CS2032 DATA WAREHOUSING AND DATA MINING

TWO MARKS WITH ANSWER

UNIT-1 DATA WAREHOUSING

1. **What are the uses of multifeature cubes? (Nov/Dec 2007)**

   multifeature cube*s*, which compute complex queries involving multiple dependent aggregates at multiple granularity. These cubes are very useful in practice. Many complex data mining queries can be answered by multifeature cubes without any significant increase in computational cost, in comparison to cube computation for simple queries with standard data cubes.

2. **Compare OLTP and OLAP Systems. (Apr/May 2008), (May/June 2010)**

   If an on-line operational database systems is used for efficient retrieval, efficient storage and management of large amounts of data, then the system is said to be on-line transaction processing.Data warehouse systems serves users (or) knowledge workers in the role of data analysis and decision-making. Such systems can organize and present data in various formats. These systems are known as on-line analytical processing systems.

3. **What is data warehouse metadata? (Apr/May 2008)**

   Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

4. **Explain the differences between star and snowflake schema. (Nov/Dec 2008)**

   The dimension table of the snowflake schema model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

5. **In the context of data warehousing what is data transformation? (May/June 2009)**

   `In *data transformation*, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following:
   Smoothing, Aggregation, Generalization, Normalization, Attribute construction

6. **Define Slice and Dice operation. (May/ June 2009)**

   The slice operation performs a selection on one dimension of the cube resulting in a sub cube.**The dice operation** defines a sub cube by performing a selection on two (or) more dimensions.

7. **List the characteristics of a data ware house. (Nov/Dec 2009)**

   There are four key characteristics which separate the data warehouse from other major operational systems:

   1. Subject Orientation: Data organized by subject
   2. Integration: Consistency of defining parameters
   3. Non-volatility: Stable data storage medium
   4. Time-variance: Timeliness of data and access terms

8. **What are the various sources for data warehouse? (Nov/Dec 2009)**

   Handling of relational and complex types of data*:* Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.
   Mining information from heterogeneous databases and global information systems:

Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases.

9. **What is bitmap indexing? (Nov/Dec 2009)**
   The bitmap indexing method is popular in OLAP products because it allows quick searching in data cubes. The bitmap index is an alternative representation of the *record ID (RID)* list.

10. **What is data warehouse? (May/June 2010)**
    A data warehouse is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making . (or)A data warehouse is a subject-oriented, time-variant and nonvolatile collection of data in support of management's decision-making process.

11. **Differentiate fact table and dimension table. (May/June 2010)**
    Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables.
    A dimension table is used for describing the dimension. (e.g.) A dimension table for item may contain the attributes item_ name, brand and type.

12. **Briefly discuss the schemas for multidimensional databases. (May/June 2010)**
    Stars schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension.
    Snowflakes schema: The snowflake schema is a variant of the star schema model, where some dimension tables are *normalized*, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.
    Fact Constellations: Sophisticated applications may require multiple fact tables to *share* dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation.

13. **How is a data warehouse different from a database? How are they similar? (Nov/Dec 2007, Nov/Dec 2010)**
    Data warehouse is a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision-making. A relational databases is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes(columns or fields) and usually stores a large set of tuples(records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. Both are used to store and manipulate the data.

14. **What is descriptive and predictive data mining? (Nov/Dec 2010)**
    Descriptive data mining, which describes data in a concise and summarative manner and presents interesting general properties of the data.

    predictive data mining, which analyzes data in order to construct one or a set of models and attempts to predict the behavior of new data sets. Predictive data mining, such as classification, regression analysis, and trend analysis.

15. **List out the functions of OLAP servers in the data warehouse architecture. (Nov/Dec 2010)**

    The OLAP server performs multidimensional queries of data and stores the results in its multidimensional storage. It speeds the analysis of fact tables into cubes, stores the cubes until needed, and then quickly returns the data to clients.

16. Differentiate data mining and data warehousing. (Nov/Dec 2011)

data mining refers to *extracting or "mining" knowledge from large amounts of data*. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as *gold* mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data,"

A **data warehouse** is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as *count* or *sales amount*.

