UNIT-1

Introduction: What is data mining? What kinds of data can be mined? What kinds of pattern can be mined? Which technologies are used? Which kinds of applications are targeted, Major Issues in Data Mining?

Introduction:

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need and how data mining can meet this need by providing tools to discover knowledge from data.

Data Mining as the Evolution of Information Technology:

Data mining can be viewed as a result of the natural evolution of information technology.

The database and data management industry evolved in the development of several critical functionalities (Figure 1.1): data collection and database creation, data management (including data storage and retrieval and database transaction processing), and advanced data analysis (involving data warehousing and data mining).

The early development of data collection and database creation mechanisms served as a prerequisite for the later development of effective mechanisms for data storage and retrieval, as well as query and transaction processing.

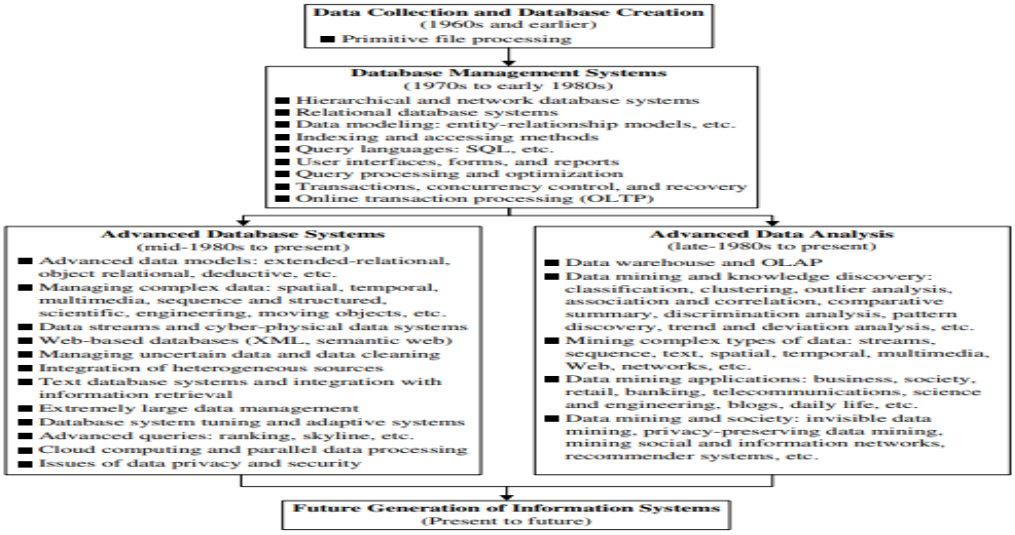


Figure 1.1 The evolution of database system technology.

After the establishment of database management systems, database technology moved toward the development of advanced database systems, data warehousing, and data mining for advanced data analysis and web-based databases.

Advanced database systems, for example, resulted from an upsurge of research from the mid-1980s onward. These systems incorporate new and powerful data models such as extended-relational, object-oriented, object-relational, and deductive models.

Application-oriented database systems have flourished, including spatial, temporal, multimedia, active, stream and sensor, scientific and engineering databases, knowledge bases, and office information bases. Issues related to the distribution, diversification, and sharing of data have been studied extensively.

One emerging data repository architecture is the data warehouse. This is a repository of multiple heterogeneous data sources organized under a unified schema at a single site to facilitate management decision making.

Data warehouse technology includes data cleaning, data integration, and online analytical processing (OLAP)—that is, analysis techniques with functionalities such as summarization, consolidation, and aggregation, as well as the ability to view information from different angles.

Although **OLAP tools** support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis—**for example**, data mining tools that provide data classification, clustering, outlier/anomaly detection, and the characterization of changes in data over time.

What Is Data Mining?

Data mining is the process of extracting knowledge or insights from large amounts of data using various statistical and computational techniques.

The data can be structured, semi-structured or unstructured, and can be stored in various forms such as databases, data warehouses, and data lakes.

The primary goal of data mining is to discover hidden patterns and relationships in the data that can be used to make informed decisions or predictions.
This involves exploring the data using various techniques such as clustering, classification, regression analysis, association rule mining, and anomaly detection.
Data mining has a wide range of applications across various industries, including marketing, finance, healthcare, and telecommunications.
For example, in marketing, data mining can be used to identify customer segments and target marketing campaigns, while in healthcare, it can be used to identify risk factors for diseases and develop personalized treatment plans.
Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data.
Data mining is also called **Knowledge Discovery in Database (KDD*).*** The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.
for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.
Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.
The knowledge discovery process is shown in Figure 1.4 as an iterative sequence of the following steps:
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)
Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.
The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data).

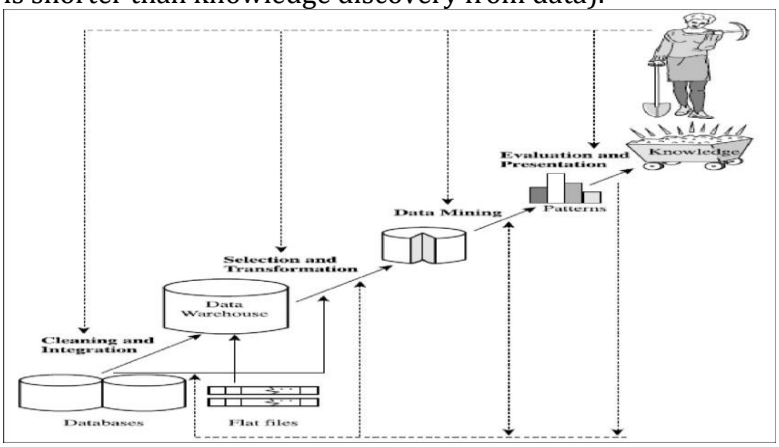


Figure 1.4 Data mining as a step in the process of knowledge discovery
Therefore, we adopt a broad view of data mining functionality: Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

What Kinds of Data Can Be Mined?
As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application.
The most basic forms of data for mining applications are

- Database data,
- Data warehouse data and
- Transactional data.

