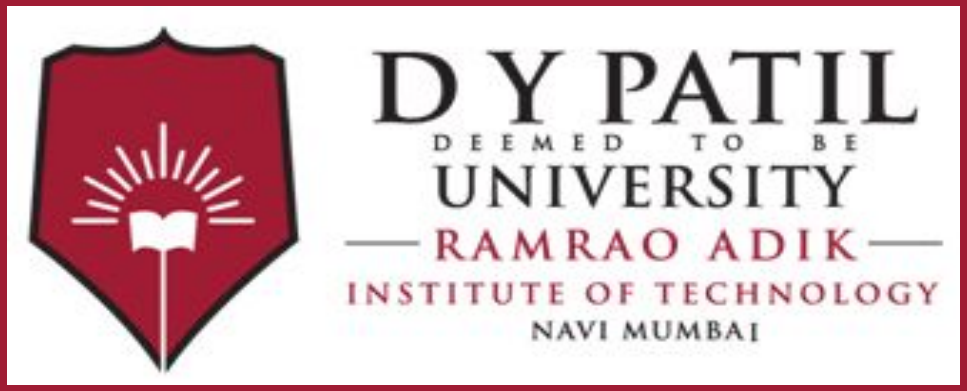


Subject Name: Data Warehousing and Mining

Unit No:2

Unit Name: Data Mining and Data Pre-processing

INDEX

Lecture No: 1

What is Data Mining; Knowledge Discovery in Database (KDD)

INTRODUCTION

Motivation: Why data mining?

What is data mining?

Data Mining: On what kind of data?

Data mining functionality

Are all the patterns interesting?

Classification of data mining systems

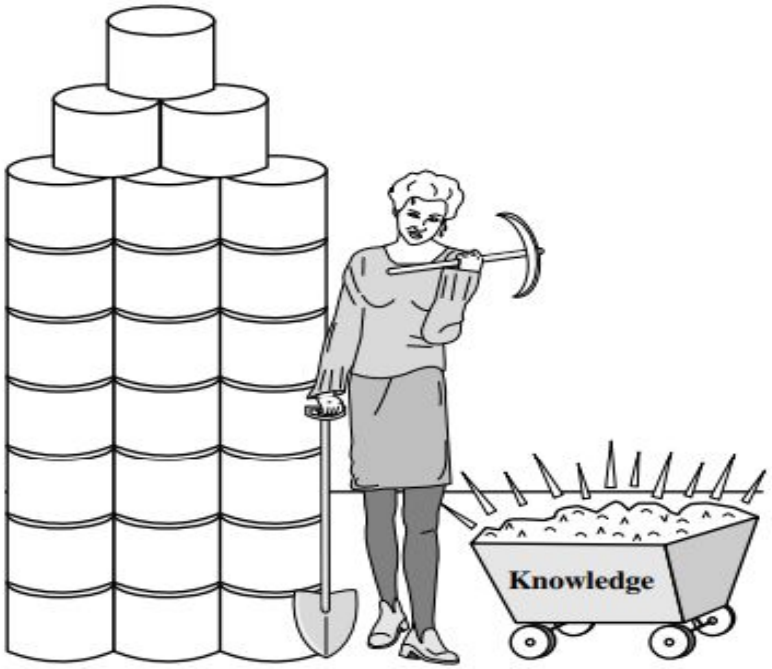
Major issues in data mining

WHY DATA MINING

world is data rich but information poor.

- Data mining is the process of searching and analyzing a large batch of raw data in order to **identify patterns and extract useful information**. Companies use data mining software to learn more about their customers. It can help them to **develop more effective marketing strategies, increase sales, and decrease costs**.
- Data mining is used to explore large data volumes to find patterns and insights that can be used for specific purposes. These purposes might include improving sales and marketing, optimizing manufacturing, detecting fraud, and enhancing security.

