frequent item sets and association rules for the following dataset: [10]

Minimum Support Count = 2

Minimum Confidence = 70%

Transaction ID	Items
100	1, 2, 5
200	2, 4
300	2, 3
400	1, 2, 4
500	1, 3
600	1, 3
700	1, 3, 2, 5
800	1, 3
900	1, 2, 3

Q6 Write Short notes on any four of the following: [20]

i. Updates to Dimension tables

ii. Metrics for Evaluating Classifier Performance

iii. FP tree

iv. Multilevel & Multidimensional Association Rule

v. Operational Vs. Decision Support System

B.E. sem (VIII) DWM computer.

Q. P. Code: 17089

Time: 3 Hours **Marks: 80**

Note: 1. Question 1 is compulsory

2. Answer any three out of remaining questions.

Q1 A) What is dimensional modelling? Design the data warehouse for wholesale furniture Company. The data warehouse has to allow analysing the company's situation at least with respect to the Furniture, Customer and Time. More over, the company needs to analyse: The furniture with respect to its type, category and material. The customers with respect to their spatial location, by considering at least cities, regions and states. The company is interested in learning the quantity, income and discount of its sales. [10]

B) Discuss different steps involved in Data Pre-processing. [10]

Q2 A) The college wants to record the Marks for the courses completed by students using the dimensions: i) Course, ii) Student, iii) Time & a measure Aggregate marks Create a Cube and describe following OLAP operations:
(i) Slice (ii) Dice (iii) Roll up (iv) Drill down (v) Pivot [10]

B) Apply the Naive Bayes classifier algorithm for buys computer classification and classify the tuple X=(age="young", income="medium", student="yes" and credit-rating="fair") [10]

Id	Age	Income	Student	Credit-rating	buys computer
1	young	high	no	fair	no
2	young	high	no	good	no
3	middle	high	no	fair	yes
4	old	medium	no	fair	yes
5	old	low	yes	fair	yes
6	old	low	yes	good	no
7	middle	low	yes	good	yes
8	young	medium	no	fair	no
9	young	low	yes	fair	yes
10	old	medium	yes	fair	yes
11	young	medium	yes	good	yes
12	middle	medium	no	good	yes
13	middle	high	yes	fair	yes
14	old	medium	no	good	no

Q3 A) Explain ETL of data warehousing in details? [10]

B) Explain types of attributes and data visualization for data exploration. [10]

Q4 A) Illustrate the architecture of Data Warehouse system. Differentiate Data warehouse and Data Mart [10]

B) Explain K-Means clustering algorithm? Apply K-Means algorithms for the following Data set with two clusters. [10]
Data Set = { 15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65}

Q5 A) Explain Updates to dimension tables in detail. [10]

B) A database has ten transactions. Let minimum support = 30% and minimum Confidence = 70% [10]
i] Find all frequent patterns using Apriori Algorithm.
ii] List strong association rules.

Transaction _Id	Items
01	A,B,C,D
02	A,B,C,D,E,G
03	A,C,G,H,K
04	B,C,D,E,K
05	D,E,F,H,L
06	A,B,C,D,L
07	B,I,E,K,L
08	A,B,D,E,K
09	A,E,F,H,L
10	B,C,D,F

Q6 Write short note on the following (Answer any FOUR) [20]

a) Major issues in Data Mining

b) Metadata in Data Warehouse

c) FP Tree

d) DBSCAN

e) Hierarchical Clustering

Time: 03 Hours **Marks: 80**

Note: 1. Question 1 is compulsory
2. Answer any three out of remaining questions.

Q1 A) i. Design star & snowflake schema for "Hotel Occupancy" considering dimensions like Time, Hotel, Room, etc. [10]

ii. Calculate the maximum number of base fact table records for the values given below:
Time period: 5 years
Hotels: 150
Rooms: 750 rooms in each Hotel (about 400 occupied in each hotel daily).

B) Explain Data mining as a step in KDD. Give the architecture of typical data mining System. [10]

Q2 A) The college wants to record the marks for the courses completed by students using the dimensions: a) Course, b) Student, c) Time & a measure d) Aggregate marks. Create a Cube and describe following OLAP operations: [10]
i) Rollup ii) Drill down iii) Slice iv) Dice v) Pivot.

B) A simple example from the stock market involving only discrete ranges has profit as categorical attribute, with values {up, down} and the training data is: [10]

Age	Competition	Type	Profit
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Apply decision tree algorithm and show the generated rules.

Q3 A) Why naive Bayesian classification is called "naive"? Briefly outline the major ideas of Naive Bayesian classification. [10]

B) Discuss different steps involved in Data Pre-processing [10]

Q4 A) Explain ETL of data warehousing in detail. [10]

B) Find clusters using k-means clustering algorithm if we have several objects (4 types of medicines) and each object have two attributes or features as shown in the table below. The goal is to group these objects into k=2 group of medicine [10]

based on the two features (pH and weight index).

Object	Attribute 1(X) Weight Index	Attribute 2 (Y) pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Q5 A) Explain Data Warehouse Architecture in detail. [10]

B) A database has five transactions. Let minimum support = 30% and minimum Confidence = 70% [10]

i. Find all frequent patterns using Apriori Algorithm.

ii. List strong association rules.

Transaction _Id	Items
A	1,3,4,6
B	2,3,5,7
C	1,2,3,5,8
D	2,5,9,10
E	1,4

Q6 Write short note on the following (Answer any **FOUR**) [20]

a) Data warehouse design strategies

b) Applications of Data Mining

c) Role of metadata

d) Multidimensional and multilevel association mining

e) Hierarchical clustering

****