Computer Engineering Department, S V N I T, Surat
End-Semester Examinations, November 2017
B Tech IV (CO) – 7th Semester
Course: Data Warehousing and Mining (CO415)

Dated: 30th Nov. 2017 Time: 15:30 hrs to 18:30 hrs Max Marks: 100

Instructions:

1. Write your B. Tech. Admission No. and other details clearly on the answer books while write your B. Tech. Admission No. on the question paper, too.
2. Assume any necessary data but give proper justifications.
3. Be precise and clear in answering the questions.

Q. 1 Answer the following [Any Five]: **[10]**

1. The candidate generation method is a method for generating apriori candidates. Which property we need to check during the whole process of generating the candidates.
2. For each of the following, say if the statement is always true, always false or sometimes true and sometimes false. S is the support and C is the confidence.
   i) Support(a → b) ≥ Support(a → b, c), ii) Confidence (a → b, c) > Confidence (a, b → c)
3. After mining a transaction database for large frequent itemsets, there is only one large frequent itemset of size 4.
   a) Let N be the total number of large itemsets (including the one of size 4). What is the minimal value of N?
   b) If large frequent itemset of size 4 is the closed itemset also. What will be the minimal and maximal value of N for closed itemset?
4. Enlist three different types of associations you can derive from the web data.
5. Enlist different criterions that are considered for the storage space of data warehouse.
6. Justify the statement: MOLAP is faster than the ROLAP.

Q. 2 1. Explain the need of other thresholds/measures where association rules' thresholds cannot be useful. Enlist some situational examples and explain one of the measures which is not affected by null transactions. **[12]**

2. Differentiate among the various data warehouse schemas. **[28]**

Q. 3 Answer the following [Any Four]:

1. Explain ECLAT Approach with its advantages and disadvantages.
2. For the given table, Prepare the expected frequency table and Check the Rules interestingness: Eat(No Cereal)→ Play(Basketball).

	BasketBall	No BasketBall	Sum
Cereal	2000		3750
No Cereal			
Sum	3000		5000

3. Explain Hybrid OLAP server architecture.
4. Explain indexing of OLAP data.
5. Explain the different heuristics of sessionization of web usage data and what data can be derived using those heuristics.

Q. 4 Answer the following [Any Three]: **[24]**

1. It is difficult to assess classification accuracy when individual data objects may belong to more than one class at a time. In such cases, comment on what criteria you would use to compare different classifiers modeled after the same data.
2. The Top Ten Data Mining Algorithms article says about SVM: "The reason why SVM insists on finding the maximum margin hyperplanes is that it offers the best generalization ability." Explain why having wide margins is good!
3. Neural network can be used to solve data mining problems. Identify which problems it is best suited for, identify which problems it has difficulties with, and describe any issues or limitations of the technique.
4. A classification model may change dynamically along with the changes of training data streams. This is known as concept drift. Explain why decision tree induction may not be a suitable method for such dynamically changing data sets. Is naïve Bayesian a better method on such data sets? Comparing with the naïve Bayesian approach, is lazy evaluation (such as the k-NN approach) even better? Explain your reasoning.

Q. 5 Answer the following: **[18]**

1. Suppose that you are to allocate a number of automatic teller machines (ATMs) in a given region so as to satisfy a number of constraints. Households or places of work may be clustered so that typically one ATM is assigned per cluster. The clustering, however, may be constrained by two factors: (1) obstacle objects (i.e., there are bridges, rivers, and highways that can affect ATM accessibility), and (2) additional user-specified constraints, such as each ATM should serve at least 10,000 households. How can a clustering algorithm such as k - means be modified for quality clustering under both constraints?
2. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
   (a) Compute the Euclidean distance between the two objects.
   (b) Compute the Manhattan distance between the two objects.
   (c) Compute the Minkowski distance between the two objects, using p = 3
3. Present conditions under which density-based clustering is more suitable than partitioning-based clustering and hierarchical clustering. Give some sample data sets to support your argument. **OR**
   Discuss the basic difference between the agglomerative and divisive hierarchical clustering algorithms and mention which type of hierarchical clustering algorithm is more commonly used.

Q.6 Subspace clustering is a methodology for finding interesting clusters in high-dimensional space. This methodology can be applied to cluster any kind of data. Outline an efficient algorithm that may extend density connectivity-based clustering for finding clusters of arbitrary shapes in projected dimensions in a high-dimensional data set. **OR** **[08]**

What are the challenges to be solved for stream processing and mining? Why does a solution using an RDBMS only, in general will not give you the required performance facing these challenges? Sketch a solution that would work for stream processing and mining.

***