WHAT IS DATA MINING

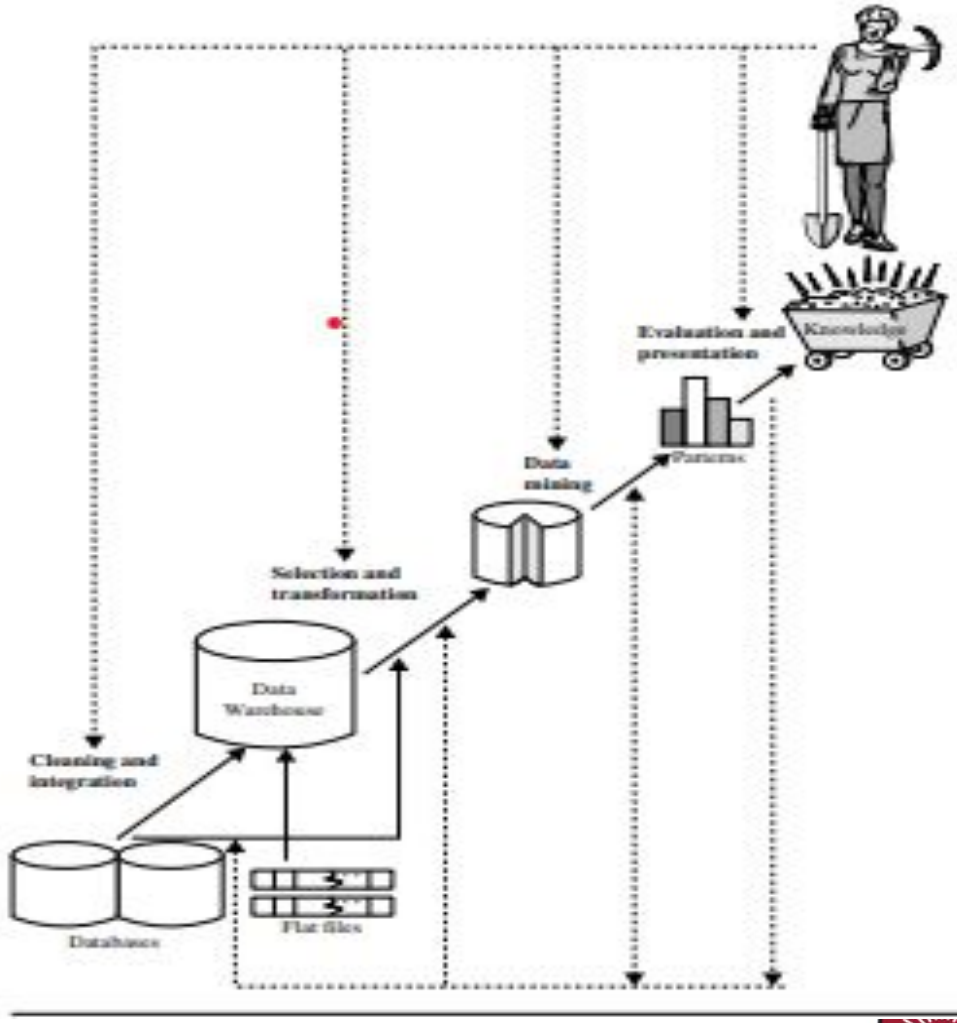


Data mining—searching for knowledge (interesting patterns) in data.

- Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information.
- Companies use data mining software to learn more about their customers.
- It can help them to develop more effective marketing strategies, increase sales, and decrease costs.

Data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD.

