Reg. No.

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent institution of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY)

END SEMESTER EXAMINATIONS, APRIL 2018

DATA WAREHOUSING AND DATA MINING [ICT 3202]

REVISED CREDIT SYSTEM

(20/04/2018)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Find frequent itemsets for the database given below using the DIC algorithm. Consider the number of rows per iteration M=3 and minimum support count as 3. Clearly show all the steps of the algorithm.
1:(M,B,E) 2:(B,S) 3:(B,C) 4:(M,B,S) 5:(C,M) 6:(C,B) 7:(C,M) 8:(M,B,C,E) 9:(M,B,C) 5

1B. Apply 3-4-5 rule to generate concept hierarchy of 3 levels for the values:
-34, -26, -22, -22, -20, -20, -18, -18, -16, 16, 19, 21, 24, 26, 26, 28, 28, 29, 32, 32, 34, 35, 42, 46, 46 3

1C. Differentiate between roll-up and roll-down OLAP operations. Consider the following cube illustrating temperature of certain days recorded weekly. Show the result of roll-up operation (Temperature) for this cube by assuming levels hot (80-85), mild (70-75), cold (64-69) for Temperature. 2

Table Q.1C

Temperature	64	65	68	69	70	71	72	75	80	81	83	85
Week 1	1	0	1	0	1	0	0	0	0	0	1	0
Week 2	0	0	0	1	0	0	1	2	0	1	0	0

2A. Use Fp-Growth algorithm to discover the frequent itemsets from the transactions below with support threshold s=2 and confidence threshold c=60%. Also mention any 4 strong association rules.
T1{H,B,K} T2{H,B,D} T3{C,D,B} T4{C,D} T5{C,K} T6{H,D,C} T7{K,D} T8{D,C,B} T9{B} 5

2B. Write the pseudocode for MFCS_gen, MFS Prune and Recovery methods of Pincer Search algorithm. 3

2C. What is a null invariant measure? Compute the correlation between two items based on any two null invariant measures. 2

Table Q.2C

	Book	$\overline{\text{Book}}$
Pen	80	18
$\overline{\text{Pen}}$	40	42

3A. Consider a group of 12 sales price records as follows:

5,10,11,13,15,35,50,55,72,92,204,215

i. Draw the boxplot for sales price records.

ii. Normalize the value 15 and 92 based on z-score normalization. 5

3B. Briefly explain the three different implementations of a warehouse server for OLAP processing. 3

3C. Draw a quantile plot for the data given below with detailed steps:

76, 92, 83, 105, 102, 109, 106, 91, 110, 89 2

4A. Apply the Partitioning Around medoids algorithm on the below mentioned data points and obtain two clusters. Let P1 and P5 be the initial cluster medoids.

P1(2,2) P2(1,14) P3(10,7) P4(1,11) P5(3,4) P6(11,8) P7(4,3) P8(12,9)

Verify whether swapping the centroid from P5 to P7 would result in better clustering? 5

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} 3

4C. Define the following terms:

i. Maximal Frequent Set ii. Closed Frequent Set 2

5A. Find the root node using Information Gain as the attribute selection measure for the data given in Table Q. 5A.

Table Q.5A

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	Young	False	False	Fair	No
2	Young	False	False	Good	No
3	Young	True	False	Good	Yes
4	Young	True	True	Fair	Yes
5	Young	False	False	Fair	No
6	Middle	False	False	Fair	No
7	Middle	False	False	Good	No
8	Middle	True	True	Good	Yes
9	Middle	False	True	Excellent	Yes
10	Middle	False	True	Excellent	Yes
11	Old	False	True	Excellent	Yes
12	Old	False	True	Good	Yes
13	Old	True	False	Good	Yes
14	Old	True	False	Excellent	Yes
15	Old	False	False	Fair	No

5

5B. What is the difference between symmetric and asymmetric binary variables? Consider the relational table where name is an object identifier, gender is a symmetric attribute, and remaining attributes are asymmetric binary. Calculate the distance between each pair of three entities given in Table Q.5B.

Table Q.5B

Name	Gender	A1	A2	A3
Ben	M	Y	N	Y
Emma	F	Y	N	Y
Joe	M	Y	Y	N

3

5C. Explain the Page rank algorithm with an example. 2