****

**Lecture Notes**

**Course Name:** Data Mining and Warehousing

**Course Code:** CSF– 334

**About Course**

* This course is introduced to 5th semester students of BE
* This course will develop understanding key concepts of data mining and obtain knowledge about how to extract useful characteristics from data using data pre-processing techniques
* It will demonstrate methods to apply and analyze relevant attributes, perform statistical measure to look for meaningful variation in data, and mine association rules for transactional datasets.

**Course Objectives**

1. Develop understanding key concepts of data mining and obtain knowledge about how to extract useful characteristics from data using data pre-processing techniques
2. Demonstrate methods to apply and analyze relevant attributes, perform statistical measure to look for meaningful variation in data, and mine association rules for transactional datasets.
3. Teach use and application of data mining techniques such as classification, decision tree, neural networks, back propagation and many more, in various applications.

**Course Outcomes**

|  |  |  |
| --- | --- | --- |
| **CO Number** | **Title** | **Level** |
| **CO1** | **Understand the basic concepts of Data Mining.** | **Understand & Remember** |
| **CO2** | **Understand warehousing architectures and tools for systematically organizing large database and use their data to make strategic decisions.** | **Understand & Remember** |
| **CO3** | **Identify the strengths and weaknesses of different methods of mining meaningful data from heap of raw data** | **Understand & Remember** |
| **CO4** | **Characterize the kinds of patterns that can be discovered by association rule mining.** | **Understand & Remember** |
| **CO5** | **Design and employ to correct data mining technique depending on the characteristics of the dataset.** | **Apply** |

**Introduction**

## DEFINITION OF DATA MINING?

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications

* Market Analysis
* Fraud Detection
* Customer Retention
* Production Control
* Science Exploration

### Major Sources of data: -

Business –Web, E-commerce, Transactions, Stocks - Science – Remote Sensing, Bio informatics, Scientific Simulation - Society and Everyone – News, Digital Cameras, You Tube \* Need for turning data into knowledge – Drowning in data, but starving for knowledge.

### Definition of Data Mining?

Extracting and ‘Mining’ knowledge from large amounts of data. “Gold Mining from rock or sand” is same as “Knowledge mining from data”

### Other terms for Data Mining:

* + Knowledge Mining
  + Knowledge Extraction o Pattern Analysis

# KNOWLEDGE DISCOVERY (KDD) PROCESS:

### Several Key Steps:

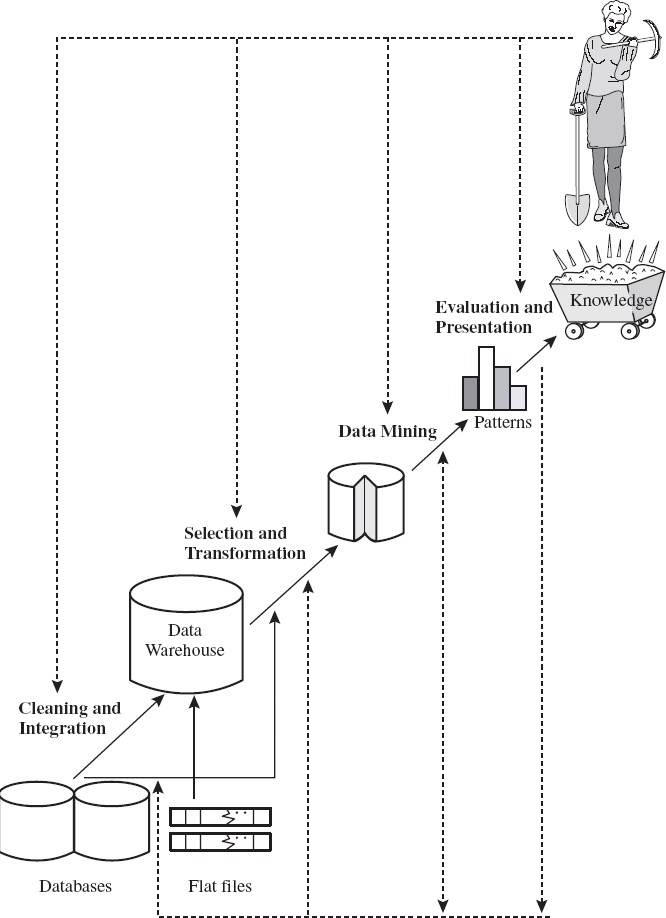
▶ Data processing

▶ **Data cleaning (**remove noise and inconsistent data)

▶ **Data integration** (multiple data sources maybe combined)

▶ **Data selection** (data relevant to the analysis task are retrieved from database)

**Data transformation** (data transformed or consolidated into forms)



▶ appropriate for mining)

(Done with data preprocessing)

▶ **Data mining** (an essential process where intelligent methods are applied to extract

data patterns)

▶ **Pattern evaluation** (identify the truly interesting patterns)

▶ **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)

# DATA MINING ON WHAT KIND OF DATA? ( TYPES OF DATA ):

## RELATIONAL DATABASES:

* + **A database system**, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
  + **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 *fields*) and usually stores a large set of tuples (*records* or *rows*).
  + Each tuple in a relational table represents an object identified by a unique *key* and described by a set of attribute values.
  + **A semantic data model**, such as an entity-relationship (ER) data model, is often constructed for relational databases.
  + An **ER data model** represents the database as a set of entities and their relationships.

# DATA MINING FUNCTIONALITIES— WHAT KINDS OF PATTERNS CAN BE MINED?:

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

## CONCEPT/CLASS DESCRIPTION: CHARACTERIZATION AND DISCRIMINATION:

* Data can be associated with classes or concepts.
* Example: *AllElectronics* store, classes of items for sale include *computers* and *printers*, and concepts of customers include ***bigSpenders*** and ***budgetSpenders*.**
* 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. These descriptions can be derived via
* ***data characterization***, by summarizing the data of the class under study (often called the target class) in general terms,
* ***data discrimination***, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or both data characterization and discrimination.

### Data characterization:

* + It 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 the output of data characterization can be presented in various forms.

**Examples** include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs.

### Data discrimination:

* It 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.
* The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.

## MINING FREQUENT PATTERNS, ASSOCIATIONS, AND CORRELATIONS:

**Frequent patterns**, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including item sets, sub sequences, and substructures.

**A *frequent itemset*** typically refers to a set of items that frequently appear together in a transactional data set, such as Computer and Software.

**Example:** Association analysis. Suppose, as a marketing manager of *AllElectronics*, you would like to determine which items are frequently purchased together within the same transactions.

**Example** of such a rule, mined from the *AllElectronics* transactional database, is ***buys*(*X*;―*computer*‖) *buys*(*X*; ―*software*‖) [*support* = 1%, *confidence* = 50%]**.

where *X* is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

# CLASSIFICATION AND PREDICTION:

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

#### “How is the derived model presented?”:

The derived model may be represented in various forms, such as ***classification (IF-THEN) rules***, *decision trees*, *mathematical formulae*, or *neural networks.*

**A decision tree** is a 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 easily be converted to classification rules.

**A neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and *k*-nearest neighbour classification.

Whereas classification predicts categorical (discrete, unordered) labels, prediction models Continuous-valued functions. That is, it is used to predict missing or unavailable *numerical data values* rather than class labels. Although the term *prediction* may refer to both numeric prediction and class label prediction,

### Cluster Analysis

Classification and prediction analyse class-labelled data objects, where as **clustering** analyzes data objects without consulting a known class label.

### 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. Most data mining methods discard outliers as noise or exceptions.

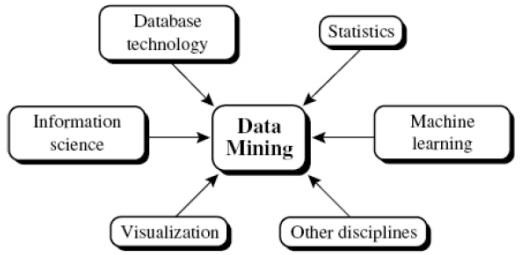
### Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of *time related* data.

## CLASSIFICATION OF DATA MINING SYSTEMS:

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science.

Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high- performance computing.



Data mining systems can be categorized according to various criteria, as follows:

**Classification according to the *kinds of databases* mined**:

A data mining system can be classified according to the kinds of databases mined. Database systems can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique.

**Classification according to the *kinds of knowledge* mined**:

Data mining systems can be categorized according to the kinds of knowledge they mine, that is, based on data mining functionalities, such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.

**Classification according to the *kinds of techniques* utilized**:

Data mining systems can be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved (e.g., autonomous systems, interactive exploratory systems, query-driven systems) or the methods of data analysis employed (e.g., database-oriented or data warehouse– oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks, and so on).

**Classification according to the *applications adapted*:**

Data mining systems can also be categorized according to the applications they adapt. **For example,** data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on. Different applications often require the integration of application-specific methods.

# DATA MINING TASK PRIMITIVES:

A data mining query is defined in terms of the following primitives:

### Task-relevant data:

This is the database portion to be investigated. For example, suppose that you are a manager of All Electronics in charge of sales in the United States and Canada. In particular, you would like to study the buying trends of customers in Canada. Rather than mining on the entire database. These are referred to as relevant attributes

### The kinds of knowledge to be mined:

This specifies the data mining functions to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis. For instance, if studying the buying habits of customers in Canada.

### Background knowledge:

Users can specify background knowledge, or knowledge about the domain to be mined. This knowledge is useful for guiding the knowledge discovery process, and for evaluating the patterns found. There are several kinds of background knowledge.

### Interestingness measures:

These functions are used to separate uninteresting patterns from knowledge. They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures.

### Presentation and visualization of discovered patterns:

This refers to the form in which discovered patterns are to be displayed. Users can choose from different forms for knowledge presentation, such as rules, tables, charts, graphs, decision trees, and cubes.

# MAJOR ISSUES IN DATA MINING:

**Mining different kinds of knowledge in databases.** - The need of different users is not the same. And Different user may be in interested in different kind of knowledge. Therefore it is necessary for data mining to cover broad range of knowledge discovery task.

**Interactive mining of knowledge at multiple levels of abstraction**. - The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on returned results.

**Incorporation of background knowledge.** - To guide discovery process and to express the discovered patterns, the background knowledge can be used.

Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple level of abstraction.

**Data mining query languages and ad hoc data mining**. - Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining. Presentation and visualization of data mining results. - Once the patterns are discovered it needs to be expressed in high level languages, visual representations. This representations should be easily understandable by the users.

**Handling noisy or incomplete data.** - The data cleaning methods are required that can handle the noise, incomplete objects while mining the data regularities. If data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

**Pattern evaluation. -** It refers to interestingness of the problem. The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

**Efficiency and scalability of data mining algorithms.** - In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

### Parallel, distributed, and incremental mining algorithms. –

The factors such as huge size of databases, wide distribution of data,and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithm divide the data into partitions which is further processed parallel. Then the results from the partitions is merged. The incremental algorithms, updates databases without having mine the data again from scratch.

**Suggestive Reading Material**

* **TEXT BOOKS**

**Introduction to Data Mining, Tan, Steinbach and Vipin Kumar, Pearson Education, 2016**

* **REFERENCE BOOKS**

**Data Mining: Concepts and Techniques, Pei, Han and Kamber, Elsevier**

* **Journals:**
  + [**http://www.ijsmsjournal.org/ijsms-v4i4p137.html**](http://www.ijsmsjournal.org/ijsms-v4i4p137.html)
  + **https://www.springer.com/journal/41060**