Name of Faculty: Prof. Puneet Nema
Designation: Assistant Professor

Department: CSE

Subject: Data mining

Unit: III

Topic: Introduction to Data Mining, Knowledge Discovery, Data Mining Functionalities, Data Mining System categorization and its Issues. Data Processing :- Data Cleaning, Data Integration and Transformation. Data Reduction, Data Mining Statistics. Guidelines for Successful Data Mining..

RAJIV GANDHI PROUDYOGIKI VISHWAVIDYALAYA, BHOPAL

New Scheme Based On AICTE Flexible Curricula

Computer Science and Engineering,VIII-Semester

CS-8203 Data Mining

UNIT-III

Topic Covered: Data Mining

Introduction to Data Mining, Knowledge Discovery, Data Mining Functionalities, Data Mining System Categorization and its Issues Data Processing :-Data Cleaning, Data Integration and Transformation. Data Reduction, Data Mining Statistics.

Introduction to Data Mining :

Data Mining is a set of method that applies to large and complex databases. This is to eliminate the randomness and discover the hidden pattern. As these *data mining methods* are almost always computationally intensive. We use**s** methodologies, and theories for revealing patterns in data. There are too many driving forces present. And, this is the reason why data mining has become such an important area of study.

We use techniques for a long process of research and product development. As this evolution was started when business data was first stored on computers. Also, it allows users to navigate through their data in real time. We use data mining in the business community because it is.

supported by three technologies that are now mature:

- Massive data collection
- Powerful multiprocessor computer
- Data mining concept

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.

The information or Knowledge extracted so can be used for any of the following application.

- Market Analysis
- Fraud Detection
- Coustmer Retention.

Knowledge Discovery:

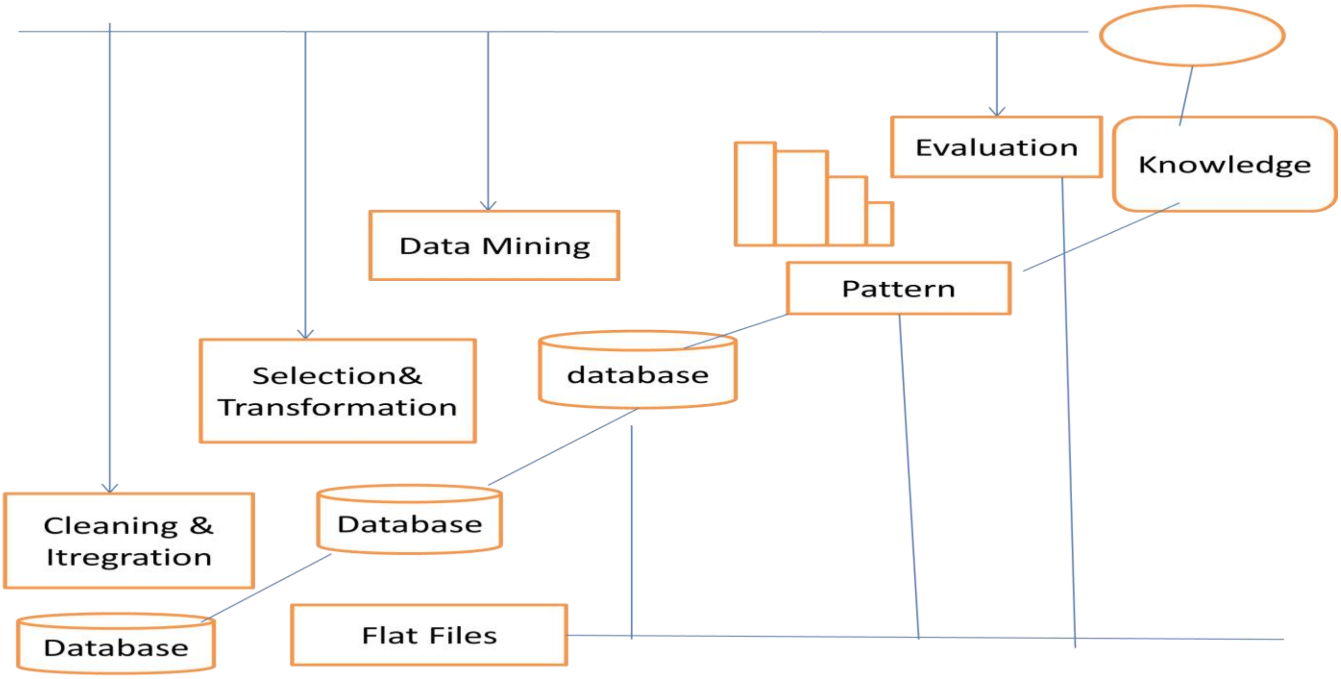
The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machin learning , pattern recognition, databases, statistics, artificial intelligence.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

