Information Security in Big Data: Privacy & Data Mining

[1]Kiran S.Gaikwad, [2]Assistant Professor. Seema Singh Solanki
[1][2]Everest College of engineering Aurangabad, Maharashtra India

Abstract: **The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy preserving data mining (PPDM). The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. Current studies of PPDM mainly focus on how to reduce the privacy risk brought by data mining operations, while in fact, unwanted disclosure of sensitive information may also happen in the process of data collecting, data publishing, and information (i.e., the data mining results) delivering. In this paper, we view the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely data provider, data collector, data miner, and decision maker. For each type of user, we discuss his privacy concerns and the methods that can be adopted to protect sensitive information**

I. INTRODUCTION

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. As a highly application-driven discipline, data mining has been successfully applied to many domains, such as business intelligence, Web search, scientific discovery, digital libraries, etc.

II. THE PROCESS OF KDD

The term ''data mining'' is often treated as a synonym for another term ''knowledge discovery from data'' (KDD) which highlights the goal of the mining process. To obtain useful knowledge from data, the following steps are performed in an iterative way (see Fig. 1):

Step1: Data preprocessing. Basic operations include data selection (to retrieve data relevant to the KDD task from the database), data cleaning (to remove noise and inconsistent data, to handle the missing data fields, etc.) and data integration (to combine data from multiple sources).

Step 2: Data transformation. The goal is to transform data into forms appropriate for the mining task, that is, to find useful features to represent the data. Feature selection and feature transformation are basic operations.

Step 3: Data mining. This is an essential process where intelligent methods are employed to extract data patterns (e.g. association rules, clusters, classification rules, etc).

Step 4: Pattern evaluation and presentation. Basic operations include identifying the truly interesting patterns which represent knowledge, and presenting the mined knowledge in an easy-to-understand fashion

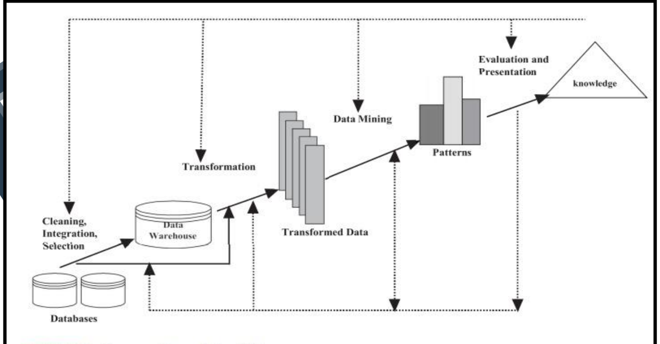


FIGURE 1. An overview of the KDD process.

III. THE PRIVACY CONCERN AND PPDM

Despite that the information discovered by data mining can be very valuable to many applications; people have shown increasing concern about the other side of the coin, namely the privacy threats posed by data mining. Individual's privacy may be violated due to the unauthorized access to personal data, the undesired discovery of one's embarrassing information, the use of personal data for purposes other than the one for which data has been collected, etc. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data. The consideration of PPDM is two-fold. First, sensitive raw data, such as individual's ID card number and cell phone number, should not be directly used for mining. Second, sensitive mining results whose disclosure will result in privacy violation should be excluded.

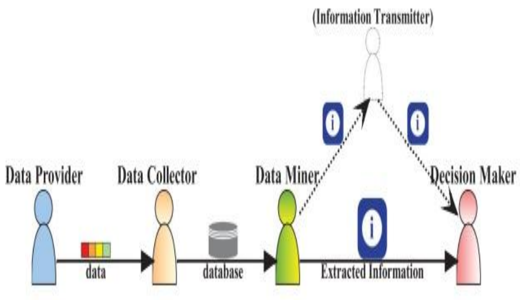


FIGURE 2. A simple illustration of the application scenario with data mining at the core.

IV. USER ROLE-BASED METHODOLOGY

Current models and algorithms proposed for PPDM mainly focus on how to hide that sensitive information from certain mining operations. However, as depicted in Fig. 1, the whole KDD process involves multi-phase operations. Besides the mining phase, privacy issues may also arise in the phase of data collecting or data preprocessing, even in the delivery process of the mining results. Here, we investigate the privacy aspects of data mining by considering the whole knowledge-discovery process. We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term ‘‘sensitive information’’ to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs. The term ‘‘sensitive data’’ refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms ‘‘privacy’’ and ‘‘sensitive information’’ are interchangeable.

