BCA Semester-6

Introduction to

Data Mining and Data Warehousing

Subject Code : CC - 308

Dr. Vimal Pandya
Ph.D. (Computer Science)
Director, Navgujarat College of Computer Applications,
Ashram Road, Ahmedabad

Dr. Shivang Patel
Ph.D. (Computer Applications),
M.Phil. (Computer Science)
National Forensic Sciences University.
Former Assistant Professor,
Shri Chimanbhai Patel Institute of Computer Applications

Dr. Sanjay B. Sonar
PhD [Computer Science]
Navgujarat College of Computer Applications,
Ashram Road, Ahmedabad.

Prof. Divita Patel
MCA, MA(English), B.Ed.
Umiya Arts and Commerce College,
Next to Umiya Dham, Sola, Ahmedabad-60

Reviewer: Dr. Bhavna Bajpai
Ph.D. (Computer Science)
Associate Professor, Dept of Computer,
Dr . C. V. Raman University, Khandwa (MP) - 01

COMPUTER® WORLD
PUBLICATION
A Division of Live Education System Pvt. Ltd.

COMPUTER WORLD
43, 5th Floor, SANIDHYA Complex, Nr. M. J. Library,
Opp. Sanyas Ashram, Ashram Road, Ahmedabad-09.
Mobile : 9725019114, 9725022917, 9725020595, 9825020595
URL : www.computerworld.Ind.In | e-Mail :Info@computerworld.Ind.In

Index

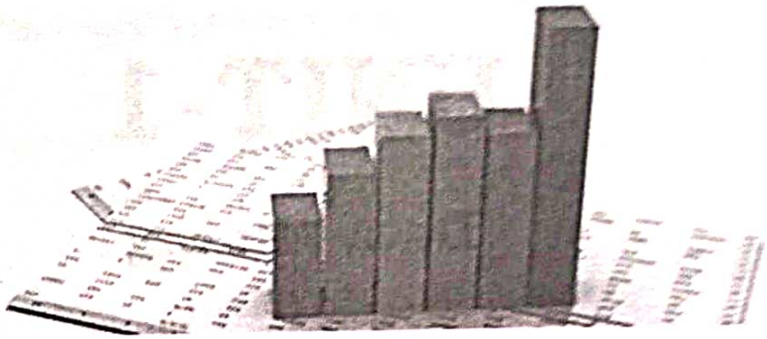
UNIT-1

Data Mining

- What is Data Mining?
- What kind of Data can be Mined?
- What Kind of patterns can be mined?
- Which technologies can be used?
- Major Issues in Data Mining

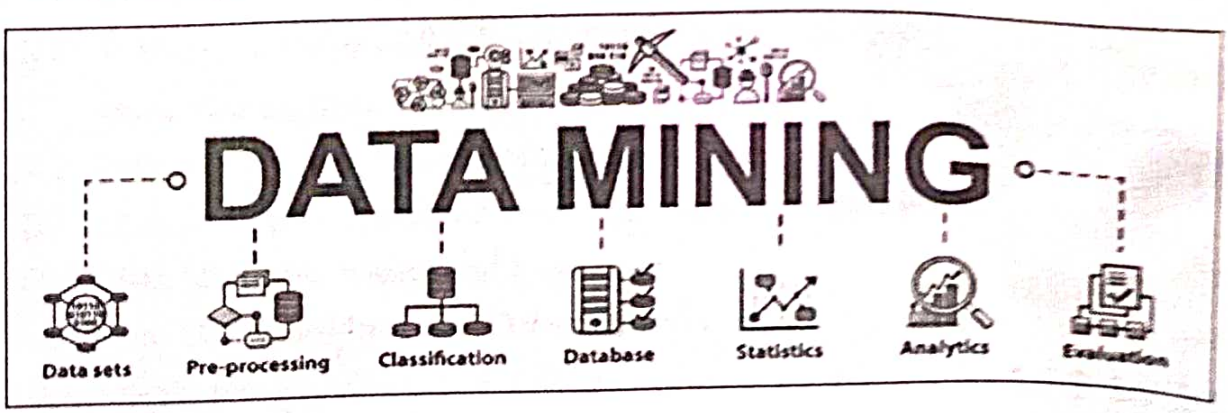
Unit –1 Data Mining

1.6 What is Data Mining?



" Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs."