17. What do you understand about knowledge discovery? (Nov/Dec 2011)

people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process and an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

UNIT-2 BUSINESS ANALYSIS

1. What is the need for preprocessing the data? (Nov/Dec 2007)

Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses. Incomplete data can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Other data may not be included simply because it was not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted. Furthermore, the recording of the history or modifications to the data may have been overlooked. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

2. What is parallel mining of concept description? (Nov/Dec 2007) (OR) What is concept description? (Apr/May 2008)

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via (1) *data characterization*, by summarizing the data of the class under study (often called the target class) in general terms, or (2) *data discrimination*, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

3. What is dimensionality reduction? (Apr/May 2008)

In *dimensionality reduction*, data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data. If the original data can be *reconstructed* from the compressed data without any loss of information, the data reduction is called lossless.

4. **Mention the various tasks to be accomplished as part of data pre-processing.**

(Nov/ Dec 2008)

1. Data cleaning
2. Data Integration
3. Data Transformation
4. Data reduction

5. **What is data cleaning? (May/June 2009)**

Data cleaning means removing the inconsistent data or noise and collecting necessary information of a collection of interrelated data.

6. **Define Data mining. (Nov/Dec 2008)**

Data mining refers to *extracting or "mining" knowledge from large amounts of data*. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as *gold* mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data,"

7. **What are the types of concept hierarchies? (Nov/Dec 2009)**

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. Concept hierarchies allow specialization, or drilling down ,where by concept values are replaced by lower-level concepts.

8. **List the three important issues that have to be addressed during data integration.**

(May/June 2009) (OR) List the issues to be considered during data integration. (May/June 2010)

There are a number of issues to consider during data integration. ***Schema integration*** and ***object matching*** can be tricky. How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem.

Redundancy is another important issue. An attribute (such as *annual revenue*, for instance) may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

A third important issue in data integration is the ***detection and resolution of data value conflicts*.** For example, for the same real-world entity, attribute values from different sources may differ. This may be due to differences in representation, scaling, or encoding. For instance, a *weight* attribute may be stored in metric units in one system and British imperial units in another.

9. **Write the strategies for data reduction. (May/June 2010)**

1. Data cube aggregation
2. Attribute subset selection
3. Dimensionality reduction
4. Numerosity reduction
5. Discretization and concept hierarchy generation.

10. **Why is it important to have data mining query language? (May/June 2010)**

The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks.

A data mining query language can be used to specify data mining tasks. In particular, we examine how to define data warehouses and data marts in our SQL-based data mining query language, DMQL.

11. **List the five primitives for specifying a data mining task. (Nov/Dec 2010)**

The set of *task-relevant data* to be mined
The *kind of knowledge* to be mined:
The *background knowledge* to be used in the discovery process

The *interestingness measures and thresholds* for pattern evaluation
The expected *representation for visualizing* the discovered pattern

12. What is data generalization? (Nov/Dec 2010)

It is process that abstracts a large set of task-relevant data in a database from a relatively low conceptual levels to higher conceptual levels 2 approaches for Generalization.

1) Data cube approach 2) Attribute-oriented induction approach

13. How concept hierarchies are useful in data mining? (Nov/Dec 2010)

A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute *age*) with higher-level concepts (such as *youth*, *middle-aged*, or *senior*). Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret.

14. How do you clean the data? (Nov/Dec 2011)

Data cleaning (or *data cleansing*) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

For Missing Values

1. Ignore the tuple
2. Fill in the missing value manually
3. Use a global constant to fill in the missing value
4. Use the attribute mean to fill in the missing value:
5. Use the attribute mean for all samples belonging to the same class as the given tuple
6. Use the most probable value to fill in the missing value

For Noisy Data

1. Binning: Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it.
2. Regression: Data can be smoothed by fitting the data to a function, such as with Regression
3. Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or "clusters.

15. What is need of GUI? (Nov/Dec 2011)

Commercial tools can assist in the data transformation step. Data migration tools allow simple transformations to be specified, such as to replace the string *"gender"* by *"sex"*. ETL (extraction/transformation/loading) tools allow users to specify transforms through a graphical user interface (GUI). These tools typically support only a restricted set of transforms so that, often, we may also choose to write custom scripts for this step of the data cleaning process.

