	Question Bank
	Sub Code & Name: IT6702 & Data Warehousing & Data Mining **Name of the faculty : Mr.K.Rajkumar** **Designation & Department : Asst Prof & CSE** **Regulation : 2013** **Year & Semester : III / 06** **Branch : CSE** **Section : A & B**

UNIT-1 DATA WAREHOUSING

Part – A

1. What is data warehouse?

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.

A **data warehouse** is a subject-oriented, time-variant and nonvolatile collection of data in support of management's decision-making process.

2. 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.

3. What is Data mart?

Data mart is a data store that is subsidiary to a data ware house of integrated data. The data mart is directed at a partition of data that is created for the use of a dedicated group of users.

4. 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.

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:

1.Smoothing,2. Aggregation,3.Generalization,4.Normalization,5.Attribute, 6.Construction.

6. List the characteristics of a data warehouse.

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

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

7. What are the various sources for data warehouse?

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.

8. What is bitmap indexing?

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.

9. Differentiate fact table and dimension table.

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.

10. 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

11. How is a data warehouse different from a database? How are they similar? (Nov/Dec 2007, Nov/Dec 2010, May/June 2012)

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.

12. 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.

13. 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*.

14. List out the logical steps needed to build a Data warehouse.

- Collect and analyze business requirements.
- Create a data model and a physical design for the Data warehouse.
- Define data source
- Choose the database technology and platform for the warehouse.
- Extract the data from the operational databases, transform it, clean it up and load it into the database.
- Choose database access and reporting tool.
- Choose database connectivity software.
- Choose data analysis and presentation software.
- Update the data warehouse

15. Write note on shared-nothing architecture.

- The data is partitioned across all disks and the DBMS is partitioned across multiple conservers.
- Each of which resides on individual nodes of the parallel system and has an ownership of its disk and thus it own database partition.
- A shared-nothing RDBMS parallelizes the execution of a SQL query across multiple processing nodes.
- Each processor has its own memory nd disk and communicates with other processors by exchanging messages and data over the interconnection network.

16. What are the access tools groups available?

- Data query and reporting tools
- Application development tools
- Executive information system(EIS) tools
- On-line analytical processing tools
- Data mining tools

17. Write down the applications of data warehousing.

- Financial services
- Banking services
- Customer goods
- Retail sectors
- Controlled manufacturing

18. What are the applications of querying tools?

- Multidimensional analysis
- Decision making
- In-depth analysis such as data classification, clustering

19. List out the two different types of reporting tools.

1. Production reporting tools – companies generate regular operational reports or support high volume batch jobs, such as calculating and printing paychecks.
2. Report writers – are inexpensive desktop tools designed for end users.

20. List the two ways the parallel execution of the tasks with in SQL statements can be done.

- Horizontal parallelism – which means that the DB partitioned across multiple disks and parallel processing with in a specific task.
- Vertical parallelism – which occurs among different tasks – all component query operations (SCAN, JOIN, SORT) are in parallel in a pipelined fashion.

21. What are the technologies included in data warehousing?

- Relational and multi-dimensional database management systems
- Client/server architecture
- Meta data modeling and repositories
- Graphical user interfaces and much more

Part-B

1 With a neat sketch, Describe in detail about Data warehouse architecture
2 List and discuss the steps involved in building a data warehouse.
3 Give detailed information about Meta data in data warehousing.
4 List and discuss the steps involved in mapping the data warehouse to a multiprocessor architecture.
5 i) Explain the role played by sourcing, acquisition, clean up and transformation tools in data warehousing. (May/June 2013)
ii) Explain about STAR Join and STAR Index. 6 Describe in detail about DBMS schemas for decision support.
7 Explain about data extraction, clean up and transformation tools.

8 Explain the following:

i) Implementation considerations in building data warehouse
ii) Database architectures for parallel processing.

UNIT-2 BUSINESS ANALYSIS

PART – A

1. What are production reporting tools? Give examples.

Production reporting tools will let companies generate regular operational reports or support high-volume batch jobs. Such as calculating and printing pay checks.

Examples:

- Third generation languages such as COBOL
- Specialized fourth generation languages such as Information builders, Inc's Focus
- High-end client/server tools such as MITI's SQL.