- Data mining is the process of analyzing a large batch of information to discern trends and patterns.
- Data mining can be used by corporations for everything from learning about what customers are interested in or want to buy to fraud detection and spam filtering.
- Data mining programs break down patterns and connections in data based on what information users request or provide.
- Social media companies use data mining techniques to commodify their users in order to generate profit.
- This use of data mining has come under criticism lately s users are often unaware of the data mining happening with their personal information, especially when it is used to influence preferences.



Data mining isn't a new invention that came with the digital age. The concept has been around for over a century, but came into greater public focus in the 1930s. One of the first instances of data mining occurred in 1936, when Alan Turing introduced the idea of a universal machine that could perform computations similar to those of modern-day computers.

We've come a long way since then. Businesses are now harnessing data mining and machine learning to improve everything from their sales processes to interpreting financials for investment purposes. As a result, data scientists have become vital to organizations all over the world as companies seek to achieve bigger goals with data science than ever before.

Data mining is the process of analyzing massive volumes of data to discover business intelligence that helps companies solve problems, mitigate risks, and seize new opportunities. This branch of data science derives its name from the similarities between searching for valuable information in a large database and mining a mountain for ore. Both processes require sifting through tremendous amounts of material to find hidden value.

Data mining can answer business questions that traditionally were too time consuming to resolve manually. Using a range of statistical techniques to analyze data in different ways, users can identify patterns, trends and relationships they might otherwise miss. They can apply these findings to predict what is likely to happen in the future and take action to influence business outcomes.

Data mining is used in many areas of business and research, including sales and marketing, product development, healthcare, and education. When used correctly, data mining can provide a profound advantage over competitors by enabling you to learn more about customers, develop effective marketing strategies, increase revenue, and decrease costs.

Knowledge from data mining can help companies and governments cut costs or increase revenue. For example, an early form of data mining was used by companies to analyze huge amounts of scanner data from supermarkets. This analysis revealed when people were most likely to shop, and when they were most likely to buy certain products, like wine or baby products. This enabled the retailer to maximize revenue by ensuring they always had enough product at the right time in the right place. One of the first best selling systems was A.C. Nielson's best-selling Spotlight, which broke down supermarket sales data into multiple dimensions including volume by region and product type (Piatesky-Shapiro et. al, 1996).

➢ Key Data Mining Concepts:

Achieving the best results from data mining requires an array of tools and techniques. Some of the most commonly-used functions include:

Data cleansing and preparation — A step in which data is transformed into a form suitable for further analysis and processing, such as identifying and removing errors and missing data.

Artificial intelligence (AI) — These systems perform analytical activities associated with human intelligence such as planning, learning, reasoning, and problem solving.

Association rule learning — These tools, also known as market basket analysis, search for relationships among variables in a dataset, such as determining which products are typically purchased together.

Clustering — A process of partitioning a dataset into a set of meaningful sub-classes, called clusters, to help users understand the natural grouping or structure in the data.

Classification — This technique assigns items in a dataset to target categories or classes with the goal of accurately predicting the target class for each case in the data.

Data analytics — The process of evaluating digital information into useful business intelligence.

Data warehousing — A large collection of business data used to help an organization make decisions. It is the foundational component of most large-scale data mining efforts.

Machine learning — A computer programming technique that uses statistical probabilities to give computers the ability to "learn" without being explicitly programmed.

Regression — A technique used to predict a range of numeric values, such as sales, temperatures, or stock prices, based on a particular data set.

➢ Advantages of Data Mining:

Data is pouring into businesses in a multitude of formats at unprecedented speeds and volumes. Being a data-driven business is no longer an option; the business' success depends on how quickly you can discover insights from big data and incorporate them into business decisions and processes, driving better actions across your enterprise. However, with so much data to manage, this can seem like an insurmountable task.

Data mining empowers businesses to optimize the future by understanding the past and present, and making accurate predictions about what is likely to happen next. For example, data mining can tell you which prospects are likely to become profitable

customers based on past customer profiles, and which are most likely to respond to a specific offer. With this knowledge, you can increase your return on investment (ROI) by making your offer to only those prospects likely to respond and become valuable customers.

