# CHAPTER 6 DATA MINING

## CONTENTS

# 1. BEGINNING

- The exponential growth of data that are stored in various forms for decades.
- The goal is to find meaningful knowledge from large amounts of data.

# 2. DEFINITION

- Comprehensive definition of Data Mining

  The process to find meaningful knowledge from large amounts of data.

- Definition of Data Mining 1

  “Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.”(Gartner Group)

- Definition of Data Mining 2

  “Data mining is a knowledge discovery process of extracting previously unknown, actionable information from very large databases." (Aaron Zornes, The META Group)

- Definition of Data Mining 3

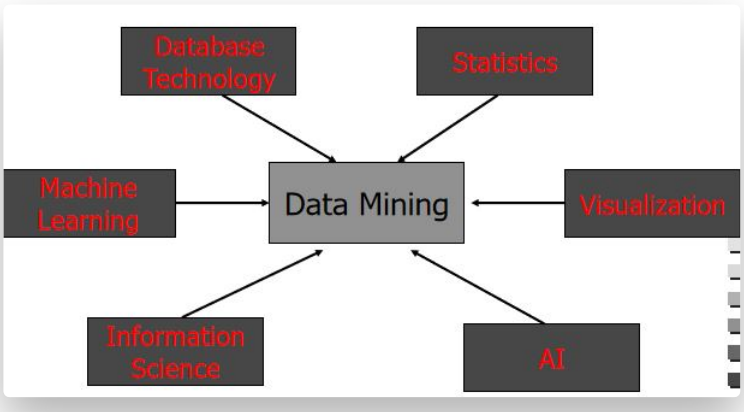
  "Data mining can discover hidden knowledge from existing databases"

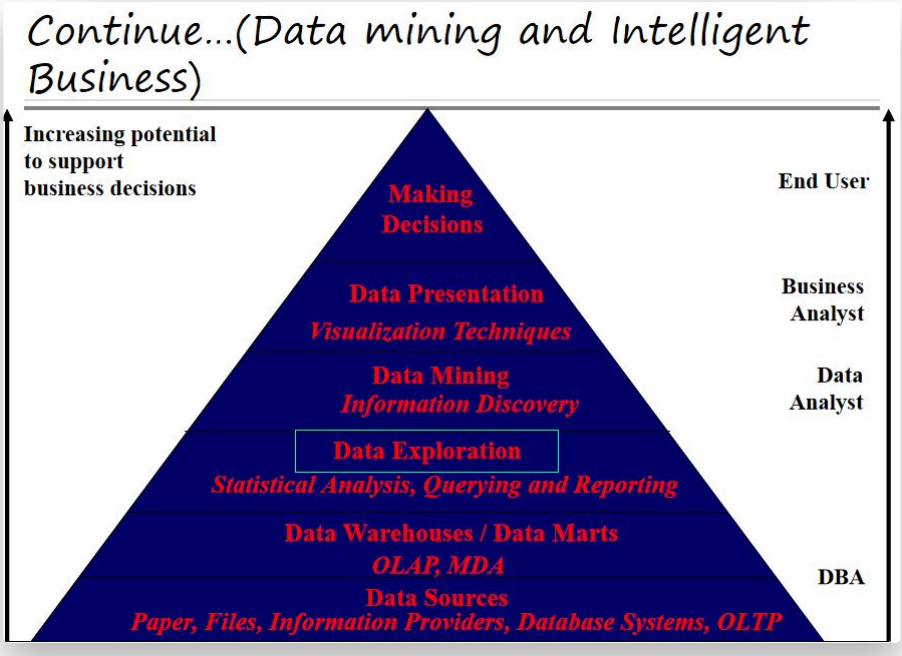
# 3. BACKGROUND

- Despite the increase in the amount of data to the lack of useful information and the difficulty of decision-making.
- The lack of data information to analyze and predict consumer buying patterns and needs.
- The spread of awareness on the need to build a data mining based on data warehouse activation
- Utilizing a wide range of large-scale transaction systems

# 4. FEATURES OF DATA MINING

- Data mining uses information from past data to analyze the outcome of a particular problem or situation that may arise.
- Practical using the processing power of the computer
- Instead of test and statistical inference, the main interest of Data Mining is a generalization of the prediction model.
- Data Mining started to be developed in the fields of statistics, computer science, artificial intelligence, and engineering.