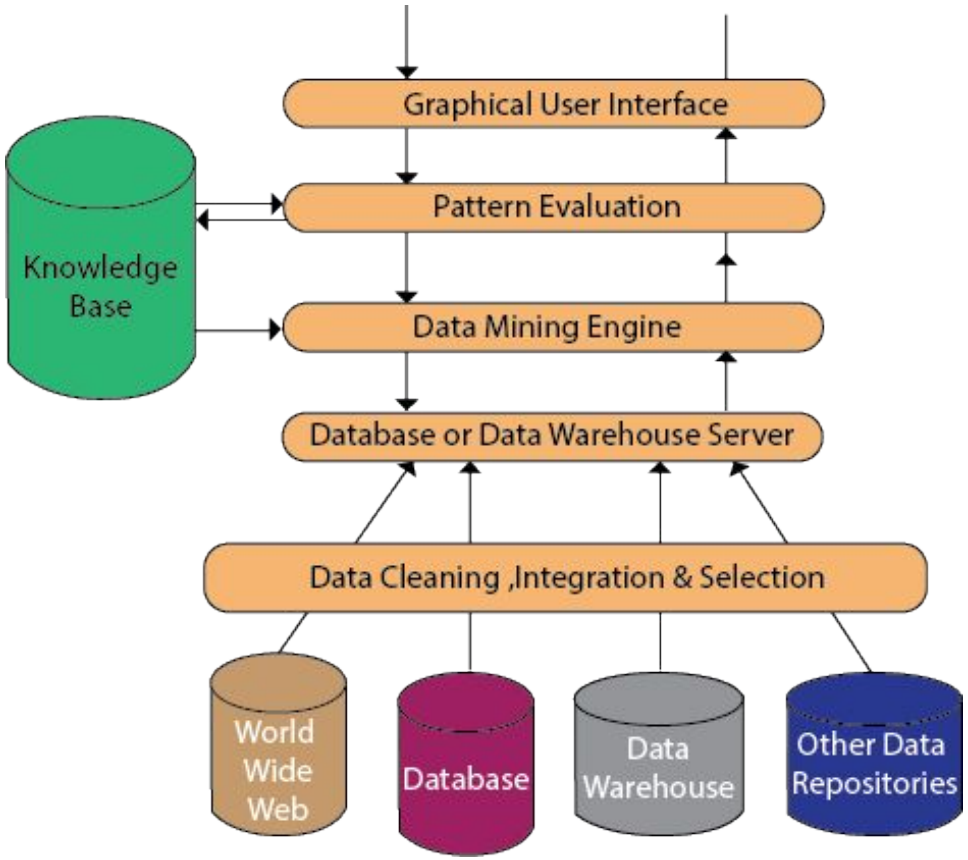
KDD PROCESS : ITERATIVE SEQUENCE OF THE FOLLOWING STEPS:



STEPS OF KDD

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

DATA MINING ARCHITECTURE



ARCHITECTURE

Data Sources

- Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data.

- Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories.

Different Processes

- first data needs to be cleaned and integrated.

- more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server.

ARCHITECTURE

Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the **server is responsible** for retrieving the relevant data based on the data mining request of the user.

Data Mining Engine

The data mining engine is the **core component** of any data mining system. It consists of a number of modules for performing **data mining tasks** including association, classification, characterization, clustering, prediction, time-series analysis etc.

Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the measure of **interestingness of the pattern** by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

ARCHITECTURE

Graphical User Interface

- The graphical user interface module **communicates between the user and the data mining system.**
- This module helps the user use the system easily and efficiently **without knowing the real complexity behind the process**.
- When the user specifies a query or a task, this module **interacts** with the **data mining system and displays the result** in an easily understandable manner.

ARCHITECTURE

Knowledge Base

- The knowledge base is **helpful in the whole data mining process**.
- It might be useful for guiding the **search or evaluating** the interestingness of the result patterns.
- The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining.
- The data mining engine might get inputs from the knowledge base to make the **result more accurate and reliable**.
- The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

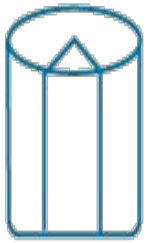
- ***NOTE*** *– Student can take any application and design a model as per the steps of Data Mining Architecture.*

DATA MINING TASK PRIMITIVES

- A data mining task can be specified in the form of a data mining query, which is input to the data mining system.
- A data mining query is defined in terms of data mining task primitives.
- These primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process or examine the findings from different angles or depths.

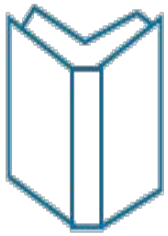
1. **Set of task-relevant data to be mined.**
2. **Kind of knowledge to be mined.**
3. **Background knowledge to be used in the discovery process.**
4. **Interestingness measures and thresholds for pattern evaluation.**
5. **Representation for visualizing the discovered patterns.**

DATA MINING TASK PRIMITIVES



Task-relevant data
Database or data warehouse name Database tables or data warehouse cubes Conditions for data selection Relevant attributes or dimensions
Data grouping criteria

Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering



Background knowledge
Concept hierarchies
User beliefs about relationships
in the data

Pattern interestingness measures
simplicity
Certainty (e.g. condifence)
Utility (e.g. support)
Novelty

Visualization of discovered patterns
Rules, tables, reports, charts, graphs , decision trees, and cubes
Drill down and roll-up

DATA MINING TASK PRIMITIVES

1. **The set of task-relevant data to be mined**

- This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (the relevant attributes or dimensions).
- In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join, and aggregation.
- The data collection process results in a new data relational called the ***initial data relation***. The initial data relation can be ordered or grouped according to the conditions specified in the query.

