IJESMR
International Journal OF Engineering Sciences & Management Research

INFORMATION PRIVACY & SECURITY IN DATA MINING

Mr. Pranav O. Chitnis *, Prof. Satish R. Todmal
*Assistant Professor, Computer Engineering Department JSPM's ICOER, Wagholi, Pune.
Professor, Computer Engineering Department JSPM's ICOER, Wagholi, Pune

Keywords: Information Security, PPDM, Privacy in data mining.

ABSTRACT

The rapid use of data mining technologies brings threat to the information security. For data processing we are using the traditional data mining algorithms, but use of traditional algorithms violate the privacy of sensitive data. To overcome on such issue, needs to modify the data in such way that it will allow extracting the knowledge discovery from the data mining process, and the result will highlight the goal of the data mining process & unwanted disclosure of sensitive data will be prohibited. How to protect sensitive data from the threats, had given rise to a new research field, known as Privacy Preserving Data Mining (PPDM). PPDM focus on how to reduce the privacy risk in Data Processing, Data Transformation, Data Mining, and Pattern Evaluation & Pattern Presentation. PPDM approaches can investigate different users involved in data mining process namely, data provider, data collector, data miner, and decision maker. For each type of user, we focus on his privacy and how to protect sensitive information.

INTRODUCTION

'Data Mining' is the process of examining large pre-existing databases in order to generate new information and the result gives direction to guide future activities. Data mining process is also used for the analysis of data for relationships that have not previously been discovered. The term data warehouse is used to store a database that is used for analysis. Warehouse should be able to tell you what type of data they want to view and at what levels relationships among data items they want to be able to view it. In the past few years, enterprises across the globe have experienced significant changes in internet, storage and data analysis **Error! Reference source not found.**.

THE PROCESS OF KDD

Generally three of the major data mining techniques are regression, classification and clustering. Data Mining also popularly known as Knowledge Discovery in Databases (KDD)[1][5]. KDD widely used data mining technique is a process that includes data preparation, selection, and generate result patterns. Some issues involved in the entire KDD process are:

- Identify the goal of the KDD process.
- Understand application domains involved and the knowledge that's required
- Select data set on which discovery is be performed.
- Alter the data as per the requirements.
- Simplify the data sets by removing unwanted variables and missing fields
- Match KDD goals with data mining methods to suggest hidden patterns.
- Choose data mining algorithms to discover hidden patterns.
- Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
- Interpret essential knowledge from the mined patterns.
- Use the knowledge and incorporate it into another system for further action.

TO solve this issue we apply following step are performed in an iterative way

Data cleaning Data cleansing is also known as data cleaning or data scrubbing. it is a Step in which irrelevant data and noise data are removed from the raw collection of data. Although data cleansing can involve deleting old, incomplete or duplicated data.

Data integration is the combination of analytical and technical processes used to combine data from distinct sources into meaningful and valuable information.

Data selection at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system.

Pattern evaluation and presentation KDD process in which discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and Interpret the data mining results.

The explosive development in KDD process leads to privacy preservation which has been one of the greater concerns in data mining and given rise to a new research field, known as Privacy Preserving Data Mining (PPDM).

PPDM mainly focus on the hiding the data in which the sensitive data like person name person identity, phone number, resident address etc., In data hiding process, we alter or block such sensitive information out from the original database, in order to preserve personal sensitive information. On the other hand, the sensitive information is extracted in data mining process. To eliminate such type of sensitive information by using association mining rule algorithm [19]. To achieve the privacy of sensitive data, user should share their sensitive information in encrypted manner with the third party or distributed environment.

PPDM is a new emerging research field. Many approaches were been developed in early years.

In the traditional approach, all the sensitive information is hided. But if we see individual concern, the data which is important to one user, here hiding rule effects positively, the same data may not be seen as important to the other user; here hiding rule has negative impact. User information store in centralized and distributed data, based on the distribution of data. In a centralized database (DB) environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases [14].

The Traditional PPDM algorithm mainly focuses on classification, association rule and clustering. In general Classification algorithms can be first divided into two step, In the first step classification based on previous data and generate the training data. In the second step, we use training data as a sample data to classify new data. Association analysis involves the discovery of associated rules, showing attribute value and conditions that occur frequently in a given set of data. Clustering Analysis means a collection of database into groups so that the data point in one group are similar to each other and are as different as possible from the data points in other groups.

