Enrollment No. _____

GANPAT UNIVERSITY

M. Tech SEMESTER II - INFORMATION TECHNOLOGY

REGULAR EXAMINATION JULY - 2013

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

(A) What is Data Mining? Explain potential applications of data mining. **4**

(B) Briefly describe the following advanced database systems and applications: object-oriented databases, spatial databases, text databases, multimedia databases, the world wide web. **4**

(C) Draw and explain multi-tiered architecture of data warehouse. **4**

OR

Q.1

(A) Explain the architecture of typical data mining system. **4**

(B) Discuss major issues in Data Mining **4**

(C) What is Online Analytical Mining? Draw and explain OLAM architecture. **4**

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. **4**

(B) What is Privacy Preserving Data Mining? Explain *k*-anonymization approach to protect respondent's data before releasing? Explain its limitations and your views on the same. **4**

(C) Explain data cleaning, data reduction, data integration & transformation in brief. **3**

OR

Q.2

(B) 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. **6**

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

(A) What is Association Rule Mining? How it differs from other data mining techniques? A dataset has six transactions. Let ***min_sup*** = 60% and ***min_conf*** = 60%. **5**

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 9.5 **6**

Instances	X	Y
1	1.3	6.8
2	1.8	9.5
3	1.8	7.4
4	2.3	2.8
5	3.9	5.4
6	2.9	4.8
7	6.7	4.3
8	2.8	0.8
9	0.8	6.0
10	5.2	5.5

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

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

Using ***Simple Bayesian Classifier***.

Training data set for a classification using Naïve Bayesian Classifier

Sample	Attr1	Attr2	Attr3	Class C
1	1	0	1	1
2	0	1	1	2
3	2	0	2	1
4	1	1	2	2
5	0	2	2	1
6	2	1	2	2
7	1	0	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 Hash-based itemset counting approaches? **6**

(B) What is web mining? How it is different from Data mining? What are the challenges for mining World Wide Web? How Web Content Mining and Web Structure Mining are different? **6**

OR

Q.4

(A) Using the data of age given in 5.A below **6**

(a) Use smoothing by bin means to smooth the above data, using a bin depth 3. Illustrate your steps.

(b) How might you determine outliers in the data?

(c) What other methods are there for data smoothing. Apply data and compare the results with smoothing by bin means.

(B) Explain following extended association rule mining techniques **6**

1. Quantitative Association Rule Mining
2. Multi-level Association Rule Mining
3. Episode Rule Mining

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	High	No	Fair	No
2	<=25	High	No	Excellent	No
3	26..50	High	No	Poor	Yes
4	>50	Medium	No	Fair	Yes
5	>50	Low	Yes	Poor	Yes
6	>50	Low	Yes	Excellent	No
7	26..50	Low	Yes	Excellent	Yes
8	<=25	Medium	No	Fair	No
9	<=25	Low	Yes	Fair	Yes
10	>50	Medium	Yes	Poor	No
11	<=25	Medium	Yes	Poor	Yes
12	26..50	Medium	No	Excellent	Yes
13	26..50	High	Yes	Fair	No
14	>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): 15, 19, 21, 21, 22, 26, 27, 34, 35, 35, 40, 44, 49, 49, 53, 55, 57, 62, 67, 70, 72, 73, 75, 78, 81, 87, 90. **5**

a) Use min-max normalization to transform the value 48 for *age* into the range [1, 5]

b) Use z-score normalization to transform the value 48 for *age*

c) Use normalization by decimal scaling to transform the value 48 for *age*.

Comment on which method you would prefer any why.

OR

Q.5

(A) Discuss following with example. Give the difference between them **6**
a) Star Schema b) Fact Constellations 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**

Q.6

(A) 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 = "Chennai"					Location = "Hyderabad"				
	Item					Item			
Time	Computer	Monitor	CPU	Keyboard	Time	Computer	Monitor	CPU	Keyboard
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 = "Delhi"					Location = "Noida"				
	Item					Item			
Time	Computer	Monitor	CPU	Keyboard	Time	Computer	Monitor	CPU	Keyboard
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

Show the slice, dice & rollup operation of the table shown in Q.6 (A)
Note: Do rollup operation on city to Mega City & Metro City.

(C) Differentiate OLAP vs OLTP **4**

--- END OF PAPER ---