Name : .....

Roll No. : .....

Invigilator's Signature : .....

CS/B.Tech (IT)/SEM-8/IT-802A/2010

2010

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 any *ten* of the following : $10 \times 1 = 10$

   i) OLAP operations are not performed on operational data because

      a) Operational data is normalized for OLTP operations

      b) Operational data needs concurrency control and logging support

      c) Typically data warehouse stores summarized data with multidimensional view

      d) all of these.

   ii) Data Warehousing is used for

      a) Decision Support System

      b) OLTP applications

      c) Database applications

      d) Data Manipulation applications.

http://qpaper.wbut.ac.in

iii) If we know exactly what information we need then ..... would suffice, but if we vaguely know the possible patterns then ..... are useful.

a) Data Warehouse, Data Mining techniques

b) DBMS query, Data Mining techniques

c) DBMS query, Data Warehouse applications

d) Data Warehouse applications, Data Mining techniques.

iv) Which of the following is TRUE?

a) Data warehouse can be used for analytical processing only

b) Data warehouse can be used for information processing (query, report) and analytical processing

c) Data warehouse can be used for data mining only

d) Data warehouse can be used for information processing (query, report), analytical processing and data mining.

v) In order to *populate* the data warehouse which of the following sets of operations are appropriate ?

a) Insert & Update

b) Refresh & Load

c) Query, Edit & Update

d) Delete, Insert & Update.

http://qpaper.wbut.ac.in

vi) A Data Warehouse is said to be contain in *time-varying* collection of data because

a) Its content vary automatically with time

b) Its life-span is very limited

c) Every key structure of data warehouse contains either implicitly or explicitly an element of time

d) Its content has explicit time-stamp.

vii) ..... is an example of predictive type of data mining whereas ..... is an example of descriptive type of data mining.

a) Association Rule, Clustering

b) Association Rule, Classification

c) Classification, Clustering

d) Clustering, Classification.

viii) Classification is an example of ..... learning whereas clustering is an example of ..... learning.

a) Supervised, Unsupervised

b) Unsupervised, Supervised

c) Machine, Supervised

d) Supervised, Machine.

ix) Which of the following is FALSE ?

a) Clustering can be done on both numeric and categorical data

b) Any subset of a frequent set is a frequent set

c) Any superset of an infrequent set is also infrequent

d) Market-basket problem is a popular example of Data warehousing application.

http://qpaper.wbut.ac.in

x) Parameters used for Association Rule Mining are

a) Confidence and Support

b) Confidence and Itemcount

c) Support and Itemcount

d) Support, Confidence and Itemcount.

xi) Decision Tree algorithm uses ..... data to determine the rules.

a) Test b) Data warehouse

c) Training d) Transaction.

xii) The algorithm which uses the concept of a train running over data to find associations of items in Association Rule mining is known as

a) Apriori Algorithm

b) Partition Algorithm

c) Dynamic Itemset Counting Algorithm

d) FP Tree growth Algorithm.

xiii) Two main types of clustering techniques in data mining are

a) Serial clustering and parallel clustering

b) Hierarchical clustering and portioning clustering

c) Homogeneous clustering and heterogeneous clustering

d) k-medoids clustering and k-means clustering.

xiv) If no hierarchy is associated with any dimension, how many cuboids would be there in an n-dimensional data cube ?

a) n^3 b) n

c) 2^n d) n^2.

http://qpaper.wbut.ac.in

WEST BENGAL UNIVERSITY OF TECHNOLOGY
Utech

GROUP – B
(Short Answer Type Questions)

Answer any *three* of the following. 3 × 5 = 15

2. State Apriori Algorithm for frequent itemset generation. 5

3. What is a Data Mart ? State the differences between Data Mart & Data Warehouse. 2 + 3

4. Describe the principle of Partitioning technique for Frequent Itemset generation and justify how it improves the efficiency of Frequent Itemset generation compared to Apriori Algorithm. 3 + 2

5. Describe the principle of Dynamic Itemset Counting technique for Frequent Itemset generation. 5

6. What is clustering ? Discuss two main methods of clustering.

GROUP – C
(Long Answer Type Questions)

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

7. a) Define Data Warehouse and briefly discuss its characteristics. 2 + 1

   b) State the difference between OLTP and OLAP systems. 4

   c) Why do we need to have separate Data Warehouse for OLAP applications ? 2

   d) A data warehouse is designed for a Sales application on 3 dimensions – *time, item and branch* and two measures *qty* and *value*. Draw ad star schema. 3

   e) Starting with the base cuboid [day, item, branch] what specific OLAP operations (e.g. slice for time = 'year') should be performed in order to list the total sales of each branch in the year 2008 ? 3

http://qpaper.wbut.ac.in

8. a) Introduce the concept of Support, Confidence and Frequent Itemset and then give a formal definition of Association Rule. 5

b) Generate all Frequent Itermsets from the following transaction data given minimum support = 0·3. 5

TID	Items	TID	Items
1	A, B, C, E	6	B, C
2	B, D, E	7	A, C, E
3	B, C	8	A, B, C, E
4	A, B, D	9	A, B, C
5	A, C	10	C, D, E

c) Find five Association Rules from the above Frequent sets at min. 50% confidence. 5

9. a) Introduce the concept of data mining and cite two application areas. 2 + 2

b) What are the different steps of a data mining task ? 2

c) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) 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 three clusters. 9

http://qpaper.wbut.ac.in

10. a) What are the uses of training data set and test data set for a decision tree classification scheme ? 2

b) Discuss the principle of FP-tree Growth algorithm. 5

c) What is an outlier data object in clustering principle ? 2

d) What is a Decision Tree ? Define Information Gain and discuss how it helps in building a Decision Tree.

2 + 2 + 2

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

a) DBMS vis-à-vis Data Mining

b) Generalized Association Rule

c) Decision Tree Construction Principle

d) PAM Clustering Technique

e) CLARANS clustering algorithm vis-à-vis PAM and CLARA.

http://qpaper.wbut.ac.in