THE PRIVACY CONCERN AND PPDM

With more and more information easily available and easily accessible in electronic forms and those electronic forms present on the web and with the increasing powerful data mining tools are developed and these tools are used in data in data mining process causes a threat to user privacy and data security. In this way, we believe that privacy concerns with unauthorized access to individual data especially focus on sensitive information for example health records, financial records, legal issue records, etc. The goal of PPDM is to protect sensitive information from unwanted or unauthorized access. The PPDM process work on two principals, first, sensitive information should not be directly used for data mining process. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded. In the other words "The Privacy and PPDM deals with obtaining valid data mining results without disclosing the sensitive information data."

USER ROLE-BASED APPROACH

Recent models and algorithms in PPDM approach mainly focus on how to hide sensitive information from data mining process. The entire KDD process involves multi-phase operations. In the data mining process, privacy issues may begin in the data collecting or data preprocessing.

User-role based approach to conduct the review of related studies. Based on the multi-phase operations in KDD process. we can identify four different types of users, namely four user roles,

- **Data Provider:** the user who owns some data that are desired by the data mining task.
- **Data Collector:** the user who collects data from data providers and then publishes the data to the data miner.
- **Data Miner:** the user who performs data mining tasks on the data.
- **Decision Maker:** the user who makes decisions based on the data mining results in order to achieve goals.

In the data mining process, a user represents either a person or an organization. Also, one user can play multiple roles at once. For example, the U.S.retailer Target once received complaints from a customer who was angry that Target sent coupons for baby clothes to his teenager daughter. However, it was true that the daughter was pregnant at that time, and Target correctly inferred the fact by mining its customer data. In this story, the customer plays the role of data provider, and the retailer plays the roles of data collector, data miner and decision maker[17].

By differentiating the four different user roles, we can explore the privacy issues in data mining in a principled way. All users care about the security of sensitive information, but each user role views the security issue from its own perspective. What we need to do is to identify the privacy problems that each user role is concerned about, and to and appropriate solutions the problems. Here we briefly describe the privacy concerns of each user role. Detailed discussions will be presented in following sections.

DATA PROVIDER

The major concern of a data provider is whether he can control the sensitivity of the data he provides to others. On one hand, the provider should be able to make his very private data, namely the data containing information that he does not want anyone else to know, inaccessible to the data collector[18]. On the other hand, if the provider has to provide some data to the data collector, he wants to hide his sensitive information as much as possible and get enough compensation for the possible loss in privacy.

DATA COLLECTOR

The data collected from data providers may contain individual's sensitive information. Directly releasing the data to the data miner will violate data providers' privacy, hence data modification is required. On the other hand, the data should still be useful after modification; otherwise collecting the data will be meaningless[22].Therefore, the major concern of data collector is to guarantee that the modified data contain no sensitive information but still preserve high utility.

DATA MINER

The data miner applies mining algorithms to the data provided by data collector, and he wishes to extract useful information from data in a privacy-preserving manner. PPDM covers two types of protections, namely the protection of the sensitive data themselves and the protection of sensitive mining results. With the user role-based methodology proposed in this paper, we consider the data collector should take the major responsibility of protecting sensitive data, while data miner can focus on how to hide the sensitive mining results from untrusted parties.

DECISION MAKER

A decision maker can get the data mining results directly from the data miner, or from some Information Transmitter. It is likely that the information transmitter changes the mining results intentionally or unintentionally, which may cause serious loss to the decision maker. Therefore, what the decision maker concerns is whether the mining results are credible[12].

In addition to investigate the privacy-protection approaches adopted by each user role, in this paper we emphasize a common type of approach, namely game theoretical approach, that can be applied to many problems involving privacy protection in data mining. The rationality is that, in the data mining scenario, each user pursues high self-interests in terms of privacy preservation or data utility, and the interests of different users are correlated.

CONCLUSION

We want to release aggregate information about the data, without leaking individual information about participants. The purpose of data mining is to identify patterns in order to make predictions from information in databases. It allows the user to be proactive in identifying and predicting trends with that information. Common uses of data mining in knowledge discovery, fraud detection, analysis of research, decision support, and website personalization. The goal of data mining ensures that the proper beneficiaries of data provider receive the correct amount of benefits. The concerns of the privacy appear to focus on the some issues such as, whether there is a clear description of a program's collection of personal information, including how the collected information will serve the program's purpose, whether information collected for one purpose will then be used for additional, secondary purposes in the future. To overcome on these issue "The Privacy and PPDM deals with obtaining valid data mining results without disclosing the sensitive information data."

