Enrollment No. _____

GANPAT UNIVERSITY

M. Tech SEMESTER II - INFORMATION TECHNOLOGY

REGULAR EXAMINATION MAY/JUNE - 2014

3IT203: Data Mining & Data Warehousing

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

Instructions:

1. Figures to the right indicate full marks
2. Each section should be written in a separate answer book
3. Be precise and to the point in your answer

SECTION-I

Q.1 Attempt ANY THREE **12**

(A) What is Data Mining? Explain potential applications of data mining.

(B) Briefly describe the following advanced database systems and applications: object-oriented databases, spatial databases, text databases, multimedia databases, the world wide web.

(C) Draw and explain generic two-level data warehousing architecture

(D) Explain data cleaning, data reduction, data integration & transformation in brief.

Q.2

(A) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. **3**

(B) What is Divide and Conquer? How it could be helpful for ***FP-growth*** method in generating frequent item sets without candidate generation? **4**

(C) Suppose that an airline company is making a loss in almost all classes for all routes for the year 2013 which cover a wide range from 12,08,234 (LOSS) to 2,50,231 (LOSS). A user wishes to have a concept hierarchy for loss automatically generated. Suppose that the data within the 5th percentile and 95th percentile are between 11,55,400 (LOSS) and 3,00,189 (LOSS). Apply 3-4-5 rule up to second level in hierarchy. **4**

OR

Q.2

(A) Explain interesting measures widely used for mining association rules. If attributes are negatively correlated then why these measures do not provide proper rules. Explain using X, Y, Z attribute set values. **5**

X	1	1	1	1	1	1	0	1
Y	1	0	1	0	0	1	1	1
Z	0	1	0	1	1	1	0	0

(B) What is Association Rule Mining? How it differs from other data mining techniques? A dataset has six transactions. Let ***min_sup*** = 50% and ***min_conf*** = 60%. **6**

TID	Items_bought
T1	K, A, D, B, C
T2	D, A, C, E, B
T3	C, A, B, E
T4	B, A, D, E
T5	C, D, K, E, F, B, A
T6	A, B, C, E
T7	A, K, C, D, B
T8	K, C, F, B, D

- Find all large item-sets in database using ***Apriori*** and ***FP-growth.***
- Compare the efficiency of the two mining process.
- Strong association rules and Exact association rules for dataset.

Q.3

(A) Apply k-means algorithms to develop **TWO** clusters with maximum squared error of cluster elements and mean should be below 10.5 **6**

Instances	X	Y
1	1.0	2.8
2	1.5	4.5
3	0.8	1.4
4	1.3	2.3
5	2.9	3.4
6	3.9	2.1
7	4.6	2.4
8	2.2	1.3
9	3.1	1.2
10	2.2	2.2

(B) For the given training set, predict classification of the sample: **6**

a) {2, 1, 1} b) {0, 1, 1}

Using ***Simple Bayesian Classifier.***

Training data set

Sample	Attr1	Attr2	Attr3	Class C
1	1	0	1	1
2	1	0	0	2
3	2	0	2	1
4	1	1	2	2
5	0	2	0	1
6	2	1	0	2
7	1	2	2	2
8	1	0	1	2

SECTION-II

Q.4

(A) What is Apriori concept? What are the disadvantages of Apriori algorithm for Association rule mining? How performance of Apriori Association Rule Mining algorithm can be improved by Dynamic Itemset Counting and partition based approaches? **6**

(B) What is web mining and how it is different from Data mining? How Web Content Mining and Web Usage Mining are different? What are the problems with web logs? Explain issues in both web taxonomies. **6**

OR

Q.4

(A) The following contingency table summarizes supermarket transaction data, where *hotdogs* refers to the transactions containing *hotdogs*, $\overline{hotdogs}$ refers to the transactions that do not contain *hotdogs*, *hamburgers* refers to the transactions containing *hamburgers*, and $\overline{hamburgers}$ refers to the transactions that do not contain hamburgers. **6**

	hotdogs	$\overline{hotdogs}$	*Sum(row)*
Hamburgers	2000	500	2500
$\overline{Hamburgers}$	1000	1500	2500
Sum(col)	3000	2000	5000

Suppose that the association rule "hotdogs → hamburgers" is mined. Given a *minsup* 25% and *minconf* 60%. Is this an association rule? If yes what is your say about strong association rule.