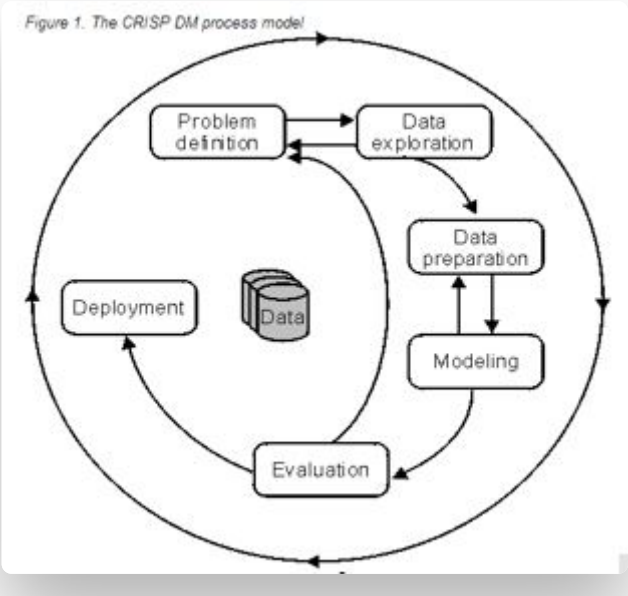

# 5. THE PROCESS OF DATA MINING

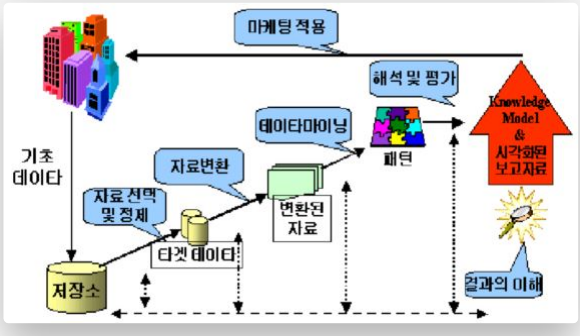
- Data mining is an iterative process that typically involves the following phases:
  - Problem Definition
    - A data mining project starts with the understanding of the business problem.
  - Data Exploration

        - Domain experts understand the meaning of the metadata.
    - Data preparation
        - Domain experts build the data model for the modeling process.
    - Modeling
        - Data Mining experts select apply various mining functions because you can use different mining functions for the same type of data mining problem.

- Data mining is an iterative process that typically involves the following phases:
    - Evaluation
        - Data mining experts evaluate the model.
    - Deployment
        - Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets.

- The following figure shows the phases of the Cross Industry Standard Process for data mining (CRISP DM) process model.



# 6. APPROACH OF DATA MINING

| Supervised Data Prediction | Unsupervised Data Prediction |
| --- | --- |
| Decision Tree Neural Network Regression Analysis Logistic Regression Case-Based Reasoning | Association Rule Discovery Market Basket Analysis Clustering |

01. **Supervised data :**
    i. **Training data includes both the input and the desired results**
    ii. **For some examples the correct results (targets) are known and are given in input to the model during the learning process**

02. **Unsupervised data :**
    i. **The model is not provided with the correct results during the training**
    ii. **The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes**

- Example
    - 01. Supervised Data for the target EnjoySoprt

      Every row's EnjoySport will supervises for every row's results.

      It says that supervised data is an evaluation data.

      For 1 row, this situation is enjoyable, so the EnjoySport is Yes.

      For 3 row, this situation is not enjoyable, so the EnjoySport is No.

| Case Index | Sky | AirTemp | Humidity | Wind | Water | Forecast | EnjoySport |
|---|---|---|---|---|---|---|---|
| 1 | Sunny | Warm | Normal | Strong | Warm | Same | Yes |
| 2 | Sunny | Warm | High | Strong | Warm | Same | Yes |
| 3 | Rainy | Cold | High | Strong | Warm | Change | No |
| 4 | Sunny | Warm | High | Strong | Cool | Change | Yes |

- 02. Unsupervised Data

  There is no evaluation data for every situation.

| TID | Items |
|---|---|
| 1 | Coke, Bread, Hamburger |
| 2 | Milk, Sandwich, Juice |
| 3 | Coke, Bread, Juice, Sandwich |
| 4 | Coke, Milk, Juice |

# 7. * ASSOCIATION RULE --- TECHNIQUES OF DATA MINING

## H3 Introduce

- Proposed by Agrawal et al in 1993.
- It is an important data mining model studied extensively by the database and data mining community
- Assume all data are categorical
- No good algorithm for numeric data
- Initially used for Market Basket Analysis to find how items purchased by customers are related

## H3 Example

Given: (1) database of transactions, (2) each transaction is a list of items (purchased by a customer in a visit)

Find: all rules that correlate the presence of one set of items with that of another set of items

- E.g., 98% of people who purchase tires and auto accessories also get automotive services done

Applications

- ? ⇒ Maintenance Agreement (What the store should do to boost Maintenance Agreement sales)
- Home Electronics ⇒ ? (What other products should the store stocks up?)
- Attached mailing in direct marketing

## H3 Sequential pattern

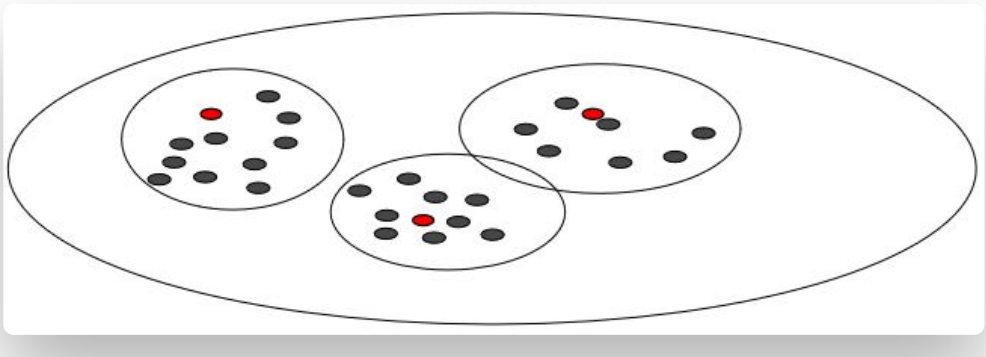
- Customer shopping sequences:
    - First buy computer, then CD-ROM, and then digital camera, within 3 months
- Medical treatments, natural disasters (e.g., earthquakes), science & eng. processes, stocks and markets, etc.
- Telephone calling patterns, Weblog click streams
- DNA sequences and gene structures
- Sequential Pattern Mining is useful in many application, e.g. weblog analysis, financial market prediction, BioInformatics, etc.
- It is similar to the frequent itemsets mining, but with consideration of ordering.

## H3 Clustering

- Clustering analysis finds clusters of data objects that are similar in some sense to one another
- The members of a cluster are more like each other than they are like members of other clusters.
- Clustering, like classification, is used to segment the data
- Clustering is useful for exploring data
- If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings
- Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models.

## H3 * K –Mean algorithm

- K-means (MacQueen, 1967) is one of the simplest unsupervised learning algorithms that solve the well known clustering problem.
- The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.
- **The algorithm is composed of the following steps:**
  - **Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.**
  - **Assign each object to the group that has the closest centroid.**
  - **When all objects have been assigned, recalculate the positions of the K centroids.**
  - **repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.**
  - **In general, the initial center values (red dots) determined randomly, as shown in the following figure. In other words, anywhere in the two-dimensional space, such as taking points K principle.**



  - **Located close to the center of each of the points that correspond to the values assigned to clusters.**
    **Calculated by the following formula:**

$$\text{Distance} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

  - **Belonging to each cluster by calculating the average of the points, calculate the value of the new center of each cluster.**