REFERENCES

[1] L. Brankovic and V. Estivill-Castro, ''Privacy issues in knowledge discovery and data mining,'' in Proc. Austral. Inst. Comput. Ethics Conf., 1999, pp. 89–99.

[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Mateo, CA, USA: Morgan Kaufmann, 2006.

[3] R. Agrawal and R. Srikant, ''Privacy-preserving data mining,'' ACM SIGMOD Rec., vol. 29, no. 2, pp. 439–450, 2000.

[4] Y. Lindell and B. Pinkas, ''Privacy preserving data mining,'' in Advances in Cryptology. Berlin, Germany: Springer-Verlag, 2000, pp. 36–54.

[5] C. C. Aggarwal and S. Y. Philip, A General Survey of PrivacyPreserving Data Mining Models and Algorithms. New York, NY, USA: Springer-Verlag, 2008.

[6] *M. B. Malik, M. A. Ghazi, and R. Ali, ''Privacy preserving data mining techniques: Current scenario and future prospects,'' in Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 26–32.*

[7] *S. Matwin, ''Privacy-preserving data mining techniques: Survey and challenges,'' in Discrimination and Privacy in the Information Society. Berlin, Germany: Springer-Verlag, 2013, pp. 209–221.*

[8] *E. Rasmusen, Games and Information: An Introduction to Game Theory, vol. 2. Cambridge, MA, USA: Blackwell, 1994.*

[9] *V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, ''Microdata protection,'' in Secure Data Management in Decentralized Systems. New York, NY, USA: Springer-Verlag, 2007, pp. 291–321.*

[10]*Tene and J. Polenetsky, ''To track or 'do not track': Advancing transparency and individual control in online behavioral advertising,'' Minnesota J. Law, Sci. Technol., no. 1, pp. 281–357, 2012.*

[11]*R. T. Fielding and D. Singer. (2014). Tracking Preference Expression (DNT). W3C Working Draft. [Online]. Available: http://www.w3.org/ TR/2014/WD-tracking-dnt-20140128/*

[12] *R. Gibbons, A Primer in Game Theory. Hertfordshire, U.K.: Harvester Wheatsheaf, 1992.*

[13] *D. C. Parkes, ''Iterative combinatorial auctions: Achieving economic and computational efficiency,'' Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA, USA, 2001.*

[14] *Verizon Communications Inc. (2013). 2013 Data Breach Investigations Report. [Online]. Available: http://www.verizonenterprise.com/ resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf*

[15]*A. Narayanan and V. Shmatikov, ''Robust de-anonymization of large sparse datasets,'' in Proc. IEEE Symp. Secur. Privacy (SP), May 2008, pp. 111–125.*

[16]*B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ''Privacy-preserving data publishing: A survey of recent developments,'' ACM Comput. Surv., vol. 42, no. 4, Jun. 2010, Art. ID 14.*

[17]*Lei Xu; Chunxiao Jiang; Jian Wang; Jian Yuan; Yong Ren, "Information Security in Big Data: Privacy and Data Mining," Access, IEEE , vol.2, no., pp.1149,1176, 2014*

[18] *R. C.-W. Wong and A. W.-C. Fu, ''Privacy-preserving data publishing: An overview,'' Synthesis Lectures Data Manage., vol. 2, no. 1, pp. 1–138, 2010.*

[19]*L. Sweeney, ''k-anonymity: A model for protecting privacy,'' Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.*

[20]*R. J. Bayardo and R. Agrawal, ''Data privacy through optimal k-anonymization,'' in Proc. 21st Int. Conf. Data Eng. (ICDE), Apr. 2005, pp. 217–228.*

[21]*K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, ''Mondrian multidimensional k-anonymity,'' in Proc. 22nd Int. Conf. Data Eng. (ICDE), Apr. 2006, p. 25.*

[22]*J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, ''Utility-based anonymization for privacy preservation with less information loss,'' ACM SIGKDD Explorations Newslett., vol. 8, no. 2, pp. 21–30, 2006.*

[23] *A. Gionis and T. Tassa, ''k-anonymization with minimal loss of information,'' IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, pp. 206–219, Feb. 2009. [24] B. Zhou, J. Pei, and W. Luk, ''A brief survey on anonymization techniques for privacy preserving publishing of social network data,'' ACM SIGKDD Explorations Newslett., vol. 10, no. 2, pp. 12–22, 2008.*