UNIT-3 DATA MINING

1. Define frequent set and border set. (Nov/Dec 2007)

A set of items is referred to as an itemset. An itemset that contains *k* items is a *k*-itemset. The set f*computer, antivirus softwar*e 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.Where each variation involves "playing" with the support threshold in a slightly different way.The variations, where nodes indicate an item or itemset that has been examined, and nodes with thick borders indicate that an examined item or itemset is frequent.

2. How are association rules mined from large databases? (Nov/Dec 2007)

Suppose, however, that rather than using a transactional database, sales and related information are stored in a relational database or data warehouse. Such data stores are multidimensional, by definition. For instance, in addition to keeping track of the items purchased in sales transactions, a relational database may record other attributes associated with the items, such as the quantity purchased or the price, or the branch location of the sale. Additional relational information regarding the customers who purchased the items, such as customer age, occupation, credit rating, income, and address, may also be stored.

3. List two interesting measures for association rules. (April/May 2008) (OR)

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule (5.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to uncover interesting statistical correlations between associated items.

4. What are Iceberg queries? (April/May 2008)

It computes an aggregate function over an attribute or set of attributes in order to find aggregate values above some specified threshold.Given relation R with attributes a1,a2,?..,an and b, and an aggregate function,agg_f, an iceberg query is the form.

Select R.a1,R.a2,?..R.an,agg_f(R,b)
From relation R
Group by R.a1,R.a2,?.,R.an
Having agg_f(R.b)>=threhold

5. What is over fitting and what can you do to prevent it? (Nov/Dec 2008)

Tree pruning methods address this problem of *overfitting* the data. Such methods typically use statistical measures to remove the least reliable branches. An unpruned tree and a pruned version of it are shown in following figure. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data (i.e., of previously unseen tuples) than unpruned trees.

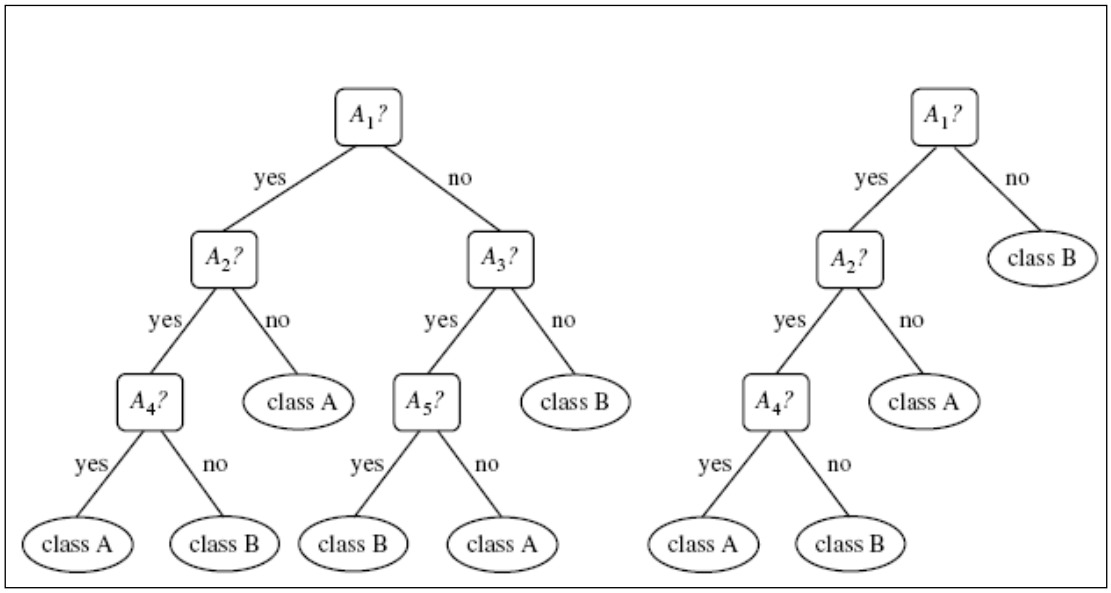

Figure: An unpruned decision tree and a pruned version of it

6. In classification trees, what are surrogate splits, and how are they used? (Nov/Dec 2008)

Decision trees can suffer from *repetition* and *replication*, making them overwhelming to interpret. Repetition occurs when an attribute is repeatedly tested along a given branch of the tree (such as *"age < 60?"* followed by *"age < 45"?* and so on). In replication, duplicate subtrees exist within the tree. These situations can impede the accuracy and comprehensibility of a decision tree. The use of multivariate splits (splits based on a combination of attributes) can prevent these problems.

