Confidentiality in Big Data using PPDM

Bincy K Sunny
PG Student
Department of Computer Science and Engineering
Mount Zion College Of Engineering, Kadammanitta,
Pathanamthitta, Kerala

Smita C Thomas
Research Scholar
Department of Computer Science and Engineering
Vels University,
Chennai

Abstract

A new research privacy preserving data mining develop due to the recent advanced in data mining, security technologies. The basic concept of PPDM is to correct the data in such a way so as to perform data mining algorithms definitely without mean the security of sensitive information contained in the data. PPDM mainly target on how to decrease the privacy danger brought by data mining operations, while in fact, undesirable acknowledgement of sensitive information may also appear in the process of data gathering, data announcing and information distributing. This paper deals with the privacy issues related to data mining from a view point and examine various approaches that can help to protect sensitive information. In particular, identify four different types of users involved in data mining applications. For each type of user, consider the privacy concerns and the methods that can be accepted to preserve sensitive information.

Keywords: Data mining, privacy preserving data mining, anonymization, provenance

I. INTRODUCTION

Data mining has bring progressively consideration in later years, apparently because of the demand of the "big data" concept. Data mining is the process of identifying impressive arrangement and learning from large number of data. As a deeply application-driven discipline, data mining has been strongly related to many domains, such as business intelligence, Web inquiry, scientific analysis, digital study, etc.

A. The Process of KDD

The word "data mining" is generally employed as a equivalent for another word "knowledge discovery from data" (KDD) which focus the objective of the mining process. To obtain proper knowledge from data, the following steps are achieved in an constant way (Fig.1):

1) Step 1: Data pre-processing. Basic activities include data selection, data cleaning, and data integration.
2) Step 2: Data transformation. The aim is to modify data into forms proper for the mining task, that is, to find useful property to represent the data. Features selection and feature transformation are basic activities.
3) Step 3: Data mining. This is a necessary process where reasonable ways are selected to abstract data arrangement.
4) Step 4: Pattern evaluation and presentation. Basic activities include describing the absolutely attractive patterns which correspond to learning, and submitting the mined knowledge in a clear-to-understand form.

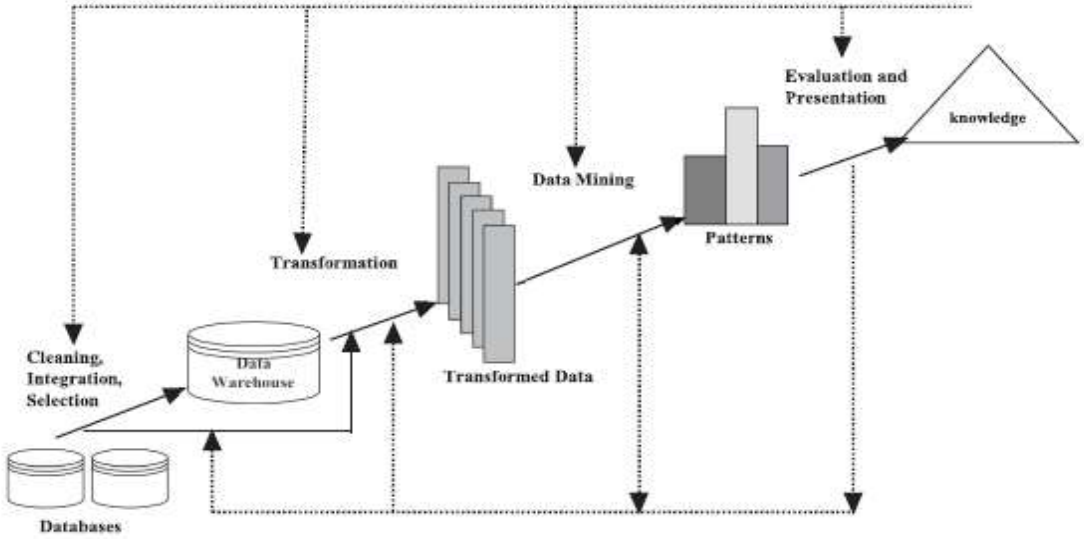

Fig. 1: An overview of the KDD process.

B. The Privacy Issues and PPDM

Personal privacy may be disrupt due to the unapproved access to individual data, the unwanted acknowledgement of one's sensitive information, the use of private data for desire other than the one for which data has been collected, etc. To deal with the confidentiality problems in data mining, known as privacy preserving data mining (PPDM) has achieved a considerable progress in later years. The intention of PPDM is to conserve knowing information from unwanted disclosure, and until, preserve the advantage of the data. The attention of PPDM is two turn. First, sensitive raw data, such as personal ID card number and cell phone number, not permitted to use directly for mining. Second, sensitive mining results whose exposure will result in privacy disrupt should be rejected.

