GANPAT UNIVERSITY

B. TECH SEMESTER - VII (IT) EXAMINATION
REGULAR EXAMINATION NOV/DEC-2011
IT 701: DATA MINING & DATA WAREHOUSING

Time: 3 Hours] **[Total Marks: 70**

Instructions:

1. **Attempt all questions.**
2. **Figures to the right indicate full marks**
3. **Each section should be written in a separate answer book**

Section 1

Q.1

(A) Explain six methods to feel Missing Values in database. (6)

(B) Architecture of a typical data mining system. (6)

OR

Q.1

(A) Explain the three Binning methods for data smoothing. (6)

(B) Explain the Balanced Iterative Reducing and Clustering Using Hierarchies. Show how effective is BIRCH? (6)

Q.2

(A) Explain Data Reduction techniques. (5)

(B) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: (6)

$A1(2, 10)$, $A2(2, 5)$, $A3(8, 4)$, $B1(5, 8)$, $B2(7, 5)$, $B3(6, 4)$, $C1(1, 2)$, $C2(4, 9)$:

The distance function is Euclidean distance. Suppose initially we assign $A1$, $B1$, and $C1$ as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*

(a) The three cluster centers after the first round execution

(b) The final three clusters

(c) Illustrate the strength and weakness of k-means algorithm in comparison with a hierarchical clustering schemes (such as AGNES).

OR

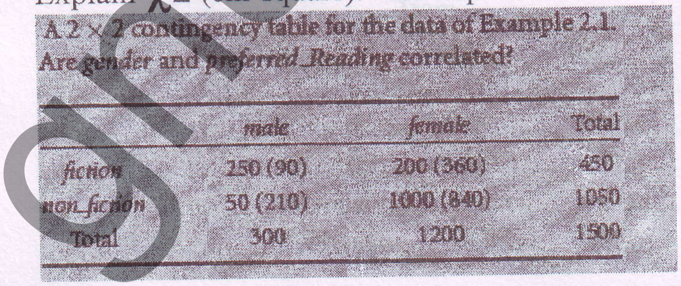
Q.2

(A) Explain the hash based technique to improve the efficiency of Apriori Algorithm with example. (5)

(B) Draw a 3-D view and Star schema of sales data for *AllElectronics*, according to the dimensions *time*, *item*, *Location and sales*. The measure displayed is *dollars sold* (in thousands). (6)

Q.3

(A) Explain $\chi 2$ (chi-square). And explain observed frequency and expected frequency. (4)



A 2 × 2 contingency table for the data of Example 2.1. Are gender and *preferred_Reading* correlated?

	male	*female*	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Find out the co relation between both the attributes.

(B) (8)

Attributes: RID, age, income, student, credit rating, Class: buys computer
1, youth, high, no, fair, no
2, youth, high, no, excellent, no
3 ,middle aged, high, no, fair ,yes
4 ,senior ,medium, no, fair, yes
5 ,senior ,low, yes, fair, yes
6 ,senior, low, yes, excellent ,no
7, middle aged, low, yes, excellent, yes
8, youth, medium, no, fair, no
9, youth ,low, yes ,fair, yes
10, senior, medium, yes ,fair ,yes
11 ,youth, medium, yes, excellent, yes
12 ,middle aged ,medium no excellent, yes
13, middle aged ,high, yes ,fair ,yes
14, senior, medium ,no, excellent, no

Predicting a class label using naïve Bayesian classification.

Where X = (age = youth, income = medium, student = yes, credit rating = fair)

P. T. O.

Section 2

Q.4

(A) Describe three challenges to data mining regarding data mining methodology. **[6]**

(B) Explain five Data transformation techniques. **[6]**

OR

Q.4

(A) **Explain following terms :** **[6]**

Roll-up ,Drill-down, Slice and dice, Pivot (rotate), drill-through, drill-across

(B) A database has four transactions. Let *min sup* = 60% and *min con f* = 80%. **[6]**

Cust_ ID TID items bought (in the form of *brand-item category*)

T100	{King's-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}
T200	{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}
T300	{Westcoast-Apple, Dairyland-Milk,Wonder-Bread, Tasty-Pie}
T400	{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}

Q.5

(A) Describe each of the following clustering algorithm **[6]**

(I)Chameleon (II)ROCK

(B) Given two objects represented by the tuples (22,1,42,10) and (20,0,36,8): **[5]**

Compute the Euclidean distance and the Manhattan distance between two objects.

OR

Q.5

(A) Explain concept characterization and concept comparison using OLAP-based approaches. **[6]**

(B) Explain the FP-Growth algorithm with example. **[5]**

Q.6

(A) A three-tier data warehousing architecture. **[4]**

(B) Discuss Lattice of Cuboids. Draw 3-D data cube of sales data, according to the dimension time, item and location for following tables. The measure is displayed in no of units sold in thousand **[8]**

Location = "Valsad"					Location = "Surat"				
	Item					Item			
Time	**Comp uter**	**Monitor**	**CPU**	**Keyboard**	**Ti me**	**Comput er**	**Monitor**	**CPU**	**Keybo ard**
Q1	201	132	123	233	Q1	238	189	143	239
Q2	520	124	435	142	Q2	508	193	434	542
Q3	234	322	532	144	Q3	238	392	532	544
Q4	433	433	534	233	Q4	439	493	544	223

Location = "Mehsana"					Location = "Patan"				
	Item					Item			
Time	**Comp uter**	**Monitor**	**CPU**	**Keyboard**	**Ti me**	**Comput er**	**Monitor**	**CPU**	**Keybo ard**
Q1	224	135	163	273	Q1	270	137	173	237
Q2	250	143	474	172	Q2	505	125	437	147
Q3	215	336	562	174	Q3	237	327	572	147
Q4	243	233	574	273	Q4	437	437	537	273