7. Explain the market basket analysis problem. (May/June 2009)

Market basket analysis, which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence). This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets". The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

8. Give the difference between Boolean association rule and quantitative association rule. (Nov/Dec 2009)

Based on the *types of values* handled in the rule: If a rule involves associations between the presence or absence of items, it is a Boolean association rule. For example, the following three rules are Boolean association rules obtained from market basket analysis.

Computer => *antivirus software* [*support* = 2%; *confidence* = 60%]
buys(*X*, "*computer*") => *buys*(*X*, "*HP printer*")
buys(*X*, "*laptop computer*") => *buys*(*X*, "*HP printer*")

Quantitative association rules involve numeric attributes that have an implicit ordering among values (e.g., *age*). If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule. In these rules, quantitative values for items or attributes are partitioned into intervals. Following rule is considered a quantitative association rule. Note that the quantitative attributes, *age* and *income*, have been discretized.

age(*X*, "30: : :39")^*income*(*X*, "42*K*....48*K*") => *buys*(*X*, "*high resolution TV*")

9. Give the difference between operational database and informational database. (Nov/Dec 2009)

Feature	Operational Database	Informational Database
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
DB design	ER based, application-oriented, current; guaranteed up-to-date	historical; accuracy maintained over time
Access	read/write	mostly read
Function	day-to-day operations	long-term informational requirements, decision support
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)

10. List the techniques to improve the efficiency of Apriori algorithm. (May/June 2010)

- Hash based technique
- Transaction Reduction
- Portioning Sampling
- Dynamic item counting

11. Define support and confidence in Association rule mining.
(May/June 2010) (Nov/Dec 2010)

Support S is the percentage of transactions in D that contain AUB.
Confidence c is the percentage of transactions in D containing A that also contain B.
Support (A=>B)= P(AUB)
Confidence (A=>B)=P(B/A)

12. What is FP growth? (May/June 2010)

FP-growth, which adopts a *divide-and-conquer* strategy as follows. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree,which retains the itemset association information. It then divides the compressed database into a set of *conditional databases* (a special kind of projected database), each associated with one frequent item or "pattern fragment," and mines each such database separately.

13. How Meta rules are useful in constraint based association mining. (May/June 2010)

Metarules allow users to specify the syntactic form of rules that they are interested in mining. The rule forms can be used as constraints to help improve the efficiency of the mining process. Metarules may be based on the analyst's experience, expectations, or intuition regarding the data or may be automatically generated based on the database schema.

14. Mention few approaches to mining Multilevel Association Rules. (Nov/Dec 2010)

Multilevel association rules can be mined using several strategies, based on how minimum support thresholds are defined at each level of abstraction, such as *uniform support*, *reduced support*, and *group-based support*. Redundant multilevel (descendant) association rules can be eliminated if their support and confidence are close to their expected values, based on their corresponding ancestor rules.

15. How rules do help in mining? (Nov/Dec 2011)

Based on the kinds of rules to be mined, categories include mining *association rules* and *correlation rules*. Many efficient and scalable algorithms have been developed for frequent itemset mining, from which association and correlation rules can be derived. These algorithms can be classified into three categories: (1) *Apriori-like algorithms*, (2) *frequent pattern growth*-based algorithms, such as FP-growth, and (3) *algorithms that use the vertical data format*.

16. What is transactional database? (Nov/Dec 2011)

A transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (*trans ID*) and a list of the items making up the transaction (such as items purchased in a store). The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the salesperson and of the branch at which the sale occurred, and so on.

