LIB
16/12/13 FN

Reg. No. : | | | | | | | | | | | | |

Question Paper Code : 81348

M.E./M.Tech. DEGREE EXAMINATION, NOVEMBER/DECEMBER 2013.

Elective

Computer Science and Engineering

CS 9264/CS 964/UCP 9164/10244 CSE 51 — DATA WAREHOUSING AND DATA MINING

(Common to M.E. Networking and Internet Engineering/M.E. Software Engineering/ M.Tech. Information Technology and M.Tech Multimedia Technologies)

(Regulation 2009/2010)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What is the use of knowledge base in data mining system?

2. How is a data warehouse different from a database?

3. Describe the method of generating frequent item-sets without candidate generation.

4. Give few techniques to improve the efficiency of Apriori algorithm.

5. How is classification different from prediction?

6. What is attribute selection measure?

7. Differentiate between agglomerative and divisive hierarchical clustering.

8. What do you mean by cluster analysis?

9. Define spatial database.

10. What kind of association can be mined from multimedia data?

PART B — (5 × 16 = 80 marks)

11. (a) With the help of neat diagrams, explain the three data warehouse schemes.

Or

(b) State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach rather than query-driven approach. Describe situations where the query-driven approach is preferable over the unpdate-driven approach.

12. (a) Find all the frequent itemsets for the following transactional database using apriori algorithm. The minimum support given is 20%. Also find strong association rules from the largest frequent itemsets with respect to 60% minimum confidence.

Transaction _id	Items in transaction
1	3,4
2	2,3
3	1,2, 3,5
4	2,5
5	1,2
6	1,3
7	2
8	1,3
9	1,2,3
10	1,3

Or

(b) For the database given below discover all frequent itemsets using FP tree algorithm with minimum support count = 2 transactions.

Tid	Items
1	2,3,5,6
2	1,4,5
3	2,3,4
4	1,3,4,5,6
5	2,3,4,5,6

13. (a) For the table given below, determine the best split attribute for first level in constructing a decision tree using information gain (entropy) approach.

rid	age	income	student	credit _ rating	Class: buys – computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-40	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>40	medium	no	excellent	no

Or

(b) Discuss the classifier which handles the non-linear data sets by enhancing margin between two classes.

14. (a) Cluster the following eight points (with (x,y) representing locations) into three clusters A1 (2, 10) A2 (2,5) A3 (8,4) A4 (5, 8) A5 (7, 5) A6 (6,4) A7, (1,2) A8 (4, 9). Initial cluster center are : A1(2,10) , A4 (5,8) and A7 (1,2).

Note : Use Manhattan distance function to find the dissimilarity.

Use k-means algorithm to find the three cluster centers after second iteration.

Or

(b) What is an outlier? Describe any three computer based outlier detection techniques.

15. (a) TF-IDF has been used as an effective measure in document classification.

(i) Give one example to show that TF-IDF may not be always a good measure in document classification. (8)

(ii) Define another measure that may overcome this difficulty. (8)

Or

(b) Write short notes on the following :

(i) Latent semantic indexing (6)

(ii) Web usage mining (5)

(iii) Mining web page layout structure. (5)

———————