IT6702 DATA WAREHOUSING AND DATA MINING

UNIT-5 CLUSTERING AND APPLICATIONS AND TRENDS IN DATA MINING

1. What do you go for clustering analysis? (Nov/Dec 2011)

Clustering can be used to generate a concept hierarchy for *A* by following either a top down splitting strategy or a bottom-up merging strategy, where each cluster forms a node of the concept hierarchy. In the former, each initial cluster or partition may be further decomposed into several subclusters, forming a lower level of the hierarchy. In the latter, clusters are formed by repeatedly grouping neighboring clusters in order to form higher-level concepts.

2. What are the requirements of cluster analysis? (Nov/Dec 2010)

x Scalability
- x Ability to deal with different types of attributes
- x Discovery of clusters with arbitrary shape
- x Minimal requirements for domain knowledge to determine input parameters
- x Ability to deal with noisy data
- x Incremental clustering and insensitivity to the order of input records
- x High dimensionality
- x Constraint-based clustering
- x Interpretability and usability

3. What is mean by cluster analysis? (April/May 2008)

A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive object.

4. Define CLARANS.

x **CLARANS(Cluster Large Applications based on Randomized Search)** to improve the quality of CLARA we go for CLARANS.
- x It Draws sample with some randomness in each step of search.
- x It overcome the problem of scalability that K-Medoids suffers from.

5. Define BIRCH,ROCK and CURE.

BIRCH(Balanced Iterative Reducing and Clustering Using Hierarchies): Partitions objects hierarchically using tree structures and then refines the clusters using other clustering methods.it defines a clustering feature and an associated tree structure that summarizes a cluster.The tree is a height balanced tree that stores cluster information.BIRCH doesn't Produce spherical Cluster and may produce unintended cluster.

ROCK(RObust Clustering using links): Merges clusters based on their interconnectivity. Great for the looseness of two clusters while emphasizing categorical data. Ignores information about interconnectivity.

CURE(Clustering Using Representatives): Creates clusters by sampling the database and shrinks them toward the center of the cluster by a specified fraction. Obviously better in runtime but lacking in precision.

6. What is meant by web usage mining? (Nov/Dec 2007)(April/May 2008)(Nov/Dec2009) (May/June 2010)

Web usage mining is the process of extracting useful information from server logs i.e. users history. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

7. What is mean by audio data mining? (Nov/Dec 2007)

Audio data mining uses audio signals to indicate the patterns of data or the features of data mining results. Although visual data mining may disclose interesting patterns using graphical displays, it requires users to concentrate on watching patterns and identifying interesting or novel features within them. This can sometimes be quite tiresome. If patterns can be transformed into sound and music, then instead of watching pictures, we can listen to pitches, rhythms, tune, and melody in order to identify anything interesting or unusual. This may relieve some of the burden of visual concentration and be more relaxing than visual mining. Therefore, audio data mining is an interesting complement to visual mining.

8. Define visual data mining. (April/May 2008)

Visual data mining discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization techniques. The human visual system is controlled by the eyes and brain, the latter of which can be thought of as a powerful, highly parallel processing and reasoning engine containing a large knowledge base. Visual data mining essentially combines the power of these components, making it a highly attractive and effective tool for the comprehension of data distributions, patterns, clusters, and outliers in data.

9. What is mean by the frequency item set property? (Nov/Dec 2008)

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset.

10. Mention the advantages of hierarchical clustering. (Nov/Dec 2008)

Hierarchical clustering (or *hierarchic clustering*) outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering. Hierarchical clustering does not require us to prespecify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic. These advantages of hierarchical clustering come at the cost of lower efficiency.

11. Define time series analysis. (May/June 2009)

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series are very frequently plotted via line charts.

12. What is mean by web content mining? (May/June 2009)

Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query.

With the massive amount of information that is available on the World Wide Web, content mining provides the results lists to search engines in order of highest relevance to the keywords in the query.

13. Write down some applications of data mining.(Nov/Dec 2009)

Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Scientific Applications, Intrusion Detection

14. List out the methods for information retrieval. (May/June 2010)

They generally either view the retrieval problem as a document selection problem or as a document ranking problem. In document selection methods, the query is regarded as specifying constraints for selecting relevant documents. A typical method of this category is the Boolean retrieval model, in which a document is

represented by a set of keywords and a user provides a Boolean expression of keywords, such as "car and repair shops," "tea or coffee" .

