DEPARTMENT OF INFORMATION TECHNOLOGY

IT6702 DATA WAREHOUSING & DATA MINING

UNIT I

PART A

1. Define data mining?

Data mining refers to extracting or “mining" knowledge from large amounts of data and another different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

2. Define data warehouse?

A data warehouse is a repository of information collected from multiple sources, stored under unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing.

3. What is OLAP?

On-Line Analytical Processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation and aggregation, as well as the ability to view information at different angles. Although 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.

4. What are the steps of data mining?

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation

 Knowledge presentation

5. What are the components of data mining?

1. Database, data warehouse, or other information repository.

2. Database or data warehouse server.

3. Knowledge base.

4. Data mining engine.

5. Pattern evaluation module.

6. Graphical user interface.

6. What are the types of data bases?

Relational databases, Transactional databases, object oriented databases, object relational databases, spatial databases, temporal and time series databases, text and multimedia databases, heterogeneous and legacy databases.

7. What is a relational database?

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 _elds) and usually stores a large number of tuples (records or rows). Each tuple in a relational table represents an object identi_ed by a unique key and described by a set of attribute values.

8. What is a transactional database?

A transactional database consists of a _le 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 sales person, and of the branch at which the sale occurred, and so on.

9. What is object oriented database?

Object oriented databases are base3d on the object oriented programming paradigm, where in

general terms, each entity considered as an object. Each object has associated with it the following:

A set of variables, a set of messages, a set of variables, a set of methods.

10. What is object relational database?

It based on an object relational data model. Relational model providing a rich data type for handling complex objects and object orientation.

11. What is a spatial database?

It contains spatial related information. Such databases include geographic databases, VLSI chip design databases and medical and satellite image databases.

Spatial data may be represented in raster format, consisting of n dimensional bit maps or pixel maps. Maps can be represented in vector format, where roads, bridges, buildings and lakes are represented as unions of basic geometric constructs.

12. What is a temporal database and time series database?

A temporal database usually stores relational data that include time-related attributes.

A time series database stores sequences of values of values that change with time, such as data collected regarding the stock exchange.

13. What is text database?

Text database are databases that contain word descriptions for objects. Text databases highly unstructured, semi structured, well structured.

14. What is multimedia database?

Multimedia database store image, audio and video data. Multimedia databases must support large objects. It predetermined rate in order to avoid picture or sound gaps and system buffer, such data are referred to as continuous media data.

15. What is mining path traversal Patterns?

Capturing user access patterns in distributed information environment is called mining path

traversal Patterns.

16. What is data mining task?

Data mining task can be classified into two types: descriptive and predictive.

Descriptive mining task characterize the general properties of the data in the databases.

Predictive mining tasks perform inference on the current data in order to make predictions.

17. What is class/concept description?

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions.

18. What is data characterization?

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query. For example, to study the characteristics of software products whose sales increased by 10% in the last year, one can collect the data related

to such products by executing an SQL query.

19. What is Data discrimination?

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.

20. What is classification?

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes 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.

21. What is prediction?

Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class

labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction.

22. Define decision tree?

A decision tree is flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules

23. What is cluster analysis?

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.

24. What is 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.

25. What is evolution analysis?

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

26. How can data mining system categorized?

According to the databases, knowledge, techniques and applications.

27. What are methodology and user interactions in data mining?

- 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 outlier or incomplete data.
- Pattern evaluation: the interestingness problem.

28. What are the Performance issues and Issues relating to the diversity of database types in data mining?

- Efficiency and scalability of data mining algorithms.
- Parallel, distributed, and incremental updating algorithms.
- Handling of relational and complex types of data.
- Mining information from heterogeneous databases and global information systems

PART B

1. What is data mining? Explain about steps and architecture of data mining?
2. Explain in detail different types of databases?
3. What kinds of pattern can be mined?
4. Explain in detail classification of data mining systems?
5. What are the major issues in data mining?

UNIT II

PART A

1. Give different data preprocessing techniques.

Data cleaning, data integration, data transformation, and data reduction

2. What is data cleaning?

It removes noise and correct inconsistencies in the data.

3. What is data integration?

The process of merging of data from multiple data stores into a coherent data store, such as data warehouse or a data cube.