(B) A public opinion poll surveyed a simple random sample of 2000 voters. Respondents were classified by gender and by voting preferences. Results are shown in the following table **6**

	Voting Preference			Row Total
	Republican	Democrat	Independent	
Male	550	400	250	1200
Female	300	300	200	800
Column Total	850	700	450	2000

Show that the attributes gender and voting preferences are strongly related to each other or not. Suppose for 2 degree of freedom, the chi-square ($\chi 2$) value needed to reject the hypothesis that they are strongly correlated at the 0.05 significance level is 14.280.

Q.5

(A) Discuss following with example. Give the difference between them **6**
a) Fact Constellations Schema b) Snowflake Schema
Why Fact Constellations Schema is called Galaxy schema?

(B) Differentiate Inter-transaction Association Rule Mining with standard Association Rule Mining? Why this approach is not applicable to market basket analysis? Explain steps to mine Inter-transaction Association Rule using Sliding window. Take suitable example for explanation. **5**

OR

Q.5

(A) Draw a decision tree for following data using the concept buys_computer. Each internal node should represent a test on an attribute. Use ***InfoGain*** to find the best split. 6

RID	Age	Income	Student	Credit Rating	Class: buys_computer
1	<=25	Low	No	Fair	No
2	<=25	High	Yes	Excellent	No
3	26..50	High	No	Poor	Yes
4	>50	Medium	No	Fair	No
5	>50	Low	No	Poor	Yes
6	26..50	Low	Yes	Excellent	No
7	<=25	Low	Yes	Excellent	Yes
8	<=25	Medium	No	Poor	Yes
9	<=25	Low	Yes	Fair	No
10	>50	Medium	No	Poor	No

(B) Suppose that the data for analysis include the attribute *age*. The *age* values for the data tuples are (in increasing order): 13, 15, 16, 18, 19, 21, 23, 24, 26, 28, 29, 32, 35, 37, 38, 39, 39, 43, 45, 48, 51, 52, 55, 57, 62 5

a) Use min-max normalization to transform the value 35 for *age* onto the range [0.25, 1.25]

b) Use z-score normalization to transform the value 35 for *age*, where the standard deviation of *age* is 14.37 years

c) Use normalization by decimal scaling to transform the value 41 for *age*.

Comment on which method you would prefer to use for the given data, giving reason as to why.

Q.6

(A) 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 thousands. 6

Location = "India"					Location = "United Kingdom"				
	Item					Item			
Time	Samsung	Nokia	HTC	Sony	Time	Samsung	Nokia	HTC	Sony
Q1	150	225	155	110	Q1	100	150	85	65
Q2	175	205	150	90	Q2	85	152	95	75
Q3	160	220	110	100	Q3	95	125	90	75
Q4	165	250	120	95	Q4	110	135	100	70

Location = "China"					Location = "Germany"				
	Item					Item			
Time	Samsung	Nokia	HTC	Sony	Time	Samsung	Nokia	HTC	Sony
Q1	200	150	100	150	Q1	65	100	60	85
Q2	150	160	125	125	Q2	70	80	55	85
Q3	160	165	110	130	Q3	60	90	55	80
Q4	165	175	110	150	Q4	55	100	55	75

Show the slice, dice & rollup operation of the table shown in Q.6 (A)

Note: Do rollup operation on Regions.

(C) Differentiate 6

i) OLAP vs. OLTP

ii) Data Warehouse vs. Data Mart

--- END OF PAPER ---