1. What are the nine decisions in the design of the data warehouse?

1. Choosing the process
2. Choosing the grain
3. Identifying and conforming the dimensions
4. Choosing the facts
5. Storing pre-calculations in the fact table
6. Rounding out the dimension tables
7. Choosing the duration of the database
8. Tracking slowly changing dimensions
9. Deciding the query priorities and the query modes

2. Define Star 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. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

3. What is Data discretization?

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels thereby reduces and simplifies the original data.

4. How is a data warehouse different from a database? How are they similar?

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 database 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.

5. Differentiate between fact table and dimension table.

2129-SJCE

A large central table (fact table) containing the bulk of the data, with no redundancy, and a set of smaller attendant tables (dimension tables), one for each dimension. The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

6. What is data warehouse metadata?

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.

7. Differentiate data mining and data warehousing.

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.

8. List out the functions of OLAP servers in the data warehouse architecture.

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.

9. What is descriptive and predictive data mining?

Descriptive data mining, this describes data in a concise and summative manner and presents interesting general properties of the data. Predictive data mining, this analyses data in order to construct one or a set of models and attempts to predict the behaviour of new data sets. Predictive data mining, such as classification, regression analysis, and trend analysis.

2129-SJCE

10. In the context of data warehousing what is data transformation?

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

11. What is data warehouse metadata?

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.

12. What are the uses of multifeature cubes?

Multifeature cubes, 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.

PART B

1. Explain seven components of Data warehouse architecture with neat diagram.
2. With a neat diagram describe the various stages of building a data warehouse?
3. Explain mapping data warehouse with multi-processor architectures with the concept of parallelism data partitioning.
4. Discuss DBMS Schemas for decision support. Describe performance problems with star schema
5. Discuss Data Extraction, Clean up and transformation tools with meta data management.

2129-SJCE

UNIT II

PART – A

1. List the distinct features of OLTP and OLAP.

Users and system
orientation Data contents
Database
design View
Access patterns

***2.*What is multidimensional data model? Give example.**

Multidimensional data model views data in the form of a data cube. A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts. A multidimensional data model is typically organized around a central theme, like sales, for instance. This theme is represented by a fact table. Facts are numerical measures. Think of themas the quantities by which we want to analyze relationships between dimensions. Examples of facts for a sales data warehouse include dollars sold (sales amount in dollars), units sold (number of units sold), and amount budgeted.

3.List the functions of OLAP servers in the data warehouse architecture.

Relational OLAP (ROLAP) servers: These are the intermediate servers that stand in between a relational back-end server and client front-end tools.
Multidimensional OLAP (MOLAP) servers: These servers support multidimensional views of data through array-based multidimensional storage engines.
Hybrid OLAP (HOLAP) servers: The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP.

4.What are the uses of multi-feature cubes?

Multi-feature cubes, 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 multi-feature cubes without any significant increase in computational cost, in comparison to cube computation for simple queries with standard data cubes.

5.List OLAP guidelines.

1. Multidimensional conceptual view
2. Transparency
3. Accessibility
4. Consistent reporting performance

219-SJCE

5. Client/server architecture
6. Generic dimensionality
7. Dynamic sparse matrix handling
8. Multi-user support
9. Unrestricted cross-dimensional operations
10. Intuitive data manipulation
11. Flexible reporting
12. Unlimited dimensions and aggregation levels

6.Comment on OLAP tools on internet.

OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time.

7.How do you clean the data?

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

13. Ignore the tuple
14. Fill in the missing value manually
15. Use a global constant to fill in the missing value
16. Use the attribute mean to fill in the missing value:
17. Use the attribute mean for all samples belonging to the same class as the given tuple
18. 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.

8.What is need of GUI?

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

212-SJCE

process.

9.How concept hierarchies are useful in data mining?

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.

10.What is data generalization?

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

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

11.Why is it important to have data mining query language?

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.

12.Write the strategies for data reduction.

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

13. List the issues to be considered during data integration.

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

2129-SJCE

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.

14. How concept hierarchies are useful in data mining?

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.

15. Write the strategies for data reduction.

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

PART B

1.Discuss different tool categories in data warehouse business analysis.

2.Descibe about Cognus Impromptu

3.Explain the OLAP operations in multidimensional model?

4.Explain Multidimensional Data Model.

5.Summarize the major differences between OLTP and OLAP system design.

