Data Mining

- Data mining overview
- Data mining processes
- How does data mining work?
- Different levels of analysis
- What technological infrastructure is required?

Data mining overview

- Data mining is a system of searching through large amounts of data for patterns.
- It is a relatively new concept which is directly related to computer science.
- Despite this, it can be used with a number of older computer techniques such as pattern recognition and statistics.

Data mining overview

- Data mining is a term that has become quite popular within certain industries.
- In a nutshell, data mining could be likened to finding a needle in a haystack.
- We live in a world that is full of information, and the biggest challenge is not only getting information,
- but searching through it to find connections and data that was not previously known.

Data mining overview

- The goal of data mining is to extract important information from data that was not previously known.
- Data mining is a technique that has a large number of applications in a wide variety of different fields.
- However, it is commonly used by businesses or organizations that need to recognize certain patterns or trends.

Data mining overview

- For example, the sales department of a company may use data mining to track the type of items a customer buys.
- As they begin observing the type of items the customer purchases,
- they may notice that the customer buys a large quantity of coffee on a regular basis.

Data mining overview

- The data mining program will automatically make a connection between that specific customer and a certain brand of coffee.
- The sales department can then use this information to launch a direct mail campaign
- which informs the customer of a sale that is being held for the brand of coffee they enjoy buying.

Data mining overview

- Because they know that the customer likes a particular brand of coffee,
- it is highly likely that they will purchace large quantities of it because it is on sale.
- In addition to this, the company can also market other products or services to the customer.
- Data mining allows the company to find out detailed information about their customer that would have been difficult to determine otherwise.

Data mining overview

- It has been said that knowledge is power, and this is exactly what data mining is about.
- It is the acquisition of relevant knowledge that can allow you to make strategic decisions which will allow your business or organization to succeed.
- Before you can efficiently use data mining tools, you must have large amounts of information in storage.

Data mining overview

- Most companies already have this information.
- A simple example of this would be a marketing list.
- A marketing list is information on a number of potential customers that you can market your products or services to.

Data mining overview

- Data mining can also be used to track behaviors within a system for a long period of time.
- For example, a large retail store chain may use data mining to analyze the type and number of items which are purchased over a five year period.
- The company may find that a certain brand of toothpaste and tooth brushes are purchased in large quantities.

Data mining overview

- Based on this information,
- the retail store chain could then proceed to take the toothpaste and tooth brushes and put them next to each other.
- This will allow their profits to greatly increase.
- The name for this sales method is Market Basket Analysis.

Data mining overview

- The increasing popularity of parallel computing has made it possible to search through massive amounts of data without the need for a theoretical framework.
- One important factor of data mining is that it will often be used to analyze information from a variety of different perspectives.
- The important information that is gained from data mining can be used to increase profits or lower costs.

Data mining overview

- There are a number of software products that have been designed for those who wish to use data mining techniques.
- Once you're able to search through large amounts of information, you will be able to analyze it in a large number of different ways.
- Once you've analyzed the information, you can make conclusions and decisions which are based on logic.

Data mining overview

- While the term data mining is a new, the concept of searching through data for patterns is not.
- Many large companies have powerful computers that allows them to search through information to analyze reports over a given period of time.

Data mining overview

- What sets data mining apart from these older research methods is that data mining is a result of the advancement of computer processing power.
- In addition to this, the storage capabilities of contemporary computers have allowed data mining to be much more accurate than techniques that were used in the past.

Data mining overview

- Because most data mining tools come in the form of software, the costs involved with searching and analyzing information have greatly dropped.
- For example, if a retail store noticed that a large number of their customers were purchasing alcoholic beverages on Thursday, this would tell them that the drinks are being purchased for the upcoming weekend.

Data mining overview

- The company could use this information by selling the drinks at full price on Thursdays in order to increase their profits.

Data mining overview

- Data mining has become a popular term among many companies and organizations.
- The reason why it has become so popular is because
- it will provide these institutions with knowledge that will allow them to make strategic decisions in a situation that is not certain.

