Unit 6: Information Privacy and Data Mining LH 3

6.1 Basic principles to Protect Information Privacy

6.2 Uses and Misuses of Data Mining

6.3 Primary Aims of data Mining

6.4 Pitfalls of Data Mining

Basic principles to Protect Information Privacy

What is information Privacy?

It is the claim of individuals, groups or institutions to determine for themselves when how and to what extent information about them is communicated to others. **It is the ability to control circulation of information.**

Information privacy can be achieved through encryption, authentication and data masking (Data masking is a way to create a fake, but a realistic version of your organizational data) to ensure that information is available only to those with authorized access.

Basic Principle to protect information. Organization for Economic Cooperation and Development (OECD has suggest following way to make data secure).

- Collection limitation
- Data quality
- Purpose specification
- Use limitation
- Security safeguards
- Openness
- Individual participation
- Accountability

Collection limitation:

- There should be limits to the collection of personal data and any such data should be obtained by **lawful** and **fair means and, where appropriate, with the knowledge or consent of the data subject.**

This principle deals with two issues.

- Firstly it limits the Collection of personal data. i.e Data can be collected as the law.
- Secondly, deals with practices used in data collection. The knowledge or consent of the data subject is essential although the principle accepts that consent of the data subject may not always be possible.

Data Quality

- Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.
- For example, data including opinions may not always be useful to collect since it may not be relevant for the purposes for which the data is to be used.

Purpose specification:

- The purpose of data collection should be identified before data collection and any changes afterwards must be specified. Also, when data no longer serves a purpose, it may be necessary to destroy the data if practicable.

Use limitation:

- Personal data should not be disclosed, made available or otherwise used for purposes other than those specified. Except with the consent of the data subject or by the authority of law.

Security safeguards:

- To enforce the principle of use limitations, it is required that there be appropriate safeguards. The safeguards should include physical safeguards as well as safeguards for the computer systems.

Openness:

- There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

Individual participation:

- An individual should have the right to obtain data from a data controller, or otherwise, confirmation or not, the data controller has data relating to him.

Accountability:

- The data controller should be accountable for complying with measures. The data collector comply with the privacy protection principles. Accountability should be supported by legal sanctions and perhaps a code of conduct.

Uses and Misuses of Data Mining

- The two primary goals of data mining tend to be **prediction and description**. Prediction involves using some variables or fields in the data set to **predict unknown or future values** of other variables of interest. Description, on the other hand, **focuses on finding patterns** describing the data that can be interpreted by humans.

- Data mining techniques are being used increasingly in a wide variety of applications.

  The applications include:

  - Fraud prevention
  - Catching tax avoidance
  - Catching drug smugglers
  - Reducing customer churn and learning more about customers behavior.

Following are primary data mining tasks:

- **Classification**: Discovery of a predictive learning function that classifies a data item into one of several predefined classes.
- **Regression**: Discovery of a predictive learning function that maps a data item to a real value prediction variable.
- **Clustering**: A common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.
- **Optimization**: enhance the use limited resources such as time space or material.

Pitfall of Data mining

Dempsey and Rosen Zweig identify five pitfalls of data mining.

These five categories are:

1. Unintentional Mistake-Mistaken identity
2. Unintentional Mistake- Faulty Inference
3. Intentional abuse
4. Security Breach
5. Mission Creep

Unintentional Mistake- Mistaken Identity:

- This kind problem arises when an innocent person shares some identifier information with one or more persons that have been identified as having a profile of interest in a data mining application.

Unintentional Mistake- Faulty inference:

- This kind of problem arises when profiles are matched and the data mining user misinterpret the result of the match.
- For example: An innocent person becoming a suspect and leading to further investigation about the person

Intentional abuse:

- People employed in running a data mining system and security organizations have access to sensitive personal information and not all of them are trust worthy. This sensitive information may be used for unauthorized purposes for personal financial gain.

Security Breach:

- Given that data mining system using personal information have sensitive information may be stolen or carelessly disclosed without strict security measures.

Mission Creep:

- A national security application, has been established and personal information from a variety of source collected, there is likely to be a misuse of information for other applications.
- Such mission creep has been reported in data mining applications.

Disadvantages of data mining

Privacy Issues

- 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.

Security issues

- Information like social security number, birthday, payroll and etc. are crucial information and 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 etc with so much personal and financial information available, the credit card stolen and identity theft has also become a big problem.

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.

Expensive:

- Advanced data mining software and Highly skilled manpower is necessary for many firms but may be expensive. Because they need to yield more useful insights, data mining often costs more than it saves for most small enterprises.

Technical Knowledge:

- Depending on how they should be used, various mining tools are available. They each have a distinctive algorithm and design. Selecting the appropriate tool will only be possible with the required technical knowledge.

Accuracy:

- Even though data mining has created a framework for simple data collection with its techniques, its accuracy is still constrained. Making decisions can be complicated by erroneous information that has been acquired.

Large databases are needed for data mining:

- Huge datasets are necessary for data mining to be effective.

Data mining methods are not perfect:

- Accurate information is only sometimes produced through data mining. There are numerous methods for analyzing data, some of which are more precise than others.