**Q1. a) What are Spatial Data Structures? Outline their importance in GIS.**

Answer:Spatial Data Structures are designated specifically to store or index special data. The techniques are

* Quad Tree: A Quad tree represents a special object by a hierarchical decomposition of the space into quadrants (cells).
* R-tree: One approach to indexing spatial data represented as MBRs is an R-tree.
* K-D tree: The K-D tree is a variation of a binary search tree where each level in the tree is used to index one of the attributes.

Spatial data are data that have spatial or location components. Spatial data can be viewed as data about objects that themselves are located in a physical space. This may be implemented with a specific location attribute(s) such as address or latitude/ longitude or by a partitioning of the database based on location. Geographic Information Systems (GIS) are used to store information related to geographic locations on the surface of the earth.

**Q1.b) What is metadata? Why do we need metadata when search engines like google seem so effective?**

Answer: Metadata is data about data. Simple search engines retrieve relevant documents usually using a keyword-based retrieval technique similar to those found in traditional IR systems.

**Q1. c) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.**

**Answer:** The various methods for handling the problem of missing values in data tuples include:

(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

(b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

(c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like “Unknown,” or −∞. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common — that of “Unknown.” Hence, although this method is simple, it is not recommended.

(d) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values: For example, suppose that the average income of AllElectronics customers is $28,000. Use this value to replace any missing values for income.

(e) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

(f) Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For example, using the other customer attributes in the data set, we can construct a decision tree to predict the missing values for income.

**Q1. d) With respect to web mining, is it possible to detect visual objects using meta-objects?**

Answer: It is possible to detect visual objects using meta-objects as visual objects are represented by small rectangle that completely contains that object which is minimum bounding rectangle (MBR).

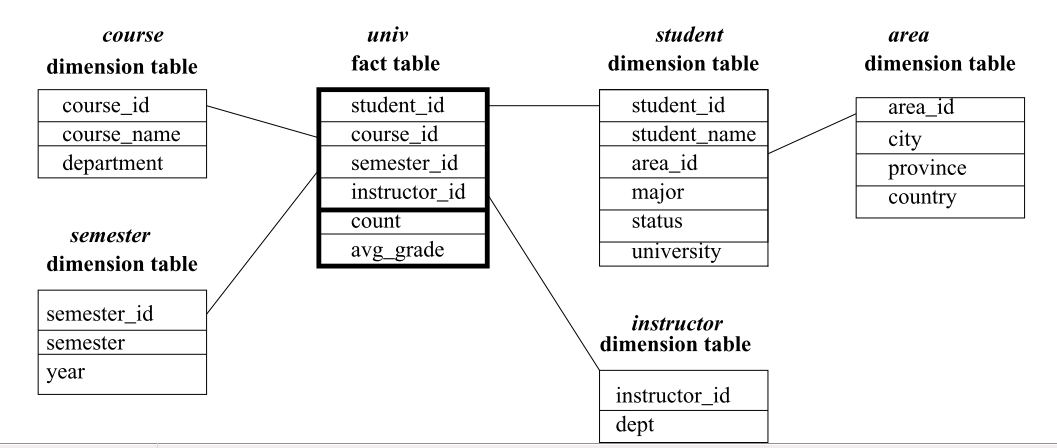
**Q2. a) Suppose that a data warehouse for DB- University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg-grade measure stores the actual course grade of the student. At higher conceptual levels, avg-grade stores the average grade for the given combination.**

**i. Draw a snowflake schema diagram for the data warehouse.**

**ii. Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year ) should one perform in order to list the average grade of CS courses for each Big University student.**

**Answer:**

**i.** A snowflake schema is as shown



**ii.** Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each DB-University student. The specific OLAP operations to be performed are:

• Roll-up on course from course id to department.

• Roll-up on student from student id to university.

• Dice on course, student with department= “CS” and university = “DB-University”.

• Drill-down on student from university to student name.