4. What is data transformation?

To improve the accuracy and efficiency of mining algorithms involving distance measurements, data are transformed into forms appropriate for mining.

5. What is data reduction?

It is a process to reduce the data size by aggregation, eliminating redundant features or clustering.

6. Give the strategies for data reduction?

Data cube aggregation, dimension reduction, data compression, numerosity reduction, and discretization and concept hierarchy generalization. Generalisation can also be used to reduce data.

7. Why we preprocess data?

Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

8. What are the basic methods in data cleaning?

Filling in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

9. How to fill the missing data?

Some methods to fill in the missing data are,

a. Ignore the tuple.

b. Fill in the missing value manually.

c. Use the attribute mean to fill in the missing value

d. Use a global constant to fill in the missing value

e. Use the attribute mean for all samples belonging to the same class as the given tuple

f. Use the most probable value to fill in the missing value.

10. What is noisy data?

Noise is a random error or variance in a measured variable.

11. List out the data smoothing techniques.

Binning, Clustering, Combined computer and human inspection, Regression.

12. What is binning?

Binning methods smooth a sorted data value in reference with its near value. The sorted values are distributed into a number of buckets or bins.

13. Give out some binning methods.

a. Smoothing by bin means.

b. Smoothing by bin medians.

c. Smoothing by bin boundaries.

14. What is linear regression?

Linear regression involves finding the best line to fit two variables, so that one variable can be used to predict the other. When more then two variables are involved and data are fit to a multidimensional surface, it can be taken as multiple regression.

15. What are the issues to be considered during data integration?

a. Entity identification problem, matching real world entities with multiple data

sources.

b. Redundancy, deriving an attribute from another table also.

c. Detection and resolution of data values conflicts.

16. List out the methodologies in data transformation.

a. Smoothing, to remove noise from data.

b. Aggregation, to construct data cube for analysis at multiple granularities.

c. Generalisation, replacing low level data with higher level concepts..

d. Normalization, scaling attribute data to fall within a small specified range.

e. Attribute construction, constructing new attributes from given set of attributes.

17. Give some data normalization methods.

a. Min-max normalization

b. z- score normalization.

c. Normalization by decimal scaling.

18. What is Min-max normalization?

It performs linear transformation on the original data. Suppose that minA and

maxA are the minimum and maximum values of an attribute A. Then Min-max normalization maps a value v of A to v' in the range [new_minA, new_maxA]. This

method preserves the relationships among the original data values.

19. What do you mean by z- score normalization?

In z- score normalization (or zero mean normalization), the values for an attribute

A are normalized based on the mean and standards deviation of A. A value v of A is

normalized to v' by computing

It is useful when the actual minimum and maximum of attribute A are unknown,

or when there are outliers that dominate the min-max normalization.

20. What is normalization by decimal scaling?

It normalizes by moving the decimal point of values of attribute A. The number of

decimal points moved depends on the maximum absolute value A.

21. Define Base cuboid and apex cuboid.

A cube created at the lowest level of abstractions is referred to as the base cuboid

whereas a cube of highest level of abstraction is apex cuboid.

22. Define Cuboid.

Data cubes created for varying levels of abstraction are referred as cuboids.

23. What is dimensionality reduction?

The methodology that reduces the data set size by removing irrelevant and redundant attributes from the data set.The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of data classes is as close as possible to original distribution obtained using all attribute.

24. List out techniques involved in attribute subset selection.

a. Stepwise forward selection: The procedure starts with an empty set of attributes.The best of the original attributes is determined and added to the set.

b. Stepwise backward elimination: The process of starting with full set of attributes and removing the worst attribute from the set.

c. Combination of forward selection and backward elimination: The stepwise forward selection and backward methods can be combined, such that at each step, the process selects best attribute and removes worst among remaining set.

25. Define decision tree Induction.

Decision tree induction constructs a flowchart like structure where internal (nonleaf) node denotes a test on an attribute, each corresponding to an outcome of the test, and each external (leaf) node denotes a class prediction.

26. What do you mean by wrapper approach?

If the mining task is classification and the mining algorithm itself is used to determine the attribute subset, then it can be taken wrapper approach; otherwise it can be taken as filter approach. In general the wrapper approach leads to greater accuracy since it optimizes the evaluation measure of the algorithm while removing attributes.

