CS1004: DATA WAREHOUSING AND MINING

TWO MARKS QUESTIONS AND ANSWERS
Unit I

1.Define Data mining.
　　　　It refers to extracting or "mining" knowledge from large amount 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

2.Give some alternative terms for data mining.
- Knowledge mining
- Knowledge extraction
- Data/pattern analysis.
- Data Archaeology
- Data dredging

3.What is KDD.

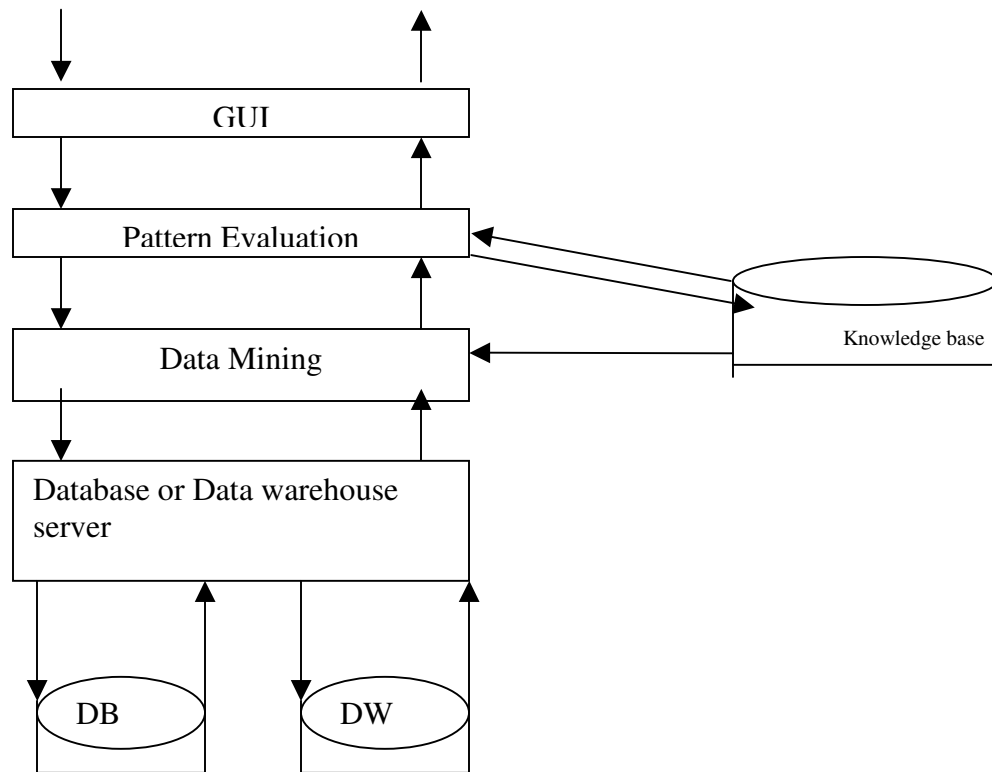　　　KDD-Knowledge Discovery in Databases.

4.What are the steps involved in KDD process.
- Data cleaning
- Data Mining
- Pattern Evaluation
- Knowledge Presentation
- Data Integration
- Data Selection
- Data Transformation

5.What is the use of the knowledge base?
　　　　Knowledge base is domain knowledge that is used to guide search or evaluate the interestingness of resulting pattern. Such knowledge can include concept hierarchies used to organize attribute /attribute values in to different levels of abstraction.

6.Arcitecture of a typical data mining system.

```
        │                    ▲
        ▼                    │
┌────────────────────────────────┐
│             GUI                │
└────────────────────────────────┘
        │                    ▲
        ▼                    │
┌────────────────────────────────┐
│      Pattern Evaluation        │◄──────────────┐
└────────────────────────────────┘               ╲
        │                    ▲              ┌──────────────┐
        ▼                    │              │              │
┌────────────────────────────────┐◄─────────│ Knowledge base│
│         Data Mining            │          └──────────────┘
└────────────────────────────────┘
        │                    ▲
        ▼                    │
┌────────────────────────────────┐
│  Database or Data warehouse    │
│  server                        │
└────────────────────────────────┘
    │        ▲        │        ▲
    ▼        │        ▼        │
  ( DB )            ( DW )
```

7.Mention some of the data mining techniques.
- Statistics
- Machine learning
- Decision Tree
- Hidden markov models
- Artificial Intelligence
- Genetic Algorithm
- Meta learning

8.Give few statistical techniques.
- Point Estimation
- Data Summarization
- Bayesian Techniques
- Testing Hypothesis
- Correlation
- Regression

9.What is meta learning.
Concept of combining the predictions made from multiple models of data mining and analyzing those predictions to formulate a new and previously unknown prediction.

10.Define Genetic algorithm.
- Search algorithm.
- Enables us to locate optimal binary string by processing an initial random population of binary strings by performing operations such as artificial mutation , crossover and selection.

11.What is the purpose of Data mining Technique?
It provides a way to use various data mining tasks.

