DATA MINING
&
WAREHOUSING

LECTURE 1: Introduction to Data Mining

DATA MINING

Introduction to Data Mining

There is huge amount of data available in Information Industry. This data is of no use until converted into useful information. Analyzing this huge amount of data and extracting useful information from it is necessary.

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

History of Data Mining

In 1960s statisticians used the terms "Data Fishing" or "Data Dredging". That was to refer to what they considered the bad practice of analyzing data. The term "Data Mining" appeared around 1990 in the database community.

What is Data Mining?

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. This information can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention

- Production Control
- Science Exploration

Advantages of Data Mining

1. Marketing / Retail

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have an appropriate approach to selling profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

2. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank, and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

3. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of the golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

4. Governments

Data mining helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

Challenges of data mining

1. **Privacy Issues**

The concerns about the personal privacy have been increasing enormously recently especially when the internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are afraid of their personal information is collected and used in an unethical way that potentially causing them a lot of troubles. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time, the personal information they own probably is sold to other or leak.

2. **Security issues**

Security is a big issue. Businesses own information about their employees and customers including social security number, birthday, payroll and etc. However how properly this information is taken care is still in questions. There have been a lot of cases that hackers accessed and stole big data of customers from the big corporation such as Ford Motor Credit Company, Sony… with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

3. **Misuse of information/inaccurate information**

Information is collected through data mining intended for the ethical purposes can be misused. This information may be exploited by unethical people or businesses to take benefits of vulnerable people or discriminate against a group of people.

In addition, data mining technique is not perfectly accurate. Therefore, if inaccurate information is used for decision-making, it will cause serious consequence.

Deriving business value from data mining

The real value of data mining comes from being able to unearth hidden gems in the form of patterns and relationships in data, which can be used to make predictions that can have a significant impact on businesses.

For example, if a company determines that a particular marketing campaign resulted in extremely high sales of a particular model of a product in certain parts of the country but not in others, it can refocus the campaign in the future to get the maximum returns.

The benefits of the technology can vary depending on the type of business and its goals. For example, sales and marketing managers in retail might mine customer information in different ways to improve conversion rates than those in the airline or financial services industries.

Regardless of the industry, data mining that's applied to sales patterns and client behavior in the past can be used to create models that predict future sales and behavior.

There's also the potential for data mining to help eliminate activities that can harm businesses. For example, you can use data mining to enhance product safety, or detect fraudulent activity in insurance and financial services transactions.

The applications of data mining

Data mining can be applied to a variety of applications in virtually every industry.

- **Retailers** can deploy data mining to better identify which products people are likely to purchase based on their past buying habits, or which goods are likely to sell at certain times of the year. This can help merchandisers plan inventories and store layouts.
- **Banks and other financial services providers** can mine data related to their clients' accounts, transactions, and channel preferences to better meet their needs. They can also gather then analyzed data from their websites and social media interactions to help increase the loyalty of existing customers and attract new ones.
- **Manufacturing companies** can use data mining to look for patterns in the production process, so they can precisely identify bottlenecks and flawed methods and find ways to increase efficiencies. They can also apply knowledge from data mining to the design of products, and make tweaks based on feedback from customer experiences.
- **Educational institutions** can benefit from data mining such as analyzing data sets to predict the future learning behaviors and performance of students, and then using this knowledge to make improvements in teaching methods or curricula.
- **Health care providers** can mine and analyze data to determine better ways of delivering care to patients and cutting costs. With the help of data mining, they can predict how many patients they will need to care for and what type of services those patients will need. In the life sciences, mining can be used to glean insights from massive biological data, to help develop new medicines and other treatments.
- In **multiple industries**, including health care and retail, you can use data mining to detect fraud and other abuses - much more quickly than with traditional methods for identifying such activities.

The key components of data mining

The process of data mining includes several distinct components that address different needs:

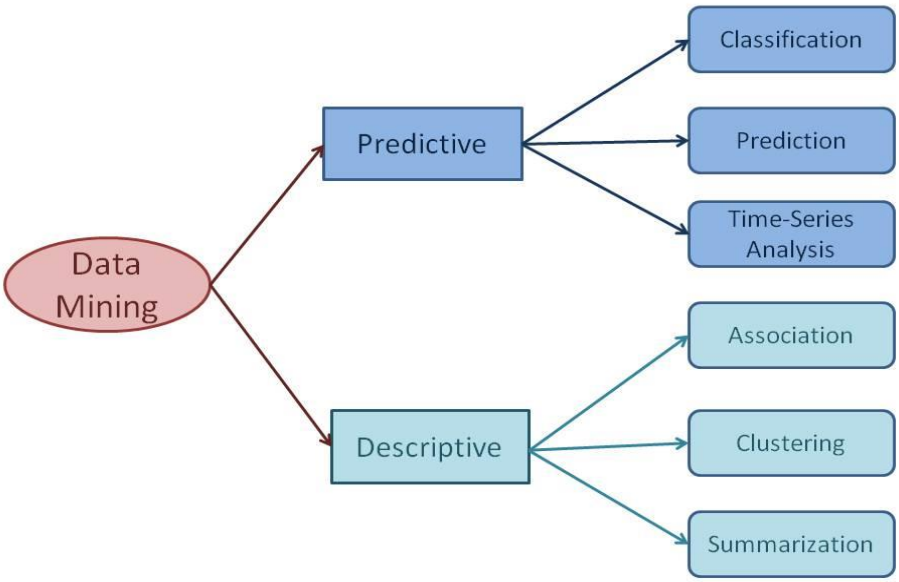
- **Preprocessing.** Before you can apply data mining algorithms, you need to build a target data set. One common source for data is a data mart or warehouse. You need to perform preprocessing to be able to analyze the data sets.
- **Data cleansing and preparation.** The target data set must be cleaned and otherwise prepared, to remove "noise," address missing values, filter outlying data points (for anomaly detection) to remove errors or do further exploration, create segmentation rules, and perform other functions related to data preparation.

- **Association rule learning** (also known as **market basket analysis**). These tools search for relationships among variables in a data set, such as determining which products in a store are often purchased together.
- **Clustering.** This feature of data mining is used to discover groups and structures in data sets that are in some way similar to each other, without using known structures in the data.
- **Classification.** Tools that perform classification generalize known structures to apply to new data points, such as when an email application tries to classify a message as legitimate mail or spam.
- **Regression.** This data mining technique tis used to predict a range of numeric values, such as sales, housing values, temperatures, or prices when given a particular data set.
- **Summarization.** This technique provides a compact representation of a data set, including visualization and report generation.

Data Mining Tasks

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.

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.



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.