You can use data mining to solve almost any business problem that involves data, including:

- ✓ Increasing revenue.
- ✓ Understanding customer segments and preferences.
- ✓ Acquiring new customers.
- ✓ Improving cross-selling and up-selling.
- ✓ Retaining customers and increasing loyalty.
- ✓ Increasing ROI from marketing campaigns.
- ✓ Detecting fraud.
- ✓ Identifying credit risks.
- ✓ Monitoring operational performance.

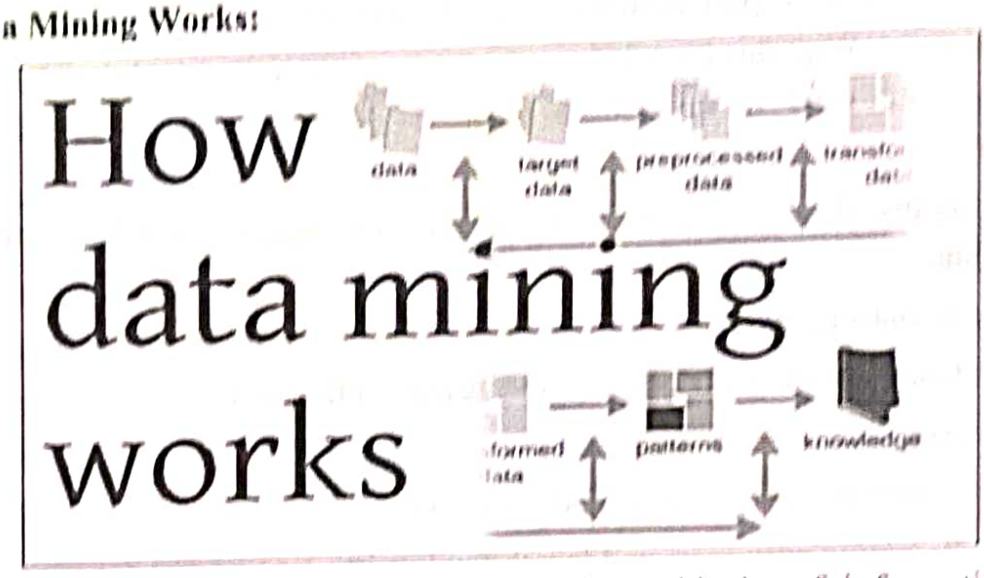
Through the application of data mining techniques, decisions can be based on real business intelligence — rather than instinct or gut reactions — and deliver consistent results that keep businesses ahead of the competition.

As large-scale data processing technologies such as machine learning and artificial intelligence become more readily accessible, companies are now able to dig through terabytes of data in minutes or hours, rather than days or weeks, helping them innovate and grow faster.

➢ Uses of Data Mining?

Data mining is primarily used by industries that cater to the consumer, like retail, financial and marketing companies. If you've ever shopped at a retail store and received customized coupons, that's a result of mining. Your individual purchase history was analyzed to find out what products you've been buying and what promotions you're likely to be interested in. Netflix uses data mining to recommend movies to its customers, Google uses mining to tailor advertisements to internet users and Walmart uses data mining to manage inventory and identify areas where new products are likely to be successful. Mining is more likely to be used by larger companies, as enormous computers are required to sift through data.

➢ **How Data Mining Works:**



Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It can be used in a variety of ways, such as database marketing, credit risk management, fraud detection, spam Email filtering, or even to discern the sentiment or opinion of users.

The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams, and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-share format, such as a graph or table.

A typical data mining project starts with asking the right business question, collecting the right data to answer it, and preparing the data for analysis. Success in the later phases is dependent on what occurs in the earlier phases. Poor data quality will lead to poor results, which is why data miners must ensure the quality of the data they use as input for analysis.

Data mining practitioners typically achieve timely, reliable results by following a structured, repeatable process that involves these six steps:

Business understanding — Developing a thorough understanding of the project parameters, including the current business situation, the primary business objective of the project, and the criteria for success.

Data understanding — Determining the data that will be needed to solve the problem and gathering it from all available sources.

Data preparation — Preparing the data in the appropriate format to answer the business question, fixing any data quality problems such as missing or duplicate data.

Modeling — Using algorithms to identify patterns within the data.

Evaluation — Determining whether and how well the results delivered by a given model will help achieve the business goal. There is often an iterative phase to find the best algorithm in order to achieve the best result.