Data mining overview

- In its core, data mining is an application of the mathematical system of statistics.
- The difference between data mining and other analytical tools is that it is not concerned with
- "why" a system behaves in a certain way.
- It is primarily concerned with "how," and
- it is used by organizations that are looking to use the information for a practical application.

Data Mining Process

- Data mining is a logical process that is used to search through large amounts of information in order to find important data.
- The goal of this technique is to find patterns that were previously unknown.
- Once you have found these patterns, you can use them to solve a number of problems.

Data Mining Process

- The goal of anyone who uses data mining should be to predict certain behaviors or patterns.
- Once you are able to predict the behavior of something you are analyzing,
- you will be able to make strategic decisions that can allow you to achieve certain goals.

Data Mining Process

- There are certain stages to data mining that you will want to become familiar with, and these are
- exploration,
- pattern identification, and
- deployment.

Exploration

- Exploration is the first stage, and as the name implies, you will want to explore and prepare data.
- You may need to clean the data you have, or it may need to be transformed into another form.
- In addition to this, you may also need to create records.

Exploration

- If you have a large number of variables to consider,
- you may need to reduce them to a range that is easy to deal with.
- Based on the problem that you are trying to solve,

Exploration

- you may need to either come up with a number of predictions,
- or you may need to use a wider selection of tools in order to analyze the data.
- An example of tools you could use are graphs and statistics.
- The goal of the exploration stage is to find important variables and determine their nature.

Pattern identification

- After you've explored, refined, and defined specific variables, move to stage 2,
- which is also called pattern identification.
- The first thing you will want to do is look for patterns and
- choose one that will allow you to make the best predictions.

Pattern identification

- This stage of data mining can be somewhat complex.
- There are a wide variety of different ways you can find the best predictive patterns.
- One of the best ways to do this is to apply different patterns to a given situation
- to determine which one performs at the highest level.

Pattern identification

- For example, if you are looking at data to find patterns that will allow your store to earn more profits,
- you could take two shopping patterns of your customers and apply them to a hypothetical strategy
- to determine which one performs the best.

Deployment

- The third stage is called deployment.
- You will not want to move to this stage until you have found a consistent pattern from stage 2
- that is highly predictive.

Deployment

- For example, if you find that many of your customers are consistently buying a specific product on a certain date,
- you will be able to predict their future behavior.
- Now that you've done this,
- you can take the pattern and apply it in order to see if you can achieve the desired outcome.

How does data mining work?

- While large-scale information technology has been evolving separate transaction and analytical systems,
- data mining provides the link between the two.
- Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries.

How does data mining work?

- Several types of analytical software are available:
- statistical, machine learning, and neural networks.
- Generally, any of four types of relationships are sought:

How does data mining work?

- **Classes**: Stored data is used to locate data in predetermined groups.
- For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order.
- This information could be used to increase traffic by having daily specials.

How does data mining work?

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

How does data mining work?

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

How does data mining work?

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

How does data mining work?

- Data mining consists of five major elements:
- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.

How does data mining work?

- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

Different levels of analysis

- **Decision trees**: Tree-shaped structures that represent sets of decisions.
- These decisions generate rules for the classification of a dataset.
- Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) .

Different levels of analysis

- CART and CHAID are decision tree techniques used for classification of a dataset.
- They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome.

Different levels of analysis

- CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits.
- CART typically requires less data preparation than CHAID.

Different levels of analysis

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset.
- Sometimes called the *k*-nearest neighbor technique.

Different levels of analysis

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data.
- Graphics tools are used to illustrate data relationships.

What technological infrastructure is required?

- Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms.
- System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest.
- Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes.

Technological drivers

- There are two critical technological drivers:
- **Size of the database**: the more data being processed and maintained, the more powerful the system required.

Technological drivers

- **Query complexity**: the more complex the queries and the greater the number of queries being processed,
- the more powerful the system required.