C. User Role Based Methodology

Recent models and algorithms allowed for PPDM mainly deals on how to hide those sensitive information from certain mining operations. However, as shown in Fig.1, the whole KDD process include multi-phase activities. Along with the mining phase, privacy problems may also occur in the phase of data collecting or data preprocessing, alike in the delivery process of the mining results. In this paper, deals the privacy manner of data mining as the whole knowledge-discovery process. This paper present an analysis of many advances which can aid to make suitable use of sensitive data and protect the security of sensitive information discovered by data mining. In this paper, form a user-role based methodology to conduct the review of related studies. Based on the stage division in KDD process (Fig 1), there are four different types of users, in a typical data mining scenario(Fig 2) :

- Data Provider: provides some data that are desired by the data mining task.
- Data Collector: collects data from data providers and then delivered the data to the data miner.
- Data Miner: the user who performs data mining tasks on the data.
- Decision Maker: makes decisions upon the data mining results.

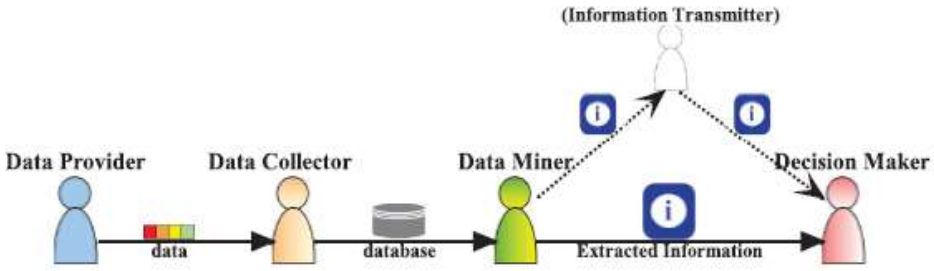

Fig. 2: A Simple Illustration of the Application Scenario with Data Mining at the Core

By understanding the four various user roles, we can search the confidentiality problems in data mining in a noble way. All users aware about the security of sensitive information, but each user role outlook the security issue from its own aspect. This paper explains the privacy involvement of respective user role.

II. DATA PROVIDER

A data provider owns some data from which relevant information can be obtained. In the data mining scenario shown in Fig2, there are two types of data providers: first, refers to the data provider that contribute data to data collector, and the other refers to the data collector contribute data to data miner

A. Approaches to Privacy Protection:

1) Limit the Access

A data provider contributes his information to the collector in an active way or passive way. When the data provider gives data actively, easily avoid the collector's need for the information that he allows very conscious. If his data are passively granted to the data collector, the data provider can take some extent to confines the collector's approach to his sensitive data.

Numerous security tools are constructed as browser extensions for comfort of use. Upon the functions, current security tools can be divided into the three types:

1) Anti-tracking extensions. Knowing that valuable information can be derived from the data composed by users' online movement. When check over the internet, a user can employ an anti-tracking extensions to block the trackers from collecting cookies. Popular anti-tracking extensions introduce Disconnect, Do Not Track me, Ghostery, etc.
2) Advertisement and script blockers. This type of providers can block publicity on the sites, and destroy scripts and device that send the user's data to some unfamiliar third party. Example tools include AdBlock Plus, NoScript, etc.

3) Encryption tools. To make sure a personal online contact between two parties cannot be catched by third parties, a user can use encryption tools such as MailCloak and TorChat, to encrypt his emails, instant messages, or other types of web traffic.

III. DATA COLLECTOR

As shown in Fig.2, a data collector assembles data from data providers in order to hold the successive data mining process. The original data published by the data providers usually contain sensitive information about person. Before releasing actual data to others the data collector should modify, so that sensitive information about data providers can neither be create in the altered data nor be implied by anyone with malicious intent. Generally, the alteration will cause a loss in data utility. The data alteration process accepted by data collector, with the goal of preserving privacy and advantage simultaneously, is usually called privacy preserving data publishing(PPDP).

B. Approaches to Privacy Protection

1) Basics of PPDM

PPDP mainly reviews anonymization approaches for delivering proper data while preserving privacy. The original data is pretended to be a private table contains multiple records. Each record contains 4 types of attributes:

- Identifier(ID) : Directly and uniquely identify an individual, such as name, ID number and mobile number.
- Quasi-identifier: Linking with external data to re- identify individual records, such as gender, age and zip code.
- Sensitive Attribute (SA): Person wants to hide, such as disease and salary.
- Non-sensitive Attribute(NSA): Attributes other than ID, QID and SA.

