JALPAIGURI GOVERNMENT ENGINEERING COLLEGE

[A GOVERNMENT AUTONOMOUS COLLEGE]

COE/B.TECH./CSE/PEC-IT601B/2021-22

2022

DATA WAREHOUSING & DATA MINING

Full Marks: 70

Times: 3 Hours

The figures in the margin indicate full marks.
Candidates are requested to write their answers in their own words as far as practicable.

GROUP-A
[OBJECTIVE TYPE QUESTIONS]

Answer all questions

5x2=10

1. Define data warehouse.
2. "FP-Tree approach is faster than Apriori algorithm for large frequent item-sets detection"- Justify.
3. Construct a lattice of cuboids for a four dimensional data warehouse?
4. What is concept hierarchy?
5. What is virtual warehouse?

GROUP-B
[LONG ANSWER TYPE QUESTIONS]

Answer any *four* questions

4x15=60

6. i) Provide the pseudo code of the k-means clustering algorithm. State the advantage and drawback of k-means algorithm. 5+2

ii) Compare the centroid update process of k-means with the medoid update process of k-medoids. 2

iii) Find two clusters from the given data (A(1,3), B(7,3), C(2,9), D(5,5), E(9,7), F(3,7), G(6,7), H(5,9), I(1,9)) when initial centroids are C and G. Show all the steps for 3 iterations. 6

7. i) State the apriori property, joining rule and pruning process for Apriori Algorithm. How does pruning step help to reduce execution time of Apriori Algorithm? 3+2

ii) Enumerate all frequent itemsets from the given database using Apriori algorithm with minimum support count S=3. List all the candidate set and large frequent itemsets for each database scan. Show the association rules along with their confidence for all the frequent itemsets in L3. 5+3

TID	Item_Codes
T1	M, O, N, K, E, Y
T2	D, O, N, K, E, Y
T3	M, A, K, E
T4	M, U, C, K, Y
T5	C, O, O, K, I, E

iii) Given frequent itemset I and subset s of I, prove that the confidence of the rule "s' ⇒ (I – s')" cannot be more than the confidence of "s ⇒ (I – s)," where s' is a subset of s. 2

8. i) Draw the decision tree of the training data given in Table 2 using information gain. {*cheat* is the class label attribute}. 10

ii) List the classification rules obtained from the decision tree. 3

iii) What are the differences between supervised and unsupervised machine learning? 2

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	10 to 15	No
2	No	Married	10 to 15	No
3	No	Single	up to 8	No
4	Yes	Married	10 to15	No
5	No	Divorced	8 to 10	Yes
6	No	Married	up to 8	No
7	Yes	Divorced	10 to 15	No
8	No	Single	8 to 10	Yes
9	No	Married	up to 8	No
10	No	Single	10 to 15	Yes

Table 2

9. i) Draw the FP-Tree for the given database with minimum support count S=2. 6

ii) Derive all the conditional FP-Tree and state the frequent itemsets. 7

iii) Write the advantage of FP-Tree Algorithm over Apriori Algorithm. 2

TranID	List of Item_IDs
T100	Butter, Milk, Rice
T200	Bread, Butter, Jam
T300	Butter, Sugar
T400	Bread, Butter, Milk
T500	Bread, Sugar, Pepsi
T600	Butter, Sugar, Curd
T700	Bread, Sugar
T800	Bread, Butter, Sugar, Jam
T900	Bread, Butter, Sugar

10. i) Briefly compare the following concepts. You may use an example to explain your point(s).

(a) Star schema and snowflake schema 3

(b) Independent and dependent data marts 3

(c) OLAP and OLTP 6

ii) Write a short note on Metadata repository. 3

11. i) Draw the 3-Tier Data Warehouse architecture and explain each tier. 8

ii) Discuss the OLAP operations which are performed in the middle tier of the data warehouse architecture on Multidimensional Data Model. 7

JALPAIGURI GOVERNMENT ENGINEERING COLLEGE

[A GOVERNMENT AUTONOMOUS COLLEGE]

COE/B.TECH./CSE/CS604C/2019-20

2020

DATA WAREHOUSING & DATA MINING

Full Marks: 70

Times: 3 Hours

The figures in the margin indicate full marks.
Candidates are requested to write their answers in their own words as far as practicable.

Answer any ***SEVEN*** questions

7x10=70

1
i) Compare classification approach with clustering. 3
ii) Illustrate the strength and weakness of k-means in comparison with k-medoids. 3
iii) Write the steps of k-means algorithm for clustering. 4

2.
i) State the apriori property. Explain the steps of apriori algorithm. 4
ii) Enumerate all frequent itemsets from the given database using Apriori algorithm with minimum support count S=3. List all the candidate set and large frequent itemsets for each database scan. 6

TID	Item_Codes
T1	M, O, N, K, E, Y
T2	D, O, N, K, E, Y
T3	M, A, K, E
T4	M, U, C, K, Y
T5	C, O, O, K, I, E

.