2. Define data cube.

Data cube consists of a large set of facts or measures and a number of dimensions. Facts are numerical measures that are quantities by which we can analyze the relationship between dimensions. Dimensions are the entities or perspectives with respect to an organization for keeping records and are hierarchical nature.

3. What is a Reporting tool? List out the two different types of reporting tools.

Reporting tools are software applications that make data extracted in a query accessible to the user. That is it used for to generate the various types of reports.

It can be divided into 2 types:

1. Production reporting tools
2. Desktop reporting tools

4. Define OLAP.

OLAP (online analytical processing) is computer processing that enables a user to easily and selectively extract and view data from different points of view.

- OLAP is becoming an architecture that an increasing number of enterprises are implementing to support analytical applications.

5. Briefly discuss the schemas for multidimensional databases.

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.

6. Define the categories of tools in business analysis.

There are 5 categories of tools in business analysis.

i) **Reporting tools** – it can be used to generate the reports.
ii) **Managed query tools** – it can be used to SQL queries for accessing the databases.
iii) **Executive information systems** – It allow developers to build customized, graphical decision support applications or "briefing books".
iv) **On-line analytical processing** – these tools aggregate data along common business subjects or dimensions and then let users navigate the hierarchies and dimensions with the click of a mouse button.
v) **Data mining –** It use a variety of statistical and artificial intelligence algorithm to analyze the correlation of variables in the data and extract interesting patterns and relationship to investigate.

8. List any four tools for performing OLAP. (Nov/Dec 2013)

Arbor Essbase Web

Information advantage web OLAP

Micro strategy DSS web

Brio technology

9. Classify OLAP Tools. (Apr/May 2011)

- MOLAP – Multidimensional Online Analytical Processing
- ROLAP – Multirelational Online Analytical Processing
- MQE – Managed Query Environment

10. Define how the complex aggregation at multiple granularities is achieved using multi-feature cubes? (May/June 2012)

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.

11. Give examples for managed query tools.

IQ software's **IQ objects**
Andyne Computing Ltd's **GQL**
IBM's **Decision server**
Oracle Corp's **Discoverer/2000**

12. What is Apex cuboid?

Apex cuboid or 0-D cuboid which holds the highest level of summarization. The Apex cuboid is typically denoted by all.

13. What is multidimensional database?

Data warehouses and OLAP tools are based on a multidimensional data model. This model is used for the design of corporate data warehouses and department data marts. This model contains a star schema, snowflake schema and fact constellation schemas. The core of multidimensional model is the data cube.

14. What are the applications of query tools?

The applications of query tools are

Multidimensional analysis

Decision making

In-depth analysis such as data classification

Clustering.

15. Compare OLTP and OLAP.

Data Warehouse (OLAP)	Operational Database (OLTP)
Involves historical processing of information.	Involves day-to-day processing.
OLAP systems are used by knowledge	OLTP systems are used by clerks,
workers such as executives, managers and	DBAs, or database professionals.
Useful in analyzing the business.	Useful in running the business.
It focuses on Information out.	It focuses on Data in.
Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
Contains historical data.	Contains current data.
Provides summarized and multidimensional	Provides detailed and flat relational
Number or users is in hundreds.	Number of users is in thousands.
Number of records accessed is in millions.	Number of records accessed is in tens.
Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
Highly flexible.	Provides high performance.

16. List out OLAP operations in multidimensional data model. (May/June 2009)

- **Roll-up** - performs aggregation on a data cube
- **Drill-down** - is the reverse operation of roll-up.
- **Slice and dice** – Slice operation selects one particular dimension from a given cube and provides a new sub-cube. Dice selects two or more dimensions from a given cube and provides a new sub-cube.
- **Pivot (or) rotate -** The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data.

17. Mention the functions of OLAP servers in the data warehousing 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.

18. What is Impromptu?

Impromptu from Cognos Corporation is positioned as an enterprise solution for interactive database reporting that delivers 1 to 100+ seat scalability.

19. Mention some supported databases of Impromptu.

ORACLE
Microsoft SQL Server
SYBASE
Omni SQL Gateway
SYBASE Net Gateway

20. What is enterprise warehouse?

An enterprise warehouse collects all the information's about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers. It contains detailed data as well as summarized data and can range in size from a few giga bytes to hundreds of giga bytes, tera bytes or beyond.

