1. Inroduction to Data Mininig

1.1 Introduction –Universe of Data

Information Technology has grown in various directions in the recent years. One natural evolutionary path has been the development of the database industry and its functionalities. Data collection, data creation, data management (including its storage and retrieval, database transaction processing) and data analysis and data understanding (involving data warehousing and data mining) has been the way in which it has progressed so far.

Lot of information is produced in this world. So much information is sometimes produced by international organizations that are difficult to read them even through a life time. Information explosion has caused a problem in many fields from medicine, to manufacturing to market.

Relational database technology allowed users convenient data access through query processing and transaction management. Methods for efficient **on line transaction processing (OLTP)** has become a major tool for storage, retrieval and management of large volumes of data. Today, the early relational model of sixties has expanded and extended to include object oriented, object, and relational, spatial, temporal, active and scientific database. The World Wide Web has emerged as a heterogeneous database system comprising of various data models.

Data can now store in a variety of ways. One possibility is the data warehouse, which is usually a repository of multiple heterogeneous sources of data. The data are normally organized with a schema, with facilitates management decision-making.

Online Analytical Processing (OLAP) is used to summarize, aggregate and consolidate data and it facilitates the view of data from different dimensions.

1.2 What is Data Mining?

Data Mining (also known as Knowledge Discovery Databases-KDD) has been defined as non trivial extraction of implicit, previously unknown, potentially useful information from data. it uses machine learning ,statistical and visualization techniques to discover and present knowledge in a form, which is easily comprehensible to humans.

Data Mining is the process of exploration and analysis, by automatic or semi automatic means, large quantities of data in order to discover meaningful patterns and rules. Meaningful patterns and rules ought to be useful-without automation it is impossible to mine large volumes of data and the same time automatic techniques by themselves enough for exploration and analysis.

Data Mining process or steps in the knowledge discovery process

1. **Data Cleaning-**The removal of noise and inconsistent data.
2. **Data Integration-**The combination of multiple sources of data.
3. **Data Selection-**The data relevant for analysis is retrieved from the database.
4. **Data Transformation-**The consolidation and transformation of data into forms appropriate for mining e.g., by performing aggregation of summary of data.
5. **Data Mining-**By use of intelligent patterns from data.
6. **Pattern Evaluation-**Identification of patterns that is interesting.
7. **Knowledge Presentation-**Visualization and Knowledge representation techniques are used to present the extracted or mined knowledge to the end user.

1.3 Components/Architecture of Data Mining

- **Databases, data warehouses or other repository information-**A set of databases such as data warehouses, spreadsheets and other kinds of information repositories where data cleaning and integration techniques may be employed.

- **Databases or data warehouses server:** Fetching data based on user's request from a data warehouse.
- **Knowledge Base-**The domain knowledge employed for finding interesting patterns.
- **Data Mining Engines-**The functional modules that are used to perform tasks such as classification, association, clusters analysis etc.
- **Pattern Evolution Module-**Interestingness measures are used to focus search towards interesting patterns.
- **Graphical User Interface-**This module interfaces the end user and the data mining System, allowing users to interact with the system by specifying a data mining task or a query through a graphical interface.

1.4 On what Kinds of Data

Relational Databases

A relational database is a collection of data tables, each of which is assigned a name, which is unique. Every table consists of a set of attributes (also known as columns or fields) and a set of tuples (known as records or rows). Each tuple represents an object identified by a unique key and a set of attributes.Relational databases can be accessed by queries written in relational query languages such as SQL.

Data warehouses

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

Transactional Databases

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number and a list of items making the transactions. The transactional databases may have additional tables associated with it, which contains other information regarding sales person, the branch, the branch at which the sale occurred etc.

Object Oriented Databases

The databases are based on the object oriented programming paradigm where each

Entity is considered as an object. Each object has associated with it following:

a) A set of variables that describe the object which correspond to attributes in an entity relationship diagram.
b) A set of messages which objects can use to communicate with other objects.
c) A set of methods which holds the code to implement a message-upon receiving a message the method return a value in response e.g., get_photo (student) returns the photograph of the student stored in the object database .

Objects that share common properties are grouped together and are called as classes. Then, each object becomes an instance of class. Objects can be organized into class/subclass hierarchy.