12.Define Predictive model.
It is used to predict the values of data by making use of known results from a different set of sample data.

13.Data mining tasks that are belongs to predictive model
- Classification
- Regression
- Time series analysis
-

14.Define descriptive model
- It is used to determine the patterns and relationships in a sample data. Data mining tasks that belongs to descriptive model:
- Clustering
- Summarization
- Association rules
- Sequence discovery

15. Define the term summarization
The summarization of a large chunk of data contained in a web page or a document.
Summarization = caharcterization=generalization

16. List out the advanced database systems.
- Extended-relational databases
- Object-oriented databases
- Deductive databases
- Spatial databases
- Temporal databases
- Multimedia databases
- Active databases
- Scientific databases
- Knowledge databases

17. Define cluster analysis

Cluster analyses data objects without consulting a known class label. The class labels are not present in the training data simply because they are not known to begin with.

18.Classifications of Data mining systems.

- Based on the kinds of databases mined:
    - According to model
        - Relational mining system
        - Transactional mining system
        - Object-oriented mining system
        - Object-Relational mining system
        - Data warehouse mining system
    - Types of Data
        - Spatial data mining system
        - Time series data mining system
        - Text data mining system
        - Multimedia data mining system
- Based on kinds of Knowledge mined
    - According to functionalities
        - Characterization
        - Discrimination
        - Association
        - Classification
        - Clustering
        - Outlier analysis
        - Evolution analysis
    - According to levels of abstraction of the knowledge mined
        - Generalized knowledge (High level of abstraction)
        - Primitive-level knowledge (Raw data level)
    - According to mine data regularities versus mine data irregularities
- Based on kinds of techniques utilized
    - According to user interaction
        - Autonomous systems
        - Interactive exploratory system
        - Query-driven systems
    - According to methods of data analysis
        - Database-oriented
        - Data warehouse-oriented
        - Machine learning
        - Statistics
        - Visualization
        - Pattern recognition
        - Neural networks
- Based on applications adopted
    - Finance

          o Telecommunication
          o DNA
          o Stock markets
          o E-mail and so on
          o

19.Describe challenges to data mining regarding data mining methodology and user interaction issues.

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualization of data mining results
- Handling noisy or incomplete data
- Pattern evaluation

20.Describe challenges to data mining regarding performance issues.
- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms

21.Describe issues relating to the diversity of database types.

- Handling of relational and complex types of data
- Mining information from heterogeneous databases and global information systems

22.What is meant by pattern?

       Pattern represents knowledge if it is easily understood by humans; valid on test data with some degree of certainty; and potentially useful, novel,or validates a hunch about which the used was curious. Measures of pattern interestingness, either objective or subjective, can be used to guide the discovery process.

23.How is a data warehouse different from a database?

       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.
       Database consists of a collection of interrelated data.

UNIT II

1. **Define Association Rule Mining.**

Association rule mining searches for interesting relationships among items in a given data set

2. When we can say the association rules are interesting?

Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Users or domain experts can set such thresholds.

3. Explain Association rule in mathematical notations.

Let I-{i1,i2,…..,im} be a set of items
Let D, the task relevant data be a set of database transaction T is a set of items
An association rule is an implication of the form A=>B where A C I, B C I, and An B=ф. The rule A=>B contains in the transaction set D with support s, where s is the percentage of transactions in D that contain AUB. The Rule A=> B has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contain B.

4. 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)

5. How are association rules mined from large databases?

- I step: Find all frequent item sets:
- II step: Generate strong association rules from frequent item sets

6. Describe the different classifications of Association rule mining.

- Based on types of values handled in the Rule
    - i. Boolean association rule
    - ii. Quantitative association rule
- Based on the dimensions of data involved
    - i. Single dimensional association rule
    - ii. Multidimensional association rule
- Based on the levels of abstraction involved
    - i. Multilevel association rule
    - ii. Single level association rule
- Based on various extensions
    - i. Correlation analysis

ii. Mining max patterns

7. What is the purpose of Apriori Algorithm?

Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

8. Define anti-monotone property.

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

9. How to generate association rules from frequent item sets?

Association rules can be generated as follows
For each frequent item set1, generate all non empty subsets of 1.
For every non empty subsets s of 1, output the rule "S=>(1-s)"if

$$\frac{\text{Support count(1)}}{\text{Support\_count(s)}} = \text{min\_conf,}$$

Where min_conf is the minimum confidence threshold.

10. Give few techniques to improve the efficiency of Apriori algorithm.

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

11. What are the things suffering the performance of Apriori candidate generation technique.

- Need to generate a huge number of candidate sets
- Need to repeatedly scan the scan the database and check a large set of candidates by pattern matching

12. Describe the method of generating frequent item sets without candidate generation.

Frequent-pattern growth(or FP Growth) adopts divide-and-conquer strategy.
Steps:

Compress the database representing frequent items into a frequent pattern tree or FP tree
Divide the compressed database into a set of conditional database
Mine each conditional database separately

13. Define Iceberg query.

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

14. Mention few approaches to mining Multilevel Association Rules

- Uniform minimum support for all levels(or uniform support)
- Using reduced minimum support at lower levels(or reduced support)
- Level-by-level independent
- Level-cross filtering by single item
- Level-cross filtering by k-item set

15. What are multidimensional association rules?

Association rules that involve two or more dimensions or predicates
- Interdimension association rule: Multidimensional association rule with no repeated predicate or dimension
- Hybrid-dimension association rule: Multidimensional association rule with multiple occurrences of some predicates or dimensions.