Data mining can also be applied to other forms of data (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).

Database Data:

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.

The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.

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.

Example 1.2 A relational database for All Electronics.

The company is described by the following relation tables: customer, item, employee, and branch.

The headers of the tables described here are shown in Figure 1.5. (A header is also called the schema of a relation.) The relation customer consists of a set of attributes describing the customer information, including a unique customer identity number (cust ID), customer name, address, age, occupation, annual income, credit information, and category.

Similarly, each of the relations item, employee, and branch consists of a set of attributes describing the properties of these entities.

Tables can also be used to represent the relationships between or among multiple entities.

In our example, these include purchases (customer purchases items, creating a sales transaction handled by an employee), items sold (lists items sold in a given transaction), and works at (employee works at a branch of AllElectronics).

customer	(*cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .*)
item	(*item_ID, brand, category, type, price, place_made, supplier, cost, . . .*)
employee	(*empl_ID, name, category, group, salary, commission, . . .*)
branch	(*branch_ID, name, address, . . .*)
purchases	(*trans_ID, cust_ID, empl_ID, date, time, method_paid, amount*)
items_sold	(*trans_ID, item_ID, qty*)
works_at	(*empl_ID, branch_ID*)

Figure 1.5 Relational schema for a relational database, *AllElectronics*.

Relational data can be accessed by database queries written in a relational query language (e.g., SQL) or with the assistance of graphical user interfaces.

A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing.

A query allows retrieval of specified subsets of the data.

Suppose that your job is to analyze the AllElectronics data.

Through the use of relational queries, you can ask things like, “Show me a list of all items that were sold in the last quarter.”

Relational languages also use aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum). Using aggregates allows you to ask: “Show me the total sales of the last month, grouped by branch,” or “How many sales transactions occurred in the month of December?” or “Which salesperson had the highest sales?”

Data Warehouses:

Suppose that AllElectronics is a successful international company with branches around the world. Each branch has its own set of databases.

The president of AllElectronics has asked you to provide an analysis of the company’s sales per item type per branch for the third quarter.

This is a difficult task, particularly since the relevant data are spread out over several databases physically located at numerous sites.

If AllElectronics had a data warehouse, this task would be easy.

A **data warehouse** is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.
Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
Figure 1.6 shows the typical framework for construction and use of a data warehouse for AllElectronics.

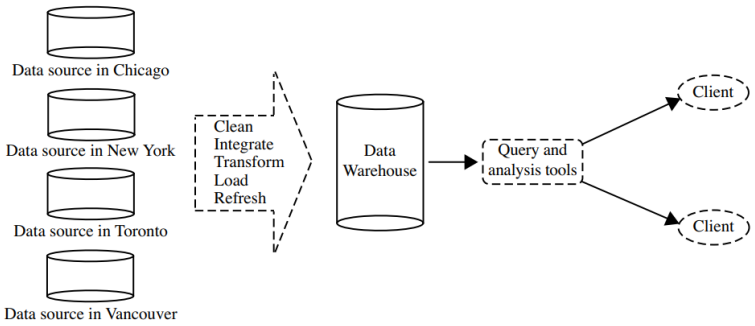
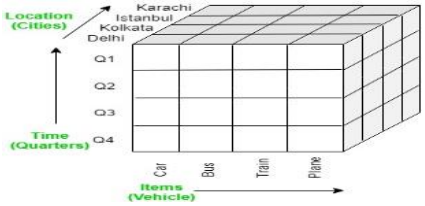


Figure 1.6 Typical framework of a data warehouse for *AllElectronics*.

ETL (Extract, Transform, and Load) Process in Data Warehouse
What is ETL?
ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.
ETL Process in Data Warehouse
ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse. The process of ETL can be broken down into the following three stages:
Extract: The first stage in the ETL process is to extract data from various sources such as transactional systems, spreadsheets, and flat files. This step involves reading data from the source systems and storing it in a staging area.
Transform: In this stage, the extracted data is transformed into a format that is suitable for loading into the data warehouse. This may involve cleaning and validating the data, converting data types, combining data from multiple sources, and creating new data fields.
Load: After the data is transformed, it is loaded into the data warehouse. This step involves creating the physical data structures and loading the data into the warehouse.
A data warehouse is usually modeled by a multidimensional data structure, called a **data cube**, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount).
A data cube provides a multidimensional view of data and allows the pre computation and fast access of summarized data.
Online Analytical Processing (OLAP): Online Analytical Processing (OLAP) consists of a type of software tool that is used for data analysis for business decisions. OLAP provides an environment to get insights from the database retrieved from multiple database systems at one time.
OLAP stands for ***Online Analytical Processing*** Server. It is a software technology that allows users to analyze information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data (eg. Delhi -> 2018 -> Sales data). OLAP databases are divided into one or more cubes and these cubes are known as *Hyper-cubes*.



OLAP operations:
Four types of analytical OLAP operations are:
- **Roll-up**
- **Drill-down**
- **Slice and dice**
- **Pivot (rotate)**

Drill down: In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:

- Moving down in the concept hierarchy
- Adding a new dimension

In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).

Consider the diagram above

Quater Q1 is drilled down to months January, February, and March. Corresponding sales are also registers.

In this example, dimension months are added

Roll up: It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:

Climbing up in the concept hierarchy

Reducing the dimensions

In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).

In the example, cities New jersey and Lost Angles and rolled up into country USA

The sales figure of New Jersey and Los Angeles are 440 and 1560 respectively. They become 2000 after roll-up. In this aggregation process, data is location hierarchy moves up from city to the country.