Deployment — Making the results of the project available to decision makers.

Throughout this process, close collaboration between domain experts and data data miners is essential to understand the significance of data mining results to the business question being explored.

➢ Data Mining Process:

Data Mining Implementation Process

1. Business understanding:

In this phase, business and data-mining goals are established.

- First, you need to understand business and client objectives. You need to define what your client wants (which many times even they do not know themselves)
- Take stock of the current data mining scenario. Factor in resources, assumption, constraints, and other significant factors into your assessment.
- Using business objectives and current scenario, define your data mining goals.
- A good data mining plan is very detailed and should be developed to accomplish both business and data mining goals.

2. Data understanding:

In this phase, sanity check on data is performed to check whether its appropriate for the data mining goals.

- First, data is collected from multiple data sources available in the organization.
- These data sources may include multiple databases, flat filer or data cubes. There are issues like object matching and schema integration which can arise during Data Integration process. It is a quite complex and tricky process as data from various sources unlikely to match easily. For example, table A contains an entity named cust_no whereas another table B contains an entity named cust-id.
- Therefore, it is quite difficult to ensure that both of these given objects refer to the same value or not. Here, Metadata should be used to reduce errors in the data integration process.

- Next, the step is to search for properties of acquired data. A good way to explore the data is to answer the data mining questions (decided in business phase) using the query, reporting, and visualization tools.
- Based on the results of query, the data quality should be ascertained. Missing data if any should be acquired.

3. Data preparation:

- In this phase, data is made production ready.
- The data preparation process consumes about 90% of the time of the project.
- The data from different sources should be selected, cleaned, transformed, formatted, anonymized, and constructed (if required).
- Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values.
- For example, for a customer demographics profile, age data is missing. The data is incomplete and should be filled. In some cases, there could be data outliers. For instance, age has a value 300. Data could be inconsistent. For instance, name of the customer is different in different tables.
- Data transformation operations change the data to make it useful in data mining. Following transformation can be applied.

4. Data transformation:

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

Smoothing: It helps to remove noise from the data.

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

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

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

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

The result of this process is a final data set that can be used in modeling.

5. Modelling:

In this phase, mathematical models are used to determine data patterns.

- Based on the business objectives, suitable modeling techniques should be selected for the prepared dataset.
- Create a scenario to test check the quality and validity of the model.
- Run the model on the prepared dataset.

- Results should be assessed by all stakeholders to make sure that model can meet data mining objectives.

6. Evaluation:

In this phase, patterns identified are evaluated against the business objectives.

- Results generated by the data mining model should be evaluated against the business objectives.
- Gaining business understanding is an iterative process. In fact, while understanding, new business requirements may be raised because of data mining.
- A go or no-go decision is taken to move the model in the deployment phase.

7. Deployment:

In the deployment phase, you ship your data mining discoveries to everyday business operations.

- The knowledge or information discovered during data mining process should be made easy to understand for non-technical stakeholders.
- A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created.
- A final project report is created with lessons learned and key experiences during the project. This helps to improve the organization's business policy.

Before the actual **data mining** could occur, there are several processes involved in **data mining implementation**. Here's how:

Step 1: Business Research – Before you begin, you need to have a complete understanding of your enterprise's objectives, available resources, and current scenarios in alignment with its requirements. This would help create a detailed **data mining plan** that effectively reaches organizations' goals.

Step 2: Data Quality Checks – As the data gets collected from various sources, it needs to be checked and matched to ensure no bottlenecks in the data integration process. The quality assurance helps spot any underlying anomalies in the data, such as missing data interpolation, keeping the data in top-shape before it undergoes mining.

Step 3: Data Cleaning – It is believed that 90% of the time gets taken in the selecting, cleaning, formatting, and anonymizing data before mining.

Step 4: Data Transformation – Comprising five sub-stages, here, the processes involved make data ready into final data sets. It involves:

- **Data Smoothing:** Here, noise is removed from the data.
- **Data Summary:** The aggregation of data sets is applied in this process.
- **Data Generalization:** Here, the data gets generalized by replacing any low-level data with higher-level conceptualizations.

- **Data Normalization:** Here, data is defined in set ranges.
- **Data Attribute Construction:** The data sets are required to be in the set of attributes before **data mining**.

