Privacy Preserving and Data Mining In Big Data

Mohammad Tarique Mohammad Saleem[1], Prof.A.P.Kankale[2],

[1]PG Scholar R.ajarshi Shahu College Of Engineering,Buldhana

[2]H.O.D. Dept.of computer science &engineering.R.S.C.E. Buldhana Maharashtra

---***---

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), has been extensively studied in recent years. 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.*

Key Words: Data mining, sensitive information, privacy-preserving data mining, anonymization, provenance, game theory, privacy auction, anti-tracking.

1.INTRODUCTION:

Data mining has attracted more and more attention in recent years, probably because of the popularity of the ``big data'' concept. Data mining is the process of discovering interest-ing patterns and knowledge from large amounts of data [1]. As a highly application-driven discipline, data mining has been successfully applied to many domains, such as busi-ness intelligence, Web search, scientific discovery, digital libraries, etc.

1.1.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):

Step 1: 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 selec-tion 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).

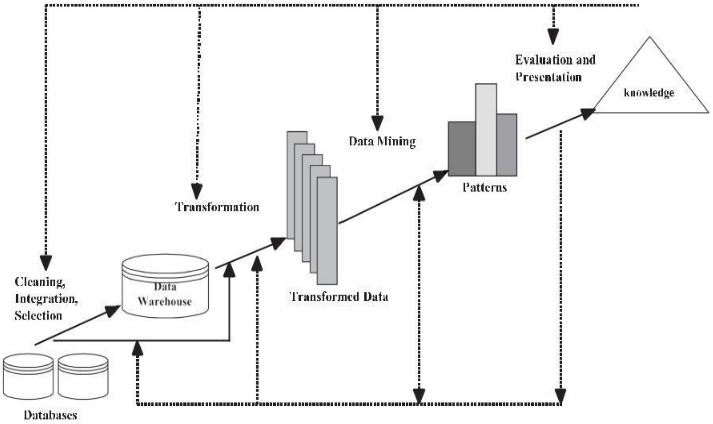

FIGURE 1. An overview of the KDD process.

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.

2.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 informa-tion, the use of personal data for purposes other than the one for which data has been collected, etc. For instance, 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. From this story, we can see that the conflict between data mining and privacy security does exist.

To deal with the privacy issues in data mining, a sub-field of data mining, referred to as *privacy preserving data mining* (PPDM) has gained a great development in recent years. The objective of PPDM is to safeguard sensitive information from unsolicited or unsanctioned disclosure, and meanwhile, preserve the utility of the data.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.

3. USER ROLE-BASED METHODOLOGY: Current models and algorithms proposed for PPDM mainly focus on how to hide those sensitive information from certain mining operations. However, as depicted in Fig. 1, the whole KDD process involve 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. In this paper, 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 infor-mation'' 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 informa-tion 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 publish 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. For example, in the Target story we mentioned above, the customer plays the role of data provider and the retailer plays the roles of data collector, data miner and decision maker.

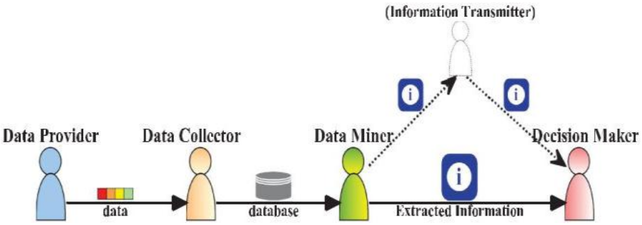

FIGURE 2. A simple illustration of the application scenario with data mining at the core.

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. Detailed discussions will be presented in following sections.

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 compensations for the possible loss in privacy.

2) DATA COLLECTOR

The data collected from data providers may contain individu-als' 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 *Informa-tion Transmitter*. It is likely that the information transmitter changes the mining results intentionally or unintentionally, which may cause serious loss to the decision maker. There-fore, what the decision maker concerns is whether the mining results are credible.

B. APPROACHES TO PRIVACY PROTECTION

1.BASICS OF PPDP: PPDP mainly studies anonymization approaches for publish-ing useful data while preserving privacy. 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.

PRIVACY-PRESERVING PUBLISHING OF SOCIAL NETWORK DATA

Social networks have gained great development in recent years. Aiming at discovering interesting social patterns, social network analysis becomes more and more important. To support the analysis, the company who runs a social net-work application sometimes needs to publish its data to a third party. However, even if the truthful identifiers of individuals are removed from the published data, which is referred to as naïve anonymized, publication of the network data may lead to exposures of sensitive information about individuals, such as one's intimate relationships with others. Therefore, the network data need to be properly anonymized before they are published.