DATA MINING TASK PRIMITIVES

2. The kind of knowledge to be mined

- This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

3. The background knowledge to be used in the discovery process

- This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found. **Concept hierarchies** are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction.
- Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level, more general concepts.

Rolling Up - Generalization of data, Drilling Down - Specialization of data

DATA MINING TASK PRIMITIVES

4. The interestingness measures and thresholds for pattern evaluation

- Different kinds of knowledge may have different interesting measures. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns.

MAJOR ISSUES IN DATA MINING (1)

Mining methodology and user interaction

- 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
- Expression and visualization of data mining results
- Handling noise and incomplete data
- Pattern evaluation: the interestingness problem

Performance and scalability

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

MAJOR ISSUES IN DATA MINING (2)

Issues relating to the diversity of data types

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

Issues related to applications and social impacts

- Application of discovered knowledge
  - Domain-specific data mining tools
  - Intelligent query answering
  - Process control and decision making
- Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
- Protection of data security, integrity, and privacy

MAJOR ISSUES IN DATA MINING (2)

Issues relating to the diversity of data types

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

Issues related to applications and social impacts

- Application of discovered knowledge
  - Domain-specific data mining tools
  - Intelligent query answering
  - Process control and decision making
- Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
- Protection of data security, integrity, and privacy

D Y PATIL DEEMED TO BE UNIVERSITY RAMRAO ADIK INSTITUTE OF TECHNOLOGY NAVI MUMBAI

WHAT IS AN ATTRIBUTE?

An attribute is a data field, representing a **characteristic** or **feature** of a data object.

The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature.

- The term **dimension** is commonly used in **data warehousing**.
- **Machine learning** literature tends to use the term **feature**,
- **statisticians** prefer the term **variable**.
- **Data mining** and database professionals commonly use the term **attribute**

Types of Attribute

1. **Nominal**
2. **Binary**
3. **Ordinal**
4. **Numeric**

NOMINAL ATTRIBUTE

- The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical.

Example:

hair color : black, brown, blond, red, gray, and white.

Marital status : single, married, divorced, and widowed.

Occupation: teacher, dentist, programmer, farmer, and so on.

nominal attribute values do not have any meaningful order about them and are **not quantitative**, it makes no sense to find the mean (average) value or median (middle) value for such an attribute.

BINARY ATTRIBUTES

- A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present.
- Binary attributes are referred to as Boolean if the two states correspond to true and false.

Example:

Patient undergoes a medical test that has two possible outcomes. The attribute medical test is binary, where a value of 1 means the result of the test for the patient is positive, while 0 means the result is negative.

BINARY ATTRIBUTE

- Symmetric: A binary attribute is **symmetric** if both of its states are equally valuable and carry the same weight. Eg: Gender
- Asymmetric: A binary attribute is **asymmetric** if the outcomes of the states are not equally important, such as the positive and negative outcomes of a medical test for Typhoid.

ORDINAL ATTRIBUTES

- An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude **(exact differences or distances between those values)** between successive values is not known.

Example: grade : A+, A, A−, B+

Professional ranks: Assistant, Associate, Professor

Customer satisfaction had the following ordinal categories: 0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied.

Note :

Nominal, binary, and ordinal attributes are qualitative. They describe a feature of an object without giving an actual size or quantity. The values of such qualitative attributes are typically words representing categories.

NUMERIC ATTRIBUTES

- A numeric attribute is **quantitative;** that is, it is a **measurable quantity**, represented in **integer or real values**.
- Numeric attributes can be

1. interval-scaled
2. ratio-scaled.