27. What do you infer from lossless and lossy compression?

If the original data can be reconstructed from the compressed data without any loss of information, the data compression technique used is lossless; if reconstruction brings out only the approximation of the original data, it is termed as lossy.

28. What is DWT or discrete wavelet Transform?

DWT is a linear signal processing technique that, when applied to a data vector D, transforms it to a numerically different vector, D' of wavelet coefficients. The two vectors are of the same length.

29. When pyramid algorithm is used?

A general procedure for applying a discrete wavelet transforms uses a hierarchical algorithm that halves the data at each iteration resulting in fast computational speed.

30. How PCA works?

Principal Component analysis searches for c k-dimensional orthogonal vectors that can be used to represent the data, where c<=k. It is a method of dimensionality reduction and original data are projected into a much smaller space, resulting in data compression.

31. How numerosity reduction can be utilized?

Numerosity reduction can be used to reduce the data volume by choosing alternative, 'smaller' forms of data representation.

32. What are the types of numerosity reduction?

Parametric reduction, a model is used to estimate the data, so that typically only the data parameters need to be stored instead of actual data. Non parametric methods is to store reduced representation of the data include histograms, clustering and sampling.

33. How regression coefficients can be solved?

It can be solved by method of least squares, which minimizes the error between the actual line separating the data and the estimate of the line.

34. Write short notes on Histogram.

Histograms use binning to approximate data distributions, which is a form of data reduction.

A histogram for an attribute A partitions the data distribution of A into disjoint subsets or buckets.The buckets are displayed on a horizontal axis, while the height of a bucket reflects the average frequency of the values represented by the bucket.

35. Define singleton buckets.

The collection of buckets where each bucket represents only a single attribute-value or frequency pair is singleton buckets.

36. How are buckets determined and the attribute values partitioned?

Some partitioning rules are,

a. Equiwidth: In an equiwidth histogram, the width of each bucket range is uniform.

b. Equidepth: The buckets are created so that the frequency of each bucket is constant.

c. V-Optimal: If all possible histograms for a given number of buckets are considered, it is the one with least variance.

d. MaxDiff: The difference between each pair of adjacent values are considered.

37. Clustering – Make short note.

Clustering takes data tuples as objects and 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. It is used to define closeness of objects in space based on a distance function.

38. Define centriod distance.

Centriod distance is an alternative measure of clusters and it is the average distance of each cluster object from the cluster centriod.

39. What do you understand by the term multidimensional index trees?

They are used for providing fast data access and they can also be used for hierarchical data reduction, providing a multiresolution clustering of the data. An index tree recursively partitions the multidimensional space for a given a set of data objects, with the root node representing the entire space, and leaf node contains pointers to the data tuples they represent.

40. Define sampling.

Sampling is used as a data reduction technique since it allows a large data set to be represented by a much smaller random sample of the data.

41. What are the types of samples that can be taken for data reduction?

a. SRSWOR(Simple Random Sample without Replacement): This is created by drawing n of the N tuples from D(n < N), where the probability of drawing any tuple in Dis 1/N.

b. SRSWR(Simle Random Sample with Replacement): This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and than replaced.

c. Cluster sample: If the tuples in D are grouped into M mutually disjoint " clusters" then a SRS of m clusters can be obtained.

d. Stratified sample: If D is divided into mutually disjoint parts called strata, a stratified sample of D is generated by obtaining an SRS at each stratum.

42. What is the advantage of Sampling?

An advantage of sampling for data reduction is that the cost of obtaining a sample is proportional to the size of the sample, n, as opposed to N, the data set size.

43. What do you mean by Concept hierarchy?

A concept hierarchy for a given numeric attribute defines a disretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low level concepts.

44. List out the methodologies of concept generation for numeric concept.

Binning, Histogram analysis, cluster analysis, entropy-based discretization, and data segmentation by natural partitioning.

PART B

1) Why is the need of Data Preprocessing?

a) Data Cleaning

b) Data Integration

c) Data transformation

d) Data Reduction

2) Explain how Data Cleaning is done.

a) Handling Missing Values

b) Handling Noisy Data –Binning, Clustering, Inspection, Regression

3) What is Data Transformation? Explain how it is done.

a) Smoothing

b) Aggregation

c) Generalization

d) Normalization

d) Max-Difference

5) What is Sampling? Explain it types.

a) Simple random sample without replacement of size n

b) Simple random sample with replacement of size n

c) Cluster Sample

d) Stratified Sample

6) Explain about Discretization and Concept Hierarchy Generation techniques can be applied for numeric data

a) Binning

b) Histogram Analysis

c) Cluster Analysis

d) Entropy-based Discretization

e) Segmentation by Natural Partitioning

UNIT III

PART A

1. What are the two steps in data classification?

a. A model is built describing a predetermined set of data classes or concepts.

b. The model is used for classification.