Step 5: Data Modelling: For better identification of data patterns, several mathematical models are implemented in the dataset, based on several conditions.

➢ **Example of Data Mining:**

Grocery stores are well-known users of data mining techniques. Many supermarkets offer free loyalty cards to customers that give them access to reduced prices not available to non-members. The cards make it easy for stores to track who is buying what, when they are buying it, and at what price. After analyzing the data, stores can then use this data to offer customers coupons targeted to their buying habits and decide when to put items on sale or when to sell them at full price.

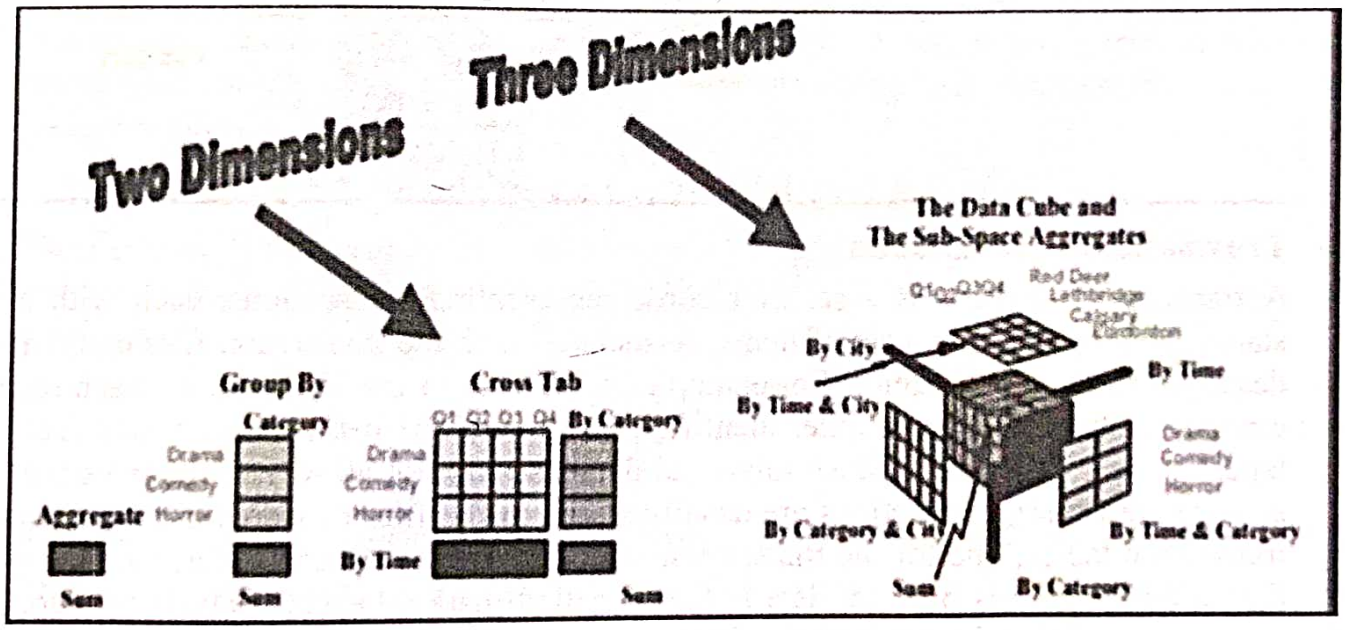
Data mining can be a cause for concern when a company uses only selected information, which is not representative of the overall sample group, to prove a certain hypothesis.

1.2 What kind of Data can be mined?

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. Here are some examples in more detail:

- **Flat files**: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific measurements, etc.
- **Relational Databases**: Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.
- **Data Warehouses**: A data warehouse as a storehouse, is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the same unified schema. A data warehouse gives the option to analyze data from different sources under the same roof. Let us suppose that OurVideoStore becomes a