INTERVAL SCALED

- Measured (thermometer0
- Ordered
- Equidistant
- It doesn’t have any meaningful zero
- Interval can be negative
- Grouping, Sorting, Arithmetic (+,-) operations can be done.
- We cannot multiply or divide the interval data as there is no meaningful zero.
- Central point(mean, median, mode), Range, Spread of data(Std.Dev, variance)

RATIO SCALED

- Measured in the form of numbers.(measuring distance with help of any measuring device)
- It has rank and order (2km is always less than 5km)
- Equidistant: Equally spaced interval.
- Meaning zero(travelled 0km)
- Eg: Distance travelled, weight, age of person
- Grouping, sorting, arithmetic (+,-,*,/) **2 year child is half the age of 4 year child**
- Central point(mean, median, mode), Range, Spread of data(Std.Dev, variance)

DIFFERENCE BETWEEN INTERVAL AND RATIO SCALED ATTRIBUTE

Features	Interval scale	Ratio scale
Variable property	All variables measured in an interval scale can be added, subtracted, and multiplied. You cannot calculate a ratio between them.	Ratio scale has all the characteristics of an interval scale, in addition, to be able to calculate ratios. That is, you can leverage numbers on the scale against 0.
Absolute Point Zero	Zero-point in an interval scale is arbitrary. For example, the temperature can be below 0 degrees Celsius and into negative temperatures.	The ratio scale has an absolute zero or character of origin. Height and weight cannot be zero or below zero.
Calculation	Statistically, in an interval scale, the arithmetic mean is calculated.	Statistically, in a ratio scale, the geometric or harmonic mean is calculated.
Measurement	Interval scale can measure size and magnitude as multiple factors of a defined unit.	Ratio scale can measure size and magnitude as a factor of one defined unit in terms of another.
Example	A classic example of an interval scale is the temperature in Celsius. The difference in temperature between 50 degrees and 60 degrees is 10 degrees; this is the same difference between 70 degrees and 80 degrees.	Classic examples of a ratio scale are any variable that possesses an absolute zero characteristic, like age, weight, height, or sales figures.

DISCRETE VERSUS CONTINUOUS ATTRIBUTES

- A **discrete attribute** has a finite or countably infinite set of values, which may or may not be represented as integers. The attributes hair color, medical test, and drink size each have a finite number of values, and so are discrete.

Note that discrete attributes may have numeric values, such as 0 and 1 for binary attributes or, the values 0 to 110 for the attribute age.

- **Continuous attributes** come from an infinite set (i.e. real numbers, you can make them as large or small as you need). Discrete attributes come from a finite or countably infinite set (i.e. integers).
- Continuous attributes would be a floating-point type, where discrete would be integers or characters.

SUMMARY

Data mining: discovering interesting patterns from large amounts of data

A natural evolution of database technology, in great demand, with wide applications

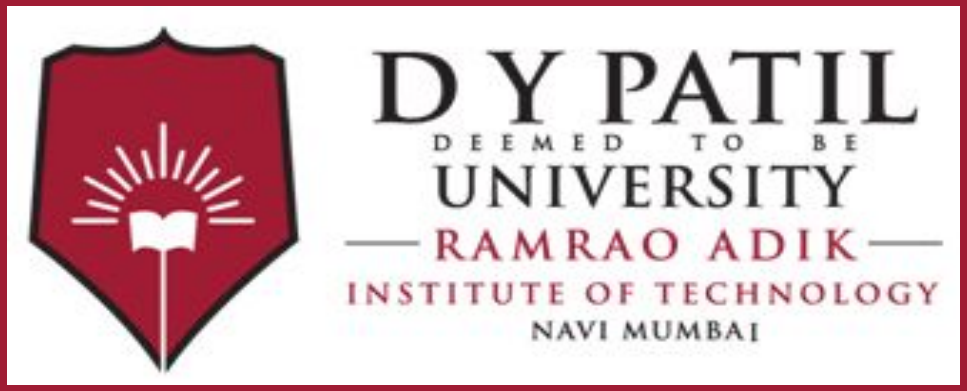
A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation

Mining can be performed in a variety of information repositories

Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

Classification of data mining systems

Major issues in data mining


D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Thank You


D Y PATIL
RAMRAO ADIK
INSTITUTE OF
TECHNOLOGY
NAVI MUMBAI