DATA MINING & DATA WAREHOUSING

Prof. Kalyani Deshpande
Dr. D.Y. Patil ACS College, Pimpri, Pune-18
INTRODUCTION TO DATA MINING

- **Definition:-**
  Data mining refers to extracting knowledge from large amount of data.
  or
  It is the process of discovering or mining knowledge from a large amount of data.

- **Need of Data Mining :-**
  It was introduced in 1990.
  Need comes from evolution in size of database.

  
  \[ \text{db} \uparrow [\text{big data}] \rightarrow \text{manual analyses} \times \rightarrow \text{need of automatic analysis} \]
**Basic Data Mining Tasks**

Data Mining Tasks

- Descriptive
  - Class/concept Description
  - Mining of Frequent Pattern
  - Mining of Associations
  - Mining of Correlations
  - Mining of Clusters

- Characterization
  - Data Discrimination

- Classification & Prediction
  - Classification
  - Decision Trees
  - Mathematical Formulae
  - Neural Networks

Frequent item set
Frequent subsequence
Frequent sub structure
DESCRIPTIVE FUNCTION

- Descriptive Function:-
  The descriptive function deals with the general properties of data in the database.
  1. Class/Concept Description
  Class/Concept refers to the data to be associated with the classes or concepts.
  For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways –
  - Data Characterization – This refers to summarizing data of class under study. This class under study is called as Target Class.
  - Data Discrimination – It refers to the mapping or classification of a class with some predefined group or class.
2. Mining of Frequent Patterns

- Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns –

- **Frequent Item Set** – It refers to a set of items that frequently appear together, for example, milk and bread.

- **Frequent Subsequence** – A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.

- **Frequent Sub Structure** – Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.
3. Mining of Association

- Associations are used in retail sales to identify patterns that are frequently purchased together.
- This process refers to the process of uncovering the relationship among data and determining association rules.
- For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.
4. Mining of Correlations

- It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

5. Mining of Clusters

- Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.
CLASSIFICATION AND PREDICTION

- Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms –
  - Classification (IF-THEN) Rules
  - Decision Trees
  - Mathematical Formulae
  - Neural Networks

The list of functions involved in these processes are as follows –
- **Classification** – It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
- **Prediction** – It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
- **Outlier Analysis** – Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
- **Evolution Analysis** – Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.
DM Versus Knowledge Discovery in Databases

- Data Mining v. Knowledge Discovery in Databases (KDD)
  - DM and KDD are often used interchangeably
  - actually, DM is only part of the KDD process

- The Knowledge Discovery Process
  - Data Mining
    - Preprocessing
      - Selection
    - Transformation
      - Target Data
      - Preprocessed Data
      - Transformed Data
    - Knowledge
      - Patterns
  - The KDD Process
What is DM & KDD?

- **DM**: “The process of identifying hidden patterns & relationships within data.”

- **KDD**: “Data mining helps & users extract useful business information from large databases.”
Some people don’t differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process −

- **Data Cleaning** – In this step, the noise and inconsistent data is removed.
- **Data Integration** – In this step, multiple data sources are combined.
- **Data Selection** – In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** – In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** – In this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** – In this step, data patterns are evaluated.
- **Knowledge Presentation** – In this step, knowledge is represented.
DATA MINING – SYSTEMS

There is a large variety of data mining systems available. Data mining systems may integrate techniques from the following –

- Spatial Data Analysis
- Information Retrieval
- Pattern Recognition
- Image Analysis
- Signal Processing
- Computer Graphics
- Web Technology
- Business
- Bioinformatics
Data Mining System Classification

A data mining system can be classified according to the following criteria –

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines
Apart from these, a data mining system can also be classified based on the kind of
(a) databases mined,
(b) knowledge mined,
(c) techniques utilized
(d) applications adapted.

Classification Based on the Databases Mined
We can classify a data mining system according to the kind of databases mined. Database system can be classified according to different criteria such as data models, types of data, etc. And the data mining system can be classified accordingly.