2. What is the difference between supervised and unsupervised learning?

The class label of each training sample is provided in supervised learning whereas in unsupervised learning the class label of each training sample is not known.

3. How prediction is different from classification?

Prediction models continuous valued functions whereas classification predicts categorical labels. Prediction is the construction and use of a model to assess the class of an unlabeled sample, or to assess the values or value ranges of an attribute. Classification and regression are two types of prediction.

4. List out the steps for preparing the data for classification and predication.

Data cleaning, Relevance analysis and Data transformation.

5. What are the criteria used for comparing classification and prediction?

Predictive accuracy, Speed, Robustness, Scalability and interoperability

6. Define a decision tree.

A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attributes, each branch represents an outcome of the test, and leaf nodes represent classes or classes or class distribution.

7. What is tree pruning?

Tree pruning methods address problem of over fitting the data when a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers.

8. What are the two approaches in tree pruning?

a. Prepruning: tree is pruned by halting its construction early.

b. Postpruning: removes branches from a fully grown tree.

9. What does Naïve Bayesian classifiers do?

It allows the representation of dependencies among subsets of attributes used for classification.

10. Give the bayes theorem. Why it is used?

Bayes theorem is useful in calculating posterior probabilities. The theorem is

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

11. What is backpropogation?

Backpropagation is a neural network learning algorithm. It performs learning on a multilayer feed forward neural network.

12. How does Backpropagation work?

Backpropagation learns by iteratively processing a set of training samples, comparing the networks prediction for each sample with the actual known class label.

13. List out the steps in Backpropagation algorithm.

a. Initialize the weights.

b. Propagation the inputs forward.

c. Backpropagate the error.

d. Terminating condition.

14. Give some classification methods.

i. k- Nearest Neighbor classifiers.

ii. Case- Based Reasoning.

iii. Genetic Algorithm.

iv. Rough set approach.

v. Fuzzy set approach.

15. Give some software to solve regression problems.

SAS, SPSS, S-plus

16. What is linear regression?

In linear regression, data are modeled using a straight line. It is the simplest form

of regression where a random variable(Y) is modeled as a linear function of another

random variable(X).

17. What is CBR? Where are they used?

Case Based Reasoning (CBR) is used for classification instantly. The samples or cases stored by CBR are complex symbolic descriptions.They are used for problem resolution for customer service, engineering field, and law where cases are legal rulings.

18. How k-Nearest Neighbor classifiers work?

When an unknown sample is given then this classifier searches pattern space for the kth

samples that are closest to the unknown sample. These k training samples are the k

“nearest neighbors” of the unknown sample.

19. How closeness in defined for k nearest neighbor?

Closeness can be defined in terms of Euclidean distance where distance between two Euclidean points is found using the distance vector formula.

Backpropagation learns by iteratively processing a set of training samples, comparing the networks prediction for each sample with the actual known class label.

13. List out the steps in Backpropagation algorithm.

a. Initialize the weights.

b. Propagation the inputs forward.

c. Backpropagate the error.

d. Terminating condition.

14. Give some classification methods.

i. k- Nearest Neighbor classifiers.

ii. Case- Based Reasoning.

iii. Genetic Algorithm.

iv. Rough set approach.

v. Fuzzy set approach.

15. Give some software to solve regression problems.

SAS, SPSS, S-plus

16. What is linear regression?

In linear regression, data are modeled using a straight line. It is the simplest form of regression where a random variable(Y) is modeled as a linear function of another random variable(X).

17. What is CBR? Where are they used?

Case Based Reasoning (CBR) is used for classification instantly. The samples or cases stored by CBR are complex symbolic descriptions.

They are used for problem resolution for customer service, engineering field, andlaw where cases are legal rulings.

