USN | | | | | | | | | | |

10IS74

Seventh Semester B.E. Degree Examination, June/July 2016
Data Warehousing and Data Mining

Time: 3 hrs. Max. Marks:100

Note: ***Answer any FIVE full questions, selecting atleast TWO questions from each part.***

PART – A

1 a. Define a data warehouse. Describe how a data warehouse is modeled and implemented using the star schema. Explain using example. **(08 Marks)**

b. What is ODS and what is it used for? Explain. **(04 Marks)**

c. What is ETL? Give three reasons for dirty data being extracted from source system. **(04 Marks)**

d. Discuss about the benefits of implementing a data warehouse. **(04 Marks)**

2 a. Define OLAP. Give two definitions. **(04 Marks)**

b. What is data cube? What are the different implementations of data cube? Explain. **(06 Marks)**

c. Explain the differences between ROLAP and MOLAP. **(06 Marks)**

d. Describe the operations of Data cube. **(04 Marks)**

3 a. What is Data Mining? Explain the four core data mining tasks with one application on each task. **(10 Marks)**

b. For the following vectors X & Y. Calculate the Cosine, Correlation, Euclidean and Jaccord similarity. X = (1, 1, 0, 1, 0, 1) ; Y = (1, 1, 1, 0, 0, 1). **(10 Marks)**

4 a. Consider the following transaction database for an supermarket Table 4.1. **(12 Marks)**

Customer	Items
C_1	Milk, egg, bread, chip
C_2	Egg, popcorn, chip, beer
C_3	Egg, bread chip
C_4	Milk, egg, bread, popcorn, chip, beer
C_5	Milk, bread, beer
C_6	Egg, bread, beer
C_7	Milk, bread, chip
C_8	Milk, egg, bread, butter, chip
C_9	Milk, egg, butter, chip

Generate all the frequent item sets. Also generate all the strong rules from the frequent itemsets by assuming the minimum support of 30% (atleast three transactions) and minimum confidence of 60%.

b. Write an algorithm to construct FP – tree, with an example. **(08 Marks)**

PART – B

5 a. Give the recursive definition of Hunts algorithm. **(04 Marks)**

b. What are the important characteristics of decision tree induction? **(06 Marks)**

c. Consider a training data set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules.
R_1 : A → + (covers 4 positive & 1 negative examples)
R_2 : B → + (covers 30 positive & 10 negative examples)
R_3 : C → + (covers 100 positive & 90 negative examples).
Determine which is the best and worst candidate rule according to : i) Rule accuracy ii) FOIL's information gain iii) The likelihood ratio statistic. **(10 Marks)**

6 a. Define Error rate. Discuss about the number of methods for estimating the accuracy of a method. **(10 Marks)**
b. List five criteria for evaluating classification methods. Discuss them briefly. **(05 Marks)**
c. Explain how bootstrapping, bagging and boosting improve the accuracy of classification. **(05 Marks)**

7 a. What is Cluster analysis? List the major issues in cluster analysis. **(05 Marks)**
b. Explain the K – means clustering method. **(05 Marks)**
c. Discuss about the hierarchical clustering method in detail. **(10 Marks)**

8 a. Explain the concept of finding similar web pages and finger printing in detail. **(10 Marks)**
b. Write short notes on :
i) Text mining ii) Spatial Data mining. **(10 Marks)**

*****