IT-704C

DATA WAREHOUSING AND DATA MINING

Time Allotted: 3 Hours Full Marks: 7

The questions are of equal value
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. Answer any *ten* questions. 10×1 = 10

(i) The 'Slice' operation deals with

(A) selecting all but 1-D of the data cube
(B) merging cells of all but 1-D
(C) merging the cells along 1-D
(D) selecting the cells of any 1-D of the data cube

(ii) A data warehouse is said to contain a 'time-varying' collection of data because

(A) its content has explicit time-stamp
(B) its life-span is very limited
(C) it contains historical data
(D) its contents vary automatically with time

(iii) The 'Dice' operation is concerned with

(A) multiple runs of slice
(B) selecting certain cells of more then 1-D
(C) slice on more than 1-D
(D) two consecutive slice operations in two different dimensions

CS/B.Tech/IT/odd/Sem-7th/IT-704C/2014-15

(iv) The 'pivot' is an OLAP operation which

(A) integrates several dimensions
(B) is not a visualization operation
(C) is a visualization operation rotating the axes for alternative presentation
(D) none of the above

(v) Which is not a schema for multidimensional database?

(A) fact constellation (B) snowflake
(C) transaction (D) star

(vi) k-means is based on

(A) hamming distance (B) euclidean distance
(C) RMS (D) none of these

(vii) 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

(viii) The first fact table index will be a B-Tree on

(A) primary key (B) secondary key
(C) foreign key (D) composite key

(ix) Association analysis is used for

(A) transaction data analysis (B) olap
(C) molap (D) none of these

(x) Consider the 3-tier architecture of the data warehouse. The OLAP engine corresponds to

(A) the first layer (B) the second layer
(C) the third layer (D) none of these

(xi) Which one is not a data mining task?

(A) indexing (B) classification (C) clustering (D) regression

(xii) An example of hierarchical clustering algorithm is

(A) clarans (B) C4.5 (C) average linkage (D) rock

HTTP://WWW.MAKAUT.COM

GROUP B
(Short Answer Type Questions)

Answer any *three* questions. 3×5 = 15

2. Explain 'Fuzzy lookup' and 'Back flushing' in Data Cleansing process. Discuss some data smoothing techniques to remove noise. 1+1+3

3. What is metadata in Data warehousing? What is metadata catalog? Discuss the different categories of metadata used in data warehouse. 1+2+2

4. What are Gain ratio, Gini index and categorical variables? 2+2+1

5. What is concept hierarchy? Explain base cuboid and apex cuboid in the context of lattice of cuboids. 3+2

6. Describe the PAM algorithm in brief. 5

GROUP C
(Long Answer Type Questions)

Answer any *three* questions. 3×15 = 45

7. (a) Discuss ROLAP, MOLAP, HOLAP and DOLAP in data warehousing technology. 6+(4+3+2)
   (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 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 IT courses for each Big University student?
   (iii) If each dimension has five levels (including all), such as "student < major <status < university<all", how many cuboids will this cube contain?

8. (a) Consider the association rule below, which was mined from student database at Big University: 4+8+3

   major (X, '*science*')→status(X, '*undergrad*')

   Suppose that the number of students at the university is 5000, that 70% of the students are majoring in science, that 64% of the students are registered in programs leading to undergraduate degrees, and that 56% of the undergraduates at the university major in science. Compute the confidence and support for the given rule.

(b) 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

(c) Write down some advantages and disadvantages of FP-Tree algorithm.

9. (a) Define with suitable examples each of the following data mining functionalities: data characterization, data association and data discrimination 5+5+5
   (b) What is the conceptual hierarchy? How many cuboids are there in n-dimensional data cube considering the hierarchies in each dimension?
   (c) In real world data, tuples with missing values for some attributes are a common occurrence. Suggest two different approaches for handling such event.

10. (a) What is clustering? What are the features of a good cluster? 3+4+8
    (b) What do you mean by *hierarchical clustering* technique?
    (c) Suppose that the data mining task is to divide 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 center of each cluster. Use K-means algorithm to determine the 3 clusters.

11. Write short notes on any *three* of the following: 3×5
    (a) Galaxy Schema
    (b) Strategic information
    (c) Metadata
    (d) Top-down approach
    (e) C 4.5