18. How k-Nearest Neighbor classifiers work?

When an unknown sample is given then this classifier searches pattern space for the kth samples that are closest to the unknown sample. These k training samples are the k “nearest neighbors” of the unknown sample.

19. How closeness in defined for k nearest neighbor?

Closeness can be defined in terms of Euclidean distance where distance between two Euclidean points is found using the distance vector formula. Combined computer and human inspection Regression Inconsistencies removal

2. Explain Data integration and Transformation process in data processing.

Data Integration

Definition

Issues

Data Transformation

Smoothing

Aggregation

Generalization of the data

Normalization

Attribute construction

3. Describe Data reduction.

Data Cube aggregation

Dimension reduction

Stepwise forward selection

Stepwise backward elimination

Combination of forward and backward elimination

Data compression

Wavelet transforms

Principal components Analysis

Numerosity reduction

Regression and log linear models

Histograms

Sampling

Clustering

Discretization and concept hierarchy generation

4. Explain in detail Discretization and concept hierarchy generation

Discretization and concept hierarchy generation for numeric data:

Binning

Histogram analysis

Cluster analysis

Entropy-based Discretization

Segmentation by Natural Partitioning

Concept hierarchy Generation for Categorical Data:

Specification of a partial ordering of attribute explicitly at the schema

level by users or experts

Specification of a portion of a hierarchy by explicit data grouping

Specification of a set of attributes, but not their partial ordering

Specification of only a partial set of attributes

5. Describe data generalization and summarization based characterization.

Attribute oriented induction:

Steps in attribute oriented induction

1. Data focusing is prior

2. Data generalization

3. Attribute generalization

Techniques in generalization

1. Attribute generalization threshold control

2. Generalized relation threshold control

Efficient implementation of attribute oriented induction

Presentation of the derived generalization

UNIT –IV

PART A

1. What is Data Warehouse?

A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile

collection of data in support of management's decision-making process.

2. What is Data warehousing?

The process of constructing and using data warehouses

3. Define the following terms: base cuboid, apex cuboid and data cube.

In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

4. Define the following terms: star schema, fact schema, fact constellations.

 Star schema: A fact table in the middle connected to a set of dimension tables

 Snowflake schema: A refinement of star schema where some dimensional

hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

 Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

5. Explain typical OLAP operations.

 Roll up (drill-up): summarize data

o by climbing up hierarchy or by dimension reduction

 Drill down (roll down): reverse of roll-up

o from higher level summary to lower level summary or detailed data, or introducing new dimensions

 Slice and dice:

o project and select

 Pivot (rotate):

o reorient the cube, visualization, 3D to series of 2D planes.

 Other operations

o drill across: involving (across) more than one fact table

o drill through: through the bottom level of the cube to its back-end relational tables (using SQL)

6. Define the following terms: Enterprise warehouse, Data Mart, Virtual warehouse

 Enterprise warehouse

o collects all of the information about subjects spanning the entire organization

 Data Mart

o a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart

 Independent vs. dependent (directly from warehouse) data mart

 Virtual warehouse

o A set of views over operational databases

o Only some of the possible summary views may be materialized

7. Define the following terms: Relational OLAP (ROLAP) , Multidimensional OLAP (MOLAP), Hybrid OLAP (HOLAP), Specialized SQL servers

 Relational OLAP (ROLAP)

o Use relational or extended-relational DBMS to store and manage warehouse data and OLAP middle ware to support missing pieces

o Include optimization of DBMS backend, implementation of aggregation navigation logic, and additional tools and services

o greater scalability

 Multidimensional OLAP (MOLAP)

o Array-based multidimensional storage engine (sparse matrix techniques)

o fast indexing to pre-computed summarized data

 Hybrid OLAP (HOLAP)

o User flexibility, e.g., low level: relational, high-level: array

 Specialized SQL servers

o specialized support for SQL queries over star/snowflake schemas

8. What is a metadata? What are its contents?

 Meta data is the data defining warehouse objects. It has the following kinds

o Description of the structure of the warehouse

 schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents

o Operational meta-data

 data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)

o The algorithms used for summarization

o The mapping from operational environment to the data warehouse

o Data related to system performance

 warehouse schema, view and derived data definitions

o Business data

 business terms and definitions, ownership of data, charging policies

9. What are the three kinds of data warehouse applications?

o Information processing

 supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs

o Analytical processing

 multidimensional analysis of data warehouse data

 supports basic OLAP operations, slice-dice, drilling, pivoting

o Data mining

 knowledge discovery from hidden patterns

 Supports associations, constructing analytical models, performing

Classification and prediction, and presenting the mining results using visualization tools.