In the roll-up process at least one or more dimensions need to be removed. In this example, Cities dimension is removed

Dice: It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:

Location = "Delhi" or "Kolkata"

Time = "Q1" or "Q2"

Item = "Car" or "Bus"

Slice: It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = "Q1".

Pivot: It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.

Example: A data cube for All Electronics.

A data cube for summarized sales data of All Electronics is presented in Figure 1.7(a).

The cube has three dimensions: address (with city values Chicago, New York, Toronto, Vancouver), time (with quarter values Q1, Q2, Q3, Q4), and item (with item type values home entertainment, computer, phone, security).

The aggregate value stored in each cell of the cube is sales amount (in thousands).

For example, the total sales for the first quarter, Q1, for the items related to security systems in Vancouver is $400,000, as stored in cell (Vancouver, Q1, security).

Additional cubes may be used to store aggregate sums over each dimension, corresponding to the aggregate values obtained using different SQL group-bys (e.g., the total sales amount per city and quarter, or per city and item, or per quarter and item, or per each individual dimension).

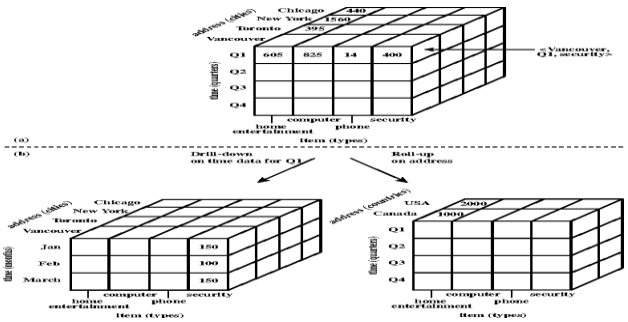


Fig: A multidimensional data cube

Transactional Data:

In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.

A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction. A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.

Example 1.4

A transactional database for All Electronics. Transactions can be stored in a table, with one record per transaction. A fragment of a transactional database for All Electronics is shown in Figure 1.8. From the relational database point of view, the sales table in the figure is a nested relation because the attribute list of item IDs contains a set of items. Because most relational database systems do not support nested relational structures, the transactional database is usually either stored in a flat file in a format similar to the table in Figure 1.8 or unfolded into a standard relation in a format similar to the items sold table in Figure 1.5.

trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

Figure 1.8 Fragment of a transactional database for sales at *AllElectronics*.

As an analyst of All Electronics, you may ask," Which items sold well together?" This kind of market basket data analysis would enable you to bundle groups of items together as a strategy for boosting sales. For example, given the knowledge that printers are commonly purchased together with computers, you could offer certain printers at a steep discount (or even for free) to customers buying selected computers, in the hopes of selling more computers (which are often more expensive than printers).

Advanced database systems and advanced database applications :

• **An objected-oriented database** is designed based on the object-oriented programming paradigm where data are a large number of objects organized into classes and class hierarchies. Each entity in the database is considered as an object. The object contains a set of variables that describe the object, a set of messages that the object can use to communicate with other objects or with the rest of the database system and a set of methods where each method holds the code to implement a message.

• **A spatial database contains** spatial-related data, which may be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps, and vector data are represented by lines, points, polygons or other kinds of processed primitives, Some examples of spatial databases include geographical (map) databases, VLSI chip designs, and medical and satellite images databases.

• **Time-Series Databases**: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time.

• **A text database** is a database that contains text documents or other word descriptions in the form of long sentences or paragraphs, such as product specifications, error or bug reports, warning messages, summary reports, notes, or other documents.

• **A multimedia database** stores images, audio, and video data, and is used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces.

• **The World-Wide Web** provides rich, world-wide, on-line information services, where data objects are linked together to facilitate interactive access. Some examples of distributed information services associated with the World-Wide Web include America Online, Yahoo!, AltaVista, and Prodigy.

What kind of patterns can be mined in data mining?

(Data Mining Functionalities):

There are a number of data mining functionalities. These include

- Class/Concept Description: Characterization and Discrimination
- Mining Of Frequent Patterns, Associations and Correlations
- Classification and Regression
- Clustering Analysis and Outlier Analysis

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

In general, such tasks can be classified into two categories: descriptive and predictive

- **Descriptive mining** tasks characterize properties of the data in a target data set.
- **Predictive mining** tasks perform induction on the current data in order to make predictions.

What is descriptive data mining?

Descriptive mining is usually used to provide correlation, cross-tabulation, frequency, etc. These techniques are used to determine the data regularities and to reveal patterns. It targets the summarization and conversion of data into meaningful data for reporting and monitoring.

As the name suggests, descriptive mining "describe" the data. Once the data is captured, we convert it into human interpretable form.

Descriptive analytics focus on answering "What has happened in the past?" Descriptive analytics is useful because it enables us to learn from the past.

How descriptive analytics is used in learning analytics

- Comparing pre-test and post-test assessments.
- Tracking course enrollments.
- Collating course survey results.
- Recording which learning resources and accessed and how often.
- Summarizing the number of times, a learner posts on a discussion board.

What is Predictive data Mining?

The term 'Predictive' means to predict something, so predictive data mining is the analysis done to predict the future event or other data or trends.

Predictive data mining can enable business analysts to make decisions and add value to the analytics team efforts.

Predictive data mining supports predictive analytics. As we know, predictive analytics is the use of information to predict outcomes.

Let's understand this concept with the help of an example;

Any retail shop may use algorithm-based tools to go through a customer database to look at the previous transactions to predict future transactions.

In other words, the previous data may enable the shopkeeper to project what will happen in future in the business, enabling business people to plan accordingly.

Class/Concept Description: Characterization and Discrimination:

Data entries can be associated with classes or concepts.

For example, in the All Electronics store,

- **classes of items** for sale include computers and printers, and
- **concepts** of customers include big Spenders and budget Spenders.

These descriptions can be derived using

- (1) **data characterization**, by summarizing the data of the class under study (often called the target class) in general terms, or
- (2) **data discrimination**, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or
- (3) **both data characterization and discrimination**.

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. The resulting descriptions can also be presented as generalized relations or in rule form (called characteristic rules)

Example: Data characterization.

A customer relationship manager at All Electronics may order the following data mining task:

Summarize the characteristics of customers who spend more than $5000 a year at All Electronics.

The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.

The data mining system should allow the customer relationship manager to drill down on any dimension, such as on occupation to view these customers according to their type of employment.

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.

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

Example: Data discrimination.

A customer relationship manager at All Electronics may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).

The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.

Drilling down on a dimension like occupation, or adding a new dimension like income level, may help to find even more discriminative features between the two classes.

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 frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures.

A **frequent itemset** refers to a set of items that often appear together in a transactional data set—
for example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers, tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.

A **substructure** can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.

If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Associations and Correlation :

Association is a data mining technique that discovers the probability of the co-occurrence of items in a collection.

The relationships between co-occurring items are expressed as **Association Rules**.

Association Rule: is learning technique that helps identify dependency between two data items

Association Analysis

- It analyses the set of items that generally occur together in a transactional dataset. It is also known as Market Basket Analysis for its wide use in retail sales.

Correlation Analysis

- Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another.

Two parameters used to find relationship between items:

- Support
- Confidence

Support refers to the relative frequency of an item set in a dataset.

Confidence is a measure of the likelihood that an itemset will appear if another itemset appears.

Support refers to the relative frequency of an item set in a dataset.

For example, if an itemset occurs in 5% of the transactions in a dataset, it has a support of 5%.

- *Support(X) = (Number of transactions containing X) / (Total number of transactions)*
- The support of an itemset is the number of transactions in which the itemset appears, divided by the total number of transactions.

For example, suppose we have a dataset of 1000 transactions, and the itemset {milk, bread} appears in 100 of those transactions. The support of the itemset {milk, bread} would be calculated as follows:

- Support({milk, bread}) = Number of transactions containing {milk, bread} / Total number of transactions = 100 / 1000 = 10%
- So the support of the itemset {milk, bread} is 10%. This means that in 10% of the transactions, the items milk and bread were both purchased.

In general, the confidence of a rule can be calculated using the following formula:

Confidence(X => Y) = (Number of transactions containing X and Y) / (Number of transactions containing X)

For example, suppose we have a dataset of 1000 transactions, and the itemset {milk, bread} appears in 100 of those transactions. The itemset {milk} appears in 200 of those transactions.

The confidence of the rule "If a customer buys milk, they will also buy bread" would be calculated as follows:

Confidence("If a customer buys milk, they will also buy bread") = Number of transactions containing {milk, bread} / Number of transactions containing {milk} = 100 / 200 = 50%

So the confidence of the rule "If a customer buys milk, they will also buy bread" is 50%. This means that in 50% of the transactions where milk was purchased, bread was also purchased.

Example: Association analysis. Suppose that, as a marketing manager at All Electronics, you want to know
which items are frequently purchased together (i.e., within the same transaction).
An example of such a rule, mined from the All Electronics 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 the transactions under analysis show that computer and software are purchased together.

This association rule involves a single attribute or predicate (i.e., buys) that repeats.

Association rules that contain a single predicate are referred to as **single-dimensional association rules.**

A data mining system may find association rules like

age(X, "20..29") ∧ income(X, "40K..49K") ⇒ buys(X, "laptop") [support = 2%, confidence = 60%].

The rule indicates that of the All Electronics customers under study,

2% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer) at All Electronics. There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., age, income, and buys). where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule.**

Classification and Regression :

There are two forms of data analysis that can be used to extract models describing important classes or predict future data trends.

These two forms are as follows:

- Classification
- Prediction

Classification is to identify the category or the class label of a new observation.

Predication is the process of identifying the missing or unavailable numerical data for a new observation.

Classification:

Classification is to identify the category or the class label of a new observation.

First, a set of data is used as training data. The set of input data and the corresponding outputs are given to the algorithm. So, the training data set includes the input data and their associated class labels. Using the training dataset, the algorithm derives a model or the classifier. The derived model can be a decision tree, mathematical formula, or a neural network. In classification, when unlabeled data is given to the model, it should find the class to which it belongs. The new data provided to the model is the test data set.

There are two main types of classification: binary classification and multi-class classification. Binary classification involves classifying instances into two classes, such as "spam" or "not spam", while multi-class classification involves classifying instances into more than two classes.

Example: The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder.

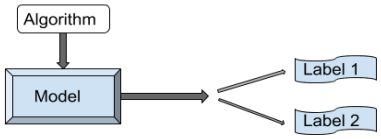


Fig: Binary classification

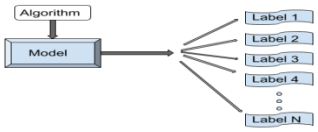


Fig: Multi class classification

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known).

The model is used to predict the class label of objects for which the the class label is unknown.

"How is the derived model presented?"

The derived model may be represented in various forms, such as

- Classification rules (i.e., if-then rules),
- Decision trees,
- Mathematical formulae, or neural networks(figure 1.9).

A decision tree is a flowchart-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 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-neighbor classification

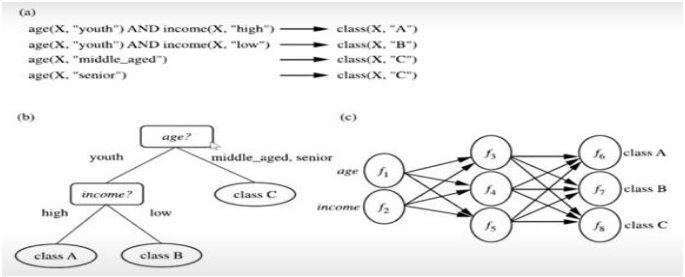


Figure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Regression is generally used for prediction. Predicting the value of a house depending on the facts such as the number of rooms, the total area, etc., is an example for prediction.

Regression analysis aims to identify the most important relationships among variables and use these relationships to make predictions.

Regression analysis is a statistical methodology that is most often used for numeric prediction

Cluster Analysis:

A cluster is a subset of similar objects

- Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups.
- Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the **principle of maximizing the intra class similarity and minimizing the interclass similarity**. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.

Example: Cluster analysis. Cluster analysis can be performed on All Electronics customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

Figure shows a 2-D plot of customers with respect to customer locations in a city. Three clusters of **data points are evident**

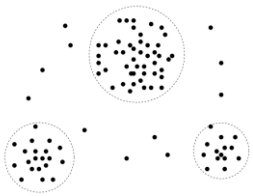


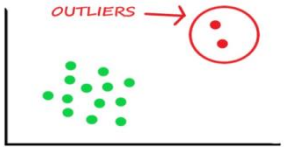
Figure: A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.

Outlier Analysis

- Outlier analysis in data mining is the process of identifying and examining data points that significantly differ from the rest of the dataset.
- An outlier can be defined as a data point that deviates significantly from the normal pattern or behavior of the data. Various factors, such as measurement errors, unexpected events, data processing errors, etc., can cause these outliers.
- Various factors, such as measurement errors, unexpected events, data processing errors, etc., can cause these outliers. For example, outliers are represented as red dots in the figure below, and you can see that they deviate significantly from the rest of the data points. Outliers are also often referred to as anomalies, aberrations, or irregularities.
- A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers. Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier analysis or anomaly mining.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers. Rather than using statistical or distance measures, density-based methods may identify outliers in a local region, although they look normal from a global statistical distribution view.

For example, outliers are represented as red dots in the figure below, and you can see that they deviate significantly from the rest of the data points. Outliers are also often referred to as anomalies, aberrations, or irregularities.



Which technologies are used?:
As a highly application-driven domain, data mining has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high performance computing, and many application domains (Figure 1.11).
In this section, we give examples of several disciplines that strongly influence the development of data mining methods

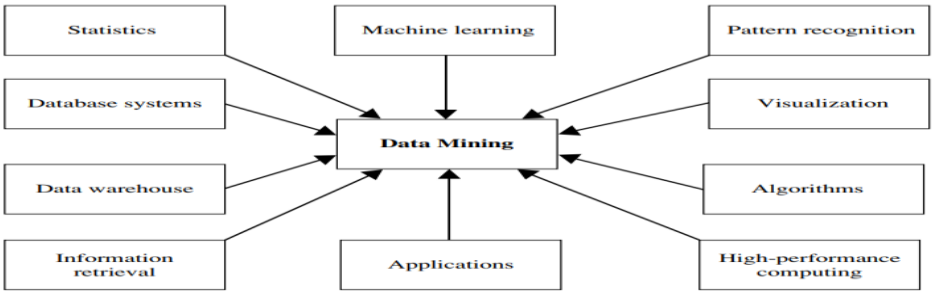


Figure 1.11 Data mining adopts techniques from many domains.

Statistics: Statistics studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics.
A statistical model is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used to model data and data classes.
For example, in data mining tasks like data characterization and classification, statistical models of target classes can be built. In other words, such statistical models can be the outcome of a data mining task.

Statistics is useful for mining various patterns from data as well as for understanding the underlying mechanisms generating and affecting the patterns.
In statistics, there are two main categories:
Descriptive Statistics: The purpose of descriptive statistics is to organize data and identify the main characteristics of that data. Graphs or numbers summarize the data. Average, Mode, SD(Standard Deviation), and Correlation are some of the commonly used descriptive statistical methods.
Inferential Statistics: The process of drawing conclusions based on probability theory and generalizing the data. By analyzing sample statistics, you can infer parameters about populations and make models of relationships within data.
Hypothesis Tests: This is a statistical procedure that was used to make inferences about population parameters based on the data. The main aim of hypothesis testing is to determine the claim or assumption supported by the sample data.
Regression Analysis: This is also a statistical inferential method used to model the relationship between dependent and independent variables. The main aim of regression analysis is to fit the mathematical equation of data which defines the relationship between variables.
Statistical methods can also be used to verify data mining results. For example, after a classification or prediction model is mined, the model should be verified by statistical hypothesis testing.
A statistical hypothesis test (sometimes called confirmatory data analysis) makes statistical decisions using experimental data. A result is called statistically significant if it is unlikely to have occurred by chance. If the classification or prediction model holds true, then the descriptive statistics of the model increases the soundness of the model
What is a Hypothesis testing?
The hypothesis can be defined as the claim that can either be related to the truth about something that exists in the world, or, truth about something that's needs to be established a fresh.
In simple words, another word for the hypothesis is the **"claim"**. Until the claim is proven to be true, it is called the hypothesis. Once the claim is proved, it becomes the new truth or new knowledge about the thing.
For **example**, let's say that a claim is made that students studying for more than 6 hours a day gets more than 90% of marks in their examination. Now, this is just a claim or a hypothesis and not the truth in the real world. However, in order for the claim to become the truth for widespread adoption, it needs to be proved using pieces

of evidence, e.g., data. In order to reject this claim or otherwise, one needs to do some empirical analysis by gathering data samples and evaluating the claim.

The process of gathering data and evaluating the claims or hypotheses with the goal to reject or otherwise (failing to reject) can be called as hypothesis testing. Note the wordings – "failing to reject". It means that we don't have enough evidence to reject the claim. Thus, until the time that new evidence comes up, the claim can be considered the truth. There are different techniques to test the hypothesis in order to reach the conclusion of whether the hypothesis can be used to represent the truth of the world.

Simply speaking, hypothesis testing is a framework that can be used to assert whether the claim or the hypothesis made about a real-world/real-life event can be seen as the truth or otherwise based on the given data (evidences).

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as:

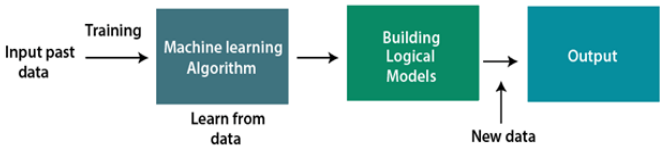
Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

How does Machine Learning work

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem.

The below block diagram explains the working of Machine Learning algorithm:



Currently, machine learning is used in self-driving cars, cyber fraud detection, face recognition, and friend suggestion by Facebook, etc. Various top companies such as Netflix and Amazon have build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly

Types of Machine Learning:

Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Semi-Supervised Machine Learning and Reinforcement Learning

Supervised machine learning:

As its name suggests, Supervised machine learning is based on supervision.

It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output.

Here, the labelled data specifies that some of the inputs are already mapped to the output.

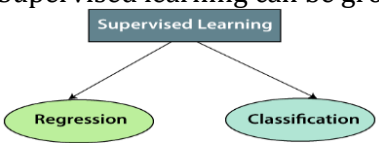
More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).

Some real-world applications of supervised learning are Risk Assessment, Fraud Detection, Spam filtering, etc.

Categories of Supervised Machine Learning

Supervised learning can be grouped further in two categories of algorithms:



a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc.

The classification algorithms predict the categories present in the dataset.

Some real-world examples of classification algorithms are Spam Detection, Email filtering, etc.

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables.

These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision.

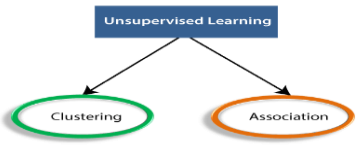
It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

Categories of Unsupervised Machine Learning

Unsupervised Learning can be further classified into two types, which are given below:



Clustering: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.

Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Association: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item.

Example of Association rule is Market Basket Analysis.

Some popular algorithms of Association rule learning are Apriori Algorithm, Eclat, FP-growth algorithm.

Semi-Supervised Learning:

Semi-supervised learning (SSL) is a machine learning technique that uses a small portion of labeled data and lots of unlabeled data to train a predictive model.

One of the simplest examples of semi-supervised learning, in general, is self-training.

Self-training is the procedure in which you can take any supervised method for classification or regression and modify it to work in a semi-supervised manner, taking advantage of labeled and unlabeled data.

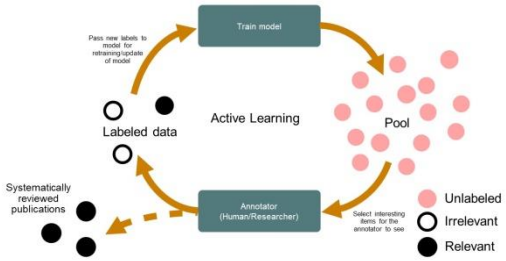
Active learning:

Active learning is a machine learning approach that lets users play an active role in the learning process. An active learning approach can ask a user (e.g., a domain expert) to label an example, which may be from a set of unlabeled examples or synthesized by the learning program. The goal is to optimize the model quality by actively acquiring knowledge from human users, given a constraint on how many examples they can be asked to label.

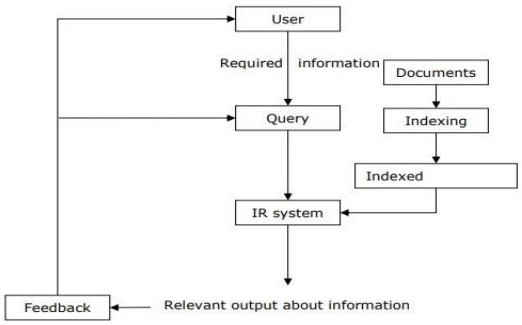
This technique is also considered in situations where labeling is difficult or time-consuming. Passive learning, or the conventional way through which a large quantity of labeled data is created by a human oracle, requires enormous efforts in terms of man hours.

In a successful active learning system, the algorithm is able to choose the most informative data points through some defined metric, subsequently passing them to a human labeler and progressively adding them to the training set.

A diagrammatic representation is shown below.



Reinforcement learning:
In this type of learning, the machine learns from the feedback it has received. It constantly learns and upgrades its existing skills by taking the feedback from the environment it is in.
Reinforcement Learning is a feedback-based Machine learning technique in which an agent learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the agent gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.
In Reinforcement Learning, the agent learns automatically using feedbacks without any labeled data, unlike supervised learning.
The agent interacts with the environment and explores it by itself. The primary goal of an agent in reinforcement learning is to improve the performance by getting the maximum positive rewards.
Database Systems and Data Warehouses:
Database systems research focuses on the creation, maintenance, and use of databases for organizations and end-users. Particularly, database systems researchers have established highly recognized principles in data models, query languages, query processing and optimization methods, data storage, and indexing and accessing methods.
Database systems are often well known for their high scalability in processing very large, relatively structured data sets. Many data mining tasks need to handle large data sets or even real-time, fast streaming data.
Therefore, data mining can make good use of scalable database technologies to achieve high efficiency and scalability on large data sets. Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities.
A data warehouse integrates data originating from multiple sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes. The data cube model not only facilitates OLAP in multidimensional databases but also promotes multidimensional data mining
Information Retrieval:
Information retrieval (IR) is the science of searching for documents or information in documents. Documents can be text or multimedia, and may reside on the Web.
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.
The differences between traditional information retrieval and database systems are two fold: Information retrieval assumes that
(1) The data under search are unstructured; and
(2) The queries are formed mainly by keywords, which do not have complex structures (unlike SQL queries in database systems). The typical approaches in information retrieval adopt probabilistic models.
With the help of the following diagram, we can understand the process of information retrieval (IR) –
It is clear from the above diagram that a user who needs information will have to formulate a request in the form of query in natural language. Then the IR system will respond by retrieving the relevant output, in the form of documents, about the required information.



Which Kinds of Applications Are Targeted?
We briefly discuss two highly successful and popular application examples of data mining:
- Business intelligence and
- Search engines.

Business Intelligence:
Business intelligence (BI) technologies provide historical, current, and predictive views of business operations. Examples include reporting, online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics.
"**How important is business intelligence**?" Without data mining, many businesses may not be able to perform effective market analysis, compare customer feedback on similar products, discover the strengths and weaknesses of their competitors, retain highly valuable customers, and make smart business decisions.

Clearly, data mining is the core of business intelligence. Online analytical processing tools in business intelligence rely on data warehousing and multidimensional data mining.
Using characterization mining techniques, we can better understand features of each customer group and develop customized customer reward programs

What Is Business Intelligence (BI)?

Business intelligence (BI) refers to the procedural and technical infrastructure that collects, stores, and analyzes the data produced by a company's activities.
BI tools and software come in a wide variety of forms such as spreadsheets, reporting/query software, data visualization software, data mining tools, and online analytical processing (OLAP).
Business Intelligence (BI) is a process of gathering, analyzing, and transforming raw data into accurate, efficient, and meaningful information which can be used to make wise business decisions and refine business strategy.

Examples of Business Intelligence Systems Used in Practice

Business Intelligence gets used in almost every business domain that requires information from the raw data sets gathered by the company. Some of the real-life examples of business intelligence are listed below.

1. Walmart BI Strategy

Walmart is a retail store that uses data analysis to study consumer behaviors. Based on this information, they decide on their online and offline strategies to attract more customers and multiply their sales. For example, if they find that the products on the shelves near the payment counter are the highest selling, they might use it to promote brands and make more profits.

2. Tesla

Tesla linked its cars to corporate offices with the help of advanced technologies. The idea was to gather data about consumer concerns and analyze it to find flaws or shortcomings. Then, they use this information to solve customer issues and earn a satisfied consumer and an excellent market reputation for the company.

Web Search Engines:

- A Web search engine is a specialized computer server that searches for information on the Web. The search results of a user query are often returned as a list (sometimes called hits).
- The hits may consist of web pages, images, and other types of files. Some search engines also search and return data available in public databases or open directories.
- A search engine is a software program that provides information according to the user query. It finds various websites or web pages that are available on the internet and gives related results according to the search.
- For example, a student wants to learn C++ language so he searches the "C++ tutorial GeeksforGeeks" in the search engine.
- So the student gets a list of links that contain the tutorial links of GeeksforGeeks. Or we can say that a search engine is an internet-based software program whose main task is to collect a large amount of data or information about what is on the internet, then categorize the data or information and then help user to find the required information from the categorized information.
- Google, Yahoo, Bing are the most popular Search Engines.

What is search engine indexing?

Step 1. Web spiders (or bots) scan all the website's known URLs. This is called crawling.
Step 2. The bots collect and store data from the web pages, which is called indexing.
Web crawlers index pages and their content, including text, internal links, images, audio, and video files. If the content is considered to be valuable and competitive, the search engine will add the page to the index, and it'll be in the "game" to compete for a place in the search results for relevant user search queries.
In short, if you want users to find your website on Google, it needs to be indexed: information about the page should be added to the search engine database.
Step 3. And finally, the website and its pages can compete in the game trying to rank for a specific query.

How do Search Engines Work?

Search engines are generally working on three parts that are crawling, indexing, and ranking
Crawling:
Search engines have a number of computers programs that are responsible for finding information that is publicly available on the internet.
These programs scan the web and create a list of all available websites. Then they visit each website and by reading HTML code they try to understand the structure of the page, the type of the content, the meaning of the content, and when it was created or updated.

Why crawling is important?

Because your first concern when optimizing your website for search engines is to make sure that they can access it correctly. If they cannot find your content you won't get any ranking or search engine traffic.

Indexing:
Information identified by the crawler needs to be organized, Sorted, and Stored so that it can be processed later by the ranking algorithm.
Search engines don't store all the information in your index, but they keep things like the Title and description of the page, The type of content, Associated keywords Number of incoming and outgoing links, and a lot of other parameters that are needed by the ranking algorithm.
Why indexing is important?
Because if your website is not in their index it will not appear for any searches this also means that if you have any pages indexed you have more chances of appearing in the search results for a related query.
3. Ranking: Ranking is the position by which your website is listed in any Search Engine. (There are three steps in which ranking works).
Major Issues in Data Mining:
Data mining is a dynamic and fast-expanding field with great strengths.
In this section, we briefly outline the major issues in data mining research, partitioning them into five groups:

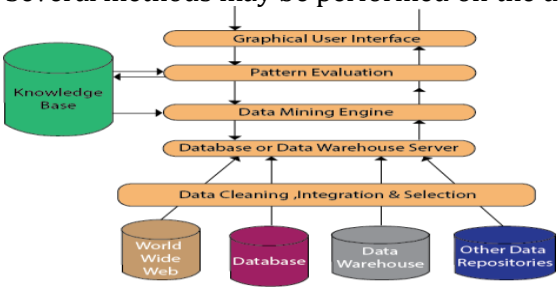
- Mining Methodology,
- User Interaction,
- Efficiency and Scalability,
- Diversity of Data Types,
- Data Mining and Society**.**