UNIT-4 ASSOCIATION RULE MINING AND CLASSIFICATION

1. What is tree pruning? (Nov/Dec 2007)

Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

2. List the requirements of clustering in data mining. (Nov/Dec 2007)

Mining data streams involves the efficient discovery of general patterns and dynamic changes within stream data. For example, we may like to detect intrusions of a computer network based on the anomaly of message flow, which may be discovered by clustering data streams, dynamic construction of stream models, or comparing the current frequent patterns with that at a certain previous time.

3. What is classification? (April/May 2008) (May/June 2009)

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

4. What is the objective function of the K-means algorithm?

The *k*-means algorithm takes the input parameter, *k*, and partitions a set of *n* objects into *k* clusters so that the resulting intracluster similarity is high but the intercluster similarity is low.Cluster similarity is measured in regard to the *mean* value of the objects in a cluster, which can be viewed as the cluster's *centroid* or *center of gravity*.

First, it randomly selects *k* of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

where *E* is the sum of the square error for all objects in the data set; ***p*** is the point in space representing a given object; and ***mi*** is the mean of cluster *Ci* (both ***p*** and ***mi*** are multidimensional).

5. The naïve Bayes classifier makes what assumption that motivates its name?

Studies comparing classification algorithms have found a simple Bayesian classifier known as the *naïve Bayesian classifier* to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called *class conditional independence*. It is made to simplify the computations involved and, in this sense, is considered "naïve."

6. What is an outlier? (May/June 2009) (OR)
Define outliers. List various outlier detection approaches. (May/June 2010)

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. These can be categorized into four approaches: the *statistical approach*, the *distance-based approach*, the *density-based local outlier approach*, and the *deviation-based approach*.

7. Compare clustering and classification. (Nov/Dec 2009)

Clustering techniques consider data tuples as objects. They partition the objects into groups or *clusters*, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how "close" the objects are in space, based on a distance function. The "quality" of a cluster may be represented by its *diameter*, the maximum distance between any two objects in the cluster.

Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.

8. What is meant by hierarchical clustering? (Nov/Dec 2009)

A hierarchical method creates a hierarchical decomposition of the given set of data objects. A hierarchical method can be classified as being either ***agglomerative*** **or** ***divisive,*** based on how the hierarchical decomposition is formed.

The ***agglomerative approach,*** also called the ***bottom-up*** approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one (the topmost level of the hierarchy), or until a termination condition holds.

The ***divisive approach***, also called the ***top-down*** approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

9. What is Bayesian theorem? (May/June 2010)

Let ***X*** be a data tuple. In Bayesian terms, ***X*** is considered "evidence." As usual, it is described by measurements made on a set of *n* attributes. Let *H* be some hypothesis, such as that the data tuple ***X*** belongs to a specified class *C*. For classification problems, we want to determine *P*(*H*j***X***), the probability that the hypothesis *H* holds given the "evidence" or observed data tuple ***X***. In other words, we are looking for the probability that tuple ***X*** belongs to class *C*, given that we know the attribute description of ***X.***

10. What is Association based classification? (Nov/Dec 2010)

Association-based classification, which classifies documents based on a set of associated, frequently occurring text patterns. Notice that very frequent terms are likely poor discriminators. Thus only those terms that are not very frequent and that have good discriminative power will be used in document classification. Such an association-based classification method proceeds as follows: First, keywords and terms can be extracted by information retrieval and simple association analysis techniques. Second, concept hierarchies of keywords and terms can be obtained using available term classes, such asWordNet, or relying on expert knowledge, or some keyword classification systems.

11. Why tree pruning useful in decision tree induction? (May/June 2010) (Nov/Dec 2010)

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of *overfitting* the data. Such methods typically use statistical measures to remove the least reliable branches.

12. Compare the advantages of and disadvantages of eager classification (e.g., decision tree) versus lazy classification (k-nearest neighbor) (Nov/Dec 2010)

Eager learners, when given a set of training tuples, will construct a generalization (i.e., classification) model before receiving new (e.g., test) tuples to classify. We can think of the learned model as being ready and eager to classify previously unseen tuples.

Imagine a contrasting **lazy approach**, in which the learner instead waits until the last minute before doing any model construction in order to classify a given test tuple. That is, when given a training tuple, a lazy learner simply stores it (or does only a little minor processing) and waits until it is given a test tuple.