Object Relational Databases

The object-relational model is the basis for constricting object relational databases. This is an extension of the relational model so that a model is provided for handling rich data type or complex objects and object orientation.

Spatial Databases

Spatial databases contain spatially related information. Such databases include geographical terrain maps, VLSI chip designs, medical and satellite image databases. Spatial data may be represented in raster format consisting of n-dimensional bitmaps or pixel maps or else in the vector format.

Temporal Databases and Time series Databases

Time related data are stored in a temporal database where the data is stored in relational database. The attributes may have several timestamps each having different semantic interpretation. A time-series database stores a sequence of data that changes with time.

Text Databases or Multimedia Databases

Where word descriptions for objects are stored it is called as text database. Long sentences, paragraphs like product specifications, error or bug reports, warning messages summary reports etc. constitute the elements of a text database.

Multimedia databases store audio, video and image data. Multimedia databases are rather large in size since gigabytes of data are needed to store video images. Unlike text searches you need specialized search techniques. The storage and search techniques used for multimedia need to be integrated with other data mining techniques.

Heterogeneous Databases and Legacy Databases

A legacy database is one that contains historical data from organizations. The legacy databases are usually heterogeneous databases i.e. database components may greatly differ making it difficult to integrate their semantics together. Different kinds of data systems may be combined such as relational or object oriented database, hierarchical databases, network database, spreadsheets and multimedia databases. It is very difficult to exchange data from one database to another since there are no precise rules to transform one data format to another.

The World Wide Web

The World Wide Web provides interactive access to users and other information services .Users normally surf through the net following one link to another. Capturing user access patterns in such distributed environments may help a more efficient system design and this process is called mining path traversal patterns. Although, it is very difficult to structure and predefine schema, type and pattern of web pages for nearly impossible for the computer to semantically interpret the knowledge of the web.

Data Mining Functionalities-What kinds of patterns can be mined?

Data mining functionalities are used to specify the patterns to be found in data mining tacks. Data mining tasks may be classified into two categories those that are descriptive and those that are prescriptive. The general properties of a database fall under the descriptive category. The prescriptive category data mining tasks perform inference on the data in order to be able to make accurate predictions.

1.5 Concepts/Class Description: Characterization and Discrimination

1. **Data characterization**- summarizing the data of the class under study (target class). Summarization of the general characteristics or features (typically collected by an SQL query)- OLAP rollup operation and OLAP drill down operation is user controlled data summarization along a specified dimension and attribute oriented induction. The output of characterized data cubes and multidimensional tables- may be expressed in rule form (characteristic rules).
2. **Data description**- comparing the general features of the target class with one or a set of other contrasting classes. Users may specify the target class. The forms of output are similar to the characteristic descriptions.
3. **Both data characterization and discrimination**- Well-defined classes, pre classified examples (training set) may be used as a model that can be used to classify unclassified data.

Examples:

1. Assigning keywords to articles as they come in off the news wire.
2. Classifying credit applicants as medium, high or low risk.
3. Assigning customers to predefined customer segments.

1.6 Association Analysis

Association analysis is the discovery of association rules showing attribute value conditions that frequently occur together in a given set of data. Normally association rules used for market basket or transaction data analysis.

Example:

Marketing managers are fond of rules like:90% of the women with red spots car and small dogs wear channel No.5

These rules may be represented as

Women (X, “red spots car”) ^ dog_owners(X, “small dogs”)=>buys(X, “Channel No.5 “)

[Support =3% Confidence =70%]

The rule indicates that of the Wear House customers 3% of the customers were women with red sports car with small dogs wearing there 70% probability that a customer of this type will buy Channel No.5. Normally attributes such as a sports car owner are called as a dimension, and the rule is treated as **multidimensional association rule**.

1.7 Review questions

1) Explain about Universe of data?
2) Expalin About architecture and components of Data Minining?
3) Explian about various types of data bases?
4) Explain about Association analysis?
5) Defferentiate Heterogeneous Databases and Legacy Databases?

1.8 References

[1]. Data Mining Techniques, Arun k pujari 1st Edition

[2] .Data warehousung,Data Mining and OLAP, Alex Berson ,smith.j. Stephen

[3].Data Mining Concepts and Techniques ,Jiawei Han and MichelineKamber

[4]Data Mining Introductory and Advanced topics, Margaret H Dunham PEA

[5] The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition