UNIT 1 Introduction to Data Mining

Data mining is the process of turning raw data into useful information. Any numbers, text, facts, web pages or documents that can be processed by a computer are considered data and mining is the process of extracting something useful. Hence, as the name indicates, data mining is the process of extracting useful information from large volumes of data. Data mining has become a buzzword for the last few years and businesses are trying to stand out from the crowd by making the best use of this technique.

Why Data Mining?

The simple answer to "Why is data mining required?" is that data which is the core of any business is anywhere and everywhere. Yes, it is a fact. We are living in a world where anything and everything is getting converted to data. Every click, tap, swipe, like, tweet, share, phone call, etc generates lots and lots of data. The amount of data getting collected and stored is exploding. Just consider the case of a telecom service provider, or a banking service provider. In short, data explosion is one of the reasons that necessitate data mining.

Secondly, the technology is so advanced today that it is really easy and cheap to collect, store and retrieve large volumes of data. Data storage costs have declined dramatically which result in big data. Also, the processing power of computers is exponentially increasing. All these technological advancements help organizations in collecting, storing and retrieving large amount of data from different sources easily and quite cheaply.

Thirdly, competition necessitates the availability of information at your finger tips in the blink of an eye. Your business might be storing terabytes of data in your databases by spending lots of effort, time and money. In addition to the data available within an organization, Internet is also a great data source. But, data in its pure form might not be useful in many situations. So in today's competitive business world, there should be processes in place in order to get useful information from raw data that might help you in critical decision making and development of new strategies.

What is Data Mining?

Data mining is the process of digging through large volumes of data and extracting previously unidentified and potentially useful information. In other words, data mining comes up with information that queries or reports cannot discover normally. By finding out useful patterns and trends about different aspects of the company, businesses can come up with new strategies that are helpful in gaining competitive advantage.

Data mining also predicts behaviors and future trends that help businesses to become more proactive and make more accurate, information-driven decisions. In short, data mining makes the whole process of information management faster, easier and efficient. It also answers business questions more accurately and efficiently.

Applications of Data Mining

Data mining helps businesses identify important facts, trends, patterns, relationships and exceptions that are normally unnoticed or hidden. Thus, data mining techniques are applied in a

wide range of industries including healthcare, insurance, finance, retail, manufacturing and so on.

Retailers make use of data mining techniques to spot sales trends. By analyzing the purchase patterns of customers, retailers can come up with smarter marketing promotions and campaigns which will in turn increase the sales. With market segmentation, retailers can identify the customers who purchase the same products. So, they can come up with new products at the right time by analyzing the interests and demographics of customers. Data mining can also be used to predict customers who are most likely start purchasing from your competitors.

Fraud detection is a major headache for finance and insurance companies. Studies show that customer demographics can be effectively used to predict their fraudulent nature. Nowadays, data mining is used to identify transactions that are most likely to be fraudulent. In the healthcare industry, data mining techniques are mainly used for most accurate disease diagnosis and most effective treatments. It is also helpful in predicting health insurance fraud, healthcare cost and length of stay (LOS) of hospitalization.

