C5-R4: DATA WAREHOUSING AND DATA MINING

NOTE:

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

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

1.

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

b) Explain the role of Starnet Query Model in querying multidimensional databases with an example.

c) Compare and contrast the incremental update operation performed in ROLAP, MOLAP and HOLAP.

d) Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.

e) What is an Iceberg Query? Can we use it in market basket analysis? Justify your answer with an example.

f) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

g) What is similarity search in time-series analysis? Explain its usefulness in various business functions.

(7x4)

2.

a) In real-world data tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

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

(9+9)

3.

a) What is a Confusion Matrix for classifier? Write a 4x4 confusion matrix for a classifier with 100% accuracy? The instances in the four classes are A(20), B(35), C(40), D(10). Also calculate classifier precision in detail?

b) What do you understand by Principal Component Portioning Algorithm? Explain the algorithm in detail.

(9+9)

4.

a) Briefly describe the following approaches to clustering: hierarchical methods, density-based methods, grid-based methods, model-based methods, methods for high-dimensional data, and constraint-based methods. Give examples in each case.

b) What are multidimensional Association Rules? Explain in brief.

(12+6)

5.

a) There are several major differences between biological sequential patterns and transactional sequential patterns. First, in transactional sequential patterns, the gaps between two events are usually nonessential. For example, the pattern “purchasing a digital camera two months after purchasing a PC" does not imply that the two purchases are consecutive. However, for biological sequences, gaps play an important role in patterns. Second, patterns in a transactional sequence are usually precise. However, a biological pattern can be quite imprecise, allowing insertions, deletions, and mutations. Discuss how the mining methodologies in these two domains are influenced by such differences.

b) 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.

   i) Draw a Star Schema for the above data warehouse.

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

(9+ 9)

6. Explain following concepts:

a) Web Data Mining

b) Attribute Oriented Induction

c) Fuzzy set theory

(3x6)

7.

a) What is Time-series Database? How to characterize the time-series data using trend analysis?

b) Explain the following methodologies for stream data processing and stream data systems:

   i) Random sampling

   ii) Sliding Windows

   iii) Sketches

(8+10)