21. Write note on Report writers.

- Report writers are inexpensive desktop tools designed for end users.
- Report writers have graphical interfaces and built-in charting functions; they can pull groups of data from variety of data sources and integrate them in a single report.
- Leading report writers include Crystal Reports, Actuate and Platinum technology, Inc's Info reports.

PART – B

1. Explain in detail about the reporting and query tools.
2. Describe in detail about COGNOS IMPROMTU.
3. Explain the categorization of **OLAP tools** with necessary diagrams.(May/June 2014)
4. i) List and explain the **OLAP operation** in multidimensional data model**.**
   ii) Differentiate between OLTP and OLAP.
5. i)List and discuss the features of Cognos Impromptu.
   ii)List and discuss the basic features data provided by reporting and query tools used for business analysis.
6. i) What is a Multidimensional data model? Explain star schema with an example.
   ii) Write the difference between multi-dimensional OLAP (**MOLAP**) and Multi-relational OLAP (**ROLAP**).

7. Explain the following:

   i) Different schemas for multidimensional databases
   ii) Different schemas for multidimensional databases. OLAP guidelines.

8. i) Write in detail about Managed Query Environment (MQE).

   ii) Explain about how to use OLAP tools on the Internet.

UNIT-3 DATA MINING

PART – A

1. Define data mining. Give some alternative terms of data mining.

Data mining refers to extracting or "mining" knowledge from large amounts of data.

Data mining is a process of discovering interesting knowledge from large amounts of data stored either, in database, data warehouse or other information repositories.

Alternative names are

- Knowledge mining
- Knowledge extraction
- Data/pattern analysis
- Data Archaeology
- Data Dredging

2. What is KDD? What are the steps involved in KDD process?

Knowledge discovery in databases *(*KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

The steps involved in KDD process are

- **Data Cleaning** − In this step, the noise and inconsistent data is removed.
- **Data Integration** − In this step, multiple data sources are combined.
- **Data Selection** − In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** − In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

- **Data Mining** − In this step, intelligent methods are applied in order to extract data patterns.

- **Pattern Evaluation** − In this step, to identify the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge Presentation** − In this step, visualization and knowledgerepresentation techniques are used to present the mined knowledge to the user.

3. What are the various forms of data preprocessing?

- Data cleaning
- Data integration
- Data transformation
- Data reduction

4. State why preprocessing an important issue for data warehousing and data mining?

In real world data tend to be ***incomplete***, ***noisy***and ***inconsistent*** data. So preprocessing is an important issue for data warehousing and data mining.

5. Write the 2 measures of association rule.

- ***Support*** – It means how often X and Y occur together as a percentage of the total transaction.
- ***Confidence*** – It measures how much a particular item is dependent on another.

6. What is descriptive and predictive data mining?

- The ***descriptive data-mining*** model is discover patterns in the data and understands the relationships between attributes represented by the data.

- In contrast, the ***predictive data-mining*** model predicts the future outcomes based on passed records present in the database or with known answers.

7. What is data transformation?

In ***data transformation***, the data are transformed or consolidated into forms appropriate for mining. Strategies for data transformation include the following:

- Smoothing
- Aggregation
- Generalization of the data
- Normalization
- Attribute construction

8. What is data cleaning?

Data cleaning is a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions. Generally data cleaning reduces errors and improves the data quality.

9. List some applications of Data Mining.

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

10. What is pattern evaluation?

This is one of the steps in the KDD process. In this step, the patterns obtained in the data mining stage are converted in to knowledge based on some interestingness measures.

11. List the primitives that specify a data mining tasks.

- Task-relevant data
- Knowledge type to be mined
- Background knowledge
- Pattern interestingness measure
- Visualization of discovered patterns

12. Differentiate between data characterization and discrimination? (Nov/Dec 2013)

Data Characterization	Data Discrimination
Characterization is a summarization of the general characteristics or features of a target class of data.	Discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

13. State why concept hierarchies are useful in data mining. (Nov/Dec 2012)

Concept hierarchies define a sequence of mappings from a set of lower-level concepts to higher-level, more general concepts and can be represented as a set of nodes organized in a tree, in the form of a lattice, or as a partial order. They are useful in data mining because they allow the discovery of knowledge at multiple levels of abstraction and provide the structure on which data can be generalized (rolled-up) or specialized (drilled-down).

