Information Privacy and Data Mining

Data privacy

Data privacy, also called information privacy, is the aspect of information technology (IT) that deals with the ability an organization or individual has to determine what data in a computer system can be shared with third parties. Information privacy is considered an important aspect of information sharing. With the advancement of the digital age, personal information vulnerabilities have increased.

Information privacy may be applied in numerous ways, including encryption, authentication and data masking - each attempting to ensure that information is available only to those with authorized access. These protective measures are geared toward preventing data mining and the unauthorized use of personal information, which are illegal in many parts of the world.

Information Protection Principles (IPPs)

1. ***Collection***
   - **Lawful**: An agency must only collect personal information for a lawful purpose. It must be directly related to the agency's function or activities and necessary for that purpose.
   - **Direct**: An agency must only collect personal information directly from you, unless you have authorized collection from someone else, or if you are under the age of 16 and the information has been provided by a parent or guardian.
   - **Open:** An agency must inform you that the information is being collected, why it is being collected, and who will be storing and using it.
   - **Relevant**: An agency must ensure that your personal information is relevant, accurate, complete, up-to-date and not excessive. The collection should not unreasonably intrude into your personal affairs.
2. ***Storage***
   - **Secure:** An agency must store personal information securely, keep it no longer than necessary and dispose of it appropriately. It should also be protected from unauthorized access, use, modification or disclosure.
3. ***Access and accuracy***

- **Transparent**: An agency must provide you with details regarding the personal information they are storing, why they are storing it and what rights you have to access it.
- **Accessible:** An agency must allow you to access your personal information without excessive delay or expense.

4. Use

- **Accurate**: An agency must ensure that your personal information is relevant, accurate, up to date and complete before using it.
- **Limited**: An agency can only use your personal information for the purpose for which it was collected.

5. Disclosure

- **Restricted:** An agency can only disclose your information in limited circumstances if you have consented or if you were told at the time they collected it that they would do so. An agency cannot disclose your sensitive personal information without your consent, for example, information about political opinions, religious or philosophical beliefs, medical conditions or trade union membership.

Applications of Data Mining

The main purpose of data mining process is to discover the records of information and summarize it in a simpler format for the purpose of others.

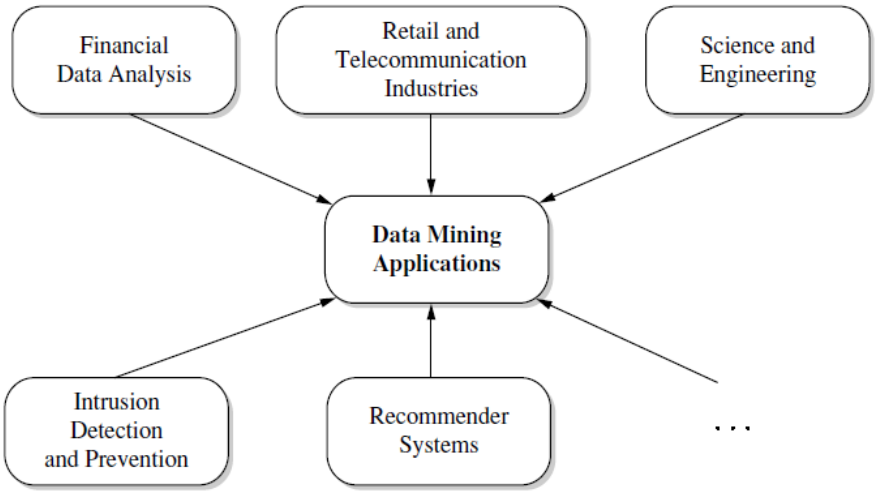

Figure 1 Common data mining application domains.

i. Data Mining for Financial Data Analysis

Most banks and financial institutions offer a wide variety of banking, investment, and credit services (the latter include business, mortgage, and automobile loans and credit cards). Some also offer insurance and stock investment services. Financial data collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. Following are a few typical cases.

- **Loan payment prediction and customer credit policy analysis**

Loan payment prediction and customer credit analysis are critical to the business of a bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors and eliminate irrelevant ones.

- **Classification and clustering of customers for targeted marketing**

Classification and clustering methods can be used for customer group identification and targeted marketing. For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similar behaviors regarding loan payments may be identified by multidimensional clustering techniques. These can help identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.

- **Detection of money laundering and other financial crimes**

To detect money laundering and other financial crimes, it is important to integrate information from multiple, heterogeneous databases (e.g., bank transaction databases and federal or state crime history databases), as long as they are potentially related to the study. Multiple data analysis tools can then be used to detect unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers.

ii. Data Mining for Retail and Telecommunication Industries

The retail industry is a well-fit application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly, especially due to the increasing availability, ease, and popularity of business conducted on the Web, or e-commerce.

Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

A few examples of data mining in the retail industry are outlined as follows:

- **Multidimensional analysis of sales, customers, products, time, and region**

The retail industry requires timely information regarding customer needs, product sales, trends, and fashions, as well as the quality, cost, profit, and service of commodities. It is therefore important to provide powerful multidimensional analysis and visualization tools, including the construction of sophisticated data cubes according to the needs of data analysis.