16. Define constraint-Based Association Mining.

Mining is performed under the guidance of various kinds of constraints provided by the user.
The constraints include the following
- Knowledge type constraints
- Data constraints
- Dimension/level constraints
- Interestingness constraints
- Rule constraints.

17. Define the concept of classification.

Two step process

- A model is built describing a predefined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes.
- The model is used for classification.

18. What is Decision tree?

A decision tree is a flow chart like tree structures, 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 root node.

19. What is Attribute Selection Measure?

The information Gain measure is used to select the test attribute at each node in the decision tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split.

20. Describe Tree pruning methods.

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outlier. Tree pruning methods address this problem of over fitting the data.
Approaches:
- Pre pruning
- Post pruning

21. Define Pre Pruning

A tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples.

22. Define Post Pruning.

Post pruning removes branches from a "Fully grown" tree. A tree node is pruned by removing its branches.
Eg: Cost Complexity Algorithm

23. What is meant by Pattern?

Pattern represents the knowledge.

24. Define the concept of prediction.

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample or to assess the value or value ranges of an attribute that a given sample is likely to have.

Unit III

1.Define Clustering?

Clustering is a process of grouping the physical or conceptual data object into clusters.

2. What do you mean by Cluster Analysis?
A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive objects.

3. What are the fields in which clustering techniques are used?
- Clustering is used in biology to develop new plants and animal taxonomies.
- Clustering is used in business to enable marketers to develop new distinct groups of their customers and characterize the customer group on basis of purchasing.
- Clustering is used in the identification of groups of automobiles Insurance policy customer.
- Clustering is used in the identification of groups of house in a city on the basis of house type, their cost and geographical location.
- Clustering is used to classify the document on the web for information discovery.

4.What are the requirements of cluster analysis?
The basic requirements of cluster analysis are
- Dealing with different types of attributes.
- Dealing with noisy data.
- Constraints on clustering.
- Dealing with arbitrary shapes.
- High dimensionality
- Ordering of input data
- Interpretability and usability
- Determining input parameter and
- Scalability

5.What are the different types of data used for cluster analysis?
The different types of data used for cluster analysis are interval scaled, binary, nominal, ordinal and ratio scaled data.

6. What are interval scaled variables?

Interval scaled variables are continuous measurements of linear scale. For example, height and weight, weather temperature or coordinates for any cluster. These measurements can be calculated using Euclidean distance or Minkowski distance.

7. Define Binary variables? And what are the two types of binary variables?

Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights.

8. Define nominal, ordinal and ratio scaled variables?

A nominal variable is a generalization of the binary variable. Nominal variable has more than two states, For example, a nominal variable, color consists of four states, red, green, yellow, or black. In Nominal variables the total number of states is N and it is denoted by letters, symbols or integers.

An ordinal variable also has more than two states but all these states are ordered in a meaningful sequence.

A ratio scaled variable makes positive measurements on a non-linear scale, such as exponential scale, using the formula

$$Ae^{Bt} \quad \text{or} \quad Ae^{-Bt}$$

Where A and B are constants.

9. What do u mean by partitioning method?

In partitioning method a partitioning algorithm arranges all the objects into various partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster. The two types of partitioning method are k-means and k-medoids.

10. Define CLARA and CLARANS?

Clustering in LARge Applications is called as CLARA. The efficiency of CLARA depends upon the size of the representative data set. CLARA does not work properly if any representative data set from the selected representative data sets does not find best k-medoids.

To recover this drawback a new algorithm, Clustering Large Applications based upon RANdomized search (CLARANS) is introduced. The CLARANS works like CLARA, the only difference between CLARA and CLARANS is the clustering process that is done after selecting the representative data sets.

11. What is Hierarchical method?

Hierarchical method groups all the objects into a tree of clusters that are arranged in a hierarchical order. This method works on bottom-up or top-down approaches.

12. Differentiate Agglomerative and Divisive Hierarchical Clustering?

Agglomerative Hierarchical clustering method works on the bottom-up approach. In Agglomerative hierarchical method, each object creates its own clusters. The single Clusters are merged to make larger clusters and the process of merging continues until all the singular clusters are merged into one big cluster that consists of all the objects.

Divisive Hierarchical clustering method works on the top-down approach. In this method all the objects are arranged within a big singular cluster and the large cluster is continuously divided into smaller clusters until each cluster has a single object.

13. What is CURE?

Clustering Using Representatives is called as CURE. The clustering algorithms generally work on spherical and similar size clusters. CURE overcomes the problem of spherical and similar size cluster and is more robust with respect to outliers.