Some essential steps in the process of knowledge discovery are as follows:

- **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.

The following diagram shows the process of knowledge discovery:



The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding Domain
   - the application domain

   - the relevant prior knowledge
   - the goals of the end-user

2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

3. Data cleaning and preprocessing.
   - Removal of noise or outliers.
   - Collecting necessary information to model or account for noise.
   - Strategies for handling missing data fields.
   - Accounting for time sequence information and known changes.

4. Data reduction and projection.
   - Finding useful features to represent the data depending on the goal of the task.
   - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

5. Choosing the data mining task
   - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

6. Choosing the data mining algorithm(s).
   - Selecting method(s) to be used for searching for patterns in the data.
   - Deciding which models and parameters may be appropriate.
   - Matching a particular data mining method with the overall criteria of the KDD process.

7. Data mining.
   - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.

8. Interpreting mined patterns.

9. Consolidating discovered knowledge.

improved without altering the external behavior or code design.

Data Mining Funtionalities:

Data mining principles are concerned with providing means to handle the complexity of the design process effectively. Effectively managing the complexity will not only reduce the effort needed for design but can also reduce the scope of introducing errors during design.

Functions are used to define the trends or correlations contained in data mining activities.

In comparison, data mining **activities** can be divided into 2 categories:

1. **Descriptive Data Mining:**
   It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set.
   For examples: count, average etc.
2. **Predictive Data Mining:**
   It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the software estimates the characteristics that are absent.
   For example: Judging from the findings of a patient's medical examinations that is he suffering from any particular disease.

Data Mining Functionality:

For small problem, we can handle the entire problem at once but for the significant problemdivide the problems and conquer the problem it means to divide the problem into smaller pieces so that each piece can be captured separately. These class or concept definitions are referred to as class/concept descriptions.

1. Class/Concept Descriptions:

Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts.

These class or concept definitions are referred to as class/concept descriptions:

Data Characterization:

This refers to the summary of general characteristics or features of the class that is under the study. For example. To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these type of data related to such products by running SQL queries,

t compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

Data Discrimination:

It compares common features of class which is under study. The output of this process can be represented in many forms. Eg., bar charts, curves and pie charts.

2. Mining Frequent Patterns, Associations, and Correlations:

Frequent patterns are nothing but things that are found to be most common in the data.

There are different kinds of frequency that can be observed in the dataset.

Frequent item set:
This applies to a number of items that can be seen together regularly for eg: milk and sugar.

Frequent Subsequence:

This refers to the pattern series that often occurs regularly such as purchasing a phone followed by a back cover.

Frequent Substructure:
It refers to the different kinds of data structures such as trees and graphs that may be combined with the itemset or subsequence.

Association Analysis:
The process involves uncovering the relationship between data and deciding the rules of the Association. It is way of discovering the relationship between various items. for example, it can be used to determine the set.

Correlation Analysis:
Correlation is a mathematical technique that can show whether and how strongly the pairs of attributes are related to each other. For example, Highted people tend to have more weight.

Data Mining System Categorization and issues:

Data Mining is the division of software into separate modules which are differently named and addressed and are integrated later on in to obtain the completely functional software. It is the only property that allows a program to be intellectually manageable. Single large programs are difficult to understand and read due to a large number of reference variables, control paths, global variables, etc.

