CS/B.Tech /IT (NEW)/SEM-7/IT-704C/2013-14

2013

DATA WAREHOUSING AND DATA MINING

lotted : 3 Hours *Full Marks* : 70

The figures in the margin indicate full marks.
***didates** are required to give their answers in their own **words** as far as practicable.*

GROUP – A

(Multiple Choice Type Questions)

Choose the correct alternatives for the following : 10 × 1 = 10

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

a) its contents vary automatically with time

b) its life-span is very limited

c) it contains historical data

d) its content has explicit time-stamp.

ii) A data warehouse is said to contain a 'subject-oriented' collection of data because

a) its contents have a common theme

b) it is built for a specific application

c) it cannot support multiple subjects

d) it is generalization of 'object-oriented'.

iii) A data warehouse is built as a separate repository c data, different from the operational data of an enterpris because

a) it is necessary to keep the operational data free c any warehouse operation

b) a data warehouse cannot afford to allow corrupte data within it

c) a data warehouse contains summarized dat whereas the operational database contair transactional data

d) it is just needed.

iv) ROLAP is preferred over MOLAP when

a) a data warehouse and relational database a inseparable

b) the data warehouse is in relational tables, but t slice and dice operations are required

c) the multidimensional model does not suppc query optimization

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

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 da cube.

i) Which of the following indexing techniques are appropriate for data warehousing ?

a) Hashing on primary key

b) Indexing on foreign keys of fact table

c) Bit-map indexing

d) Join indexing.

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) supports multidimensional OLAP operations.

d) MOLAP is an ROLAP with a supporting multidimensional model.

iii) The advantage of FP-tree Growth Algorithm is

a) it counts the support values of the item sets in the dashed structure as it moves along from one stop point to another

b) it avoids the generation of large numbers of candidate sets

c) to update the association rules when the database discover the set of frequent items sets

d) to prune the item sets which are not frequent.

ix) The ID3 generates a

a) binary decision tree

b) a decision tree with as many branches as there ar distinct values of the attribute

c) a tree with a variable number of branches, no related to the domain of the attributes

d) a tree with an exponential number of branches.

x) An oblique tree is relevant when

a) the attributes are correlated

b) the attributes are independent

c) there are only two attributes

d) all attributes are categorical.

GROUP – B

(Short Answer Type Questions)

Answer any *three* of the following. 3 × 5 = 1

2. Differentiate among Enterprise Warehouse, Data mart an Virtual warehouse.

3. Distinguish between OLTP and OLAP systems.

4. Explain support, confidence, frequent itemset and give formal definition of association rule.

5. Compare between HOLAP, ROLAP and MOLAP.

6. Describe the basic algorithm for decision tree induction.

GROUP – C

(Long Answer Type Questions)

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

a) How is data warehouse different from a database ?

b) What is the significance of a multi-dimensional data model in data-warehousing ? Briefly compare the snowflake schema and fact constellation concepts with a suitable example.

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

i) Draw a star schema for the above warehouse.

ii) Starting with the base cuboid (month, doctor, patient), what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2012 ? 3 + 6 + 6

a) What is FP-tree ?

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

c) Consider the following transaction database T, which contains 15 records :

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	1	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

The set of items, *A* = { *A*1, *A*2, *A*3, *A*4, *A*5, *A*6, *A*7, *A*8, *A*9 }.

Assume σ = 20 %.

Illustrate the working of a FP-tree growth algorithm for the above database. 2 + 4 + 9

9. a) Define with suitable examples of each of the following data mining functionalities : data characterization, data association and data discrimination.

b) What is conceptual hierarchy ? How many cuboids are there in an *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. 5 + 5 + 5

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

b) What do you mean by *hierarchical* clustering technique ?

c) Suppose that the data mining task is to divide the following eight points representing locations into 3 clusters : *A*1 (2, 10), *A*2 (2, 5), *A*3 (8, 4), *B*1 (5, 8), *B*2 (7, 5), *B*3 (6, 4), *C*1 (1, 2), *C*2 (4, 9). The distance function is Euclidian distance. Initially, we assign *A*1, *B*1 and *C*1 as the center of each cluster. Use *k*-means algorithm to determine the 3 clusters. 3 + 4 + 8

11. a) What is tree pruning ? What are the different tree pruning techniques ?

b) Describe PAM algorithm in brief.

c) Evaluate Information Gain and Gain Ratio with suitable example. 5 + 5 + 5