14. Define Chameleon method?

Chameleon is another hierarchical clustering method that uses dynamic modeling. Chameleon is introduced to recover the drawbacks of CURE method. In this method two clusters are merged, if the interconnectivity between two clusters is greater than the interconnectivity between the objects within a cluster.

15. Define Density based method?

Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.

16. What is a DBSCAN?

Density Based Spatial Clustering of Application Noise is called as DBSCAN. DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. DBSCAN defines the cluster as a maximal set of density connected points.

17. What do you mean by Grid Based Method?

In this method objects are represented by the multi resolution grid data structure. All the objects are quantized into a finite number of cells and the collection of cells build the grid structure of objects. The clustering operations are performed on that grid structure. This method is widely used because its processing time is very fast and that is independent of number of objects.

18. What is a STING?

Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

19. Define Wave Cluster?

It is a grid based multi resolution clustering method. In this method all the objects are represented by a multidimensional grid structure and a wavelet transformation is

applied for finding the dense region. Each grid cell contains the information of the group of objects that map into a cell. A wavelet transformation is a process of signaling that produces the signal of various frequency sub bands.

20. What is Model based method?

For optimizing a fit between a given data set and a mathematical model based methods are used. This method uses an assumption that the data are distributed by probability distributions. There are two basic approaches in this method that are
1. Statistical Approach
2. Neural Network Approach.

21. What is the use of Regression?

Regression can be used to solve the classification problems but it can also be used for applications such as forecasting. Regression can be performed using many different types of techniques; in actually regression takes a set of data and fits the data to a formula.

22. What are the reasons for not using the linear regression model to estimate the output data?

There are many reasons for that, One is that the data do not fit a linear model, It is possible however that the data generally do actually represent a linear model, but the linear model generated is poor because noise or outliers exist in the data.
Noise is erroneous data and outliers are data values that are exceptions to the usual and expected data.

23. What are the two approaches used by regression to perform classification?

Regression can be used to perform classification using the following approaches
1. **Division:** The data are divided into regions based on class.
2. **Prediction:** Formulas are generated to predict the output class value.

24. What do u mean by logistic regression?

Instead of fitting a data into a straight line logistic regression uses a logistic curve. The formula for the univariate logistic curve is

$$P = \frac{e^{(C0+C1X1)}}{1+e^{(C0+C1X1)}}$$

The logistic curve gives a value between 0 and 1 so it can be interpreted as the probability of class membership.

25. What is Time Series Analysis?

A time series is a set of attribute values over a period of time. Time Series Analysis may be viewed as finding patterns in the data and predicting future values.

26. What are the various detected patterns?

Detected patterns may include:

- *Trends :* It may be viewed as systematic non-repetitive changes to the values over time.
- *Cycles :* The observed behavior is cyclic.
- *Seasonal :* The detected patterns may be based on time of year or month or day.
- *Outliers :* To assist in pattern detection , techniques may be needed to remove or reduce the impact of outliers.

27. What is Smoothing?

Smoothing is an approach that is used to remove the nonsystematic behaviors found in time series. It usually takes the form of finding moving averages of attribute values. It is used to filter out noise and outliers.

28. Give the formula for Pearson's *r*

One standard formula to measure correlation is the correlation coefficient *r*, sometimes called Pearson's *r*. Given two time series, X and Y with means X' and Y', each with n elements, the formula for *r* is

$$\frac{\Sigma\, (x_i - X')\, (y_i - Y')}{(\Sigma\, (x_i - X')^2\, \Sigma(y_i - Y')^2)^{1/2}}$$

29. What is Auto regression?

Auto regression is a method of predicting a future time series value by looking at previous values. Given a time series $X = (x_1, x_2, \ldots x_n)$ a future value, x_{n+1}, can be found using

$$x_{n+1} = \xi + \varphi_n x_n + \varphi_{n-1} x_{n-1} + \ldots\ldots + \varepsilon_{n+1}$$

Here ε_{n+1} represents a random error, at time n+1. In addition, each element in the time series can be viewed as a combination of a random error and a linear combination of previous values.

UNIT-IV

1. Define 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 .

(or)

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 operational databases?

Organizations maintain large database that are updated by daily transactions are called operational databases.

3.Define OLTP?

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.

4.Define OLAP?

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.

5.How a database design is represented in OLTP systems?

Entity-relation model

6. How a database design is represented in OLAP systems?

Star schema
Snowflake schema
Fact constellation schema

7.Write short notes on multidimensional data model?

Data warehouses and OLTP 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 the multidimensional model is the data cube.

8.Define data cube?

It consists of a large set of facts (or) measures and a number of dimensions.

9.What are facts?

Facts are numerical measures. Facts can also be considered as quantities by which we can analyze the relationship between dimensions.

10.What are dimensions?

Dimensions are the entities (or) perspectives with respect to an organization for keeping records and are hierarchical in nature.

11.Define dimension table?

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.Define fact table?

Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables.

13.What are lattice of cuboids?

In data warehousing research literature, a cube can also be called as cuboids. For different (or) set of dimensions, we can construct a lattice of cuboids, each showing the data at different level. The lattice of cuboids is also referred to as data cube.

14.What is apex cuboid?

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

15.List out the components of star schema?

> A large central table (fact table) containing the bulk of data with no redundancy.
> A set of smaller attendant tables (dimension tables), one for each dimension.

16.What is snowflake schema?

The snowflake schema is a variant of the star schema model, where some dimension tables are normalized thereby further splitting the tables in to additional tables.

17.List out the components of fact constellation schema?

This requires multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars and hence it is known as galaxy schema (or) fact constellation schema.

18.Point out the major difference between the star schema and the snowflake schema?

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.

19.Which is popular in the data warehouse design, star schema model (or) snowflake schema model?

Star schema model, because the snowflake structure can reduce the effectiveness and more joins will be needed to execute a query.

20.Define concept hierarchy?

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level concepts.

21.Define total order?

If the attributes of a dimension which forms a concept hierarchy such as "street<city< province_or_state <country", then it is said to be total order.
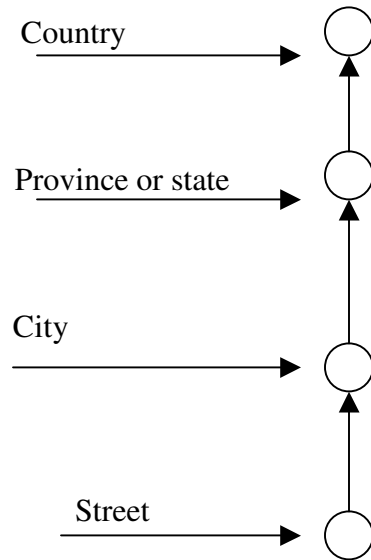
Fig: Partial order for location

22. Define partial order?

If the attributes of a dimension which forms a lattice such as "day<{month<quarter; week}<year, then it is said to be partial order.

23. Define schema hierarchy?

A concept hierarchy that is a total (or) partial order among attributes in a database schema is called a schema hierarchy.

24. List out the OLAP operations in multidimensional data model?

- Roll-up
- Drill-down
- Slice and dice
- Pivot (or) rotate

25. What is roll-up operation?

The roll-up operation is also called drill-up operation which performs aggregation on a data cube either by climbing up a concept hierarchy for a dimension (or) by dimension reduction.

26. What is drill-down operation?

Drill-down is the reverse of roll-up operation. It navigates from less detailed data to more detailed data. Drill-down operation can be taken place by stepping down a concept hierarchy for a dimension.

27.What is slice operation?

The slice operation performs a selection on one dimension of the cube resulting in a sub cube.

28.What is dice operation?

The dice operation defines a sub cube by performing a selection on two (or) more dimensions.

29.What is pivot operation?

This is a visualization operation that rotates the data axes in an alternative presentation of the data.

30.List out the views in the design of a data warehouse?
- Top-down view
- Data source view
- Data warehouse view
- Business query view

31.What are the methods for developing large software systems?
- Waterfall method
- Spiral method

32.How the operation is performed in waterfall method?

The waterfall method performs a structured and systematic analysis at each step before proceeding to the next, which is like a waterfall falling from one step to the next.

33.How the operation is performed in spiral method?

The spiral method involves the rapid generation of increasingly functional systems, with short intervals between successive releases. This is considered as a good choice for the data warehouse development especially for data marts, because the turn around time is short, modifications can be done quickly and new designs and technologies can be adapted in a timely manner.

34.List out the steps of the data warehouse design process?
- Choose a business process to model.
- Choose the grain of the business process
- Choose the dimensions that will apply to each fact table record.
- Choose the measures that will populate each fact table record.

35.Define ROLAP?

The ROLAP model is an extended relational DBMS that maps operations on multidimensional data to standard relational operations.

36.Define MOLAP?

The MOLAP model is a special purpose server that directly implements multidimensional data and operations.

37.Define HOLAP?
The hybrid OLAP approach combines ROLAP and MOLAP technology, benefiting from the greater scalability of ROLAP and the faster computation of MOLAP,(i.e.) a HOLAP server may allow large volumes of detail data to be stored in a relational database, while aggregations are kept in a separate MOLAP store.

38.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. An enterprise data warehouse may be implemented on traditional mainframes, UNIX super servers (or) parallel architecture platforms. It requires business modeling and may take years to design and build.

39.What is data mart?
Data mart is a database that contains a subset of data present in a data warehouse. Data marts are created to structure the data in a data warehouse according to issues such as hardware platforms and access control strategies. We can divide a data warehouse into data marts after the data warehouse has been created. Data marts are usually implemented on low-cost departmental servers that are UNIX (or) windows/NT based. The implementation cycle of the data mart is likely to be measured in weeks rather than months (or) years.

40.What are dependent and independent data marts?
Dependent data marts are sourced directly from enterprise data warehouses. Independent data marts are data captured from one (or) more operational systems (or) external information providers (or) data generated locally with in particular department (or) geographic area.

