KUMARAGURU college of technology
character is life

Register Number:……..…

M.E DEGREE EXAMINATIONS: JUNE 2015

(Regulation 2014)

Second Semester

COMPUTER SCIENCE AND ENGINEERING

P14CST201: Data Warehousing and Data Mining

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 1 = 10 Marks)

1. Which of the following is the most important when deciding on the data structure of a data mart? [K_2]

   (i) XML data exchange standards
   (ii) Data access tools to be used
   (iii) Metadata naming conventions
   (iv) Extract, Transform, and Load (ETL) tool to be used

   a) i and iv only
   b) ii only
   c) iii and iv only
   d) i, ii, iii and iv

2. Identify the true statement from the following regarding the fact table of a data warehouse . [K_1]

   a) A fact table describes the granularity of data held in a data warehouse
   b) The fact table is the main store of descriptions of the transactions stored in a data warehouse
   c) A fact table describes the transactions stored in a data warehouse
   d) The fact table of a data warehouse is the main store of all of the recorded transactions over time

3. Which of the following is not an ETL tool? [K_1]

   a) Informatica
   b) Oracle warehouse builder
   c) Datastage
   d) Visual studio

4. Match list I with list II and select the correct answer for the given below table [K_2]

List I	List II
A. Data Management	i. The attributes within a given entity are fully dependent on the entire primary key of the entity.
B. Sampling	ii. Modeling and designing the warehouse
C. Data Enhancement	iii. Used to enable rapid election polling
D. Functional dependency of data	iv. The process of cleaning valid data to make it more meaningful

	A	B	C	D
a)	i	ii	iii	iv
b)	ii	iii	iv	i
c)	iv	iii	ii	i
d)	ii	iv	iii	i

5. Assertion (A): The data mining environment is the part of an organization whose core competency is data mining [K_2]
Reason (R): Organization have different sources of business intelligence data

a) both A and R are individually true and R is the correct explanation of A
b) both A and R are individually true but R is not the correct explanation of A
c) A is true but R is false
d) A is false but R is true.

6. Which one of the following is not correct statement? [K_2]

a) Entropy is one of the dimensionality reduction technique
b) Attribute subset selection reduces the data set size by removing irrelevant or relevant attributes
c) Smoothing works to remove noise from the data
d) Histograms use binning to approximate data distribution and are a popular form of data reduction

7. The basic features of clustering are _____ [K_2]
i) The best number of clusters is not known
ii) There may not be any a priori knowledge concerning the clusters
iii) Cluster results are dynamic

a) i and ii only
b) i and iii only
c) ii and iii only
d) i, ii and iii

8. Assertion (A) : Data warehouse refers to a data repository that is maintained separately from an organization's operational database [K_2]
Reason (R) : Data warehouse stores information from present data.

a) both A and R are individually true and R is the correct explanation of A
b) both A and R are individually true but R is not the correct explanation of A
c) A is true but R is false
d) A is false but R is true.

9. Document Clustering Analysis is one of the_____ approaches. [K_2]

a) Text Mining
b) WWW Mining
c) Spatial data Mining
d) Multimedia data Mining

10. Sequence the operations for cleansing the data [K_2]
(i) Locates and identifies individual data elements in the source files and then isolates these data elements in the target files
(ii) Transform data into its consistent format using both standard and custom business rules
(iii) Corrects individual data components using sophisticated data algorithms and secondary data sources
(iv) Searching and matching records within and across the standardized data, based on predefined business rules to eliminate duplications

a) i, iii, ii and iv
b) i, ii, iii and iv
c) i, ii, iv and iii
d) iv, ii, iii and i

PART B (10 x 2 = 20 Marks)

11. Give reasons, why it is necessary to separate data warehouse from operational database. [K_3]
12. Distinguish the features between OLAP and OLTP. [K_3]
13. What is the recommended method for data warehouse development? Explain with a sketch. [K_3]
14. State the drawbacks of virtual data warehouse. [K_2]
15. How can data preprocessing technique improves the quality of the data? [K_3]
16. Specify the advantages of generating concept hierarchies in data mining applications. [K_3]
17. Calculate the distance between two objects A= (22,2,54,10) and B=(21,0,43,9) using Mahalanobis distance and Minkowski distance for p=4. [K_2]
18. State any two major strengths of decision tree method. [K_2]
19. Bring out any two salient points with respect to social impact of data mining. [K_3]
20. Differentiate between web content mining and web structure mining. [K_2]

PART C (6 x 5 = 30 Marks)

21. Write the importance of metadata in a data warehouse environment. Also explain how data mining system can be integrated with datawarehouse system. [K_3]
22. Enumerate the importance of visualization of discovered patterns. What are the various visualization techniques in KDD. [K_3]
23. Describe the five primitives for specifying a data mining task. [K_2]
24. Suppose that the academic data warehouse for an University consists of the following dimensions: *student, course, semester and coordinator*. Along with this data two more measures *count* and *avg-grade* are provided. When at the lowest conceptual level (e.g.,for a given *student, course, semester and coordinator* combination), the *avg-grade* measure stores the actual course grade of the student and the *count* measure has value 1. At higher conceptual levels, *avg-grade* stores average grade for the given combination, and *count* stores the appropriate count value. [K_3]
    (i) Draw a snowflake schema diagram for the above data warehouse.
    (ii) Starting with the base cuboid (student, course, semester, coordinator) what specific OLAP operations should one perform in order to list the average grade of specific course for each student.
25. Illustrate how Back Propagation in Neural Network works? [K_2]
26. Briefly explain the text data analysis and information retrieval. [K_2]

PART D (4 x 10 = 40 Marks)

27. Write Apriori algorithm and using the algorithm find all frequent itemsets for the follwoing database (min_sup=20%) [K_4]

List of item IDs									
Transaction ID	A1	A2	A3	A4	A5	A6	A7	A8	A9
T1	1	0	0	0	1	1	0	1	0
T2	0	1	0	1	0	0	0	1	0
T3	0	0	0	1	1	0	1	0	0
T4	0	0	1	0	0	0	0	0	0
T5	0	0	0	0	1	1	1	0	0
T6	0	1	1	1	0	0	0	0	0
T7	0	1	0	0	0	1	1	0	1
T8	0	0	0	0	1	0	0	0	0

28. Write the advantages and disadvantages of k-means clustering as against model-based clustering. For the data set {2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377}, use the following techniques to form two clusters. [K_4]
    (i) k- means with initial centroids {1} and {378}
    (ii) k - means with initial centroids {21} and {34}

29. Discuss and elabrate the current trends in data mining. Apprise the applications of data mining for financial data analysis. [K_3]

30. Explain the three tier data warehousing architecture with a neat diagram. [K_3]

****