CE3-R3: DATA WAREHOUSING AND 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) What is Data Mining? Why should data be preprocessed? List various data preprocessing techniques and explain the purpose of each technique.

b) List four conditions that a distance function used for clustering must satisfy. Show that Minskowski function satisfies all these condition.

c) Describe the process of designing a data warehouse?

d) What are metadata? Explain metadata repository.

e) What is web content mining? What are different techniques of web content mining?

f) Give mathematical formulation of association rule mining problem.

g) What are Bayesian classifiers? Explain the theorem on which Bayesian classification is based.

(7x4)

2.

a) What are the different criteria on which classification and prediction methods can be evaluated? Explain the major steps of decision tree classification.

b) How does backpropagation work? Write the neural network learning algorithm for classification, using the backpropogation.

c) Explain the following: -

i) K-Nearest Neighbor Classifier algorithm

ii) Case-based Reasoning algorithm

(6+6+6)

3.

a) Compare and contrast OLTP and OLAP systems. What is the reason for constructing separate data warehouse to perform online analytical processing?

b) Suppose that a data warehouse consists of four dimensions date, viewer, cinema hall and movie and two measures count and charge, where charge is the ticket fee that the viewer pays for watching the movie on a given date. The viewers can be children below 5, above 5, adults or seniors with each category having its own charge rate.

i) Draw a star schema diagram for data warehouse.

ii) Starting with the base cuboid [date, viewer, cinema hall, movie], what specific OLAP operations one should perform in order to list the total charge paid by adults at the cinema hall 'Paradise' in 2004?

iii) Bitmap indexing is useful in data warehousing. Taking this cube as an example, briefly discuss advantages and problems of using a bitmap index structure.

(6+[4+4+4])

4.

a) What are different types of association rules? Give one example of each type.

b) Explain Apriori Algorithm with an example.

(6+12)

5.

a) What is an outlier? How are they caused? Why is it important to mine for outliers? What are the applications of outlier mining?

b) Assuming that there are n objects in a cluster each with dimension 'd', illustrate how is the centroid of this cluster computed.

c) Assuming there are K classes, N objects in the data set, each with d dimensions, how are the following computed?
   i) Information Gain
   ii) Gain Ratio
   iii) Gini Index

(6+6+6)

6.

a) Differentiate between direct query answering and intelligent query answering. Suppose that a web-based on-line shopping center maintains several databases for its business, which include on-line catalog database, a transaction history database and a weblog database. Explain by giving examples, how intelligent query answering may improve online shopping services by incorporating data mining techniques.

b) Mining weblog access sequences may help prefetch certain Web pages into a Web server buffer such as those pages that are likely to be requested in the next several clicks. Design an algorithm for mining such access sequences.

c) Given the following distance matrix for five objects, cluster them using single link distance. Specify the thresholds used. Represent the clustering as a dendrogram.

	a	b	c	d	e
a	0				
b	10	0			
c	5	2	0		
d	25	20	35	0	
e	30	30	25	5	0

(6+6+6)

7. Explain the following in detail (**any three**):

a) Iceberg Queries

b) Text Mining

c) Data Compression

d) Snowflake Schema

(3x6)