In this paper, we develop a user-role based methodology to conduct the review of related studies. Based on the stage division in KDD process (see Fig. 1), we can identify four different types of users, namely four user roles, in a typical data mining scenario (see Fig. 2):

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 certain goals.

In the data mining scenario depicted in Fig. 2, a user represents either a person or an organization. Also, one user can play multiple roles at once. The customer plays the role of data provider, and the retailer plays the roles of data collector, data miner and decision maker. 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 find appropriate solutions the problems. Here we briefly describe the privacy concerns of each user role.

1) 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. 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.

2) Data Collector: The data collected from data providers may contain individuals’ 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. Therefore, the major concern of data collector is to guarantee that the modified data contain no sensitive information but still preserve high utility.

3) 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. As introduced in Section I-B, 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.

4) *Decision Maker:* As shown in Fig. 2, 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. 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

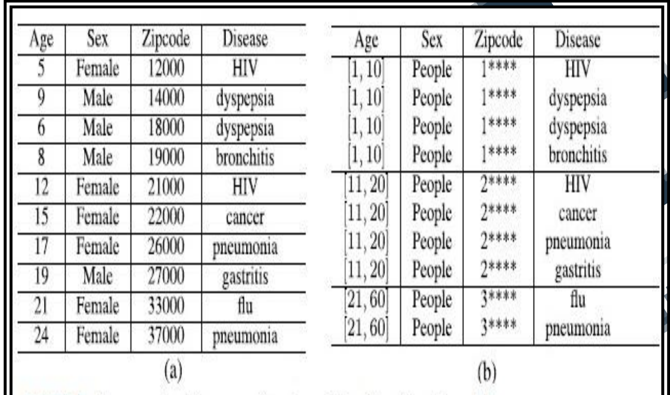
IFERP connecting engineers... developing research

the interests of different users are correlated. Hence the interactions among different users can be modeled as a game. By using methodologies from game theory, we can get useful implications on how each user role should behavior in an attempt to solve his privacy problems.

V. APPROACHES TO PRIVACY PROTECTION BASICS OF PPDP

PPDP mainly studies anonymization approaches for publishing useful data while preserving privacy. The original data is assumed to be a private table consisting of multiple records. Each record consists of the following 4 types of attributes:

- Identifier (ID): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile number.
- Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.
- Sensitive Attribute (SA): Attributes that an individual wants to conceal, such as disease and salary.
- Non-sensitive Attribute (NSA): Attributes other than ID, QID and SA. Before being published to others, the table is anonym zed, that is, identifiers are removed and quasi-identifiers are modified. As a result, individual's identity and sensitive attribute values can be hidden from adversaries.



Age	Sex	Zipcode	Disease
5	Female	12000	HIV
9	Male	14000	dyspepsia
6	Male	18000	dyspepsia
8	Male	19000	bronchitis
12	Female	21000	HIV
15	Female	22000	cancer
17	Female	26000	pneumonia
19	Male	27000	gastritis
21	Female	33000	flu
24	Female	37000	pneumonia

(a)

Age	Sex	Zipcode	Disease
[1, 10]	People	1****	HIV
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	dyspepsia
[1, 10]	People	1****	bronchitis
[11, 20]	People	2****	HIV
[11, 20]	People	2****	cancer
[11, 20]	People	2****	pneumonia
[11, 20]	People	2****	gastritis
[21, 60]	People	3****	flu
[21, 60]	People	3****	pneumonia

(b)

FIGURE 3. An example of 2-anonymity, where QID= {*Age*, *Sex*, *Zipcode*}. (a) Original table. (b) 2-anonymous table.

How the data table should be anonymized mainly depends on how much privacy we want to preserve in the anonymized data. Different privacy models have been proposed to quantify the preservation of privacy. Based on the attack model which describes the ability of the adversary in terms of identifying a target individual, privacy models can be roughly classified into two categories. The first category considers that the adversary is able to identify the record of a target individual by linking the record to data from other sources, such as liking the record to a record in a published data table (called *record linkage*), to a sensitive attribute in a published data table (called *attribute linkage*), or to the published data table itself (called *table linkage*). The second category considers that the adversary has enough background knowledge to carry out a *probabilistic attack*, that is, the adversary is able to make a confident inference about whether the target's record exist in the table or which value the target's sensitive attribute would take. Typical privacy models include *k*-anonymity (for preventing record linkage), *l*-diversity (for preventing record linkage and attribute linkage), *t*-closeness (for preventing attribute linkage and probabilistic attack), *epsilon*-differential privacy (for preventing table linkage and probabilistic attack), etc.

