Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL 576104

(Constituent College of Manipal University)

SIXTH SEMESTER B.Tech. (IT) DEGREE END SEMESTER EXAMINATION MAY – 2016
SUBJECT: DATA WAREHOUSING AND DATA MINING – ICT 306
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS **09/05/2016** **MAX. MARKS: 50**

Instructions to candidates

- Answer any **FIVE FULL** questions.
- Missing data, if any, may be suitably assumed.

1A. Find the median, first quartile (Q1), third quartile (Q3), variance and standard deviation of the audit time considering the frequency distribution of the time in days required to complete year-end audits as given below.

Audit Time (days)	Frequency
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1

1B. Briefly describe the following advanced database systems and applications: object-relational databases, spatial databases, temporal and time series database.

1C. Mean of a dataset containing 15 data values is 200. If one of the data value is excluded, their mean is 198. Find the value of excluded data value. . [5+3+2]

2A. Find all frequent itemsets from the transaction data set given below using Pincer Search algorithm with minimum support =40%. Indicate all the steps.
Data Set = T1:{A,B,C,D,E,F}; T2:{A,B,C,G}; T3:{A,B,D,H}; T4:{B,C,D,E,I}; T5:{A,B,C;}

2B. What is the use of bitmap indexing? Write the bitmap indices for the customer data given below.

ID	MARITAL_STATUS	DEPARTMENT	INCOME_LEVEL
70	single	HR	190,000 - 249,999
80	married	Admin	150,000 - 169,999
90	single	Admin	150,000 - 169,999
100	married	Marketing	170,000 - 189,999
110	married	Sales	130,000 - 149,999
120	single	Marketing	170,000 - 189,999
130	single	Sales	130,000 - 149,999
140	married	Sales	130,000 - 149,999

2C. Consider a fact table Sales(saleID, itemID, color, size, qty, unitPrice), and the following three queries:
Q1: Select itemID, color, size, Sum(qty*unitPrice) From Sales Group By itemID, color, size
Q2: Select itemID, size, Sum(qty*unitPrice) From Sales Group By itemID, size
Q3: Select itemID, size, Sum(qty*unitPrice) From Sales Where size < 10 Group By itemID, size
Depending on the order in which we execute two of these queries, the pair of actions may be viewed as an example of roll-up, drill-down or slicing. Which of the following statements is correct? Justify.
(i) Going from Q2 to Q1 is an example of roll-up (iii) Going from Q2 to Q3 is an example of slicing
(ii) Going from Q2 to Q3 is an example of roll-up (iv) Going from Q3 to Q2 is an example of drilldown
[5+3+2]

3A. Find all the frequent items for the transactions given below by using Dynamic Itemset Counting algorithm. Assume minimum support count >= 3 and number of stops = 2.
T1: {1,3,4}; T2: {2,3,5}; T3: {1,2,3,5}; T4: {2,5}; T5: {1,3,5}

3B. Write the pseudo code for Apriori algorithm. Differentiate between upward and downward closure properties with an example.

3C. Draw a quantile plot for the data given below with detailed steps:
76 92 83 105 102 109 106 91 110 89 [5+3+2]

4A. Consider the set of transactions given below and generate the frequent patterns using PC tree.
T1:{A,B,D}; T2:{A,C,D,E}; T3:{A,D,E,F}; T4:{B,E,F}; T5:{A,B,D,E,F} [min_sup= 60%]

4B. Consider the following transaction database and list all the association rules along with their confidence for the frequent itemsets: (i) A,B,C (ii) A,B,E
T1:{A,B,C,D}; T2:{A,B,C,E}; T3:{A,B,E,F,H}; T4: {A,C,H}

4C. Calculate the dependent/correlated measures lift and cosine from the information given below.

	Basketball	Non-Basketball	Sum(rows)
Cereal	2000	1750	3750
Non-cereal	1000	250	1250
Sum(Col.)	3000	2000	5000

[5+3+2]

5A. Apply k-means algorithm on the dataset given below and generate three clusters. Let the initial centroid be C1=(185,72) , C2=(170,56) and C3=(168,60).
Dataset: { X1(185,72) ; X2(170,56) ; X3(168,60) ; X4(182,72) ; X5(188,77); }

5B. Find the root node using Information Gain as the attribute selection measure for the data given below.

Can Fly	Live in Water	Have legs	Class
No	No	Yes	Mammals
No	No	No	Non-mammals
No	Yes	No	Non-mammals
No	Yes	No	Mammals
No	Sometimes	Yes	Non-mammals
No	No	Yes	Non-mammals
Yes	No	Yes	Mammals
Yes	No	Yes	Non-mammals
No	No	Yes	Mammals
No	Yes	No	Non-mammals
No	Sometimes	Yes	Non-mammals
No	Sometimes	Yes	Non-mammals
No	No	Yes	Mammals

5C. Explain two main approaches in web usage mining driven by the application of the discoveries.
[5+3+2]

6A. Explain how CLARANS is different from other partitioning methods PAM and CLARA in cluster analysis? Write the pseudo code for CLARANS algorithm.

6B. Find the dissimilarity matrix for the following dataset.

Object id	Band (Nominal)	Position (Ordinal)	Salary (Numeric)
1	Red	Senior	50000
2	Green	Junior	12000
3	Blue	Mid	30000
4	Green	Senior	45000

6C. Consider the following set of transactions and calculate support and confidence for the association rules:
(i) A➔ B (ii)B➔ A
T1:{A,B,D}; T2:{A,C,D}; T3:{A,D,E}; T4:{B,E,F}; T5:{B,C,D,E,F} [5+3+2]