41.What is virtual warehouse?
A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized. A virtual warehouse is easy to build but requires excess capability on operational database servers.

42.Define indexing?
Indexing is a technique, which is used for efficient data retrieval (or) accessing data in a faster manner. When a table grows in volume, the indexes also increase in size requiring more storage.

43.What are the types of indexing?
  ➢ B-Tree indexing

> ➢ Bit map indexing
> ➢ Join indexing

44.Define metadata?

Metadata is used in data warehouse is used for describing data about data. (i.e.) meta data are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse.

45.Define VLDB?

Very Large Data Base. If a database whose size is greater than 100GB, then the database is said to be very large database.

UNIT – V

1.What are the classifications of tools for data mining?

- Commercial Tools
- Public domain Tools
- Research prototypes

2.What are commercial tools?

Commercial tools can be defined as the following products and usually are associated with the consulting activity by the same company:
1. 'Intelligent Miner' from IBM
2. 'SAS' System from SAS Institute
3. 'Thought' from Right Information Systems. etc

3. What are Public domain Tools?

Public domain Tools are largely freeware with just registration fees: 'Brute' from University of Washington. 'MC++' from Stanford university, Stanford, California.

4. What are Research prototypes?

Some of the research products may find their way into commercial market: 'DB Miner' from Simon Fraser University, British Columbia, 'Mining Kernel System' from University of Ulster, North Ireland.

5.What is the difference between generic single-task tools and generic multi-task tools?

Generic single-task tools generally use neural networks or decision trees. They cover only the data mining part and require extensive pre-processing and post-processing steps.

Generic multi-task tools offer modules for pre-processing and post-processing steps and also offer a broad selection of several popular data mining algorithms as clustering.

6. What are the areas in which data warehouses are used in present and in future?

The potential subject areas in which data ware houses may be developed at present and also in future are

1.Census data:

The registrar general and census commissioner of India decennially compiles information of all individuals, villages, population groups, etc. This information is wide ranging such as the individual slip. A compilation of information of individual households, of which a database of 5%sample is maintained for analysis. A data warehouse can be built from this database upon which OLAP techniques can be applied, Data mining also can be performed for analysis and knowledge discovery

2.Prices of Essential Commodities

The ministry of food and civil supplies, Government of India complies daily data for about 300 observation centers in the entire country on the prices of essential commodities such as rice, edible oil etc, A data warehouse can be built for this data and OLAP techniques can be applied for its analysis

7. What are the other areas for Data warehousing and data mining?
- Agriculture
- Rural development
- Health
- Planning
- Education
- Commerce and Trade

8. Specify some of the sectors in which data warehousing and data mining are used?
- Tourism
- Program Implementation
- Revenue
- Economic Affairs
- Audit and Accounts

9. Describe the use of DBMiner.

Used to perform data mining functions, including characterization, association, classification, prediction and clustering.

10. Applications of DBMiner.

The DBMiner system can be used as a general-purpose online analytical mining system for both OLAP and data mining in relational database and datawarehouses.
Used in medium to large relational databases with fast response time.

11. Give some data mining tools.
DBMiner
GeoMiner
Multimedia miner
WeblogMiner

12. Mention some of the application areas of data mining
DNA analysis
Financial data analysis
Retail Industry
Telecommunication industry
Market analysis
Banking industry
Health care analysis.

13. Differentiate data query and knowledge query
A data query finds concrete data stored in a database and corresponds to a basic retrieval statement in a database system.
A knowledge query finds rules, patterns and other kinds of knowledge in a database and corresponds to querying database knowledge including deduction rules, integrity constraints, generalized rules, frequent patterns and other regularities.

14.Differentiate direct query answering and intelligent query answering.
Direct query answering means that a query answers by returning exactly what is being asked.
Intelligent query answering consists of analyzing the intent of query and providing generalized, neighborhood, or associated information relevant to the query.

15. Define visual data mining
Discovers implicit and useful knowledge from large data sets using data and/ or knowledge visualization techniques.
Integration of data visualization and data mining.

16. What does audio data mining mean?
Uses audio signals to indicate patterns of data or the features of data mining results.
Patterns are transformed into sound and music.
To identify interesting or unusual patterns by listening pitches, rhythms, tune and melody.

Steps involved in DNA analysis
Semantic integration of heterogeneous, distributed genome databases
Similarity search and comparison among DNA sequences
Association analysis: Identification of co-occuring gene sequences
Path analysis: Linking genes to different stages of disease development
Visualization tools and genetic data analysis

17.What are the factors involved while choosing data mining system?
Data types
System issues
Data sources
Data Mining functions and methodologies
Coupling data mining with database and/or data warehouse systems
Scalability
Visualization tools
Data mining query language and graphical user interface.

18. Define DMQL
Data Mining Query Language
It specifies clauses and syntaxes for performing different types of data mining
tasks for example data classification, data clustering and mining association
rules. Also it uses SQl-like syntaxes to mine databases.

19. Define text mining
Extraction of meaningful information from large amounts free format textual
data.
Useful in Artificial intelligence and pattern matching
Also known as text mining, knowledge discovery from text, or content
analysis.