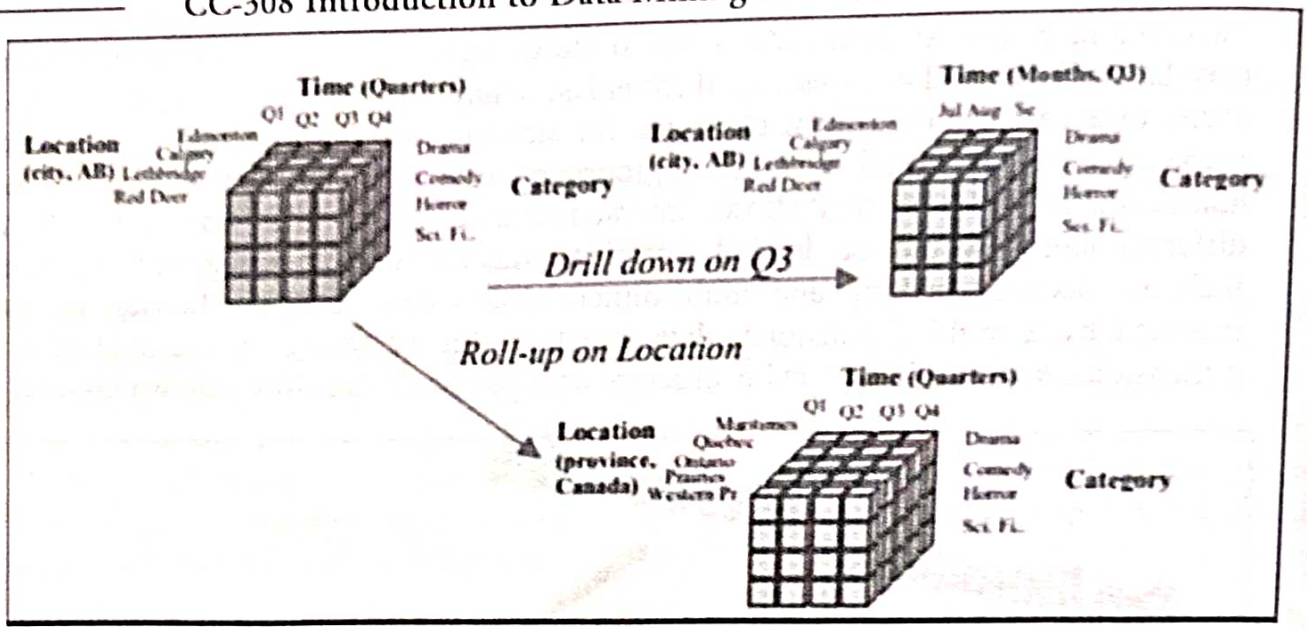
franchise in North America. Many video stores belonging to OurVideoStore company may have different databases and different structures. If the executive of the company wants to access the data from all stores for strategic decision-making, future direction, marketing, etc., it would be more appropriate to store all the data in one site with a homogeneous structure that allows interactive analysis. In other words, data from the different stores would be loaded, cleaned, transformed and integrated together. To facilitate decision-making and multi-dimensional views, data warehouses are usually modeled by a multi-dimensional data structure. Figure shows an example of a three dimensional subset of a data cube structure used for OurVideoStore data warehouse.



The figure shows summarized rentals grouped by film categories, then a cross table of summarized rentals by film categories and time (in quarters).

The data cube gives the summarized rentals along three dimensions: category, time, and city. A cube contains cells that store values of some aggregate measures (in this case rental counts), and special cells that store summations along dimensions. Each dimension of the data cube contains a hierarchy of values for one attribute.

Because of their structure, the pre-computed summarized data they contain and the hierarchical attribute values of their dimensions, data cubes are well suited for fast interactive querying and analysis of data at different conceptual levels, known as On-Line Analytical Processing (OLAP). OLAP operations allow the navigation of data at different levels of abstraction, such as drill-down, roll-up, slice, dice, etc. Figure illustrates the drill-down (on the time dimension) and roll-up (on the location dimension) operations.



Transaction Databases:

A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items. Associated with the transaction files could also be descriptive data for the items. For example, in the case of the video store, Each record is a rental contract with a customer identifier, a date, and the list of items rented (i.e. video tapes, games, VCR, etc.). Since relational databases do not allow nested tables (i.e. a set as attribute value), transactions are usually stored in flat files or stored in two normalized transaction tables, one for the transactions and one for the transaction items. One typical data mining analysis on such data is the so-called market basket analysis or association rules in which associations between items occurring together or in sequence are studied.

- **Multimedia Databases**: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational or object-oriented databases, or simply on a file system. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation, and natural language processing methodologies.