**Q2. b**) **What is the relationship between data warehousing and data replication? Which form of replication (synchronous or asynchronous) is better suited for data warehousing? Why? Explain with appropriate example.**

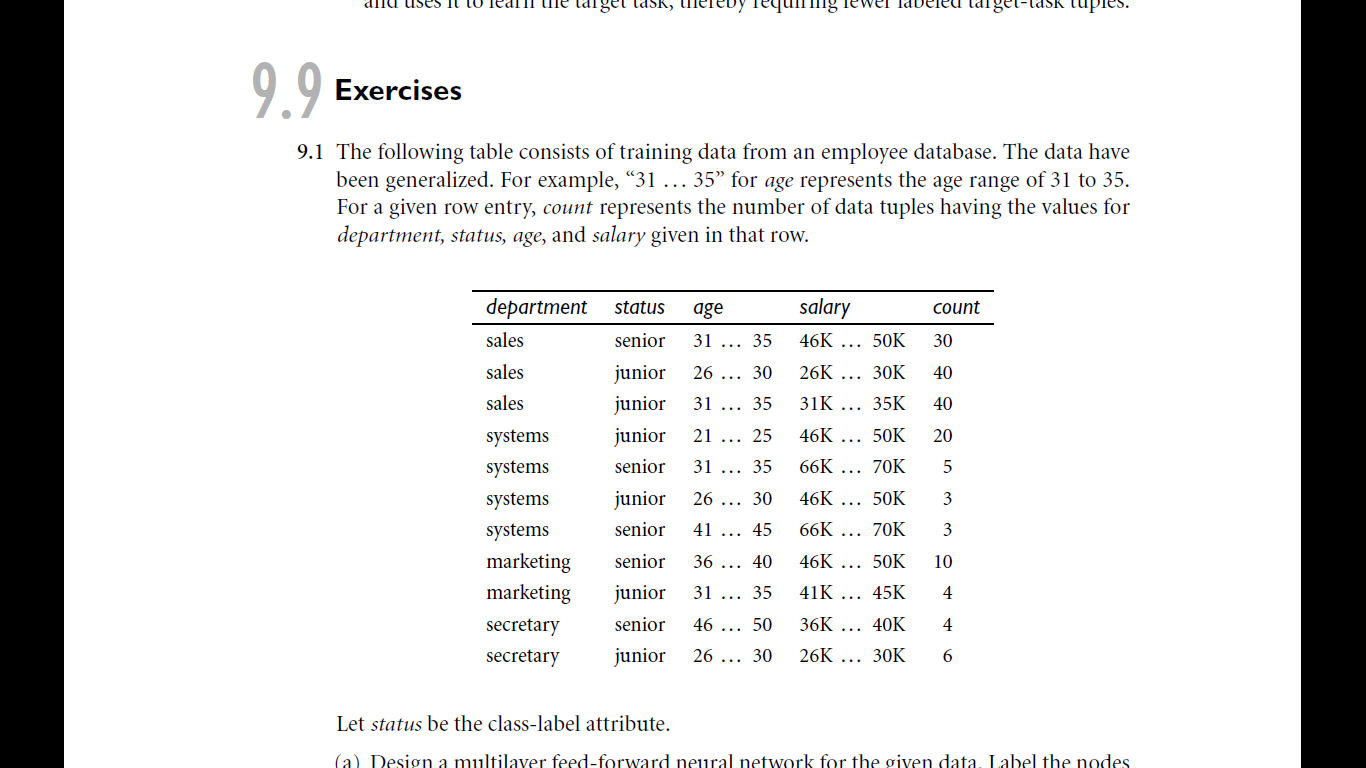
Answer: Data replication is simply a method for creating copies of data in a distributed environment.

Replication technology can be used to capture changes to source data.

* Synchronous replication: Synchronous replication is used for creating replicas in real time. In synchronous replication data is written to primary storage and the replica is done simultaneously. Primary copy and the replica should always remain synchronized.
* Asynchronous replication: It is used for creating time delayed replicas. In asynchronous replication data is written to the primary storage first and then copy data to the replica.