2.PRIVACY-PRESERVING PUBLISHING OF TRAJECTORY DATA

Driven by the increased availability of mobile communication devices with embedded positioning capabilities, location-based services (LBS) have become very popular in recent years. By utilizing the location information of individuals, LBS can bring convenience to our daily life. For example, one can search for recommendations about restaurant that are close to his current position, or monitor congestion levels of vehicle traffic in certain places. However, the use of private location information may raise serious privacy problems. Among the many privacy issues in LBS , here we focus on the privacy threat brought by publishing trajec-tory data of individuals. To provide location-based services, commercial entities (e.g. a telecommunication company) and public entities (e.g. a transportation company) collect large amount of individuals' trajectory data, i.e. sequences of consecutive location readings along with time stamps. If the data collector publish such spatio-temporal data to a third party (e.g. a data-mining company), sensitive information about individuals may be disclosed. For example, an adver-tiser may make inappropriate use of an individual's food preference which is inferred from his frequent visits to some restaurant. To realize a privacy-preserving publication, anonymization techniques can be applied to the trajectory data set, so that no sensitive location can be linked to a spe-cific individual. Compared to relational data, spatio-temporal data have some unique characteristics, such as time depen-dence, location dependence and high dimensionality. There-fore, traditional anonymization approaches cannot be directly applied.

SUMMARY

Privacy-preserving data publishing provides methods to hide identity or sensitive attributes of original data owner. Despite the many advances in the study of data anonymization, there remain some research topics awaiting to be explored. Here we highlight two topics that are important for devel-oping a practically effective anonymization method, namely personalized privacy preservation and modeling the background knowledge of adversaries.

Current studies on PPDP mainly manage to achieve privacy preserving in a statistical sense, that is, they focus on a univer-sal approach that exerts the same amount of preservation for all individuals. While in practice, the implication of privacy varies from person to person. For example, someone consid-ers salary to be sensitive information while someone doesn't; someone cares much about privacy while someone cares less.

DECISION MAKER:

A. CONCERNS OF DECISION MAKER

The ultimate goal of data mining is to provide useful information to the decision maker, so that the decision maker can choose a better way to achieve his objective, such as increas-ing sales of products or making correct diagnoses of diseases. At a first glance, it seems that the decision maker has no responsibility for protecting privacy, since we usually inter-pret privacy as sensitive information about the original data owners (i.e. data providers). Generally, the data miner, the data collector and the data provider himself are considered to be responsible for the safety of privacy. However, if we look at the privacy issue from a wider perspective, we can see that the decision maker also has his own privacy concerns. The data mining results provided by the data miner are of high impor-tance to the decision maker. If the results are disclosed to someone else, e.g. a competing company, the decision maker may suffer a loss. That is to say, from the perspective of deci-sion maker, the data mining results are sensitive information. On the other hand, if the decision maker does not get the data mining results directly from the data miner, but from someone else which we called *information transmitter*, the decision maker should be skeptical about the credibility of the results, in case that the results have been distorted. Therefore, the privacy concerns of the decision maker are twofold: how to prevent unwanted disclosure of sensitive mining results, and how to evaluate the credibility of the received mining results.

APPROACHES TO PRIVACY PROTECTION

To deal with the first privacy issue proposed above, i.e. to prevent unwanted disclosure of sensitive mining results,

usually the decision maker has to resort to legal measures. For example, making a contract with the data miner to forbid the miner from disclosing the mining results to a third party. To handle the second issue, i.e. to determine whether the received information can be trusted, the decision maker can utilize methodologies from data provenance, credibility anal-ysis of web information, or other related research fields. In the rest part of this section, we will first briefly review the studies on data provenance and web information credibility, and then present a preliminary discussion about how these studies can help to analyze the credibility of data mining results.

1) DATA PROVENANCE

If the decision maker does not get the data mining results directly from the data miner, he would want to know how the results are delivered to him and what kind of modification may have been applied to the results, so that he can determine whether the results can be trusted. This is why ``provenance'' is needed. The term *provenance* originally refers to the chronology of the ownership, custody or location of a historical object. In information science, a piece of data is treated as the historical object, and *data provenance* refers to the information that helps determine the derivationThe following aspects are used to capture the characteristics of a provenance system:

Application of provenance. Provenance systems may be constructed to support a number of uses, such as estimate data quality and data reliability, trace the audit trail of data, repeat the derivation of data, etc.

Subject of provenance. Provenance information can be collected about different resources present in the data processing system and at various levels of detail.

Representation of provenance. There are mainly two types of methods to represent provenance information, one is annotation and the other is inversion. The anno-tation method uses metadata, which comprise of the derivation history of the data, as annotations and descrip-tions about sources data and processes. The inversion method uses the property by which some derivations can be inverted to nd the input data supplied to derive the output data.

Provenance storage. Provenance can be tightly coupled to the data it describes and located in the same data storage system or even be embedded within the data le. Alternatively, provenance can be stored separately with other metadata or simply by itself

CONCLUSION

How to protect sensitive information from the security threats brought by data mining has become a hot topic in recent years. In this paper we review the privacy issues related to data mining by using a user-role based methodology. We differ-entiate 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 compensations 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 tech-niques to the data.For data miner, his privacy-preserving objective is to get correct data mining results while keep sensitive informa-tion 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 sensi-tive information contained in the learned model.For decision maker, his privacy-preserving objective is to make a correct judgement 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 classi er to discriminate true information from false informatio

REFERENCE

[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 discov-ery 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 chal-lenges,'' 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.