13. What is called Bayesian classification? (Nov/Dec 2011)

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, described below. Studies comparing classification algorithms have found a simple Bayesian classifier known as the *naïve Bayesian classifier* to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

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)

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

- **CLARANS(Cluster Large Applications based on Randomized Search)** to improve the quality of CLARA we go for CLARANS.
- It Draws sample with some randomness in each step of search.
- 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 categorical data. Ignores information about the looseness of two clusters while emphasizing 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.

16 MARK QUESTIONS

UNIT-1

1. Write in detail about the architecture and implementation of the data warehouse. (Nov/Dec '07) (OR) Diagrammatically illustrate and discuss the three tier data warehousing architecture. (May/June 2009).

(OR) Write a detailed diagram describe the general architecture of data warehouse. (Nov/Dec 2010). (OR) Describe the data warehouse architecture with a neat diagram. (May/June 2010)

2. **List and discuss the major features of a data warehouse. (May/June 2009) (OR)**
3. **Discuss the various types of warehouse schema with suitable example. (Nov/Dec'09) (OR) What do you understand about database schemas? Explain. (Nov/Dec 2011)**
4. **Describe OLAP operations in multidimensional data model. (Nov/Dec 2011)**
5. **Explain the types of OLAP server in detail. (Nov/Dec 2009)**
6. **Enumerate the building blocks of a data warehouse. Explain the importance of metadata in a data warehouse environment. What are the challenges in metadata management? (Nov/Dec '08).**
7. **Compare and contrast the data warehouse and operational DB with various features.(Nov/Dec 2011). Explain in detail about the different kinds of data on which data mining can be applied. (Nov/Dec '07).**

UNIT-2

1. **Describe the various Reporting and Query Tools and Application.**
2. **What are differences between three main types of data usage: information processing, analytical processing and data mining? Discuss the motivation behind OLAP mining. (Apr/May '08)**

UNIT-3

1. **Describe the architecture of typical data mining system with neat Sketch.**
2. **Explain the Steps of Knowledge Discovery in Databases with neat Sketch.**
3. **Describe the data mining functionality and examine. What kinds of patterns can be mined.**
4. **Explain the classification of Data Mining Systems.**
5. **Describe the various issues in datamining techniques.**
6. **Discuss the various data mining techniques.**
7. **Explain the need and steps involved in data preoprocessing.**
8. **List out the primitives for specifying a data mining task.**
9. **Describe how concepts hierarchies are usuful in data mining.**
10. **What are the various issued addressed during data integration?**

UNIT-4

1. **Describe the various techniques for data preprocessing with examples.**
2. **Explain the various primitives for specifying a data mining task.**
3. **Describe the various descriptive statistical measures for data mining.**
4. **Briefly explain the data mining functionalities and examine,what kinds of patterns can be mined?**
5. **Elaborately explain the discretization and concepts hierarchy generation for numeric data and categorical data.**
6. **Write the algorithm to discover frequent itemsets without candidate generation and explain it with an example.**

7. **Discuss Apriori Algorithm with a suitable example and explain how its efficiency can be improved.**
8. **Discuss mining of multi-level association rules from transactional databases.**
9. **What are classification rules? How is regression related to classification?**
10. **Explain with example the various steps in Decision tree induction.**
11. **State Bayes theorem and discuss how Bayesian classifiers work.**
12. **What back propagation? How does it work?**
13. **Describe the various techniques for improving classifiers accuracy.**

UNIT-5

1. **Discuss the different types of clustering methods.**
2. **Discuss the working of PAM algorithm.**
3. **Describe K-means clustering with an example.**
4. **Explain hierarchical methods of clustering.**
5. **Explain the various methods for detecting outliers.**
6. **Explain the mining of spatial databases.**
7. **Discuss the mining of text data mining.**
8. **What are the salient features of times series data ming?**
9. **What is web mining? Discuss the various web mining techniques.**
10. **Discuss in detail the application of Data mining for financial data analysis?**
11. **Discuss the application of data mining in business.**
12. **Discuss in detail of applications of data mining for biomedical and DNA data analysis and telecommunication industry.**
13. **Discuss the social impacts of data mining systems.**