- **Spatial Databases**: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. Such spatial databases present new challenges to data mining algorithms.

- **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.

- **World Wide Web**: The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily. Data in the World Wide Web is organized in inter-connected documents. These documents can be text, audio, video, raw data, and even applications. Conceptually, the World Wide Web is comprised of three major components: The content of the Web, which encompasses documents available; the structure of the Web, which covers the hyperlinks and the relationships between documents; and the usage of the web, describing how and when the resources are accessed. A fourth dimension can be added relating the dynamic nature or evolution of the documents. Data mining in the World Wide Web, or web mining, tries to address all these issues and is often divided into web content mining, web structure mining and web usage mining.

1.3 What kind of patterns can be mined?

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. data mining tasks can be classified into two categories: descriptive and predictive.

Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

- **Concept/Class Description: Characterization and Discrimination**

Data can be associated with classes or concepts. For example, in the *AllElectronics* store, classes of items for sale include *computers* and *printers*, and concepts of customers include ***bigSpenders*** and ***budgetSpenders***. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via.

data characterization, by summarizing the data of the class under study (often called the target class) in general terms,

data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

- **Data characterization** is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query the output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs.

- **Data discrimination** is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target

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

Discrimination descriptions expressed in rule form are referred to as discriminate rules.

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 itemsets, subsequences, and substructures.

A ***frequent itemset*** typically refers to a set of items that frequently appear together in a transactional data set, such as Computer and Software. A frequently occurring subsequence, such as thepattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (*frequent*) *sequential pattern.*

- **Classification and Prediction:**

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

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

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can easily be converted to classification rules.

A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classification models, such as naïve.

Bayesian classification, support vector machines, and *k*-nearest neighbor classification. Whereas classification predicts categorical (discrete, unordered) labels, prediction models Continuous-valued functions.

That is, it is used to predict missing or unavailable *numerical data values* rather than class labels. Although the term *prediction* may refer to both numeric prediction and class label prediction.

- **Cluster Analysis:**

Classification and prediction analyze class-labeled data objects,
heres **clustering** analyzes data objects without consulting a known class label.

- **Outlier Analysis:**

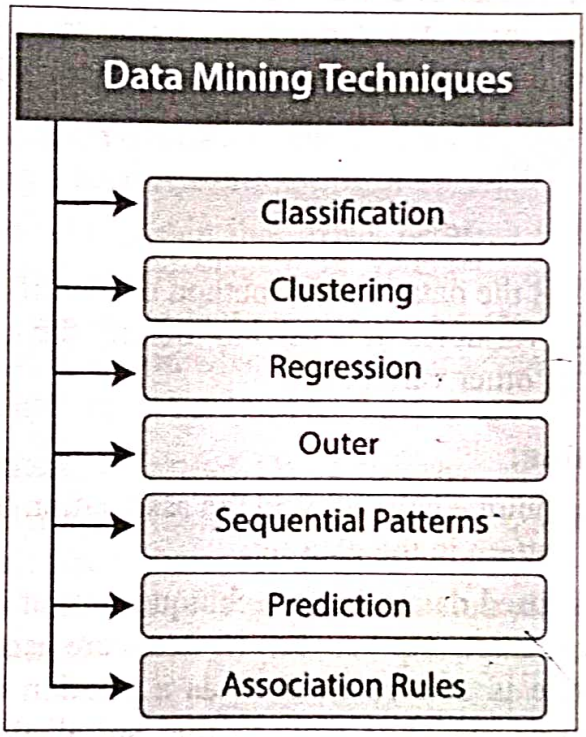
A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard

outliers as noise or exceptions. However, in some applications such as 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 mining.

- **Evolution Analysis:**

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of *time related* data, distinct features of such an analysis include time-series data analysis,Sequence or periodicity pattern matching, and similarity-based data analysis.

1.4 Which technologies can be used?



1.Classification:

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

This technique finds its origins in machine learning. It classifies items or variables in a data set into predefined groups or classes. It uses linear programming, statistics, decision trees, and artificial neural network in data mining, amongst other techniques. Classification is used to develop software that can be modelled in a way that it becomes capable of classifying items in a data set into different classes.

For instance, we can use it to classify all the candidates who attended an interview into two groups – the first group is the list of those candidates who were selected and the

second is the list that features candidates that were rejected. Data mining software can be used to perform this classification job.