6.Explain different categories of OLAP tools with diagram

2129-SJCE

UNIT III

PART A

1. Why we need data transformation? Mention the ways by which data can be transformed.

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

2. List the five primitives for specification of a data mining task.

- ✓ 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 patterns

3. How concept hierarchies are useful in data mining?

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).

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

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) Data preprocessing is important when mining image data and can include data cleaning, data transformation, and feature extraction.

5. Give an example of outlier analysis for the library management system.

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. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

2129-SJCE

6. What are the different steps in Data transformation?

- ✓ Smoothing
- ✓ Aggregation
- ✓ Generalization
- ✓ Normalization
- ✓ Attribute construction

7. How rules do help in mining?

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.

8. What is transactional database?

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.

9. Mention few approaches to mining Multilevel Association Rules.

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.

10. How Meta rules are useful in constraint based association mining.

2129-SJCE

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.

11. What is FP growth?

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.

12. Define support and confidence in Association rule mining.

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)

13. What Is Data Mining?

Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories. Data mining refers to extracting or "mining" knowledge from large amounts of data. Other terms for data mining are knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Popularly used term, Knowledge Discovery from Data, or KDD.

2129-SJCE

PART B

1.Explain the Evolution of Database System Technology?

2.Explain the Steps of Knowledge Discovery in Databases with neat Sketch?

3.Describe the architecture of typical data mining system with neat Sketch?

4.Explain about the types of data?

5.Describe the data mining functionality and examine. What kinds of patterns can be mined?

6.State and explain the various classification of data mining systems with example.

7.What are the various issued addressed during data integration?

8.Describe with suitable examples, the forms of data pre-processing: data cleaning, data integration, data transformation and data reduction.

9.Explain different strategies of Data Reduction

10. Describe Data Discretization and Concept Hierarchy Generation. State why concept hierarchies are useful in data mining.

UNIT – IV

PART A

1. **What is tree pruning?**

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

2. **What is classification?**

   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).

3. **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*.

2129-SJCE

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).

4. 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."

5. What is an outlier? (OR)
Define outliers. List various outlier detection approaches.

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*.

6. Compare clustering and classification.

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

219-SJCE

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.

7. What is meant by hierarchical clustering?

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.

8. What is Bayesian theorem?

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.***

9. What is Association based classification?

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.

10. Why tree pruning useful in decision tree induction?

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.

2129-SJCE

11. Compare the advantages of and disadvantages of eager classification (e.g., decision tree) versus lazy classification (k-nearest neighbor)

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.

12. What is called Bayesian classification?

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.

14.List the techniques to improve the efficiency of Apriori algorithm.

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

15.What is decision tree? Mention two phases in decision tree induction.

- A decision tree is a graph that uses a branching method to illustrate every possible outcome of a decision. Decision tree software is used in data mining to simplify complex strategic challenges and evaluate the cost-effectiveness of research and business decisions.
- The two phases are Decision tree Construction and Tree Pruing.

PART B

1.Discuss the single dimensional Boolean association rule mining for transaction database?

2129-SJCE SCAD

2.Discuss about constraint based association rule mining with examples and state how association mining to correlation analysis is dealt with?

3.Explain Naive Bayesian classifications with algorithm and sample example?

4.Find all frequent items sets for the given training set using Apriori and FP growth, respectively. Compare the efficiency of the two mining processes

TID items_bought

T100 (M,O,N,K,E,Y)

T200 (D,O,N,K,E,Y)

T300 (M,A,K,E)

T400 (M,U,C,K,Y)

T500 (C,O,O,K,I,E)

5.Explain how support vector machines (SVM) can be used for classification?

6.Explain as to how neural networks are used for classification of data?

7.Explain various attribute selection measure in classification?

2129-SJCE

UNIT V

PART A

1. **What do you go for clustering analysis?**

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?**
   - 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?

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?

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?

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

2020-SJCE

mining is an interesting complement to visual mining.

8. Define visual data mining.

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?

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.

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.

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?

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

2129-SJCE

search engines in order of highest relevance to the keywords in the query.

13. Write down some applications of data mining.

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

14. List out the methods for information retrieval.

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?

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?

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?

The diversity of data, data mining tasks, and data mining approaches

2129-SJCE

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?

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.

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.

20. Give the reason on why clustering is needed in data mining?

Clustering is needed to identify set of similar data objects. The objects are similar in terms of multiple dimensions. By considering only similar data objects for further mining improves the interestingness of retrieved knowledge and accuracy

2129-SJCE

PART B

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?

2129-SJCE