CE3-R3: DATA WAREHOUSING AND DATA MINING

NOTE:

1. **Answer question 1 and any FOUR questions from 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) Describe the differences between the following architectures for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semi-tight coupling, and tight coupling. Also mention which architecture is the most popular one and why.

b) Discuss various types of concept hierarchies by providing two examples for each type?

c) Discuss the differences between canned queries and ad hoc queries.

d) Illustrate the typical requirements of clustering data mining.

e) State various evaluation criteria that are essential for classification and prediction methods.

f) What are the difficulties that can arise with hieratical clustering?

g) State the differences between data quality and data accuracy.

(7x4)

2. Suppose that a data warehouse consists of four dimensions date, spectator, location, and game, and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors. With each category having its own charge rate.

a) Draw a star schema diagram for the data warehouse.

b) How many cuboids are needed to build the data cube? List them

c) Starting with base cubiod, what specific OLAP operations should one need to perform in order to list the total charge paid by student spectators at New Delhi in the year 2004?

(6+6+6)

3.

a) What are the advantages and limitations of snowflake schema design?

b) What is meant by data reduction? Discuss any two data reduction strategies for obtaining a reduced data representation.

(9+9)

4.

a) What does hierarchical clustering mean? In what way it is different from partition-based methods.

b) Discuss the functionality of Chameleon's clustering method with an example.

c) What is concept hierarchy? Explain its importance in Data Mining.

(6+8+4)

5.

a) Give an example to show that items in a strong association rule may actually be negatively correlated.

b) What is meant by Multi level association rule? Discuss any two approaches for mining multi level association rules with examples.

(6+12)

6.

a) Discuss how to develop an efficient implementation of data mining system for mining weblog access sequences.

b) An e-mail database is a database that stores a large number of electronic mail messages. Such a database is one kind of semi-structured database consisting of textual data.

i) How can you structure such an email database in order to facilitate multidimensional search.

ii) What can be mined from such an email database?

iii) Suppose email messages were classified as junk, unimportant, normal or important, describes how a data mining system may take this as the training set to automatically classify new email messages or unclassified ones.

(6+12)

7.

a) Define a spatial data cube. Discuss different types of dimensions in a spatial data cube.

b) State the salient differences between data query and knowledge query?

c) An object cube can be constructed by generalization of an object-oriented database into relatively structured data prior to performing multidimensional generalization. Discuss how to handle set-oriented data in an object cube.

(5+5+8)