2. Clustering:

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

This technique creates meaningful object clusters that share the same characteristics. People often confuse it with classification, but if they properly understand how both these techniques work, they won't have any issue. Unlike classification that puts objects into predefined classes, clustering puts objects in classes that are defined by it.

Let us take an example. A library is full of books on different topics. Now the challenge is to organize those books in a way that readers don't have any problem in finding out books on a particular topic. We can use clustering to keep books with similarities in one shelf and then give those shelves a meaningful name. Readers looking for books on a particular topic can go straight to that shelf. They won't be required to roam the entire library to find their book.

3. Regression:

Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

It is one of the most used data mining techniques out of all the others. In this technique, a transaction and the relationship between its items are used to identify a pattern. This is the reason this technique is also referred to as a relation technique. It is used to conduct market basket analysis, which is done to find out all those products that customers buy together on a regular basis.

This technique is very helpful for retailers who can use it to study the buying habits of different customers. Retailers can study sales data of the past and then lookout for products that customers buy together. Then they can put those products in close proximity of each other in their retail stores to help customers save their time and to increase their sales.

5. Outer detection:

This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a

variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns:

This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

This technique aims to use transaction data, and then identify similar trends, patterns, and events in it over a period of time. The historical sales data can be used to discover items that buyers bought together at different times of the year. Business can make sense of this information by recommending customers to buy those products at times when the historical data doesn't suggest they would. Businesses can use lucrative deals and discounts to push through this recommendation.

7. Prediction:

Prediction has used a combination of the other techniques of data mining like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

This technique predicts the relationship that exists between independent and dependent variables as well as independent variables alone. It can be used to predict future profit depending on the sale. Let us assume that profit and sale are dependent and independent variables, respectively. Now, based on what the past sales data says, we can make a profit prediction of the future using a regression curve.

➢ **Challenges of Implementation of Data mine:**

- Skilled Experts are needed to formulate the data mining queries.
- Overfitting: Due to small size training database, a model may not fit future states.
- Data mining needs large databases which sometimes are difficult to manage
- Business practices may need to be modified to determine to use the information uncovered.
- If the data set is not diverse, data mining results may not be accurate.
- Integration information needed from heterogeneous databases and global information systems could be complex.

1.5 What are the issues in Data Mining?

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way.

- **Security and social issues:**

  Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behaviour understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

- **User interface issues:**

  The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge? The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

- **Mining methodology issues:**

  These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently.

  Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the

accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information.

More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality*. This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

- **Performance issues:**

  Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming? There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

- **Data source issues:**

  There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and

methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

Exercises

❖ Answer the following Questions in brief.

Q.1 What is Data Mining?

Q.2 Discuss Key Data Mining Concepts.

Q.3 What is Advantages of Data Mining?

Q-4 Who uses Data Mining?

Q-5 How Data Mining Works?

Q-6 Explain Data Mining Process.

Q-7 What kind of Data can be mined?

Q-8 What kind of patterns can be mined?

Q-9 Explain Data Mining Techniques.

Q-10 What is Major issues in Data Mining?

❖ Indicate whether the following statements are true or false

1. Social media companies use data mining techniques to commodity their users in order to generate profit. (_____)
2. Data mining is a new invention that came with the digital age. (_____)
3. Achieving the best results from data mining requires an array of tools and techniques. (_____)

4. Data mining empowers businesses to optimize the future by understanding the past and present, and making accurate predictions about what is likely to happen next. (_____)
5. The data mining process breaks down into two steps. (_____)
6. Data cleaning is a process to "clean" the data by smoothing noisy data and filling in missing values. (_____)
7. Data mining should be not applicable to any kind of information repository. (_____)
8. Flat files are actually the most common data source for data mining algorithms, especially at the research level. (_____)
9. Multimedia databases include video, images, audio and text media. (_____)
10. Data characterization is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. (_____)
11. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. (_____)
12. Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. (_____)
13. Clustering analysis is the data mining method of identifying and analyzing the relationship between variables. (_____)
14. This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period. (_____)
15. Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. (_____)

Answer:

1. True	2. False	3. True	4. True	5. False
6. True	7. False	8. True	9. True	10. False
11. True	12. True	13. False	14. True	15. True