Roll No. | | | | | | | | | | | NCS066

B.TECH.

THEORY EXAMINATION (SEM–VI) 2016-17

DATA WAREHOUSING & DATA MINING

Time : 3 Hours ***Max. Marks : 100***

Note : *Be precise in your answer. In case of numerical problem assume data wherever not provided.*

SECTION – A

1. Explain the following: **10 x 2 = 20**

(a) How is the data warehouse different from a database?

(b) Distinguish the snowflake model from the fact constellation model.

(c) Mention the characteristics of a data warehouse.

(d) State the use of meta data in the context of data warehousing.

(e) Give a precise definition of the term 'concept hierarchy'.

(f) Why data cleaning routines are needed?

(g) Give the definition of the terms 'frequent itemset', 'support' and 'confidence'.

(h) Represent a decision tree for a student record database.

(i) Classify OLAP tools.

(j) Bring out any two points with respect to spatial mining.

SECTION – B

2. Attempt any five of the following questions: **5 x 10 = 50**

(a) "A data warehouse can be modeled by either a star schema or a snowflake schema". With relevant examples discuss the two types of schema.

(b) Enumerate the steps involved in mapping the data warehouse to a multiprocessor architecture.

(c) Describe challenges to data mining regarding data mining methodology and user interaction issues.

(d) Summarize the smoothing techniques followed in data cleaning process.

(e) How data mining systems are classified? Describe each classification with example.

(f) Discuss issues that are important to consider when employing a decision tree – based classification algorithm. Explain the decision tree induction algorithm with appropriate examples. Discuss the disadvantages of this approach? What is over fitting, and how can it be prevented for decision trees?

(g) Diagrammatically illustrate and discuss the architecture of MOLAP and ROLAP.

(h) What is web mining? Differentiate between web content mining, web structure mining and web usage mining.

SECTION – C

Attempt any two of the following questions: **2 x 15 = 30**

3 Suppose that a data warehouse for a University consists of the following four dimensions: student, course, semester and instructor, and two measures such as count and avg_grade.

When at the lowest conceptual level (e.g., for a given student, course, semester and instructor combination) the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.

(i) Draw a snowflake scheme diagram for the data warehouse.

(ii) Starting with th t, course, semester, instructor], what

specific OLAP operations (e.g. roll-up from semester to year) should one perform in order to list the average grade of CS courses for each student of the University.

4 Consider five points $\{X_1, X_2, X_3, X_4, X_5\}$ with the following coordinates as a two dimensional sample for clustering: $X_1=(0,2.25)$; $X_2=(0,0.25)$; $X_3=(1.25,0)$; $X_4=(4.5,0)$; $X_5=(4.5,2.5)$; Illustrate the K-means partitioning algorithm (clustering algorithm) using the above data set.

5 Compare and contrast spatial, temporal mining with relevant examples.

quescol.com