10. Differentiate between Data Warehouse vs. Heterogeneous DBMS

 Traditional heterogeneous DB integration:

o Build wrappers/mediators on top of heterogeneous databases

o Query driven approach

 When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneoussites involved, and the results are integrated into a global answer set.

 Complex information filtering, compete for resources

 Data warehouse: update-driven, high performance

o Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

11.. What are the three categories of measures that based on the kind of aggregate functions

used?

 Distributive: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning.

 E.g., count(), sum(), min(), max().

 Algebraic: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function.

 E.g., avg(), min_N(), standard deviation().

 Holistic: if there is no constant bound on the storage size needed to describe a sub

aggregate.

 E.g., median(), mode(), rank().

12. What are the four views regarding the design of a data warehouse?

o Top-down view

 allows selection of the relevant information necessary for the data warehouse

o Data source view

 exposes the information being captured, stored, and managed by

operational systems

o Data warehouse view

 consists of fact tables and dimension tables

o Business query view

 sees the perspectives of data in the warehouse from the view of end-user

13. What are the efficient cube computation methods available?

 ROLAP-based cubing algorithms

 Array-based cubing algorithm

 Bottom-up computation method

14. Write short notes on ROLAP-based cubing algorithms.

 Sorting, hashing, and grouping operations are applied to the dimension attributes

in order to reorder and cluster related tuples

 Grouping is performed on some sub aggregates as a “partial grouping step”

 Aggregates may be computed from previously computed aggregates, rather than

from the base fact table

15. Write short notes on Bitmap Index.

 Index on a particular column

 Each value in the column has a bit vector: bit-op is fast

 The length of the bit vector: # of records in the base table

 The i-th bit is set if the i-th row of the base table has the value for the indexed column

 not suitable for high cardinality domains

16. What are the Data Warehouse Back-End Tools and Utilities?

 Data extraction:

o get data from multiple, heterogeneous, and external sources

 Data cleaning:

o detect errors in the data and rectify them when possible

 Data transformation:

o convert data from legacy or host format to warehouse format

 Load:

o sort, summarize, consolidate, compute views, check integrity, and build indices

and partitions

 Refresh

o propagate the updates from the data sources to the warehouse

PART B

1. Explain about star, snowflake and Fact Constellation schemas for Multidimensional

Databases with suitable examples.

Star, snowflake and Fact Constellation schemas:

2. a) What are the measures that can be organized into three categories based on the kind of

aggregate functions?

3. What is a concept Hierarchy? Explain with example.

Concept Hierarchy :

4. a) Explain typical OLAP operations on multidimensional data with examples.

Roll-up, Drill-down, Slice and dice, Pivot, drill-across, drill-through:

b) Explain OLAP systems versus Statistical Databases.

OLAP systems versus Statistical Databases:

5. a) Describe about Starnet Query Model for Querying Multidimensional databases.

Starnet Query Model for Querying Multidimensional databases:

b) What does the data warehouse provide for business analysts?

c) What are the four views regarding the design of a data warehouse?

Top-down view, Data source view, and Data warehouse view, business query view:

5. a) How do you design a data warehouse?

Design a data warehouse:

b) Describe a three-tier data warehousing architecture with suitable diagram.

Three-tier data warehousing architecture:

6. a) What kinds of OLAP Servers exist? Explain it.

ROLAP, MOLAP, HOLAP servers:

b) What are the efficient cube computation methods available? Explain it.

a. ROLAP-based cubing algorithms

b. Array-based cubing algorithm

c. Bottom-up computation method

7. a) Write short notes on Indexing OLAP Data.

Indexing OLAP Data :

b) How will you efficiently process OLAP Queries?

8. a) What is a metadata? What are its contents?

Metadata contents:

o Description of the structure of the warehouse

o Operational meta-data

o The algorithms used for summarization

o The mapping from operational environment to the data warehouse

o Data related to system performance

o Business data

b) What are the Data Warehouse Back-End Tools and Utilities? Describe it.

