## 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) What is Concept Hierarchy? Describe why Concept Hierarchies are useful in data mining?
b) Can we construct a data cube for multimedia data analysis? Justify your answer. What are multidimensional association rules?
c) What is data mining? Explain how the evolution of database technology led to data mining?
d) How can you measure whether the pattern is interesting or useful for taking any decision? Explain each measurement in brief.
e) What are measures for assessing quality of text retrieval mining system?
f) Update-driven approach is chosen rather than query driven approach while integrating multiple heterogeneous information sources. Justify whether true or false.
g) Differentiate data query and knowledge query.

**(7x4)**

**2.**

a) Star schema or a snowflake schema is used to model data warehouse. What are the similarities and the differences of the two models, and then state their advantages and disadvantages.
b) Briefly outline how to compute the dissimilarity between objects described by the following types of variables
   i) Interval-scaled variables
   ii) Asymmetric Binary variables
   iii) Categorical variables
   iv) Ratio-scaled variables
   v) Non-metric vector objects

**(6+12)**

**3.**

a) What is the purpose of Apriori Algorithm? How to generate association rules from frequent item sets? Explain.
b) What are multidimensional association rules? Explain in brief.
c) List out the OLAP operations in multidimensional data model? What is roll-up and drill-down operation?

**(8+6+4)**

**4.**

a) What is a confusion matrix for a 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 and recall.
b) What do understand by principal component partitioning algorithm? Explain the algorithm in detail.

**(9+9)**

**5.**

a) A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive objects. What are the requirements of cluster analysis? Explain each of them in brief.

b) Describe each of the following clustering algorithms in term of the following criteria:
   - i) Shapes of clustering that can be determined.
   - ii) Input parameters that must be specified; and
   - iii) Limitations
     - i.i) K-means
     - i.ii) K-medoids
     - i.iii) CLARA

c) What do you understand by spatial database and spatial data mining? Can we construct a spatial data warehouse?

**(6+6+6)**

**6.**

a) What are Bayesian classifiers? Explain briefly Baye's theorem. Also explain how Naïve Bayesian classifier works?

b) What are the issues related to data integration of pre-processing steps.

**(10+8)**

**7.**

a) What is time-series database? How to characteristics the time series data using tread analysis.

b) Explain the following methodologies for stream data processing and stream data systems:
   - i) Random sampling
   - ii) Sliding Windows
   - iii) Sketches

**(6+12)**