14. What do data mining functionalities include? (Apr/May 2011)

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified in to 2 categories: Descriptive and Predictive.

15. What is classification? (May/June 2011)

Classification involves finding rules that partition the data into disjoint groups. The

input for the classification is the training data set, whose class labels are already known. Classification analyzes the training data set and contracts a model based on the class label and aims to assign a class label to the future unlabeled records.

16. Describe challenges to data mining regarding performance issues.

- Efficiency and scalability of data mining algorithms
- Parallel, distributed and incremental mining algorithms

17. What is prediction?

Prediction is used to predict missing or unavailable data values rather than class labels. Prediction refers to both data value prediction and class label prediction.

18. What are outliers?

Data objects which differ significantly from the remaining data objects are referred to as outliers.

19. What are the two steps using in data cleaning as a process?

- Discrepancy detection
- Data transformation

20. List out the issues of data integration.

- Schema integration and object matching
- Redundancy
- Detection and resolution of data value conflict

21. List out data reduction strategies.

- Data cube aggregation
- Attribute subset selection
- Dimensionality reduction
- Numerosity reduction
- Discretization and concept hierarchy generation

PART-B

1. i) Explain various methods of data cleaning in detail.

   ii) How a data mining system can be integrated with a data warehousing? Discuss with example.

2. What is the use of data mining task? What are the basic types of data mining tasks?

3. i) Explain with diagrammatic illustration data mining as a step in the process of knowledge discovery.

   ii) What is evaluation analysis? Give example.

4. i) Explain with diagrammatic illustration data mining as a confluence of multiple disciplines.

   ii) Explain with diagrammatic illustration the primitives for specifying a data mining task.

5. Explain the various data mining issues and functionalities in detail.

6. State and explain the various classifications of data mining systems with examples.

7. i) Describe the various descriptive statistical measures for data mining.

   ii) What are the major issues in data mining? Explain.

8. Write short notes on the various preprocessing tasks.

UNIT – 4 ASSOCIATION RULE MINING AND CLASSIFICATION

PART-A

1. What is decision tree method?

A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and leaf nodes represent classes or class distributions. The top most in a tree is the node.

2. Distinguish between classification and clustering.

classification	Clustering
Supervised learning	Unsupervised learning
Class label of each training sample is provided.	Class label of each training sample is not known
The set of classes are not known in advance.	The number or set of classes to be learned advance.
Learning by example.	Learning by observation.

3. **List out the major strength of decision tree method.**

- Construction of decision tree classifiers does not require any domain knowledge or parameter setting and therefore is appropriate for exploratory knowledge discovery.
- Handle high dimensional data.
- Simple and fast.
- Used for many applications such as medicine, manufacturing, financial analysis, astronomy, etc.
- Basis of several commercial role induction systems.

4. **How do you choose best split while constructing a decision tree.**

By using **attribute selection measures** we choose best split while constructing a decision tree. The measures are

- Information gain
- Gain ratio
- Gini index

5. **What are tree pruning methods?**

Tree pruning use statistical measures to remove the least reliable branches. Pruned trees tend to be smaller and less complex and easier to comprehend. The tree pruning methods are

- Prepruning
- Post pruning
- Pessimistic pruning

6. **With an example explain correlation analysis?**

A correlation measure can be used to augment the support and confidence but also for association rules. This leads to correlation rules of the form

A=> B [support, confidence, correlation]

7. **Give examples for binary and multidimensional association rules.**

- Binary or Single dimensional association rule
  buys(X="computer") => buys(X="HP Printer")
- Multidimensional association rules
  Age(X,"30..39") ∧ income(X,"2000..40000") => buys(X,"LCD TV")

8. **List the 2 interesting measures of an association rule.**

There are 2 interesting measures od an association rule. They are
Support (A=>B)=P(A U B) Confidence (A=>B)=P(B / A)

9. What is a support vector machine?

Decision tree induction algorithms function recursively. First, an attribute must be selected as the root node. In order to create the most efficient (smallest) tree, the root node must effectively split the data. Each split attempts to pare down a set of instances (the actual data) until they all have the same classification. The split is the one that provides what is termed the most information gain.

