Data Mining and Big Data

m5211143 - Tomohiro Saito

Database Systems Laboratory, 141-F

Hello everyone, I’m Tomohiro Saito.
Today, I would like to show you presentation about data mining and bigdata.

Background - Data mining (DM)

> Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
>
> Quoted from: Data mining - wikipedia

Variations

- Text mining -> text
- Web mining -> web resource

At first, I search about Data mining.
data mining is the computing process of …

For example, there are some variations of data mining, such as text mining and web mining.

Background - Big data

> Big data is a term for data sets
> that are so large or complex that traditional data
> processing application software is inadequate
> to deal with them.
>
> Quoted from: Big data - wikipedia

customer data

DB, server data

SNS

Next I search about Big data
Big data is a term for data sets…

For example, a customer data, Data base, server data, and SNS data.

Data mining and its advantage

Lots of data exist today, for example,

- Customer data in company
- Web data

Advantages to apply data mining to such big data is

- Reduce the cost for analyzing data
- Saving time and get more information from data

There is a lot of data today.
for example, customer data in company, web data as I mentioned before slide.

An advantages to apply data mining to such big data is

Data mining techniques

- Frequent pattern extraction

  → A method for extracting patterns appearing in a data set

- Regression analysis

  → A statistical process for estimating the relationships among variables

- Clustering

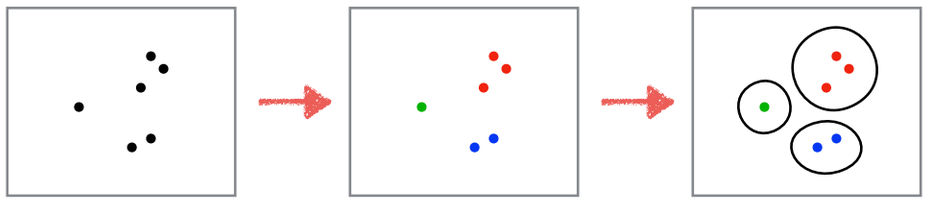
  → A method for extracting patterns appearing in a data set

There are several techniques for data mining.
Frequent pattern extraction is a method for extracting patterns…

What is Clustering?

- Clustering

  → A method for extracting patterns appearing in a data set



For example, three clusters are created according to Euclidian distance in above dataset.

In this time, I focused on clustering.
If we run clustering, the dataset will be separated in some clusters.
For example, in this example, there are three clusters.

Example. Clustering information resource in AOJ

There are many source codes in AOJ. By using Data mining method, we can analyze

- Characteristics of source codes
  - typical mistakes for solving problems
  - writing habit of user
  - etc…
- Problems in AOJ
  - Data manipulation
  - String processing
  - etc…

I also give an example for data mining.
My laboratory manage online judge system known as AOJ.
there are many information resource in AOJ.
If we use data mining such as clustering for datum in AOJ, we can know…

Problems applying data mining to big data

When we use data mining, we need to decide before clustering:

- Suitable method
  - Which algorithm is applicative for the dataset?
- Data selection
  - Which data would be usable for analysis?
- Accuracy
  - How to calculate the correctness of results which obtained by data mining?

At last, I will talk about problems when we use data mining method.
When we use data mining, we need to decide below term before clustering.

Thank you for listening.

This is all for my presentation about data mining and bigdata.
Thank you for your kind attention.