Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY)

END SEMESTER EXAMINATIONS, APR/MAY 2017

SUBJECT: DATA WAREHOUSING AND DATA MINING [ICT 3202]

REVISED CREDIT SYSTEM
(25/04/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer ALL the questions.
- Missing data if any, may be suitably assumed.

1A. Consider the following sorted data points: 1,1,1,2,3,6,6,6,6,11,11,13,15,15,16. Partition them into three bins and plot the histogram for the V-optimal method. 5

1B. Illustrate stratified sampling using the following data points for age: 15,15,15,18,30,31,32,33,34,35,36,40,65,70. Assume sample size=7 and the age strata as youth (13-25), Middle-aged (30-52) and senior (65 and above). 3

1C. The intervals for the age values and the corresponding frequencies are as given in Table Q.1C. Compute the approximate median value.

Table Q.1C

Age	Frequency
[10-15)	9
[15-20)	16
[20-25)	22
[25-30)	8
[30-35)	5

2

2A. A data warehouse consists of 4 dimensions time, location, salesperson and customer, and two measures count and charge, where charge is the price that a salesperson charges for a product.

a) Draw a snowflake schema for the above data warehouse.

b) Starting with the base cuboid [day,salesperson,customer], what specific OLAP operations should be performed in order to list the total amount collected by each salesperson in 1990? 5

2B. Explain the various ways of measuring central tendency of data. Graphically represent their relationship with respect to symmetric and skewed data. 3

2C. What is data cube materialization? Discuss the various data cube materialization techniques. 2

3A. Define gain ratio. For the data-set given in Table Q.3A calculate information gain for the attributes Outlook, Temp., Humidity and Windy.

Table Q.3A

ID	Outlook	Temp.	Humidity	Windy	Play
A	Sunny	Hot	High	False	No
B	Sunny	Hot	High	True	No
C	Overcast	Hot	High	False	Yes
D	Rainy	Mild	High	False	Yes
E	Rainy	Cool	Normal	False	Yes
F	Rainy	Cool	Normal	True	No
G	Overcast	Cool	Normal	True	Yes
H	Sunny	Mild	High	False	No
I	Sunny	Cool	Normal	False	Yes
J	Rainy	Mild	Normal	False	Yes
K	Sunny	Mild	Normal	True	Yes
L	Overcast	Mild	High	True	Yes
M	Overcast	Hot	Normal	False	Yes
N	Rainy	Mild	High	True	No

5

3B. Given the data-set and initial centroids as C1 and C2, determine the data objects in clusters after 2 iterations by applying the k-means algorithm.

C1(1,1) C2(2,1) C3(4,3) C4(5,4) 3

3C. Briefly discuss any four data reduction techniques. 2

4A. Write the pseudo code for Partition algorithm. Find all the frequent item-sets from the transaction data-set given below using Partition algorithm with number of partitions=2 and minimum support =50%. Indicate all the steps.

T1: {8, 9,7,10} T2: {7, 10,11,8} T3: {11,12,8} T4:{7, 10} 5

4B. Explain the following terminologies used in DBSCAN algorithm with an example.

a) Directly density Reachable
b) Density Reachable
c) Density Connected 3

4C. Explain Page rank algorithm with an example. 2

5A. Find the frequent patterns and generate any two strong association rules for the set of transactions using Pincer Search algorithm: (min_sup>=3 and confidence =60%)

T1: {A,B,C,E} T2:{A,E} T3:{A,B} T4:{A,B,C} 5

5B. Construct a PC Tree for the given set of transactions:

T1: {1,2,3,4}; T2: {1,2,4}; T3: {1,2}; T4: {2,3,4}; T5: {2,3}; T6: {3,4}; T7: {2,4} 3

5C. Differentiate between symmetric and asymmetric binary variables. Discuss the distance measure used to assess the dissimilarity for symmetric and asymmetric binary data type. 2