Total No. of printed pages = 7

63/2 (SEM-3) CSIT 3.3

2021

(held in 2022)

CSIT

(Theory Paper)

Paper Code : CSIT-3.3

(Data Mining and Warehousing)

Full Marks – 80

Time – Two hours

The figures in the margin indicate full marks for the questions.

1. Answer the following questions : 1×10=10

(a) Cluster is ———

(i) Group of similar objects that differ significantly from the other objects.

(ii) Operations on a database to transform or simplify data in order to prepare it for a machine learning algorithm.

(iii) Symbolic representation of facts or ideas from which information can potentially be extracted.

(iv) None of these.

(b) Data warehouse architecture is based on

(i) DBMS (ii) RDBMS

(iii) SYBASE (iv) SQL Server

(c) Which of the following is not belong to data mining ?

(i) Knowledge extraction

(ii) Data transformation

(iii) Data exploration

(iv) Data archaeology

(d) What is the output of KDD ?

(i) Query (ii) Useful information

(iii) Data (iv) Information

(e) Which of the following is a good alternative to the star schema ?

(i) Snowflake schema

(ii) Star schema

(iii) Star snowflake schema

(iv) Fact constellation

(f) Removing loopholes and deficiencies in the data is said to be

(i) Data aggregation

(ii) Extraction of data

(iii) Compression of data

(iv) Cleaning of data.

(g) Which of the following statements is true ?

(i) Data mining can be referred to as the procedure of mining knowledge from data.

(ii) Data mining can be defined as the procedure of extracting information from a set of the data.

(iii) The procedure of data mining also involves several other process like data cleaning, data transformation and data integration.

(iv) All of the above.

(h) —— may be defined as the data objects that do not comply with the general behaviour or model of the data available.

(i) Outlier analysis (ii) Evolution analysis

(iii) Prediction (iv) Classification

(i) The Apriori property means

(i) If a set cannot pass a test all of its supersets will fail the same test as well.

(ii) To improve the efficiency the level-wise generation of frequent item sets.

(iii) If a set can pass a test, all of its supersets will fail the same test as well.

(iv) To decrease the efficiency the level-wise generation of frequent item sets.

(j) A data warehouse is said to contain a "Subject oriented" collection of data because

(i) It contents have a common theme

(ii) It is built for a specific application

(iii) It cannot support multiple subjects

(iv) It is a generalization of "Object-oriented".

2. Answer any *five* of the following questions : 2×5=10

(a) What is fact table ? Give an example.

(b) Explain major requirements and challenges in data mining.

(c) What is apex of cuboid, define with example.

(d) Define frequent set. Define an association rule.

(e) What do you mean by classification and prediction ?

(f) Why we need data warehouse ?

3. Answer any *six* of the following questions : 5×6=30

(a) Discuss the classification by decision-tree induction.

(b) Explain OLAP operations in the multidimensional data model.

(c) What do you mean by data mart ? Why do we need a data mart ?

(d) Write down the Apriori Algorithm.

(e) Describe the different syntax for various schemas of a data warehouse.

(f) Describe briefly the major features between OLAP and OLTP.

(g) The following table consists of training data. Construct a decision-tree based on this data using the basic algorithm for decision-tree induction. Classify the records by the status attributes. Write down the rules that can be generated from the obtained decision-tree.

Department	Age-range	Salary-class	Status
Sales	Middle-aged	High	Senior
Sales	Young	Low	Junior
Sales	Middle-aged	Low	Junior
Systems	Young	High	Junior
Systems	Middle-aged	High	Senior
Systems	Young	High	Junior
Systems	Senior	High	Senior
Systems	Middle-aged	High	Senior
Marketing	Middle-aged	Average	Junior
Marketing	Senior	Average	Senior
Secretary	Young	Low	Junior

4. What is the difference between data mining and knowledge discovery ? What is the need of data mining ? Explain the stages involved in KDD process. 10

5. Draw the diagram and explain the architecture of data mining. 10

6. Explain the algorithm for mining frequent item. Sets without candidate generation for the given set (min support = 03) 10

TID	List of items
1	Milk, Bread, Eggs
2	Bread, Sugar
3	Bread, Cereal
4	Milk, Bread, Sugar
5	Milk, Cereal
6	Bread, Cereal
7	Milk, Bread, Cereal, Eggs
8	Milk, Bread, Cereal.