10. State the need for pruning phase in decision tree construction?

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

11. Define frequent patterns.

Frequent patterns are patterns (such as item sets, subsequences or substructures) that appear in a dataset frequently.

12 . Define Market basket analysis.

A typical example of frequent item set mining is market basket analysis. This process analyzes customer buying habits by fining associations between the different items that customers place in their shopping baskets.

13. What is STRONG?

The rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong.

14. Define Apriori property.

All nonempty subsets of a frequent itemset must also be frequent.

15. What is Anti-monotone?

If a set cannot pass a test, all of its subsets will fail the same test as well.

16. Define Association rule.

It searches for interesting relationships among items in a given data set.

17. What are the constraints can include the constraint-based association mining?

- Knowledge type constraints
- Data constraints
- Dimension/level constraints
- Interestingness constraints
- Rule constraints

18. What are the steps involved in data classification?

Data classification is a 2 step process. They are

1) Learning step 2) Classification step.

19. What are the preprocessing steps can be used in classification or prediction process.

- Data cleaning
- Relevance analysis
- Data transformation and reduction

20. What is ID3?

It is a decision tree algorithm used by decision tree induction.

21. List some popular attribute selection measures.

- Information gain
- Gain ratio
- Gini index

22. Define Bayesian Belief networks.

- It specifies joint conditional probability distributions they allow class conditional independencies to be defined between subsets of variables.
- A belief network is defined by 2 components:

  - Directed acyclic graph
  - Set of conditional probability tables.

PART-B

1. Write and explain the algorithm for mining frequent item sets **with** candidate generation. Give relevant example.
2. Write and explain the algorithm for mining frequent item sets **without** candidate generation. Give relevant example.
3. Discuss the approaches foe mining **multi-level** and **multi-dimensional association rules** from the transactional databases. Give relevant example.
4. i) Explain the algorithm for constructing a decision tree from training samples.
   ii) Explain about **Bayes Theorem**.

5. .Explain the following:

   i) Constraint-based association mining

   ii) Classification by backpropagation

6 i) Apply the **Apriori algorithm** for discovering frequent item sets of the following. Use 0.3 for minimum support value.

TID	Items purchased
101	milk,bread,eggs
102	milk,juice
103	juice,butter
104	milk,bread,eggs
105	coffee,eggs
106	coffee
107	coffee,juice
108	milk,bread,cookies,eggs
109	cookies,butter
110	milk,bread

ii) Write short notes on: **Prediction.**

7. i) Explain the 2 steps for data classification.
   ii) Explain about Bayesian classification.

8. Describe in detail about classification by decision induction.

9. Explain the following:
   i) Rule-based classification
   ii)Lasy learners.

10. Explain the following:
   i) Support vector machines
   ii) Associative classification

UNIT-5 CLUSTERING AND APPLICATIONS AND TRENDS IN DATA MINING

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 sub clusters, 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: Outlier Analysis.

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. The analysis of outliers data is referred to as outlier analysis.

5. 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 overcomes the problem of scalability that K-Medoids suffers from.

6. 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.

7. 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.

8. 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.

9. 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

10. 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.

11. 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.

12. 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.

13. 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.

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

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

15. 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" .

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. Define: Dendrogram.

1. A tree structure called a dentrogram is commonly used to represent the process of hierarchical clustering.
2. Decompose data objects into a several levels of nested partitioning (tree of clusters) called a dendrogram.

PART – B

1. i) What is cluster analysis? Explain about requirements of clustering in data mining.

   ii) Explain about data mining applications.
2. Describe in detail about types of data in cluster analysis.

3. i) Write note on: categorization of major clustering methods.
ii) Explain the following clustering methods in detail:

i) BIRCH

ii) CURE

4. Explain in detail about partitioning methods.
5. Describe in detail about hierarchical methods.
6. Explain the following:

i) Density based - methods
ii) Constraint based cluster analysis

7. Explain the following:

i) Grid based methods
ii) Model based clustering methods

8. Explain about clustering high dimensional data
9. Explain about outlier analysis.