Many of these issues have been addressed in recent data mining research and development to a certain extent and are now considered data mining requirements; others are still at the research stage.
Mining Methodology:
Researchers have been vigorously developing new data mining methodologies. Let's have a look at these various aspects of mining methodology.
Mining various and new kinds of knowledge: Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis.
For example, for effective knowledge discovery in information networks, integrated clustering and ranking may lead to the discovery of high-quality clusters and object ranks in large networks.
Mining knowledge in multidimensional space: When searching for knowledge in large data sets, we can explore the data in multidimensional space.
That is, we can search for interesting patterns among combinations of dimensions (attributes) at varying levels of abstraction. Such mining is known as (exploratory) multidimensional data mining. In many cases, data can be aggregated or viewed as a multidimensional data cube.
Mining knowledge in cube space can substantially enhance the power and flexibility of data mining.
Data mining—an interdisciplinary effort: The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines.
For example to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing.
As another example, consider the mining of software bugs in large programs.
This form of mining, known as bug mining, benefits from the incorporation of software engineering knowledge into the data mining process.
User Interaction:
The user plays an important role in the data mining process. Interesting areas of research include how to interact with a data mining system, how to incorporate a user's background knowledge in mining, and how to visualize and comprehend data mining results. We introduce each of these here.
Interactive mining: The data mining process should be highly interactive. Thus, it is important to build flexible user interfaces and an exploratory mining environment, facilitating the user's interaction with the system.
A user may like to first sample a set of data, explore general characteristics of the data, and estimate potential mining results. Interactive mining should allow users to dynamically change the focus of a search, to refine mining requests based on returned results, and to drill, dice, and pivot through the data and knowledge space interactively, dynamically exploring "cube space" while mining.
Incorporation of background knowledge: Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated into the knowledge discovery process. Such knowledge can be used for pattern evaluation as well as to guide the search toward interesting patterns.

Ad hoc data mining and data mining query languages: Query languages (e.g., SQL) have played an important role in flexible searching because they allow users to pose ad hoc queries.
Ad hoc query refers to user-defined searches that are used to gain insight into a given data set without requiring any predefined dashboards.
Similarly, high-level data mining query languages or other high-level flexible user interfaces will give users the freedom to define ad hoc data mining tasks.
Presentation and visualization of data mining results: How can a data mining system present data mining results, vividly and flexibly, so that the discovered knowledge can be easily understood and directly usable by humans?
This is especially crucial if the data mining process is interactive. It requires the system to adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.
Efficiency and scalability of data mining algorithms: Data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams. In other words, the running time of a data mining algorithm must be predictable, short, and acceptable by applications.
Parallel, distributed, and incremental mining algorithms: The humongous size of many data sets, the wide distribution of data, and the computational complexity of some data mining methods are factors that motivate the development of parallel and distributed data-intensive mining algorithms.
Cloud computing and cluster computing, which use computers in a distributed and collaborative way to tackle very large-scale computational tasks, are also active research themes in parallel data mining.
Diversity of Database Types:
The wide diversity of database types brings about challenges to data mining. These include
Handling complex types of data: Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi-structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, biological sequences, sensor data, spatial data, hypertext data, multimedia data, software program code, Web data, and social network data..
Mining dynamic, networked, and global data repositories: Multiple sources of data are connected by the Internet and various kinds of networks, forming gigantic, distributed, and heterogeneous global information systems and networks.
Data Mining and Society:
How does data mining impact society? What steps can data mining take to preserve the privacy of individuals? Do we use data mining in our daily lives without even knowing that we do? These questions raise the following issues
Social impacts of data mining: How does data mining impact society? What steps can data mining take to preserve the privacy of individuals? Do we use data mining in our daily lives without even knowing that we do? These questions raise the following issues:
Social impacts of data mining: With data mining penetrating our everyday lives, it is important to study the impact of data mining on society. How can we use data mining technology to benefit society? How can we guard against its misuse? The improper disclosure or use of data and the potential violation of individual privacy and data protection rights are areas of concern that need to be addressed.
Privacy-preserving data mining: Data mining will help scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyber attacks). However, it poses the risk of disclosing an individual's personal information.
Invisible data mining: We cannot expect everyone in society to learn and master data mining techniques. More and more systems should have data mining functions built within so that people can perform data mining or use data mining results simply by mouse clicking, without any knowledge of data mining algorithms. Intelligent search engines and Internet-based stores perform such invisible data mining by incorporating data mining into their components to improve their functionality and performance.
Data Mining Architecture
The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.
Data Source:
The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses.

Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires be cleaning and unifying. Several methods may be performed on the data as part of selection, integration, and cleaning.



Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Graphical User Interface:

After communicating data with engines and various pattern evaluation modules, it's necessary to share several components and make it user-friendly. Therefore, the need for a graphical user interface popularly known as GUI to effectively and efficiently use all the present components with the data mining system.

Data Mining Examples

The predictive capacity of data mining has changed the business strategies design. Below listed are some examples in the current industry.

Marketing: In marketing, data mining is used to explore large databases and improve market segmentation. It analyses various parameters like customers, age, gender, etc., to guess their behavior and direct personalized loyalty campaigns.

Retail: Supermarkets are well-known users of data mining techniques. It analyses the purchasing patterns of customers to identify product associations. Also detects the offers which are most valued by customers or increase sales at the checkout queue.

Banking: In identifying market risks, banks use data mining techniques. For credit ratings and anti-fraud systems to analyze customer purchasing patterns, card transactions, and more. It helps banks learn more about user online preferences to improve return on their marketing campaigns, sales performance, and manage regulatory compliance obligations.

UNIT-WISE QUESTIONS:

1. Compare and contrast operational data base systems with data warehouse.
2. Discuss number of data repositories on which mining can be performed
3. Define data mining and explain architecture of data mining system
4. Describe data mining functionalities and kinds of patterns they can be mined.
5, Write a brief note on relational databases and data warehouses
6. Describe the various phases in knowledge discovery process with a neat diagram.
7. Explain how the evolution of database technology led to data.
8. What are the major issues in data mining and explain in detail
9. Explain the techniques used in data mining
10. Discuss major application s targeted in data mining
11. Discuss the kinds of data can be mined in data mining