The desirable properties of a Categorization system are:

- Each category is a well-defined system that can be used with other applications.
- Each module has single specified objectives.
- Modules can be separately compiled and saved in the library.

Data Mining Issues:

It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to on the basis of a training set of data containing observations and whose categories membership is known.

Data mining systems face a lot of challenges and issues in today's world some of them are:

1 **Mining methodology and user interaction issues:**

- It allows large programs to be written by several or different people
- It encourages the creation of commonly used routines to be placed in the library and used by other programs.
- It simplifies the overlay procedure of loading a large program into main storage.
- It provides more checkpoints to measure progress.
- It provides a framework for complete testing, more accessible to test
- It produced the well designed and more readable program.

2 **Performance issues**

- Execution time maybe, but not certainly, longer
- Storage size perhaps, but is not certainly, increased
- Compilation and loading time may be longer
- Inter-module communication problems may be increased
- More linkage required, run-time may be longer, more source lines must be written, and more documentation has to be done
-

3.**Issues relating to the diversity of database types**

Data Minig design reduces the design complexity and results in easier and faster implementation by allowing parallel development of various parts of a system. We discuss a different section of modular design in detail in this section:

1. Functional Independence:

Functional independence is achieved by developing functions that perform only one kind of task and do not excessively interact with other modules. Independence is important because it makes implementation more accessible and faster. The independent modules are easier to maintain, test, and reduce error propagation and can be reused in other programs as well. Thus, functional independence is a good design feature which ensures software quality.

2. Different user –

Different knowledge - different way.That means different client want a different kind of information so it becomes difficult to cover vast range of data that can meet the client requirement.
Interactive mining allows users to focus the search for patterns from different angles.The data mining process should be interactive because it is difficult to know what can be discovered within a databases.

3. Information hiding: The fundamental of Information hiding suggests that modules can be characterized by the design decisions that protect from the others, i.e., In other words.

modules should be specified that data include within a module is inaccessible to other modules that do not need for such information.

Data Processing:

The use ofprocessing as design criteria for modular system provides the most significant benefits when modifications are required during testing's and later during software maintenance. This is because as most data and procedures are hidden from other parts of the software, inadvertent errors introduced during modifications are less likely to propagate to different locations within the software.

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares ra Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues:

Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).

Data goes through a series of steps during preprocessing:

In this phase, data is made production ready.

The data preparation process consumes about 90% of the time of the project. data from different

sources should be selected, cleaned, transformed, formatted, anonymized, and constructed .

Data Cleaning:

Data cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
The data from different sources should be selected, cleaned, transformed, formatted, anonymized.
Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values.

Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not. Here Metadata should be used to reduce errors in the data integration process.

Next, the step is to search for properties of acquired data. A good way to explore the data is to answer The data mining questions (decided in business phase) using the query, reporting, and visualization tools.

