Government Polytechnic College, Jodhpur
Department of Computer Science (NBA Accredited)

Programme: Diploma **Class Test**: II **Session**: 2017-18
Course: DATA WAREHOUSE AND MINING **Year**: IIIrd
Course CODE: CS-307 **Time:** 14:45 to 15:45
Max.Marks : 15 **Date:** 25-01-2018
Instructions to candidates: Attempt Any Three Questions

Sl#	**Question**	**Marks**	**CO** MAPPING
1	Explain Clustering technique.	5	CO3
2	What are the Characteristics of data warehouse?	5	CO4
3	Explain the technique of Decision Tree.	5	CO3
4.	Explain the difference between Data Warehouse (OLAP) and Database(OLTP).	5	CO4

SOLUTION:

Q1. Explain Clustering technique.

Sol: Clustering is the process of making a group of abstract objects into classes of similar objects.A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Clustering Methods

(1) Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements –

- Each group contains at least one object.
- Each object must belong to exactly one group.

(2) Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here –

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

(3) Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

(4) Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

(5) Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

(6) Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

Q2. What are the Characteristics of data warehouse?

Sol: There are basically four characteristics of data warehouse:

(i) Subject Oriented:

Data are organized according to subject instead of application. The data organized by subject only contain the information necessary for decision support processing. In other words data warehouse are designed to help you analyze data.

(ii) Integrated:

(a) Data that is gathered into the data warehouse from variety of sources and merged into a coherent whole.
(b) Integration is closely related to subject orientation.
(c) Data Warehouses must put data from disparate sources into consistent format.
(d) They must resolve such problems as naming conflicts and inconsistencies among units of measure.
(e) There is no consistency in encoding, naming convention etc.
(f) Heterogeneous data sources.
(g) When data is moved to data warehouse it is converted.

(iii) Non Volatile:

It means data are not updated or changed in any way once they enter the data warehouse , but are only loaded and accessed.

(iv) Time Variant:

All data in data warehouse is identified with a particular time period, to be used for comparisons, trends and forecasting. These data are not updated.

Data warehouse time variance:

-The time horizon for the data warehouse is significantly longer than that of operational systems.
- Operational databases: current value data.
-Data warehouse data: a snapshot taken of at some moment of time.
-The key structure of operational data may or may not contain some element of time. The key structure of data warehouse always contains some element of time.

Q3. Explain the technique of Decision Tree.

Sol: Decision tree is one of the classification technique used in decision support system and machine learning process. A decision tree is a predictive modeling technique that used in classification, clustering and predictive task. Decision tree uses a divide-conquer technique to split the problem search space into subsets. The most important feature of decision tree classifier is their ability to break down a complex decision making process into collection of simpler decision, thus providing solution which is easier to interpret.

A Decision tree is a tree where root and each internal node are labeled with question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration. The constructions of decision tree classifier don't require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery. Decision tree can handle high dimensional data.

Benefits of decision trees in data mining :

- Self-explanatory and easy to follow when compacted.
- Able to handle a variety of input data: nominal, numeric and textual.
- Able to process datasets that may have errors or missing values.
- High predictive performance for a relatively small computational effort .
- Available in many data mining packages over a variety of platforms.
- Useful for various tasks, such as classification, regression, clustering and feature selection.

The construction of the decision tree involves the following three main phases.

A. Construction Phase: The initial decision tree is constructed in this phased, based on the entire training data-set. It requires recursively partitioning the training set into two or more, sub-partition using a splitting criterion, until a stopping criteria is met.

B. Pruning Phase: the tree constructed in the previous phase may not result in the best possible set of rules due to over-fitting. The pruning phase removes some of the lower branches and nodes to improve its performance.

C. Processing the Pruned tree: to improve understandability.

DECISION TREE ALGORITHM

A decision tree (DT) model is a computational model consisting of three parts:

1) A decision tree is defined.

2) An algorithm to create the tree.

3) An algorithm that applies the tree to data and solves the problem under consideration.

```
Algorithm:
Input: T// Decision Tree
       D// Input Database
Output: M// Model Prediction
DT Proc algorithm:
// simplest algorithm to illustrate prediction technique using DT.
For each t in D do
    n = root node of T;
    While n not leaf node do
        Obtain answer to question on n applied to t;
        Indentify arc from t, which contains correct answer;
        n = node at end of this arc;
    Make prediction for t based on label of n;
```

Q4. Explain the difference between Data Warehouse(OLAP) and Database(OLTP).

Sol:

The major differences between OLTP and OLAP system design.

	OLTP System Online Transaction Processing (Operational System)	***OLAP System Online Analytical Processing (Data Warehouse)***
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method