3.
i) Draw the decision tree of the training data given in Table 2 using information gain. {***cheat*** is the class label attribute}. 10

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	10 to 15	No
2	No	Married	10 to 15	No
3	No	Single	up to 8	No
4	Yes	Married	10 to15	No
5	No	Divorced	8 to 10	Yes
6	No	Married	up to 8	No
7	Yes	Divorced	10 to 15	No
8	No	Single	8 to 10	Yes
9	No	Married	up to 8	No
10	No	Single	10 to 15	Yes

Table 2

4. Discuss the 3-tier architecture of data warehouse with a suitable diagram. 10

5. Briefly explain any two concepts of data warehouse schema with example. 10

6. Discuss the OLAP operations which are performed in the middle tier of the data warehouse architecture on Multidimensional Data Model. 10

7.
i) Explain any two algorithms for parent selection. 6
ii) Discuss different types of crossover operations in GA. 4

8. Write the differences between OLAP and OLTP. 10

9. Find all frequent itemsets using FP-Tree and derive the strong association rules for the given dataset. Let the min_support = 60% and min_conf = 70%. 10

Trans. ID	*List of Item ID's*
T1	*f, a, c, d, g, i, m, p*
T2	*a, b, c, f, l, m, o*
T3	*b, f, h, j, o*
T4	*b, c, k, s, p*
T5	*a, f, c, e, l, p, m, n*

10 Write short notes on any ***three*** **2**x5=10 10

a) Social impacts of data mining, b) Knowledge Discovery in Database, c) Multidimensional Data, d) Optimization of view materialization

----

4
3
3
3
2

2

6

5

2

10

3

2

No

JALPAIGURI GOVERNMENT ENGINEERING COLLEGE

[A GOVERNMENT AUTONOMOUS COLLEGE]

COE/B.TECH./CSE/CS604A/2018-19

2019

DATA WAREHOUSING & DATA MINING

Full Marks: 70

Times: 3 Hours

The figures in the margin indicate full marks.
Candidates are requested to write their answers in their own words as far as practicable.

GROUP-A
[OBJECTIVE TYPE QUESTIONS]

Answer all questions 5x2=10

1. How is a data warehouse different from a database?
2. "FP-Tree approach is faster than Apriori algorithm for large frequent item-sets detection"- Justify.
3. Construct a lattice of cuboids for a four dimensional data warehouse?
4. What is concept hierarchy?
5. What is virtual warehouse?

GROUP-B
[LONG ANSWER TYPE QUESTIONS]

Answer any ***four*** questions 4x15=60

6. i) Provide the pseudo code of the object reassignment step of the PAM algorithm. 4
ii) Illustrate the strength and weakness of k-means in comparison with k-medoids. 3
iii) Compare the PAM algorithm with CLARA method for clustering. 3
iv) Write short note on Clustering Large Applications based upon RANdomized Search (CLARANS) algorithm. 3
v) Define agglomerative and divisive hierarchical clustering method. 2

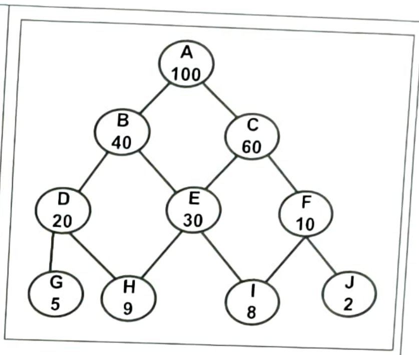
7. i) State the apriori property. 2
ii) Enumerate all frequent itemsets from the given database using Apriori algorithm with minimum support count S=3. List all the candidate set and large frequent itemsets for each database scan. 6
iii) Draw the FP-Tree for the same database with minimum support count S=3. 5
iv) Given frequent itemset l and subset s of l, prove that the confidence of the rule "s' ⇒ (l – s')" cannot be more than the confidence of "s ⇒ (l – s)," where s' is a subset of s. 2

TID	Item_Codes
T1	M, O, N, K, E, Y
T2	D, O, N, K, E, Y
T3	M, A, K, E
T4	M, U, C, K, Y
T5	C, O, O, K, I, E

8. i) Draw the decision tree of the training data given in Table 2 using information gain. {***cheat*** is the class label attribute}. 10
ii) List the classification rules obtained from the decision tree. 3
iii) What are the differences between supervised and unsupervised machine learning? 2

TID	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	10 to 15	No
2	No	Married	10 to 15	No
3	No	Single	up to 8	No
4	Yes	Married	10 to15	No
5	No	Divorced	8 to 10	Yes
6	No	Married	up to 8	No
7	Yes	Divorced	10 to 15	No
8	No	Single	8 to 10	Yes
9	No	Married	up to 8	No
10	No	Single	10 to 15	Yes