For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.
Classification Based on the kind of Knowledge Mined

We can classify a data mining system according to the kind of knowledge mined. It means the data mining system is classified on the basis of functionalities such as –

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Outlier Analysis
- Evolution Analysis
Classification Based on the Techniques Utilized
We can classify a data mining system according to the kind of techniques used. We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.

Classification Based on the Applications Adapted
We can classify a data mining system according to the applications adapted. These applications are as follows –
- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail
Integrating a Data Mining System with a DB/DW System

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

The list of Integration Schemes is as follows –

- **No Coupling** – In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.

- **Loose Coupling** – In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data. It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.

- **Semi-tight Coupling** – In this scheme, the data mining system is linked with a database or a data warehouse system and in addition to that, efficient implementations of a few data mining primitives can be provided in the database.

- **Tight coupling** – In this coupling scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.
DATA MINING ISSUES

Issues In Data Mining

- Mining Methodology & user interaction
- Performance issues
- Diverse data types issues
MINING METHODOLOGY & USER INTERACTION

- Mining different kind of knowledge in data base.
- Interactive mining of knowledge at multiple level of abstraction.
- Pattern evolution.
- Presentation & visualization of data mining result.
- Handling noisy or incomplete data.
PERFORMANCE ISSUES

- Efficiency & scalability of data mining algorithm.
- Parallel distributed & incremental mining algorithms.
DIVERSE DATA TYPES ISSUES

- Handling complex data.
- Mining information for different heterogeneous information system.
Data Mining Metrics

Data Mining has emerged at the confluence of artificial intelligence, statistics, and databases as a technique for automatically discovering summary knowledge in large datasets. Data mining first requires understanding the data available, developing questions to test, and finally drawing conclusions from data analytic results. Metrics are some parameters or measures of quantitative assessment used for measurement or comparison in a given context. A metric for all practical purpose is just a variable. It needs to be clearly defined. The number of metrics needs to be kept under control to ensure that the measuring task is achievable. It is thus reasonable to expect that as the context changes, the metrics would change. Literature has not defined Data mining metrics as such. Data mining metrics may be defined as a set of measurements which can help in determining the efficacy of a Data mining Method / Technique or Algorithm. They are important to help take the right decision as like as choosing the right data mining technique or algorithm.
Data mining metrics generally fall into the categories of accuracy, reliability, and usefulness.

Accuracy is a measure of how well the model correlates an outcome with the attributes in the data that has been provided. There are various measures of accuracy, but all measures of accuracy are dependent on the data that is used. In reality, values might be missing or approximate, or the data might have been changed by multiple processes. Particularly in the phase of exploration and development, we might decide to accept a certain amount of error in the data, especially if the data is fairly uniform in its characteristics. For example, a model that predicts sales for a particular store based on past sales can be strongly correlated and very accurate, even if that store consistently used the wrong accounting method. Therefore, measurements of accuracy must be balanced by assessments of reliability.

Reliability assesses the way that a data mining model performs on different data sets. A data mining model is reliable if it generates the same type of predictions or finds the same general kinds of patterns regardless of the test data that is supplied. For example, the model that we generate for the store that used the wrong accounting method would not generalize well to other stores, and therefore would not be reliable.

Usefulness includes various metrics that tell us whether the model provides useful information. For example, a data mining model that correlates store location with sales might be both accurate and reliable, but might not be useful, because you cannot generalize that result by adding more stores at the same location. Moreover, it does not answer the fundamental business question of why certain locations have more sales. We might also find that a model that appears successful in fact is meaningless, because it is based on cross-correlations in the data.
OVERVIEW OF APPLICATIONS OF DATA MINING

 Customer segmentation: Business use data mining techniques to understand customer.
 Market basket analysis: Retail out of 10 customers, if 8 are purchasing both, then both will be kept near.
 Risk management: Insurance companies uncover risks associated with potential customers.
 Fraud Detection: Credit card companies detect abnormal spending.
 Demand Prediction: Retail & online demand of products.
Thank You