Document ranking methods use the query to rank all documents in the order of relevance. For ordinary users and exploratory queries, these methods are more appropriate than document selection methods.

15. What is the categorical variable? (Nov/Dec 2010)

A categorical variable is a generalization of the binary variable in that it can take on more than two states. For example, *map color* is a categorical variable that may have, say, five states: *red, yellow, green, pink,* and *blue*. Let the number of states of a categorical variable be *M*. The states can be denoted by letters, symbols, or a set of integers, such as 1, 2, …, *M*. Notice that such integers are used just for data handling and do not represent any specific ordering.

16. What is the difference between row scalability and column scalability? (Nov/Dec 2010)

Data mining has two kinds of scalability issues: row (or database size) scalability and column (or dimension) scalability. A data mining system is considered row scalable if, when the number of rows is enlarged 10 times, it takes no more than 10 times to execute the same data mining queries. A data mining system is considered column scalable if the mining query execution time increases linearly with the number of columns (or attributes or dimensions). Due to the curse of dimensionality, it is much more challenging to make a system column scalable than row scalable.

17. What are the major challenges faced in bringing data mining research to market? (Nov/Dec 2010)

The diversity of data, data mining tasks, and data mining approaches poses many challenging research issues in data mining. The development of efficient and effective data mining methods and systems, the construction of interactive and integrated data mining environments, the design of data mining languages, and the application of data mining techniques to solve large application problems are important tasks for data mining researchers and data mining system and application developers.

18. What is mean by multimedia database? (Nov/Dec 2011)

A multimedia database system stores and manages a large collection of ***multimedia data,*** such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages. Multimedia database systems are increasingly common owing to the popular use of audio, video equipment, digital cameras, CD-ROMs, and the Internet.

19. Define DB miner. (Nov/Dec 2011)

DBMiner delivers business intelligence and performance management applications powered by data mining. With new and insightful business patterns and knowledge revealed by DBMiner. DBMiner Insight solutions are world's first server applications providing powerful and highly scalable association, sequence and differential mining capabilities for Microsoft SQL Server Analysis Services platform, and they also provide market basket, sequence discovery and profit optimization for Microsoft Accelerator for Business Intelligence.

Part B & Part C Questions

1. **How to implement types of data in cluster analysis**
   - Interval-Scaled Variables
   - Binary Variables
   - Categorical, Ordinal, and Ratio-Scaled Variables
   - Variables of Mixed Types
   - Vector Objects
2. **Discuss the Partitioning Methods**
   - Classical Partitioning Methods: *k*-Means and *k*-Medoids
   - Partitioning Methods in Large Databases: From *k*-Medoids to CLARANS
3. **Brief explain about the Hierarchical Methods**
   - Agglomerative and Divisive Hierarchical Clustering
   - BIRCH: Balanced Iterative Reducing and Clustering Using Hierarchies
   - ROCK: A Hierarchical Clustering Algorithm for Categorical Attributes
   - Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling
4. **Explain about Density-Based Methods**
   - DBSCAN: A Density-Based Clustering Method Based on Connected Regions with Sufficiently High Density
   - OPTICS: Ordering Points to Identify the Clustering Structure
   - DENCLUE: Clustering Based on Density Distribution Functions
5. **Explain about the Grid-Based Methods**
   - STING: STatistical INformation Grid
   - WaveCluster: Clustering Using Wavelet Transformation
6. **Discuss about the Model-Based Clustering Methods**
   - Expectation-Maximization
   - Conceptual Clustering
   - Neural Network Approach
7. **Explain about the Clustering High-Dimensional Data**
   - CLIQUE: A Dimension-Growth Subspace Clustering Method
   - PROCLUS: A Dimension-Reduction Subspace Clustering Method
   - Frequent Pattern–Based Clustering Methods
8. **Discuss about the Constraint-Based Cluster Analysis**
   - Clustering with Obstacle Objects
   - User-Constrained Cluster Analysis
   - Semi-Supervised Cluster Analysis
9. **Explain about the Outlier Analysis**
   - Statistical Distribution-Based Outlier Detection
   - Distance-Based Outlier Detection
   - Density-Based Local Outlier Detection
   - Deviation-Based Outlier Detection