- Data extraction:
- Data cleaning:
- Data transformation:
- Load:

UNIT V

PART-A

1)What is Precision?

It is percentage of retrieved documents that are relevant to the query.

2) What is Recall?

It is percentage of documents that are relevant to the query and retrieved

3) What are the different methods of IR?

1) Keyword based IR

2) Similarity based IR

4) What is synonymy problem?

A keyword may not appear in the document even if it is closely related to it.

5) What is polysemy problem?

The same keyword may mean different things in different context.

6) What is term frequency?

It is no of occurrences of the term in the document

7) What is relative term frequency?

This is term frequency versus the total no of occurrences of all the terms in the document

8) Give the formula for consine measure?

Sim(v1,v2)=V1.V2/ |v1||v2|

9) What is latent Singular value decomposition?

SVD is a technique in matrix theory to reduce the size of the term frequency matrix. It removes rows and columns to reduce the matrix to size KXK, where K is usually taken to be around a few hundred for large document collections.

10) What is inverted index?

An inverted index is an index structure that maintains two hash indexed or B+-tree indexed tables.

11) What is keyword based association analysis?

It collects sets of keywords that occur frequently together and then finds the association or correlation relationship between them.

12) What are the challenges in mining the WWW?

a) It is too he for effective data warehousing and mining

b) Complexity of WebPages is greater than of any traditional text document collection

c) Web is a highly dynamic information source

d) Web serves a broad diversity of user communities

e) Only a small portion of the information on the Web is truly relevant or useful

13) What are the Web Mining tasks?

a) Web Structure Mining

b) web Usage Mining

14) Give the factors to consider in choosing a Data Mining System?

Data-Types, System Issues, Data Sources, Data Mining Functions and methodologies, Coupling data mining with database and data warehouse systems

15) Give examples of Commercial Data Mining Systems

Intelligent Miner, enterprise Miner, MineSet,Clementine,DBMiner

16) Differentiate Decision Tree and regression trees?

It is similar to decision tree except for at the leaf level the mean of the objective attribute is computed and used as predicted value in case of regression tree instead of majority voting.

17) Differentiate between 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 inducing deduction rules, integrity constraints, generalized rules, frequent patterns etc

18) What do you mean by Direct Query Answering?

It means a query is answered by returning exactly what is being asked

19) What do you mean by Intelligent Query Answering?

It consists if analyzing the intent of the query and providing generalized, neighborhood or associated information relevant to the query.

20)Name the different phases in Data Mining Life Cycle of technology adoption

Innovation, Early adopters, Chasm, Early Majority, Late Majority, Laggards

21) Define CRM.

Customer Relationship Management helps companies provide more customized, personal service to their customer in lieu of mass marketing.

22) What are the principles of far information practices?

a) Purpose Specification and use limitation

b) Openness

23) Give some trend in data mining.

a) Application Exploration

b) Scalable data mining methods

c) Integration of data mining with database systems, data warehouse systems and Web database systems

d) Standardization of data mining language

e) Visual Data mining

24) Which are the tasks supported by DBMiner?

a) OLAP Analyzer

b) Association

c) Classification

d) Clustering

e) Prediction

f) Time-Series Analysis

25) What are the preprocessing tasks done by DBMiner?

Data Cleaning, Data Integration and Data Consolidation

PART B

1. Explain about mining the spatial databases?

Key points:

Spatial data cube construction and spatial OLAP-spatial association analysis-spatial clustering method-spatial classification-spatial trend analysis –mining raster database

2. How to mining the text databases?

Text data analysis and informational retrieval-text mining: keyword based association and document classification

3. How to mining the World Wide Web?

Mining the web link-automatic classification of web documents-construction of a multilayered web information base-web usage mining

4. Explain In detail about data mining applications.

Data mining for biomedical and DNA data analysis-Financial data analysis-Retail industry-telecommunication industry

5. How to choose an data mining system. Explain with suitable example.

Data types-system issues-data sources-function and methodologies-scalability- visualization tools

6. What are the themes used in data mining. Explain briefly.

Visual and audio data mining-scientific and statistical, foundation of data mining-data mining and intelligent query

7. What are the social impact of data mining? Explain in detail.

Data mining –hype or persistent-threat to privacy and data security

8. Explain in detail about DB miner architecture and task supported by the system.

System architecture-Input and Output-Tasks