20. What does web mining mean
Technique to process information available on web and search for useful data.
To discover web pages, text documents , multimedia files, images, and other
types of resources from web.
Used in several fields such as E-commerce, information filtering, fraud
detection and education and research.

21.Define spatial data mining.
Extracting undiscovered and implied spatial information.
Spatial data: Data that is associated with a location
Used in several fields such as geography, geology, medical imaging etc.

22. Explain multimedia data mining.
Mines large data bases.
 Does not retrieve any specific information from multimedia databases
Derive new relationships , trends, and patterns from stored multimedia data
mining.

Used in medical diagnosis, stock markets ,Animation industry, Airline industry, Traffic management systems, Surveillance systems etc.

16 MARKS QUESTIONS AND ANSWERS

UNIT-I

1. Explain the evolution of Database technology?

➢ Data collection and Database creation
➢ Database management systems
➢ Advanced database systems
➢ Data warehousing and Data Mining
➢ Web-based Database systems
➢ New generation of Integrated information systems

2.Explain the steps of knowledge discovery in databases?

➢ Data cleaning
➢ Data integration
➢ Data selection
➢ Data transformation
➢ Data mining
➢ Pattern evaluation
➢ Knowledge presentation

3. Explain the architecture of data mining system?

➢ Database, datawarehouse, or other information repository
➢ Database or data warehouse server
➢ Knowledge base
➢ Data mining engine
➢ Pattern evaluation module
➢ Graphical user interface

4.Explain various tasks in data mining?
(Or)

Explain the taxonomy of data mining tasks?
➢ Predictive modeling
  • Classification
  • Regression

- Time series analysis
- Descriptive modeling
  - Clustering
  - Summarization
  - Association rules
  - Sequence discovery

5.Explain various techniques in data mining?
- Statistics (or) Statistical perspectives
- Point estimation
  - Data summarization
  - Bayesian techniques
  - Hypothesis testing
  - Correlation
- Regression
- Machine learning
- Decision trees
- Hidden markov models
- Artificial neural networks
- Genetic algorithms
- Meta learning

UNIT-II

6.Explain the issues regarding classification and prediction?
- Preparing the data for classification and prediction
  - Data cleaning
  - Relevance analysis
  - Data transformation
- Comparing classification methods
  - Predictive accuracy
  - Speed
  - Robustness
  - Scalability
  - Interpretability

7.Explain classification by Decision tree induction?
- Decision tree induction
- Attribute selection measure.
- Tree pruning
- Extracting classification rules from decision trees

8.Write short notes on patterns?
- Pattern definition
- Objective measures
- Subjective measures

- ✓ Can a data mining system generate all of the interesting patterns?
- ✓ Can a data mining system generate only interesting patterns?

9.Explain mining single –dimensional Boolean associated rules from transactional databases?

- ✓ The apriori algorithm: Finding frequent itemsets using candidate generation
- ✓ Mining frequent item sets without candidate generation

10.Explain apriori algorithm?

- ✓ Apriori property
- ✓ Join steps
- ✓ Prune step
- ✓ Example
- ✓ Algorithm

11.Explain how the efficiency of apriori is improved?

- ❑ Hash-based technique (hashing item set counts)
- ❑ Transaction reduction (reducing the number of transactions scanned in future iteration)
- ❑ Partitioning (Partitioning the data to find candidate item sets)
- ❑ Sampling (mining on a subset of the given data)
- ❑ Dynamic item set counting (adding candidate item sets at different points during a scan)

12.Explain frequent item set without candidate without candidate generation?

- ❑ Frequent patterns growth (or) FP-growth
- ❑ Frequent pattern tree (or) FP-tree
- ❑ Algorithm

13. Explain mining Multi-dimensional Boolean association rules from transaction databases?

- ❑ Multi-dimensional (or) Multilevel association rules
- ❑ Approaches to mining Multilevel association rules
  - • Using uniform minimum support for all levels
  - • Using reduced minimum support at lower levels
    - o Level-by-level independent
    - o Level-cross filtering by single
    - o Level- cross filtering by k-item set

- ❑ Checking for redundant Multilevel association rules

14.Explain constraint-based association mining?

      ✓   Knowledge type constraints
      ✓   Data constraints
      ✓   Dimension/level constraints
      ✓   Interestingness constraints
      ✓   Rule constraints
      ✓   Metarule-Guided mining of association of association rules
      ✓   Mining guided by additional rule constraints

Unit –III

15.Explain regression in predictive modeling?

    ✓ Regression definition
    ✓ Linear regression
    ✓ Multiple regression
    ✓ Non-linear regression
    ✓ Other regression models

16.Explain statistical perspective in data mining?