Synchronous replication is best suited for data warehouse as it creates replicas in real time.

**Q3. a) The following table consists of training data from an employee database. The data have been generalized. For example, “31: : : 35” for *age* represents the age range of 31 to 35. For a given row entry, *count* represents the number of data tuples having the values for *department, status, age*, and *salary* given in that row.**

****

**Let *status* be the class label attribute.**

1. **How would you modify the basic decision tree algorithm to take into consideration the *count* of each generalized data tuple (i.e., of each row entry)?**

Answer:

The basic decision tree algorithm should be modified as follows to take into consideration the count of each generalized data tuple.

• The count of each tuple must be integrated into the calculation of the attribute selection measure (such as information gain).

• Take the count into consideration to determine the most common class among the tuples.

**ii.** **Use your algorithm to construct a decision tree from the given data.**

The resulting tree is:

(salary = 26K...30K

: junior

= 31K...35K

: junior

= 36K...40K

: senior

= 41K...45K

: junior

= 46K...50K (department = secretary

: junior

= sales

: senior

= systems

: junior

= marketing

: senior)

= 66K...70K:

senior)

**Q3. b) Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? Given a decision tree, you have the option of (i) converting the decision tree to rules and then pruning the resulting rules, or (ii) pruning the decision tree and then converting the pruned tree to rules. What advantage does (i) have over (ii)?**

Answer: The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers. Tree pruning addresses this issue of overfitting the data by removing the least reliable branches (using statistical measures). This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data.

The drawback of using a separate set of tuples to evaluate pruning is that it may not be representative of the training tuples used to create the original decision tree. If the separate set of tuples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree’s classification accuracy. Furthermore, using a separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree. While this is considered a drawback in machine learning, it may not be so in data mining due to the availability of larger data sets.

If pruning a subtree, we would remove the subtree completely with method (b). However, with method (a), if pruning a rule, we may remove any precondition of it. The latter is less restrictive.

**Q4. a) Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters. A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only (i) The three cluster centers after the first round of execution and (ii) The final three clusters**

Answer: (i) After the first round, the three new clusters are:

(1) {A1},

(2) {B1, A3, B2, B3, C2},

(3) {C1, A2}, and their centers are (1) (2, 10), (2) (6, 6), (3) (1.5, 3.5).

**(ii)** The final three clusters are:

(1) {A1, C2, B1},

(2) {A3, B2, B3},

(3) {C1, A2}

**Q4. b) Briefly outline with example, how to compute the dissimilarity between objects described by the following:**

**i. Nominal Attributes ii. Asymmetric binary attributes**

Answer:

i. Nominal Attributes: Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. Suppose that hair color and marital status are two attributes describing person objects. In our application, possible values for hair color are black, brown, blond, red, auburn, gray, and white. The attribute marital status can take on the values single, married, divorced, and widowed. Both hair color and marital status are nominal attributes. Another example of a nominal attribute is occupation, with the values teacher, dentist, programmer, farmer, and so on.

The dissimilarity computed between objects described by nominal attributes is that the dissimilarity between two objects i and j can be computed based on the ratio of mismatches: d(i, j) = (p – m)/p , where m is the number of matches (i.e., the number of attributes for which i and j are in the same state), and p is the total number of attributes describing the objects.

ii. Asymmetric binary attributes: A binary attribute is asymmetric if the outcomes of the states are not equally important, such as the positive and negative outcomes of a medical test for HIV. By convention, we code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive) and the other by 0 (e.g., HIV negative).

Object dissimilarity can be computed for objects described by nominal attributes and by asymmetric binary attributes are as follows:

Measures of dissimilarity are used in data mining applications such as clustering, outlier analysis, and nearest-neighbour classification. Examples include the Jaccard coefficient for asymmetric binary attributes and Euclidean, Manhattan, Minkowski, and supremum distances for numeric attributes.

