Name : ....

Roll No. : ....

Invigilator's Signature : ....

CS/B.Tech/IT/SEM-8/IT-802A/2013

2013

DATA WAREHOUSING & DATA MINING

Time Allotted : 3 Hours *Full Marks* : 70

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable.

GROUP – A
(Multiple Choice Type Questions)

1. Choose the correct alternatives for any *ten* of the following :

10 × 1 = 10

i) Data Warehousing is used for

a) Decision Support System

b) OLAP application

c) Database application

d) Data manipulation applications.

ii) The algorithm which uses the concept of a train running over data to find associations of items in data mining is known as

a) Apriori Algorithm

b) Partition Algorithm

c) DIC Algorithm

d) FP-Tree growth Algorithm.

iii) A star schema has what type of relationship between a dimension and fact table ?

a) Many-to-Many b) One-to-One

c) One-to-Many d) All of these.

iv) Suppose there is 1,00,000 no. of transactions, out of which 2,000 transactions contain both *A* and *B*, 800 no. of transactions contain only *C*. The support of *C* when *A* and *B* is purchased on the same trip is

a) 0·6% b) 0·8%

c) 40% d) 50%.

v) What is Metadata ?

a) Summarized data

b) Data used only by IS organization

c) Definitions of data elements

d) Any business data occurring in large volumes.

vi) A goal of data mining includes which of the following ?

a) To explain some observed event or condition

b) To confirm the data exists

c) To create hidden patterns

d) To create a new data warehouse.

vii) Which of the following is false ?

a) Any superset of an infrequent set is also infrequent

b) Any subset of a frequent set is infrequent

c) Data mining is one of the steps in KDD

d) *K*-means is a clustering based algorithm.

viii) Decision Tree uses .... data to determine the rules.

a) Test b) Data Warehouse

c) Training d) Transaction.

ix) FP tree algorithm is

a) Frequent Position tree

b) Frequent Pattern tree

c) Frequent Pairwise tree

d) Frequent Parameter tree.

x) A data mart differs from a data warehouse in that the

a) data mart has a smaller scope

b) data mart may be restricted to a particular type of data

c) data mart may be restricted to a particular business function

d) data mart may be restricted to a particular business unit or location

e) all of these.

xi) The slice operation deals with

a) selecting all but one dimension of the data cube

b) merging cells of all but one dimension

c) merging the cells along one dimension

d) selecting the cells of any one dimension of the data cube.

xii) ROLAP is preferred over MOLAP when

a) a data warehouse and relational database are inseparable

b) the data warehouse is in relational tables, but no slice and dice operations are required

c) the multidimensional model does not support query optimization

d) a data warehouse contains many fact tables and many dimension tables.

GROUP – B

(Short Answer Type Questions)

Answer any *three* of the following. 3 × 5 = 15

2. What are the differences between data warehouse and data mart ? What is virtual warehouse ? 3 + 2

3. What is metadata ? What is the advantage of metadata ? What are the typical contents of metadata ? 1 + 2 + 2

4. Describe the principle of partitioning technique for frequent itemset generation and justify how it proves the efficiency of frequent itemset generation compared to Apriori Algorithm.

   3 + 2

5. a) What is the difference between ER Modelling and Dimensional Modelling ? 2

   b) Why is Data Modelling important ? 3

6. What are the differences between OLAP & OLTP ?

GROUP – C

(Long Answer Type Questions)

Answer any *three* of the following. 3 × 15 = 45

7. a) Define data warehouse. What are the characteristics of data warehouse ? 2 + 3

   b) Discuss the three-tier architecture of data warehouse. 5

   c) The weather data is stored for different locations in a warehouse. The warehouse data consists of 'temperature', 'pressure', 'humidity' and 'wind velocity'. The location is defined in terms of 'latitude', 'longitude', 'altitude' and 'time'. Assume that nation ( ) is a function that returns the name of the country for a given latitude and longitude. Assume that this information is stored in a data warehouse as a data cube.

i) Write the type of schema suitable for the warehouse.

ii) Write sequence of operations to get average temperature and maximum pressure for all locations year-wise. 2 + 3

8. a) Write Apriori algorithm for frequent set generation. 5

b) Define a boarder set. Show that every subset of any itemset must contain either a frequent set or a boarder set. Define confidence of an item set. 2 + 2 + 1

c) Generate frequent item set using FP growth algorithm considering minimum count 3. 5

TID	Items bought
1	*f, a, c, d, g, i, m, p*
2	*a, b, c, f, l, m, o*
3	*b, f, h, j, o*
4	*b, c, k, s, p*
5	*a, f, c, e, l, p, m, n*

9. a) Define FP tree. Discuss the method of computing FP tree. 1 + 4

b) Introduce the concept of Splitting attribute and Splitting criterion. 2 + 2

c) What are the uses of Training data set and Test data set for a decision tree classification scheme ? 2

d) Define information gain and discuss how it helps in building a Decision Tree. 4

10. a) What are the Data Mining application areas ? 2

b) What is the difference between DBMS and Data Mining ? 3

c) Suppose the data mining task is to cluster the following 8 points (with (x, y) representing locations. Suppose 3 clusters are to be formed.

$A1$ (2, 10), $A2$ (2, 5), $A3$ (8, 4), $B1$ (5, 8), $B2$ (7, 5), $B3$ (6, 4), $C1$(1, 2), $C2$(4,9)

Euclidian distance is used as the distance function. Initially $A1$, $B1$ and $C1$ are assigned as the centre of each cluster. Use K-means algorithm to determine the 3 clusters. 10

11. Write short notes on any *three* of the following : 3 × 5

a) HOLAD

b) Pruning

c) Temporal mining

d) WUM

e) Decision Tree constructing principle.

=============