M.TECH/IT/2ND SEM /INFO 5257/2016
2016

DATA WAREHOUSING AND DATA MINING
(INFO 5257)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

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

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternatives for the following: **10 x 1=10**

(i) ______ is the technique which is used for discovering patterns in dataset at the beginning of data mining process.
(a) Kohenon map (b) Visualization.
(c) OLAP (d) SQL.

(ii) In the Star schema, 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) Which table is more stable and less volatile?
(a) fact table (b) dimension table
(c) factless fact table (d) none of these.

(v) If 25% of the records change daily, then which option is preferred?
(a) update (b) refresh
(c) initial load (d) none of these.

(vi) In KDD and data mining, noise is referred to as ________.
(a) repeated data (b) complex data.
(c) meta data (d) random errors in database.

(vii) Shannon's notation of information content of message is_______.
(a) Log 1divided by n equals log n
(b) log n equals log 1divided by n.
(c) log 1divided by n equals minus log n
(d) log minus n =log 1divided by n.

(viii) Association rules are always defined on________.
(a) binary attribute (b) single attribute.
(c) relational database (d) multidimensional attribute.

(ix) Which of the following is the not a types of clustering?
(a) K-means (b) Hiearachical.
(c) Partitional (d) Splitting.

(x) Classification rules are extracted from__________.
(a) root node (b) decision tree
(c) siblings (d) branches.

Group – B

2. (a) What is data granularity and how it is applicable to the data warehouse?

(b) As far as users and system orientation, data contents, database designs and access patterns are concerned, what are the differences between OLTP and OLAP systems?

(c) While designing for data warehouse, when should you use star schema and when you should be using snowflake schema?

(4) + (4) + (4) = 12

3. (a) Suppose that a data warehouse for *Big University* consists of the following four dimensions: *student, course, semester*, and *instructor*, and two measures *count* and *avg_grade*. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination),

the *avg_grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg_grade* stores the average grade for the given combination.

(i) Draw a *snowflake schema* diagram for the data warehouse.

(ii) Starting with the base cuboid [*student, course, semester, instructor*], what specific *OLAP operations* (e.g., roll-up from *semester* to *year*) should one perform in order to list the average grade of *CS* courses for each *Big University* student.

(b) What is Key Restructuring? Explain with an example why it is needed?

(4+4) + 4= 12

Group – C

4. (a) Suppose that a data cube consists of the three dimensions: time, location, and product. The dimension hierarchies considered for the data cube are time: (month<quarter<year); location: (city<state<country). Let's say a cuboid C[product, quarters, city] consists of products – TV, Audio, Computer, Mobile Phones; quarters- Q1, Q2, Q3, Q4; and cities – Vijaywada, Mumbai, Chennai, Pune and Hyderabad. Every cell of the cube contains the sales amount in rupees.

(i) You want to know the total sales for TV in Q3. What OLAP operation(s) need to be performed? Explain.

(ii) You want to know the total sales for TV in Q1 and Q2 for locations – Hyderabad and Vijaywada. What OLAP operation(s) need to be performed? Explain.

(iii) You want to know the total sales of TV in Q4 in the state of Maharashtra. What OLAP operation(s) need to be performed? Explain.

(b) Why feeding data into the OLAP system directly from the source operational system is not recommended?

(2+3+2) + 5 = 12

5. (a) Consider a multi-national company having business in North America, Europe, Asia, Africa. In the North America it operates in Canada, USA and Mexico. The product it deals with are Cell phones, Modems, Wireless mouse, Radar Detector. The Cell phones it carries in its stores are Nokia, Motorola, Ericsson and LG. It is doing business from year 2006. Build a OLAP cube for multidimensional analysis. If you want to

compare the sales of Nokia phones in Canada in Quarter 1 for the years 2006 thru 2009, what operations would you perform on the cube?

(b) What is Nesting? Explain with an example how it helps in multi-dimensional analysis?

(c) If you have been asked to implement OLAP systems for a heavy chemicals company, which options will you prefer- ROLAP or MOLAP? Justify your answer. Explain which platform will you use for implementing the OLAP system?

4+3+5 = 12

Group – D

6. (a) How can you link data mining with DBMS? What is the difference between maximal frequent set and boarder set?

(b) Find all frequent itemsets or frequent patterns in the following database using FP-growth algorithm. Take minimum support as 30%.

Tid	Items
1	E, A,D,B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

(4+2) + 6 = 12

7. (a) Differentiate between K-Means and fuzzy C Means algorithm.

(b) Apply the above two algorithms to cluster the following items into 2 clusters.

{2, 4, 10, 12, 3, 20, 30, 11, 25, 98}

(c) How outliers are handled in the above mentioned algorithms?

4+6+2 = 12

Group – E

8. (a) For the following trainging data set explain in detail the different steps of decision tree construction with presorting (choose the apt splitting attributes, stopping criteria, etc).

Age	Car Type	Spent
20	M	$200
30	M	$150
25	T	$300
30	S	$220
40	S	$400
20	T	$80
30	M	$100
25	M	$125
40	M	$500
20	S	$420

(b) Define page rank. How it can be used in social network analysis?

8 + (1+3) = 12

9. Write short notes (any two):

(a) Hadoop

(b) Web structure mining

(c) PCA

(d) Text mining

6 x 2 = 12