**Q5. a) Frequent pattern mining algorithms considers only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transactional data analysis. How can one mine frequent itemsets efficiently considering multiple occurrences of items? Generate Frequent Pattern Tree for the following transaction with 30% minimum support:**

|  |  |
| --- | --- |
| **Transaction ID** | **Items** |
| **T1** | **E, A, D, B** |
| **T2** | **D, A, C, E, B** |
| **T3** | **C, A, B, E** |
| **T4** | **B, A, D** |
| **T5** | **D** |
| **T6** | **D, B** |
| **T7** | **A, D, E** |
| **T8** | **B, C** |

Answer:

Frequent pattern tree:

Minimum support count = (30\* 8)/100 = 2.4

Null ()

|  |  |  |
| --- | --- | --- |
| B | 6 |  |
| D | 6 |  |
| A | 5 |  |
| E | 4 |  |
| C | 3 |  |

B:6

D:4 A:1 C:1 D:2

A:3 E:1 A:1   
 E:2 C:1 E:1

**Q5. b) Differentiate between simple linkage, average linkage and complete linkage algorithms. Use complete linkage algorithm to find the clusters from the following dataset.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **X** | **4** | **8** | **15** | **24** | **24** |
| **Y** | **4** | **4** | **8** | **4** | **12** |

Answer:

Single linkage: Minimum distance is considered

Complete linkage: Maximum distance is considered

Average linkage: Average distance is considered

Calculate Distance Matrix using Euclidean distance formula

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| P1 | 0 |  |  |  |  |
| P2 | 4 | 0 |  |  |  |
| P3 | 11.704 | 8.06 | 0 |  |  |
| P4 | 20 | 16 | 9.848 | 0 |  |
| P5 | 21.54 | 17.888 | 9.848 | 8 | 0 |

Distance matrix after merging P1 and P2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (P1,P2) | P3 | P4 | P5 |
| (P1,P2) | 0 |  |  |  |
| P3 | 8.06 | 0 |  |  |
| P4 | 16 | 9.848 | 0 |  |
| P5 | 17.888 | 9.848 | 8 | 0 |

Distance matrix after merging P4 and P5

|  |  |  |  |
| --- | --- | --- | --- |
|  | (P1,P2) | P3 | (P4,P5) |
| (P1,P2) | 0 |  |  |
| P3 | 8.06 | 0 |  |
| (P4,P5) | 16 | 9.848 |  |

Distance matrix after merging (P1,P2) and P3

|  |  |  |
| --- | --- | --- |
|  | (P1,P2,P3) | (P4,P5) |
| (P1,P2,P3) | 0 |  |
| (P4,P5) | 9.848 | 0 |

P1 P2 P3 P4 P5

**Q6. a) Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.**

Answer: Other dimensions that can be used to assess the quality of data include timeliness, believability, value added, interpretability and accessibility, described as follows:

• Timeliness: Data must be available within a time frame that allows it to be useful for decision making.

• Believability: Data values must be within the range of possible results in order to be useful for decision making.

• Value added: Data must provide additional value in terms of information that oﬀsets the cost of collecting and accessing it.

• Interpretability: Data must not be so complex that the eﬀort to understand the information it provides exceeds the beneﬁt of its analysis.

• Accessibility: Data must be accessible so that the eﬀort to collect it does not exceed the beneﬁt from its use.

**Q6. b) Present an example where data mining is crucial to the success of a business. What data mining functions does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?**

Answer:

A department store, for example, can use data mining to assist with its target marketing mail campaign. Using data mining functions such as association, the store can use the mined strong association rules to determine which products bought by one group of customers are likely to lead to the buying of certain other products. With this information, the store can then mail marketing materials only to those kinds of customers who exhibit a high likelihood of purchasing additional products. Data query processing is used for data or information retrieval and does not have the means for finding association rules. Similarly, simple statistical analysis cannot handle large amounts of data such as those of customer records in a department store.