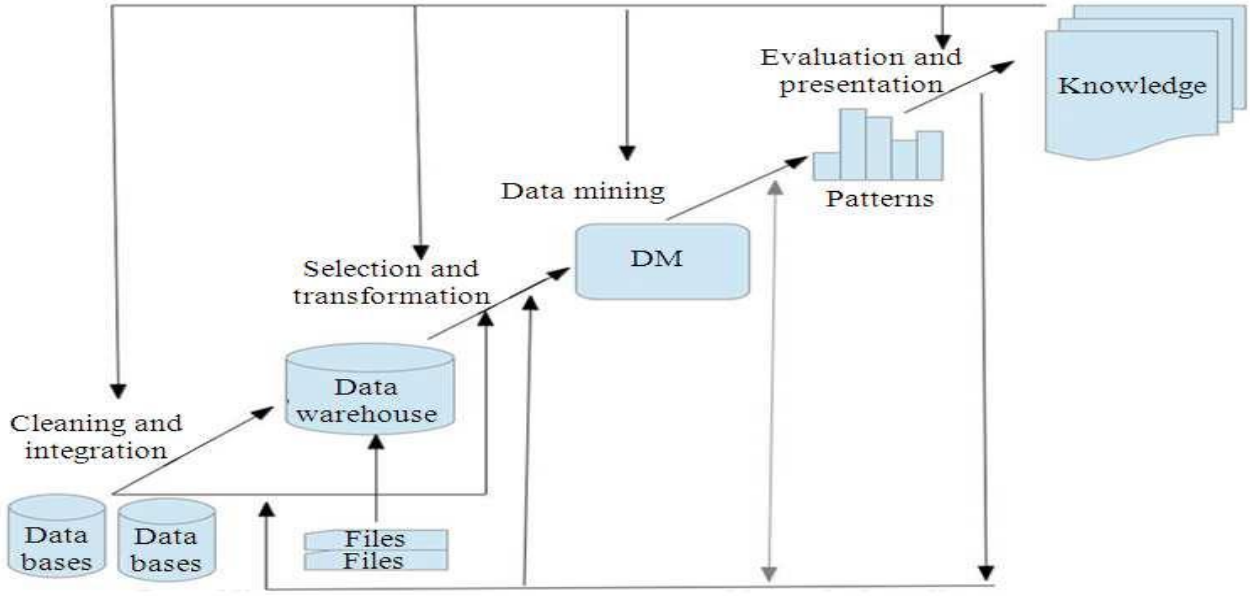


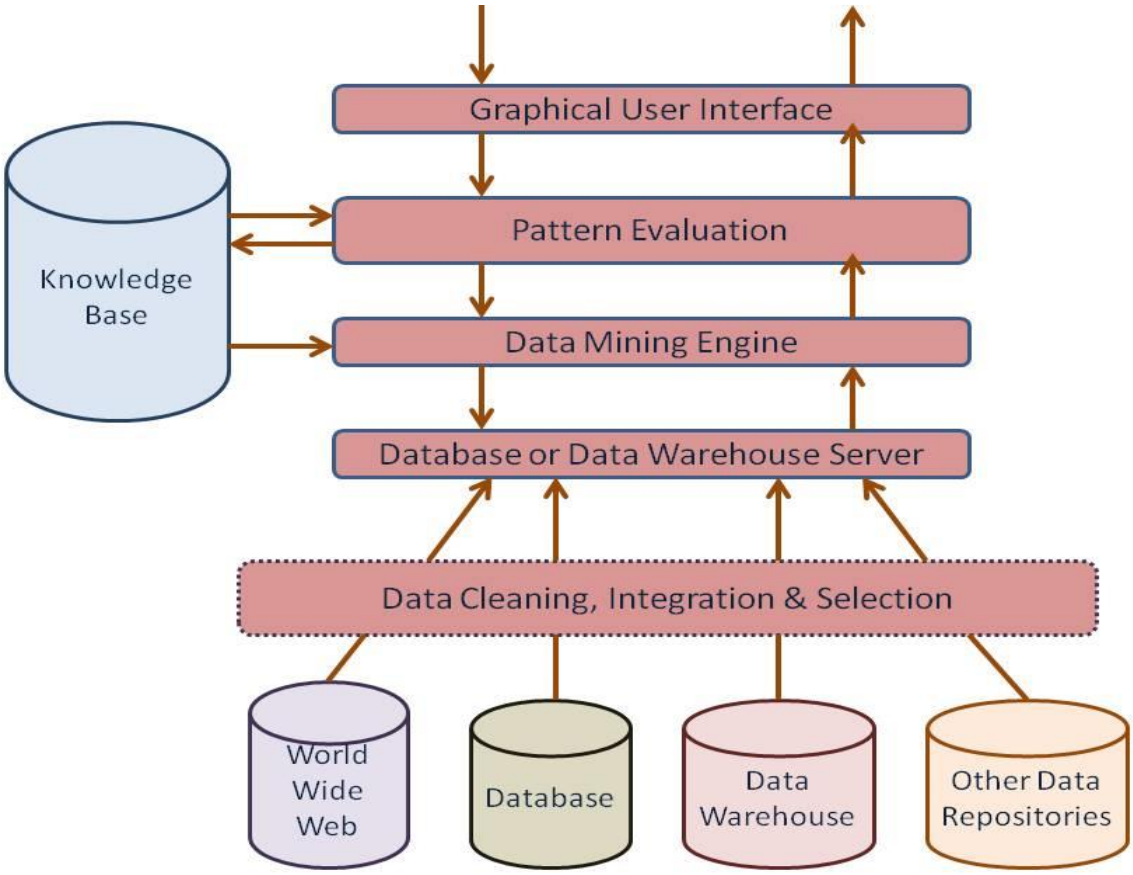
Fig1 Data mining as a step in process of knowledge discovery.

