

here is a huge

amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

What is 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 – Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining –

- Descriptive
- Classification and Prediction

Descriptive Function

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions –

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

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.

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.

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.

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.

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.

Data Mining Task Primitives

- We can specify a data mining task in the form of a data mining query.
- This query is input to the system.
- A data mining query is defined in terms of data mining task primitives.

Note – These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives –

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

Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following –

- Database Attributes
- Data Warehouse dimensions of interest

Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are –

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Background knowledge

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

Interestingness measures and thresholds for pattern evaluation

This is used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge.

Representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed. These representations may include the following. –

- Rules
 - Tables
 - Charts
 - Graphs
 - Decision Trees
 - Cubes
-
- Market Analysis
 - Fraud Detection
 - Customer Retention
 - Production Control
 - Science Exploration

Data Mining Applications

Data mining is highly useful in the following domains –

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid

Market Analysis and Management

Listed below are the various fields of market where data mining is used –

- **Customer Profiling** – Data mining helps determine what kind of people buy what kind of products.

- **Identifying Customer Requirements** – Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** – Data mining performs Association/correlations between product sales.
- **Target Marketing** – Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern** – Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** – Data mining provides us various multidimensional summary reports.

Corporate Analysis and Risk Management

Data mining is used in the following fields of the Corporate Sector –

- **Finance Planning and Asset Evaluation** – It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** – It involves summarizing and comparing the resources and spending.
- **Competition** – It involves monitoring competitors and market directions.

Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

Data Mining Tasks

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining –

- Descriptive
- Classification and Prediction

Descriptive Function

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions –

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

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.

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.

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.

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.

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.

Data Mining Task Primitives

- We can specify a data mining task in the form of a data mining query.
- This query is input to the system.
- A data mining query is defined in terms of data mining task primitives.

Note – These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives –

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

Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following –

- Database Attributes
- Data Warehouse dimensions of interest

Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are –

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Background knowledge

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

Interestingness measures and thresholds for pattern evaluation

This is used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge.

Representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed. These representations may include the following. –

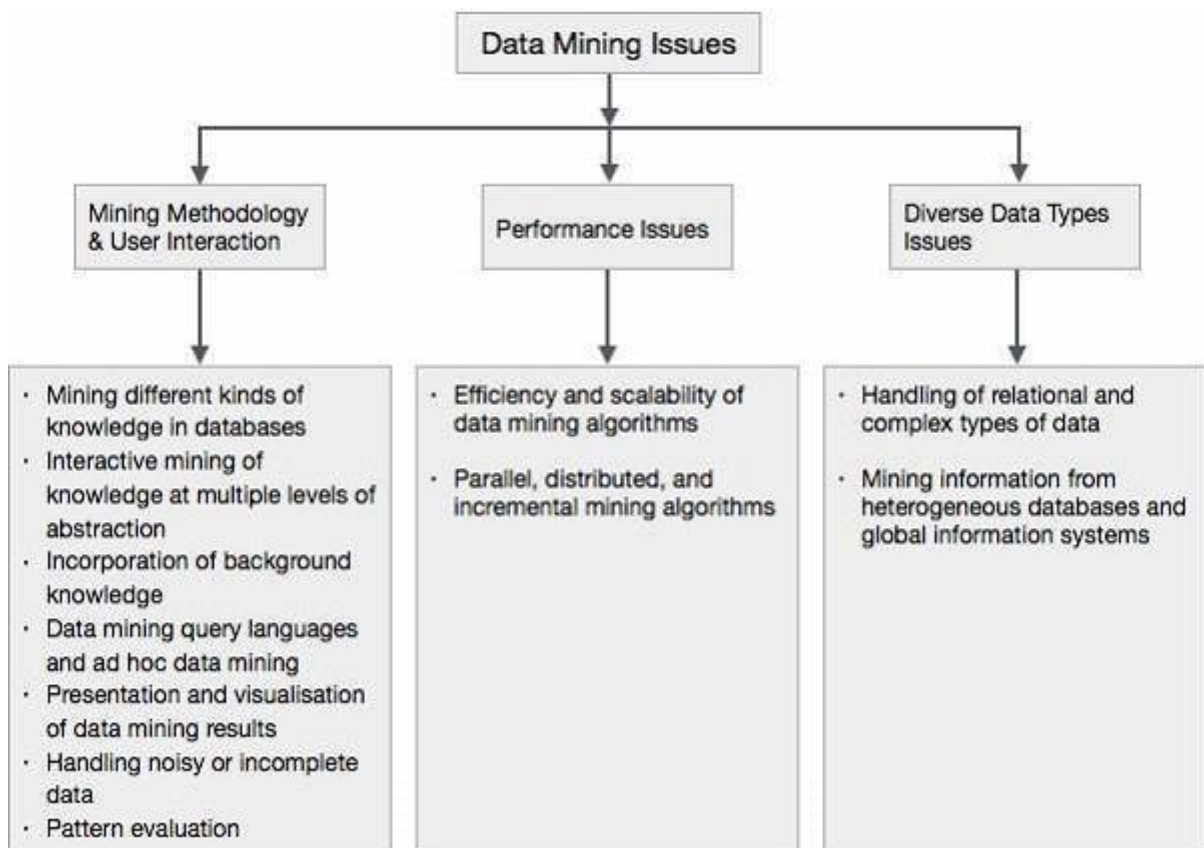
- Rules
- Tables
- Charts
- Graphs
- Decision Trees
- Cubes

Data Mining - Issues

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding –

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues –

- **Mining different kinds of knowledge in databases** – Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a 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 the 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 levels 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, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** – The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** – The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows –

- **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 algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** – The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.

- **Mining information from heterogeneous databases and global information systems** – The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

What is Knowledge Discovery?

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.

The following diagram shows the process of knowledge discovery –