- **Repeat steps 4, 5.**

  **When the center does not change the value of cluster**

  **-> Formation of the final cluster**

## H3 * Association Rule

- In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.
- support and confidence as in apriori:

  an arbitrary combination of supported interest measures can be used
- Apriori is the best-known algorithm to mine association rules
- 연관규칙은 항목의 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙이다. 가장 먼저 개발 되었으며 또 가장 많이 쓰이는 알고리즘이다

| | hamburger | coke | both |
|---|---|---|---|
| { coke, bread, hamburger } | o | o | o |
| { coke, hamburger , juice} | o | o | o |
| { milk, sandwich, juice } | x | x | x |
| { sandwich, milk, juice, bread } | x | x | x |
| { hamburger, juice, coke} | o | o | o |
| { coke, bread, hamburger } | o | o | o |
| { coke, hamburger , juice} | o | o | o |
| { hamburger, juice } | o | x | x |
| { milk, hamburger, sweater } | o | x | x |
| { coke, milk, juice } | x | o | x |
| | 7 | 6 | 5 |

For an assoication rule
{coke} --> { hamburger },

support : 5 out of 10 = 50 %
confidence : 5 out of 6 = 83 %

- Description

  support = ( both / all ) * 100%

  confidence = (both / coke) * 100%  (Because coke is key in {coke} --> {hamburger})

## H4 * Two steps for Association Rule

- Determining **"large itemsets"**
- Find all combinations of items that have transaction support **above minimum support**
- Researches  have been  focused on this phase.
- Generating rules

```
for each large itemset L do
   for each subset c of L do
         if (support(L) / support(L - c) ≥ minimum confidence) then
            output the rule (L - c) → c,
                  with confidence = support(L)/support(L - c)
                  and support = support(L);
```

- Example

**Database D**

| TID | Items |
|---|---|
| 100 | A C D |
| 200 | B C E |
| 300 | A B C E |
| 400 | B E |

K=3, MinS = 2

**$L_3$**

| Item set | Support |
|---|---|
| {BCE} | 2 |

$L_3 = \{\{B\,C\,E\}\}$, $c$ *is Subset of* $B\,C\,E$

$If\ c = \{B, C\}, minCon = 60\ then\ S(L_3) = 50,\ S(L_3 - c) = 75$

$Therefore\ S(L3)\ /S(L_3 - c) = C(c) that\ is\ equal\ to\ 66$

$If\ C(c) >= minCon\ then$

**R : E -> B, C (50, 66)**

## H3 Algorithm Apriori

$L_k$: Set of Large k-itemsets

$C_k$:Set of Candidate k-itemsets

step; $C_1 \rightarrow L_1$, $C_2 \rightarrow L_2$, ..., $C_k \rightarrow L_k$

nput File: Transaction File, Output: Large itemsets

$L_1$ = {large 1-itemset}
**for** ( k=2; $L_{k-1} \neq \emptyset$; k++) **do begin**
$C_k$= apriori-gen($L_{k-1}$);
**forall** transactions t $\in$ D **do begin**
$C_t$ = subset($C_k$, t);
**forall** candidates c $\in$ $C_t$ **do**
c.count++;
**end**
$L_k$= {c $\in$ $C_k$| c.count $\geq$ minsup}
**end**
Answer = $U_k$ $L_k$;

## H3 Apriori-gen(Lk-1)

Join step

**insert into** $C_k$
**select** $p.item_1$, $p.item_2$, ..., $p.item_{k-1}$, $q.item_{k-1}$
**from** $L_{k-1}$ p, $L_{k-1}$ q
**where** $p.item_1 = q.item_1$, ..., $p.item_{k-2} = q.item_{k-2}$,
$p.item_{k-1} < q.item_{k-1}$

Prune step

**forall** itemsets c $\in$ $C_k$ **do**
**forall** (k-1)-subsets s of c **do**
**if** ( s $\notin$ $L_{k-1}$ ) **then**
**delete** c from $C_k$;