Among the many privacy models, *k*-anonymity and its variants are most widely used. The idea of *k*-anonymity is to modify the values of quasi-identifiers in original data table, so that every tuple in the anonymized table is indistinguishable from at least *k*□1 other tuples along the quasi-identifiers. The anonymized table is called a *k*-anonymous table. Fig. 3 shows an example of 2-anonymity. Intuitionally, if a table satisfies *k*-anonymity and the adversary only knows the quasi- identifiers values of the target individual, then the probability that the target's record being identified by the adversary will not exceed 1=*k*. To make the data table satisfy the requirement of a specified privacy model, one can apply the following anonymization operations:

Generalization. : This operation replaces some values with a parent value in the taxonomy of an attribute. Typical generalization schemes including full-domain generalization, subtree generalization, multidimensional generalization, etc.

Suppression. : This operation replaces some values with a special value (e.g. a asterisk `*'), indicating that the replaced values are not disclosed. Typical suppression schemes include record suppression, value suppression, cell suppression, etc.

Anatomization. : This operation does not modify the quasi-identifier or the sensitive attribute, but de-associates the relationship between the two. Anatomization-based method releases the data on QID and the data on SA in two separate tables.

Permutation. : This operation de-associates the relationship between a quasi-identifier and a numerical sensitive attribute by partitioning a set of data records into groups and shuffling their sensitive values within each group.

Perturbation. : This operation replaces the original data values with some synthetic data values, so that the statistical information computed from the perturbed data does not differ significantly from the statistical information computed from the original data. Typical perturbation methods include adding noise, swapping data, and generating synthetic data.

The anonymization operations will reduce the utility of data. The reduction of data utility is usually represented by *information loss*: higher information loss means lower utility of the anonymized data. Various metrics for

IFERP connecting engineers... developing research

measuring information loss have been proposed, such as minimal distortion, discernibility metric, the normalized average equivalence class size metric, weighted Certainty penalty, information-theoretic metrics, etc.

A fundamental problem of PPDP is how to make a tradeoff between privacy and utility. Given the metrics of privacy preservation and information loss, current PPDP algorithms usually take a greedy approach to achieve a proper tradeoff: multiple tables, all of which satisfy the requirement of the specified privacy model, are generated during the anonymization process, and the algorithm outputs the one that minimizes the information loss.

CONCLUSION

In this paper we review the privacy issues related to data mining by using a user-role based methodology. We differentiate four different user roles that are commonly involved in data mining applications, i.e. data provider, data collector, data miner and decision maker. Each user role has its own privacy concerns; hence the privacy-preserving approaches adopted by one user role are generally different from those adopted by others:

For data provider, his privacy-preserving objective is to effectively control the amount of sensitive data revealed to others. To achieve this goal, he can utilize security tools to limit other's access to his data, sell his data at auction to get enough compensation for privacy loss, or falsify his data to hide his true identity.

For data collector, his privacy-preserving objective is to release useful data to data miners without disclosing data providers' identities and sensitive information about them. To achieve this goal, he needs to develop proper privacy models to quantify the possible loss of privacy under different attacks, and apply anonymization techniques to the data.

For data miner, his privacy-preserving objective is to get correct data mining results while keep sensitive information undisclosed either in the process of data mining or in the mining results. To achieve this goal, he can choose a proper method to modify the data before certain mining algorithms are applied to, or utilize secure computation protocols to ensure the safety of private data and sensitive information contained in the learned model.

Decision maker, his privacy-preserving objective is to make a correct judgment about the credibility of the data mining results he's got. To achieve this goal, he can utilize provenance techniques to trace back the history of the received information, or build classifier to discriminate true information from false information.

To achieve the privacy-preserving goals of different users’ roles, various methods from different research fields are required.

REFERENCES

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

[2] 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.

[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 Privacy-Preserving 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] O. 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 ef_ciency," Ph.D. dissertation, Univ. Pennsylvania,Philadelphia, PA, USA, 2001.

[14] S. Carter, ``Techniques to pollute electronic pro_ling," U.S. Patent 11/257 614, Apr. 26, 2007. [Online]. Available: https://www.google.com/patents/US20070094738

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

[16] A. Narayanan and V. Shmatikov, ``Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 111_125.

[17] 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.

[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.*

IFERP
connecting engineers... developing research