Table 2

9.	i) What is the goal of optimization of view materialization problem? ii) Define base and apex cuboids with example. iii) Consider the following lattice of views (Fig. 1) along with a representation of the number of rows in each view where A is the base cuboids. Consider that view A is already materialized. Find another three views for materialization from B-J views which provide maximum benefit. iv) Compute the overall benefit achieved after materialization of the views.  A 100 B 40 C 60 D 20 E 30 F 10 G 5 H 9 I 8 J 2	2 2 9 2
10.	i) Briefly compare the following concepts. You may use an example to explain your point(s). (a) Star schema, snowflake schema and fact constellation (b) Independent and dependent data marts (c) OLAP and OLTP	6 3 6
11.	i) Discuss the OLAP operations which are performed in the middle tier of the data warehouse architecture on Multidimensional Data Model. ii) Write a short note on Metadata repository.	10 5
12	i) What is genetic algorithm? What are the common steps of genetic algorithm? ii) Explain any two algorithms for parent selection. iii) Discuss different types of crossover operations in GA.	2+2 6 5

JALPAIGURI GOVERNMENT ENGINEERING COLLEGE

[A GOVERNMENT AUTONOMOUS COLLEGE]

COE/B.TECH./CSE/CS604C/2015-16

2016

DATA WAREHOUSING & DATA MINING

Full Marks: 70

Times: 3 Hours

The figures in the margin indicate full marks.
Candidates are requested to write their answers in their own words as far as practicable.

GROUP-A
[OBJECTIVE TYPE QUESTIONS]

Answer **all** questions

5x2=10

1. What is strong association rule?
2. What are the various steps of data mining?
3. Write the difference between database and knowledge base.
4. Compare clustering and classification techniques.
5. What is fact constellation?

GROUP-B
[LONG ANSWER TYPE QUESTIONS]

Answer any ***Four*** of the following

4x15=60

6. a) Write down the k-means clustering algorithm. State the strong point and limitation of this algorithm.
b) Cluster the following items into three (03) clusters using k-means algorithm and Euclidean distance.
Items: **A1**(2,10); **A2**(2,5); **A3**(8,4); **A4**(9,4); **A5**(5,8); **A6**(7,5); **A7**(6,4); **A8**(1,2); **A9**(4,9); **A10**(6,10).

Suppose that the initial seeds (centroid of each cluster) are **A1, A4** and **A9**.

Run the k-means algorithm for three iterations and at the end of each iteration, show:
i) The new clusters (i.e. items belonging to each cluster)
ii) Centre of the new cluster 4+2+9=15

7. a) Write the main steps of Apriori algorithm. Find the frequent itemsets in the transaction database given in Table 1 using Apriori algorithm [Min_sup= 2 and min_conf= 70%].

b) Write at least two strong association rules for the records given in Table 1.

Table 1

Tran_ Id	List of items
T001	a, b, e
T002	b, d
T003	b, c
T004	a, b, d
T005	a, c
T006	b, c
T007	a, c
T008	a, b, c, e
T009	a, b, c

(2+10)+3=15

8. Suppose that a data warehouse for big-bazar consists of the four dimensions customer, city, product, and time, and two measures count and sales-amount. At the lowest conceptual level (i.e., for a given customer, city, product and time combination), the sales-amount measure stores the actual purchase amount of the customer. At higher conceptual levels, sales-amount stores the total purchase amount for the given combination.

a) Draw a schema diagram for modeling the above data warehouse. State clearly the tables, facts & keys.

b) Starting with the base cuboid [customer, city, product and time], what specific OLAP operations should you perform in order to list the total sales amount of "product=computer" for each country.

10+5=15

9. Construct the FP-Tree for the given database (Table 2) and state all conditional FP-Tree. (min_sup = 3).

Table 2

Tran_Id	**List of items**
T1	F, A, C, D, G, I, M, P
T2	A, B, C, F, L, M, O
T3	B, F, H, J, O
T4	B, C, K, S, P
T5	A, F, C, E, L, P, M, N

15

10 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* attribute. Write down the rules that can be generated from the obtained Decision Tree.

TID	**Dept.**	**Age-group**	**Salary-class**	**Status**
1	Sales	Middle	High	Senior
2	Sales	Young	Low	Junior
3	Sales	Middle	Low	Junior
4	Systems	Young	High	Junior
5	Systems	Middle	High	Senior
6	Systems	Young	High	Junior
7	Systems	Senior	High	Senior
8	Marketing	Middle	High	Senior
9	Marketing	Middle	Average	Junior
10	Secretary	Senior	Average	Senior
11	Secretary	Young	Low	Junior

12+3=15

11 Write short notes on any *three*3x5=15

a) Snowflake Schema, b) Social impacts of data mining, c) Data warehouse architecture, d) Knowledge Discovery in Database, e) Multidimensional Data,f) Optimization of view materialization

OR

Compare OLTP with OLAP systems. Discuss the various OLAP operations in the multidimensional data model. 7+8=15

"END"