Exam. Code: 0924
Sub. Code: 6516

2054

B.E. (Information Technology)
Sixth Semester
IT-601: Data Warehouse & Data Mining

Time allowed: 3 Hours **Max. Marks: 50**

NOTE: ***Attempt five questions in all, including Question No. I (Section-A) which is compulsory and selecting two questions each from Section B-C. Any missing or misprinted data may be assumed suitably.***

x-x-x

Section-A

I. [2 x 5 =10]

a. Is Data Mining another hype in CS/IT/Statistics field? Comment.

b. Define Apriori Property. Explain with the help of an example.

c. What is data mart?

d. Compare characterization and discrimination.

e. Discuss the significance of Numerosity reduction.

Section B

II. a) How is a data warehouse different from a database? How are they similar? [5]

b) Compare OLAP and OLTP [5]

III. Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. [10]

(i) Enumerate three classes of schemas that are popularly used for modeling data warehouses.

(ii) Draw a schema diagram for the above data warehouse using one of the schema classes listed in (i).

(iii) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?

(iv) To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema *fee* (day, month, year, doctor, hospital, patient, count, charge).

IV. a) Discuss the properties of parallel RDBMS. [5]

b) Describe the steps involved in data mining when viewed as a process of knowledge discovery. [5]

Section C

V. A database has 5 transactions. Let min sup = 60% and min conf = 80%. [10]

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

Also Explain "support" and "confidence" with suitable example.

P.T.O.

VI. a) Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how the assessment of data quality can depend on the intended use of the data, giving examples. Propose two other dimensions of data quality. [6]

b) Use the methods below to normalize the following group of data: [4]

200, 300, 400, 600, 1000

(i) min-max normalization by setting min = 0 and max = 1

(ii) z-score normalization

(iii) z-score normalization using the mean absolute deviation instead of standard deviation

(iv) normalization by decimal scaling

VII. a) Discuss the relevance of Data Mining in business intelligence. [5]

b) Explain Mining class comparison with the help of suitable example. [5]

x-x-x