Name : .....

Roll No. : .....

Invigilator's Signature : .....

CS / B.TECH (IT) / SEM-8 / IT-802A / 2011

2011

DATA WAREHOUSING AND 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 the following : $10 \times 1 = 10$

i) The most distinguishing characteristic of DSS data is

a) Granularity b) Timespan

c) Dimensionality d) Data currency.

ii) The fact table is related to each dimension table in a

a) 1 : 1 relationship b) 1 : M relationship

c) M : 1 relationship d) M : M relationship.

iii) To optimize data warehouse design, which one is done ?

a) Normalization of fact tables and denormalization of dimension tables

b) Normalization of fact tables and dimension tables

c) Denormalization of fact tables and dimension tables

d) Normalization of dimension tables and denormalization of fact tables.

iv) Data warehouse architecture is just an overall guideline. It is not a blueprint for the data warehouse.

a) True b) False.

v) The major drawback of CLARANS algorithm is

a) it cannot handle very large volumes of data

b) it assumes that all objects fit into the main memory, and the result is very sensitive to input order

c) it cannot find the best clustering if any sampled medoid is not among the best *k* medoids

d) it is time inefficient.

vi) 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) Dynamic Item-set Counting Algorithm

d) FP-Tree growth Algorithm.

vii) While ..... is a clustering technique, ... is one of the most popular methods of building decision trees.

a) DIC, CLARA b) CLARA, CLARANS

c) CLARA, CART d) CLARA, DBSCAN.

viii) The mining activity which mines web log records to discover user access patterns of web pages is

a) web content mining b) web usage mining

c) web structure mining d) web search mining.

ix) *K*-means is based on

a) Euclidian distance b) Hamming distance

c) RMS d) none of these.

x) Gain ratio has an advantage over Gain when an attribute is

a) categorical with only two possible values

b) numerical with two possible values

c) categorical with large number of distinct values

d) numerical with large number of distinct values.

GROUP – B
(Short Answer Type Questions)

Answer any *three* of the following. $3 \times 5 = 15$

2. What is a Data Mart ? When is a Data Mart appropriate ? 2 + 3

3. What are the differences between OLAP & OLTP ?

4. What is Knowledge Discovery in Database ? How does it relate to data mining ? 2 + 3

5. Differentiate between CF tree operations and B^+ tree operations.

6. What is sequence mining ? How is it related with temporal mining ? 2 + 3

GROUP – C
(Long Answer Type Questions)

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

7. a) Explain what is an OLAP cube.

   b) Suppose a data warehouse consists of three dimensions : doctor, time, patient and two measures count and charge where charge is the fee of a doctor charges for a patient for a visit.

      i) Draw the Star schema diagram for the above data warehouse.

      ii) Starting with the base cuboid [day, doctor, patient] what specific OLAP operations (e.g., Slice for Time = Year) should be performed in order to list the total fee collected by each doctor in the year 2000 ?

   c) Draw the OLAP system architecture and explain the functioning of OLAP. 2 + 6 + 7

8. a) Draw the Star schema for Sales fact where dimensions are location, time, product, person with meaningful attributes.

   b) What is a factless table ?

   c) What are the attribute hierarchies and aggregation levels in the star schema context and what is their purpose ?

   d) While designing for data warehouse, when should you use star schema and when should you be using snowflake schema ? 5 + 2 + 5 + 3

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

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

b) Consider the following table of transactions. Each row represents a transaction and each column represents an item. If an item is present in a transaction, it is marked as '1', else it is marked as '0'. Determine the Frequent Itemsets using apriori algorithm. Use intervals of 5 transactions and min_support = 20%. 10

A1	A2	A3	A4	A5	A6	A7	A8	A9
1	0	0	0	1	1	0	1	0
0	1	0	1	0	0	0	1	0
0	0	0	1	1	0	1	0	0
0	0	1	0	0	0	0	0	0
0	0	0	0	1	1	1	0	0
0	1	1	1	0	0	0	0	0
0	1	0	0	0	1	1	0	1
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
0	0	1	0	1	0	1	0	0
0	0	1	0	1	0	1	0	0
0	0	0	0	1	1	0	1	0
0	1	0	1	0	1	1	0	0
1	0	1	0	1	0	1	0	0
0	1	1	0	0	0	0	0	1

10. a) There are 5 documents in a text database – *A, B, C, D* and *E*. The interdocument distance matrix is shown in the form of the following table. Using an agglomerative hierarchical clustering algorithm, build and draw the dendrogram. You should use a step of 0·5. 9

Document	***A***	***B***	***C***	***D***	***E***
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

b) There are two clusters C1 and C2 formed from a dataset. The Clustering Feature (CF) vectors of these two clusters are : CF1 = (2, 8, 18) and CF2 = (3, 6, 14). Determine the following : 6

i) Centroids of C1 and C2
ii) Radii of C1 and C2
iii) Diameters of C1 and C2
iv) Average inter-cluster distance between C1 and C2 defined as :

$$\frac{1}{n_1 n_2} \sum_{i=C_1} \sum_{j=C_2} (O_i - O_j)^2$$

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

a) MOLAP
b) Web-Enabled Data Warehouse
c) ROCK *vs.* CACTUS
d) GSP *vs.* SPADE
e) Decision tree construction with presorting.

==========