- **Analysis of the effectiveness of sales campaigns**

The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers. Careful analysis of the effectiveness of sales campaigns can help improve company profits. Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions containing the sales items during the sales period versus those containing the same items before or after the sales campaign. Moreover, association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

- **Customer retention—analysis of customer loyalty**

We can use customer loyalty card information to register sequences of purchases of particular customers. Customer loyalty and purchase trends can be analyzed systematically. Goods purchased at different periods by the same customers can be grouped into sequences. Sequential pattern mining can then be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing and variety of goods to help retain customers and attract new ones.

- **Product recommendation and cross-referencing of items**

By mining associations from sales records, we may discover that a customer who buys a digital camera is likely to buy another set of items. Such information can be used to form product recommendations. Product recommendations can also be advertised on sales receipts, in weekly flyers, or on the Web to help improve customer service, aid customers in selecting items, and

increase sales. Similarly, information, such as "hot items this week" or attractive deals, can be displayed together with the associative information to promote sales.

- **Fraudulent analysis and the identification of unusual patterns**

Fraudulent activity costs the retail industry millions of dollars per year. It is important to (1) identify potentially fraudulent users and their atypical usage patterns; (2) detect attempts to gain fraudulent entry or unauthorized access to individual and organizational accounts; and (3) discover unusual patterns that may need special attention. Many of these patterns can be discovered by multidimensional analysis, cluster analysis, and outlier analysis.

iii. Data Mining in Science and Engineering

In the past, many scientific data analysis tasks tended to handle relatively small and homogeneous data sets. Such data were typically analyzed using a "formulate hypothesis, build model, and evaluate results" paradigm. In these cases, statistical techniques were typically employed for their analysis. Massive data collection and storage technologies have recently changed the landscape of scientific data analysis.

Today, scientific data can be amassed at much higher speeds and lower costs. This has resulted in the accumulation of huge volumes of high-dimensional data, stream data, and heterogenous data, containing rich spatial and temporal information. Consequently, scientific applications are shifting from the "hypothesize-and-test" paradigm toward a "collect and store data, mine for new hypotheses, confirm with data or experimentation" process. This shift brings about new challenges for data mining.

Following are examples of data mining in science and engineering:

- **Mining complex data types**

Scientific data sets are heterogeneous in nature. They typically involve semi-structured and unstructured data, such as multimedia data and georeferenced stream data, as well as data with sophisticated, deeply hidden semantics (e.g., genomic and proteomic data). Robust and dedicated analysis methods are needed for handling spatiotemporal data, biological data, related concept hierarchies, and complex semantic relationships.

- **Graph-based and network-based mining**

It is often difficult or impossible to model several physical phenomena and processes due to limitations of existing modeling approaches. Alternatively, labeled graphs and networks may be used to capture many of the spatial, topological, geometric, biological, and other relational characteristics present in scientific data sets.

Data mining in engineering shares many similarities with data mining in science. Both practices often collect massive amounts of data, and require data preprocessing, data warehousing, and scalable mining of complex types of data. Both typically use visualization and make good use of graphs and networks. Moreover, many engineering processes need real-time responses, and so mining data streams in real time often becomes a critical component.

iv. Data Mining for Intrusion Detection and Prevention

The security of our computer systems and data is at continual risk. The extensive growth of the Internet and the increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection and prevention to become a critical component of networked systems. An intrusion can be defined as any set of actions that threaten the integrity, confidentiality, or availability of a network resource (e.g., user accounts, file systems, system kernels, and so on). Intrusion detection systems and intrusion prevention systems both monitor network traffic and/or system executions for malicious activities.

v. Data Mining and Recommender Systems

Today's consumers are faced with millions of goods and services when shopping online. Recommender systems help consumers by making product recommendations that are likely to be of interest to the user such as books, CDs, movies, restaurants, online news articles, and other services. An advantage of recommender systems is that they provide personalization for customers of e-commerce, promoting one-to-one marketing.

Primary Aims of Data Mining

In practice, 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.

Following are primary data mining tasks:

1. Classification: Discovery of a predictive learning function that classifies a data item into one of several predefined classes.

2. Regression: Discovery of a predictive learning function that maps a data item to a real value prediction variable.

3. Clustering: A common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data.

4. Optimization: enhance the use limited resources such as time space or material

Disadvantages of data mining

i. 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 etc. 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.

ii. 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 etc with so much personal and financial information available, the credit card stolen and identity theft has also become a big problem.

iii. 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.

iv. Other Disadvantages:

- Excessive work intensity may require investment in high performance teams and staff training.
- The difficulty of collecting the data. Depending on the type of data that you want to collect can be a lot of work.
- Although less and less, the requirement of a large investment can also be considered an inconvenience. Sometimes, the necessary technologies to carry out the data collection, is not an easy task and consumes many resources that could suppose a high cost.
- Data mining is not a perfect process, if the information is inaccurate, it would affect the outcome of the decision making process.

Pitfalls of Data Mining

1. The amount of available and relevant data may be much less than initially supposed
2. Data mining can occasionally take place with no clear goals and no idea of how results will be used. So early planning and definition of business goal is required
3. Insufficient business and data knowledge
4. The more factors in a data set the system considers, the more likely the program is to find the relation, valid or not.
5. Faulty assumptions such as no customer can hold different kinds of bank account.

References

[1] J. Han and K. Micheline, Data Mining: Concepts and Techniques, San Francisco: Elsevier Inc., 2006.