MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL

Paper Code : IT-704C

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 important aspect of the data warehouse environment is that data found within the data warehouse is
   a) subject-oriented b) time-variant
   c) integrated d) all of these.

ii) The data is stored, retrieved & updated in
   a) OLAP b) OLTP
   c) SMTP d) FTP.

iii) What is bit-map indexing ?
   a) It is a method of indexing a relational table where each tuple is mapped to a binary string
   b) It is a method of indexing fact tables based or signatures
   c) It is a method of mapping a set of tuples to a bucket, based on a single attribute
   d) It is a code whether a particular value of a selected attributed is present in a tuple or not.

iv) Data warehousing is used for
   a) decision support system
   b) OLTP applications
   c) Database applications
   d) Data manipulation applications.

v) The slice operation deals with
   a) selecting all but one dimension of the data cube
   b) merging the cells along one dimension
   c) merging cells of all but one dimension
   d) selecting the cells of any one dimension of the data cube.

vi) A data warehouse is said to contain a "time-varying" collection of a data because
   a) its contents vary automatically with time
   b) it's life span is very limited
   c) it contains historical data
   d) it content has explicit time-stamp.

vii) What is MOLAP ?

a) MOLAP is an OLAP engine for (i) relational models and (ii) multidimensional OLAP operations

b) MOLAP is an OLAP engine for (i) multidimensional models and (ii) SQL based OLAP operations

c) MOLAP is an OLAP engine for (i) multidimensional models and (ii) support dimensional OLAP operations

d) MOLAP is an ROLAP with a supporting multidimensional.

viii) One of the techniques of implementing the OLAP engine is a "specialized SQL server". This server exhibits which of the following properties ?

a) It assumes that the data warehousing a multidimensional model and is implemented in a relational DBMS

b) It facilitates SQL queries for the data warehousing that is physically organized as a multidimensional model

c) It facilitates OLAP operation in SQL

d) It facilitates OLAP operation in SQL when the data warehouse is organized as relational tables.

ix) A 'drill down' operation is concerned with

a) which merge cells of two dimentions

b) which merges calls of any one dimension based on the characteristic of the dimension

c) which splits cells of two dimensions

d) which splits cell of any one dimension based on the characteristics of the dimension.

x) In order to populate the data warehouse which of the following set of operations are appropriate ?

a) Refresh and load b) Create and edit

c) Insert and delete d) Query and update.

GROUP – B

(Short Answer Type Questions)

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

2. Introduce the idea of Data Mining with example(s). What are the steps involved in Knowledge discovery in Database (KDD) process ? $2\frac{1}{2} + 2\frac{1}{2}$

3. What is decision tree ? What is pre-pruning and post-pruning ? 2 + 3

4. How is a data warehouse different from a database ? What are the issues relating to the diversity of database types ? 3 + 2

5. What is snowflake schema ? What are its limitations ? 3 + 2

6. What is OLAP ? How is different that from OLTP ? 2 + 3

GROUP – C
(Long Answer Type Questions)

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

7. a) How is datawarehouse different from a database ?

   b) Suppose that a data warehouse for Big University consists of the following four dimensions : student, course, semester and instructor and two measures : count and average_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 the snowflake schema diagram for the data warehouse.

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

      iii) If each dimension has five levels (including all), such as student<major<status<university<all, how many cuboids will this cube contain ?

   3 + (5 + 4 + 3)

CS/B.TECH/IT/ODD SEM/SEM-7/IT-704C/2016-17

8. a) What is the use of Regression ? What may be the reasons for not using the linear regression model to estimate the output data ?

   b) How is time series data used in pattern analysis ? Give the formula for Pearson's *r*.

   c) Explain Bayesian classification.

   (2 + 3) + (3 + 2) + 5

9. a) What is clustering ? What are the features of a good cluster ?

   b) What do you mean by hierarchical clustering technique ?

   c) Suppose that the data mining task is to devide the following eight points into 3 clusters :

   A1 (2, 10), A2 (2, 5), A3 (8, 4), B1 (5, 8), B2 (7, 5), B3 (6, 4), C1 (1, 2), C2 (4, 9). The distance function is Euclidean distance. Initially, we assign A1, B1 and C1 as the centre of each cluster. Use k-means algorithm to determine the 3 clusters.

   3 + 4 + 8

10. a) What is FP-tree ?

b) Discuss the different phases of FP-tree growth algorithm.

c) Consider the 9 transactions given below. If minimum support count is 2, determine the frequent itemsets by using the Apriori Algorithm.

Transaction	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

2 + 5 + 8

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

a) Backpropagation algorithm

b) Supervised *vs.* unsupervised learing

c) Strategic information

d) Metadata

e) Hierarchical clustering.