    ✓ Point estimation
    ✓ Data summarization
    ✓ Bayesian techniques
    ✓ Hypothesis testing
    ✓ Regression
    ✓ Correlation

17. Explain Bayesian classification.

    ✓ Bayesian theorem
    ✓ Naïve Bayesian classification
    ✓ Bayesian belief networks
    ✓ Bayesian learning

18. Discuss the requirements of clustering in data mining.

    ✓ 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

- ✓ Insensitivity to the order of input records
- ✓ High dimensionality
- ✓ Interpretability and usability
- ✓ Interval scaled variables
- ✓ Binary variables
  - o Symmetric binary variables
  - o Asymmetric binary variables
- ✓ Nominal variables
- ✓ Ordinal variables
- ✓ Ratio-scaled variables

20. Explain the partitioning method of clustering.
K-means clustering
K-medoids clustering

21. Explain Visualization in data mining.
Various forms of visualizing the discovered patterns
- ✓ Rules
- ✓ Table
- ✓ Crosstab
- ✓ Pie chart
- ✓ Bar chart
- ✓ Decision tree
- ✓ Data cube
- ✓ Histogram
- ✓ Quantile plots
- ✓ q-q plots
- ✓ Scatter plots
- ✓ Loess curves

UNIT IV

22. Discuss the components of data warehouse.
- ✓ Subject-oriented
- ✓ Integrated
- ✓ Time-Variant
- ✓ Non-volatile

23. List out the differences between OLTP and OLAP.
- ✓ Users and system orientation
- ✓ Data contents
- ✓ Database design

- ✓ View
- ✓ Access patterns

24.Discuss the various schematic representations in multidimensional model.
- ✓ Star schema
- ✓ Snow flake schema
- ✓ Fact constellation schema

25. Explain the OLAP operations I multidimensional model.
- ✓ Roll-up
- ✓ Drill-down
- ✓ Slice and dice
- ✓ Pivot or rotate

26. Explain the design and construction of a data warehouse.
- ▪ Design of a data warehouse
  - • Top-down view
  - • Data source view
  - • Data warehouse view
  - • Business query view
- ▪ Process of data warehouse design

27.Expalin the three-tier data warehouse architecture.
- ✓ Warehouse database server(Bottom tier)
- ✓ OLAP server(middle tier)
- ✓ Client(top tier)

28. Explain indexing.
- ✓ Definition
- ✓ B-Tree indexing
- ✓ Bit-map indexing
- ✓ Join indexing

29.Write notes on metadata repository.
- ✓ Definition
- ✓ Structure of the data warehouse
- ✓ Operational metadata
- ✓ Algorithms used for summarization
- ✓ Mapping from operational environment to data warehouse
- ✓ Data related to system performance
- ✓ Business metadata

30. Write short notes on VLDB.
- ✓ Definition
- ✓ Challenge related to database technologies
- ✓ Issues in VLDB

UNIT V

31.Explain data mining applications for Biomedical and DNA data analysis.
- ✓ Semantic integration of heterogeneous, distributed genome databases

- ✓ Similarity search and comparison among DNA sequences
- ✓ Association analysis.
- ✓ Path analysis
- ✓ Visualization tools and genetic data analysis.

32. Explain data mining applications fro financial data analysis.
- ✓ Loan payment prediction and customer credit policy analysis.
- ✓ Classification and clustering of customers fro targeted marketing.
- ✓ Detection of money laundering and other financial crimes.

33. Explain data mining applications for retail industry.
- ✓ Multidimensional analysis of sales, customers, products, time and region.
- ✓ Analysis of the effectiveness of sales campaigns.
- ✓ Customer retention-analysis of customer loyalty.
- ✓ Purchase recommendation and cross-reference of items.

34. Explain data mining applications for Telecommunication industry.
- ✓ Multidimensional analysis of telecommunication data.
- ✓ Fraudulent pattern analysis and the identification of unusual patterns.
- ✓ Multidimensional association and sequential pattern analysis
- ✓ Use of visualization tools in telecommunication data analysis.

35. Explain DBMiner tool in data mining.
- ✓ System architecture
- ✓ Input and Output
- ✓ Data mining tasks supported by the system
- ✓ Support of task and method selection
- ✓ Support of the KDD process
- ✓ Main applications
- ✓ Current status

36. Explain how data mining is used in health care analysis.
- ✓ Health care data mining and its aims
- ✓ Health care data mining technique
- ✓ Segmenting patients into groups
- ✓ Identifying patients into groups
- ✓ Identifying patients with recurring health problems
- ✓ Relation between disease and symptoms
- ✓ Curbing the treatment costs
- ✓ Predicting medical diagnosis
- ✓ Medical research
- ✓ Hospital administration
- ✓ Applications of data mining in health care
- ✓ Conclusion

37. Explain how data mining is used in banking industry.

- ✓ Data collected by data mining in banking
- ✓ Banking data mining tools
- ✓ Mining customer data of bank
- ✓ Mining for prediction and forecasting
- ✓ Mining for fraud detection
- ✓ Mining for cross selling bank services
- ✓ Mining for identifying customer preferences
- ✓ Applications of data mining in banking
- ✓ Conclusion

38. Explain the types of data mining.
- ✓ Audio data mining
- ✓ Video data mining
- ✓ Image data mining
- ✓ Scientific and statistical data mining