Data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process in depicted in fig 1 & consists of an iterative sequence of the following steps :

1. **Data cleaning** (to remove noise & 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 database)

4. **Data transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns)
6. **Pattern evaluation**(to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. **Knowledge presentation** (where visualization & knowledge representation techniques are used to present the mined knowledge to the user)

Architecture of Data Mining system



The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

a) Data Sources

Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

Different Processes

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server. These processes are not as simple as we think. A number of techniques may be performed on the data as part of cleaning, integration and selection.

b) Database or Data Warehouse Server

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

c) Data Mining Engine

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

d) Pattern Evaluation Modules

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

e) Graphical User Interface

The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

f) Knowledge Base

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data

mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

Difference Between Data mining and Machine learning

Basic for comparison	Data mining	Machine learning
Meaning	Extracting knowledge from a large amount of data	Introduce new algorithm from data as well as past experience
History	Introduce in 1930, initially referred as knowledge discovery in databases	introduce in near 1950, the first program was Samuel's checker-playing program
Responsibility	Data mining is used to get the rules from the existing data.	Machine learning teaches the computer to learn and understand the given rules.
Origin	Traditional databases with unstructured data	Existing data as well as algorithms.
Implementation	We can develop our own models where we can use data mining techniques for	We can use machine learning algorithm in the decision tree, neural networks and some other area of artificial intelligence.
Nature	Involves human interference more towards manual.	Automated, once design self-implemented, no human effort
Application	used in cluster analysis	used in web search, spam filter, credit scoring, fraud detection, computer design

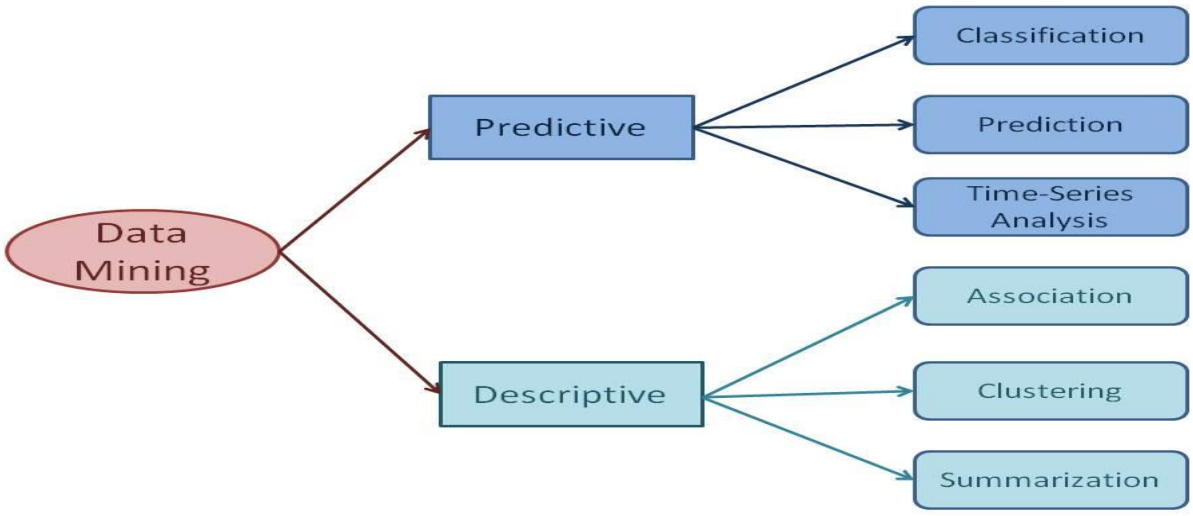
Abstraction	Data mining abstract from the data warehouse	Machine learning reads machine
Techniques involve	Data mining is more of a research using methods like machine learning	Self-learned and trains system to do the intelligent task.
Scope	Applied in the limited area	Can be used in a vast area.

Data Mining Tasks/ Models

The data mining tasks can be classified generally into two types based on what a specific task tries to achieve. Those two categories are descriptive tasks and predictive tasks. The descriptive data mining tasks characterize the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave.

Different Data Mining Tasks/ Models

There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining.



Predictive data mining tasks come up with a model from the available data set that is helpful in predicting unknown or future values of another data set of interest. A medical practitioner trying to diagnose a disease based on the medical test results of a patient can be considered as a predictive data mining task. Descriptive data mining tasks usually finds data describing patterns and comes up with new, significant information from the available data set. A retailer trying to identify products that are purchased together can be considered as a descriptive data mining task.

a) Classification

Classification derives a model to determine the class of an object based on its attributes. A collection of records will be available, each record with a set of attributes. One of the attributes will be class attribute and the goal of classification task is assigning a class attribute to new set of records as accurately as possible.

