Printed Pages : 4 TCS-031

(Following Paper ID and Roll No. to be filled in your Answer Book)

PAPER ID : 0105 Roll No. □□□□□□□□□□

B. Tech.

(SEM. VII) EXAMINATION, 2007-08

DATA MINING AND DATA WAREHOUSING

Time : 3 Hours] *[Total Marks : 100*

*Note : Attempt **all** questions as per directions given thereof.*

1 Attempt any **two** parts of the following :

(a) What is data mining ? In your answer, address the following :

(i) Is it another hype ?

(ii) Is it a simple transformation of technology developed from databases, statistics and machine learning ?

(iii) Explain how the evolution of database technology led to data mining.

(iv) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

Present an example where data mining is crucial to the success of business. What data mining function does this business need ? Can they be performed alternatively by data query processing or simple statistical analysis ?

(b) How is a data warehouse differing from a database ? How are they similar to each other ? Describe different challenges to data mining regarding data mining methodologies and user interactions.

(c) In both data warehousing and data mining, it is important to have some hierarchical information associated with each dimension. If such a hierarchy is not given, discuss how to generate such hierarchy automatically for the first case of dimension containing only numerical data and also for the second case of a dimension containing only categorical data.

2 Attempt any **two** parts of the following :

(a) What are the differences between the three main types of data warehouse usage : information processing, analytical processing and data mining ? Discuss the motivation behind OLAP mining.

(b) Propose an algorithm, in pseudo code or in your favourite language you know, for the automatic generation of a concept hierarchy for numerical data base on the equi-depth partitioning rule.

(c) If your data set contains missing values, discuss the basic analyses and corresponding decisions you will take in the preprocessing phase of the data-mining process. Develop a software tool for the detection of outliers if the data for preprocessing are given in the form of a flat file with n-dimensional samples.

3 Attempt any **two** parts of the following :

(a) List and describe the different primitive for specifying a data mining task. Many authors include OLAP tools as a standard data-mining tool. Give the arguments for and against this classification.

(b) Suppose that university course database for UPTU contains the following atrributes : name, address, status, major of each student and their comulative grade point average (GPA), propose a concept hierarchy for the atrributes status, major GPA and address.

(c) Why is the validation of a clustering process highly subjective ? What increases the complexity of clustering algorithms ?

4 Attempt any **two** parts of the following :

(a) Explain the concept of a data cube and where it is used for visulization of large data sets. Use examples to discuss the differences between icon-based and pixel-oriented visualization techniques.

(b) Why is the text-refining task very important in text-mining process ? What are results of text refining ?

(c) Implement and *Apriori* algorithm and discover large itemsets in transactional database.

5 Attempt any **two** parts of the following :

(a) What is clustering ? How is it different from classification ?

(b) Data cubes and multidimensional database contains catergorical, ordinal and numerical data in hierarchical or aggregate form, design a clustering method which finds clusters in large data cubes effectively and efficiently.

(c) Human eyes are fast and effective at judging the quality of clustering methods for two dimensional data. Can you design a data visualization method which can help humans visualize data clusters and judge the clustering quality for three dimensional data ? What about even higher dimensional data ?