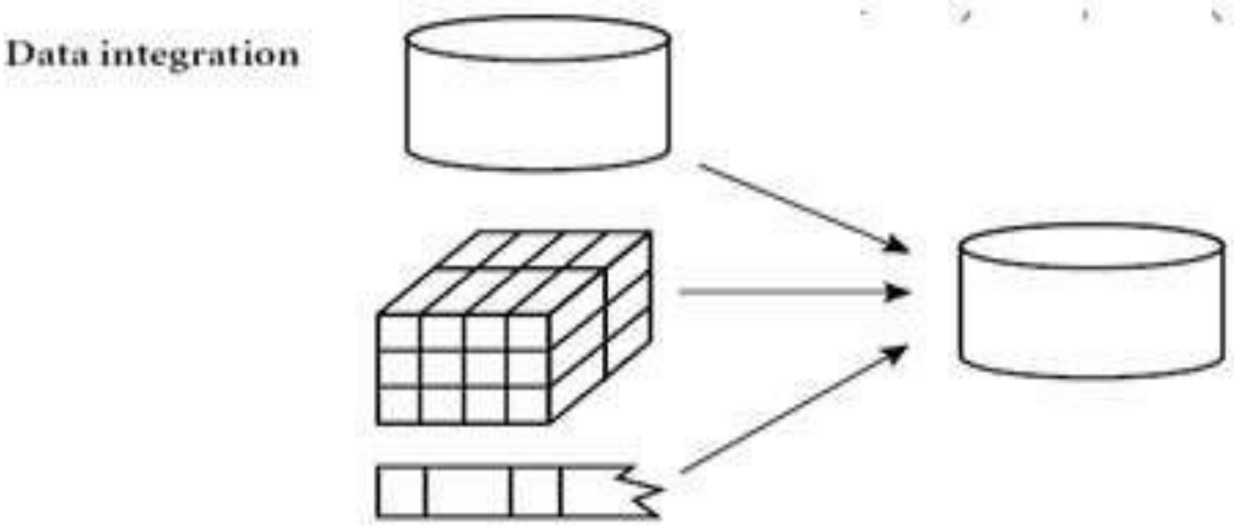
Data preparation:

In Data Mining, the Preparation databases is the degree of interdependence between software modules. Two modules that are tightly coupled are strongly dependent on each other. However, two modules that are loosely coupled are not dependent on each other. **Uncoupled modules** have no interdependence at all within them.

These data sources may include multiple databases, flat filer or data cubes. There are issues like object matching and schema integration which can arise during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily. For example, table A contains an entity named cust_no whereas another table B contains an entity named cust-id.

A good design is the one that has low quality. processing is measured by the number of relations between the modules. That is, the coupling increases as the number of calls between modules increase or the amount of shared data is large. Thus, it can be said that a design with high coupling will have more errors.

Data Intregration:



Data integration is one of the steps of data pre-processing that involves combining data residing in different sources and providing users with a unified view of these data.

The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.

- It merges the data from multiple data stores (data sources)
- It includes multiple databases, data cubes or flat files.
- Metadata, Correlation analysis, data conflict detection, and resolution of semantic Heterogeneity contribute towards smooth data integration.
- There are mainly 2 major approaches for data integration - commonly known as "tight coupling approach" and "loose coupling approach".

In this case, modules are subordinates to different modules. Therefore, no direct coupling.

1. **Tight Coupling:** When data of one module is passed to another module, this is called data coupling.

   - Here data is pulled over from different sources into a single physical location through the process
   - The single physical location provides an uniform interface for querying the data.
   - ETL layer helps to map the data from the sources so as to provide a uniform data warehouse. This approach is called tight coupling since in this approach the data is tightly coupled the physical repository at the time of query.

2. **Loose Coupling:**

   - Here a virtual mediated schema provides an interface that takes the query from the user,
   - Transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.
   - In this approach, the data only remains in the actual source databases. However, mediated schema contains several "adapters" or "wrappers" that can connect back to the source systems in order to bring the data to the front end.

DATA TRANSFORMATION:

Data transformation operations would contribute toward the success of the mining process.

Aggregation:
Summary or aggregation operations are applied to the data. I.e., the weekly sales data is aggregated to calculate the monthly and yearly total.

Generalization: In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.

Normalization: Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

Attribute construction: these attributes are constructed and included the given set of attributes helpful for data mining.

Data Reduction

In data mining, reduction defines to the degree to which the elements of a module belong together. Thus, cohesion measures the strength of relationships between pieces of functionality within a given module. For example, in highly cohesive systems, functionality is strongly related.

The method of data reduction may achieve a condensed description of the original data which is Much smaller in quantity but keeps the quality of the original data.

Methods of data reduction:

1. **Data Cube Aggregation:**

This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months. They involve you in the annual sales, rather than the quarterly average, So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

2. **Dimension reduction:**

Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant feature.

Step-wise Forward Selection –

The selection begins with an empty set of attributes later on we decide best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.

Step-wise Backward Selection:

This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.