G 6174

M.E. DEGREE EXAMINATION, MAY/JUNE 2007.

Elective

Computer Science and Engineering/Software Engineering

CS 1634 — DATA WAREHOUSING AND DATA MINING

(Regulation 2005)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What is outlier analysis?

2. Describe the role of DBMS in data mining.

3. What are the means to improve the performance of association rule mining algorithms?

4. What is Concept Hierarchy? Give example.

5. State the advantages of the decision tree approach over other approaches for performing classification.?

6. What is supervised learning? Give example.

7. What is a data mark? Give example.

8. Write short notes on data warehouse meta data.

9. Give some examples of data mining tools.

10. What is spatial data mining? Give examples.

PART B — (5 × 16 = 80 marks)

11. (a) (i) Diagrammatically illustrate and discuss the architecture of a typical data mining system. (10)

(ii) What is a knowledge base? Explain the need of knowledge base in data mining. (6)

Or

(b) (i) Describe the various issues in data mining. (8)

(ii) Tabulate and briefly discuss the various techniques. (8)

12. (a) Suppose that the data for analysis include the attribute age. The age values for the data tuples are 13, 15, 16, 19, 20,20,21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(i) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps.

(ii) How will you determine outliers in the data.

(iii) What other methods are there for data smoothing?

Or

(b) Write the algorithm to discover frequent itemsets without candidate generation and explain it with an example. (8 + 8)

13. (a) (i) Why is tree pruning useful in decision tree induction? Give example. Also state the draw back of using a separate set of samples to evaluate pruning? (8)

(ii) Why is naïve Baysesian classification called 'naïve"? Briefly outline the major ideas of naïve Bayesian classification. (8)

Or

(b) Explain the following clustering methods in detail.

(i) BIRCH (8)

(ii) CURE (8)

14. (a) Suppose that a data warehouse for big-university consist of the following four dimensions: Student, Course, Semester and Instructor, and two measures count and avg_grade. When at the lowest conceptual level (Ex. 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 schema diagram for the data warehouse. (8)

(ii) Starting with the base cuboid (Student, Course, Semester Instructor), what specific OLAP operations (eg Roll-up from Semester to Year) should one perform in order to list the average grade of CS courses for each big-university students. (4)

(iii) If each dimension has five levels(including all) such as Student < major < status<university < all, how many cuboids will this cube contains (including the base and Apex cuboids). (4)

Or

(b) (i) Explain the data model which is suitable for data warehouse with examples. (8)

(ii) Compare OLAP with OLTP. (8)

15. (a) (i) Discuss the social impacts of data mining. (10)

(ii) Write the difference between direct query processing and intelligent query processing. Give examples. (6)

Or

(b) (i) Discuss the application of data mining in business. (8)

(ii) What is web mining? Discuss the various web mining techniques. (8)