



Name :

Roll No. :

Invigilator's Signature :

CS/B.TECH(IT)/SEM-8/IT-802A/2012

2012

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 × 1 = 10

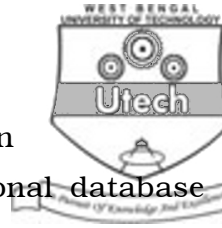
- i) A data warehouse is built as a separate repository of data, different from the operational data of an enterprise because
 - a) It is necessary to keep the operational data free of any warehouse operations.
 - b) A data warehouse cannot afford to allow corrupted data within it
 - c) A data warehouse contains summarized data whereas the operational database contains transactional data
 - d) None of these.



- ii) 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.
- iii) Data Warehousing is used for
 - a) decision support system
 - b) OLTP applications
 - c) database applications
 - d) data manipulation applications.
- 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) Dimension data within a warehouse exhibits which one of the following properties ?
- Dimension data consists of the minor part of the warehouse
 - The aggregated information is actually dimension data
 - It contains historical data
 - Dimension data is the information that is used to analyze the elemental transaction.
- vi) If we know exactly what information we need then would suffice, but if we vaguely know the possible patterns then are useful.
- Data Warehouse, Data Mining techniques
 - DBMS Query, Data Mining techniques
 - DBMS Query, Data Warehouse applications
 - Data Warehouse applications, Data Mining techniques.
- vii) The slice operation deals with
- selecting all but one dimension of the data cube
 - merging the cells along one dimension
 - merging cells of all but one dimension
 - selecting the cells of any one dimension of the data cube.
- viii) In order to populate the data warehouse which of the following sets of operations are appropriate ?
- Insert and Update
 - Refresh and Load
 - Query, Edit and Update
 - Delete, Insert and Update.



- ix) ROLAP is preferred over MOLAP when
 - a) A data warehouse and relational database are inseparable
 - b) The data warehouse is in relational tables, but no slice and dice operations are required
 - c) The multidimensional model does not support query optimization
 - d) A data warehouse contains many fact tables and many dimension tables.
- x) Consider the 3-tier architecture of the data warehouse. The OLAP engine corresponds to
 - a) the first layer of the architecture
 - b) second layer
 - c) third layer.
- xi) is an example of predictive type of data mining whereas is an example of descriptive type of data mining.
 - a) Association Rules, Clustering
 - b) Association rule, Classification
 - c) Classification, Clustering
 - d) Clustering, Classification.
- xii) The advantage of FP-tree Growth Algorithm is
 - a) it counts the support values of the itemsets 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 itemsets.



GROUP – B

(Short Answer Type Questions)

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

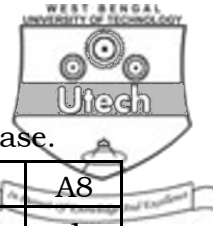
2. Describe the principle of Partitioning technique for frequent itemset generation and justify how it improves the efficiency of frequent itemset generation compared to a priori algorithm.
3. What is metadata in Data Warehousing ? Discuss the different categories of metadata used in Data Warehouse.
4. What are the different methods of computing the best split ? What are entropy gain and gain ratio ?
5. How is CLARANS different from CLARA ? Illustrate this using a small example.
6. State the main features of GSP algorithm. Explain the main difference between GSP algorithm and a priori algorithm with an example.

GROUP – C

(Long Answer Type Questions)

Answer any *three* of the following. $3 \times 15 = 45$

7.
 - a) What are the shortcomings of a priori algorithm ?
 - b) What is FP-tree ?
 - c) Discuss the different phases of FP-tree growth algorithm.



d) Consider the following transaction database.

A1	A2	A3	A4	A5	A6	A7	A8
1	0	0	0	1	1	0	1
0	1	0	1	0	0	0	1
0	0	0	1	1	0	1	0
0	1	1	0	0	0	0	0
0	0	0	0	1	1	1	0
0	1	1	1	0	0	0	0
0	1	0	0	0	1	1	0
0	0	0	0	1	0	0	0
0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0

Assuming $\sigma = 20\%$, find out the all possible frequent itemsets using a priori algorithm. 2 + 2 + 4 + 7

8. a) What are the uses of training data set and test data set for a decision tree classification scheme ?
- b) Define the entropy gain and gini's index.

age	income	student	credit_rating	buys_computer
< = 30	high	no	fair	no
< = 30	high	no	excellent	no
31 ... 40	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	yes	fair	yes
> 40	low	yes	excellent	no
31 ... 40	low	yes	excellent	yes
< = 30	medium	no	fair	no
< = 30	low	yes	fair	yes
> 40	medium	yes	fair	yes
< = 30	medium	yes	excellent	yes
31 ... 40	medium	no	excellent	yes
31 ... 40	high	yes	fair	yes
> 40	medium	no	excellent	no



- c) Generate classification rules from a decision tree for the above database using entropy gain computation.

2 + 4 + 9

9. a) Introduce the concept of data mining and cite two application area.
- b) What are the different steps of a data mining task ?
- c) Suppose that the data mining task is to cluster the following ten points (with (x, y) representing location) into two clusters :

X1	2	6
X2	3	4
X3	3	8
X4	4	7
X5	6	2
X6	6	4
X7	7	3
X8	7	4
X9	8	5
X10	7	6

The distance function is defined as $|x_i - x_j| + |y_i - y_j|$.

Use k -means or k -medoid algorithm to determine the two clusters.

(2 + 2) + 2 + 9

10. The following table contains five sample data items with the distance between the elements indicated in the table entries. Suppose that the two medoids A and B are initially chosen. Form two cluster based on the distance between the



elements with the medoids A and B and also obtain the new two clusters after replacing the medoid A by one of the non-medoids using PAM algorithm.

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

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

- a) Text Mining
- b) ROLAP
- c) ROCK
- d) Arbor Essbase Web
- e) WUM.