Classification can be used in direct marketing, that is to reduce marketing costs by targeting a set of customers who are likely to buy a new product. Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past. Hence, {purchase, don't purchase} decision forms the class attribute in this case. Once the class attribute is assigned, demographic and lifestyle information of customers who purchased similar products can be collected and promotion mails can be sent to them directly.

b) Prediction

Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

c) Time - Series Analysis

Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series reflects the process being measured and there are certain components that affect the behavior of a process. Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.

d) Association

Association discovers the association or connection among a set of items. Association identifies the relationships between objects. Association analysis is used for commodity management, advertising, catalog design, direct marketing etc. A retailer can identify the products that normally customers purchase together or even find the customers who respond to the promotion

of same kind of products. If a retailer finds that beer and nappy are bought together mostly, he can put nappies on sale to promote the sale of beer.

e) Clustering

Clustering is used to identify data objects that are similar to one another. The similarity can be decided based on a number of factors like purchase behavior, responsiveness to certain actions, geographical locations and so on. For example, an insurance company can cluster its customers based on age, residence, income etc. This group information will be helpful to understand the customers better and hence provide better customized services.

f) Summarization

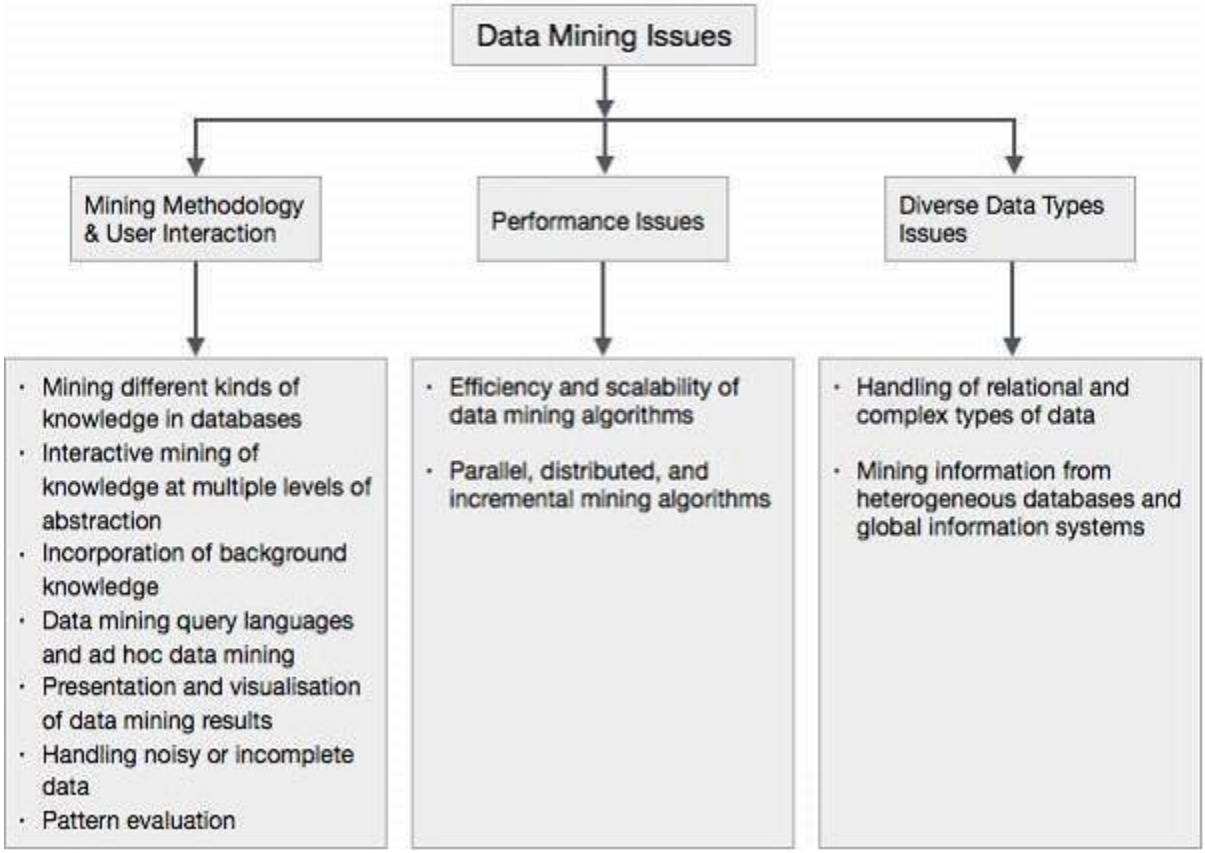
Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data. For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

Data Mining issues

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

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

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.