Previous being delivered to others, the table is anonymized, that is, identifiers are detached and quasi-identifiers are altered. As a result, personal identity and sensitive attribute values can be hidden from attackers.

With the various privacy models, k-anonymity and its alternatives are greatly used. The concept of k-anonymity is to alter the values of quasi-identifiers in actual data table, so that every tuple in the anonymized table is identical from at least k-1 other tuples forward the quasi-identifiers. The anonymized table is called a k- anonymous table. Intentionally, if a table suit k-anonymity and the rival only knows the quasi-identifier values of the target individual, then the possibility that the target's record being seen by the rival will not rise above 1/k.

To make the data table satisfy the requirement of a specified privacy model, one can apply the following anonymization operations:

1) Generalization
2) Suppression
3) Anatomization
4) Permutation
5) Perturbation

IV. DATA MINER

The primary concern of data miner is how to prevent sensitive information from appearing in the mining results. To perform a privacy preserving data mining, the data miner usually needs to modify the data be got from the data collector.

A. Approaches to Privacy Protection

Define the privacy-preserving goal of data miner as preventing sensitive information from being revealed by the data mining results, in this section, classify PPDM approaches according to the type of ddata mining tasks like privacy-preserving clustering.

1) Privacy-Preserving Clustering

Cluster analysis is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity. De and Tripathy recently develop a secure algorithm for clustering over vertically partitioned data. There are two parties involved in the computation. . In the proposed algorithm, each party first computes k clusters on their own private data set. Then, both parties compute the distance between each data point and each of the k cluster centers. The resulting distance matrices along with the randomized cluster centers are exchanged between the two parties. Based on the information provided by the other party, each party can compute the final clustering result.

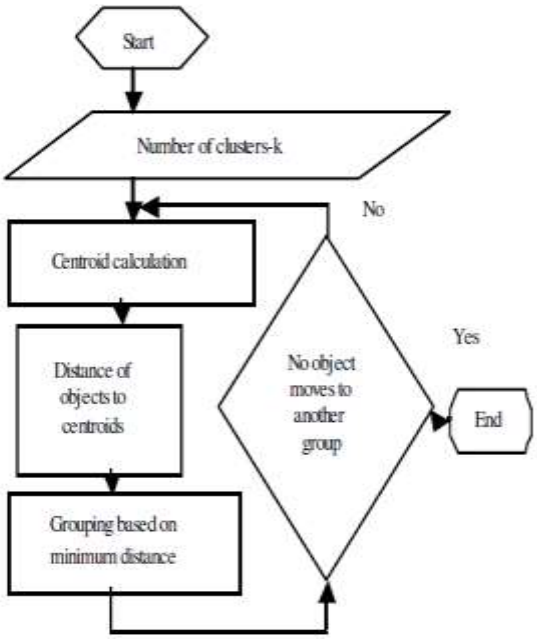

Fig. 4: Clustering Process

V. DECISION MAKER

The data mining results provided by the data miner are of high importance to the decision maker. 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 received mining results

A. Approaches to Privacy Protection

To deal with the privacy issue, i.e. to prevent unwanted disclosure of sensitive mining results, usually the decision maker has to resort to legal measures. 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 analysis of web information, or other related research fields.

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. It helps to determine the derivation history of the data, starting from the original source.

2) Web Information Credibility

Due to the lack of delivering data privacy, the minimum cost of dissemination, and the lax control of quality, judging of information in web has become a serious problem. The five criteria can be used to understand the false information from the truth they are:

- Authority
- Accuracy
- Objectivity
- Currency
- Coverage

VI. CONCLUSION

In this paper reviews the privacy problems on data mining by using a user-role based methodology. Here understand four various user roles that included in data mining applications they are data provider, data collector, data miner and decision maker. All user role has its own view in privacy, hence the privacy preserving way accepted by one user role are commonly differ from those accepted by others.

REFERENCES

[1] LEI XU, CHUNXIAO JIANG, (Member, IEEE), "Information Security in Big Data : Privacy and data mining". JIAN WANG, (Member, IEEE), JIAN YUAN,(Member, IEEE), and YONG REN, (Member, IEEE)
[2] J.Han. M.Kamber, and J.Pei, "Data Mining : Concepts and Techniques".San Mateo, CA,USA :Morgan Kaufmann,2006.
[3] Isha K.Gayki, Arvind S.Kapse1ME (CSE) Scholar, Assistant Professor Department of CSE, P R Patil College of Engg. & Tech. "Privacy Preservation of Published Data Using Anonymization Techniques" .
[4] Pingshui WANG College of Management Science and Engineering, Anhui University of Finance & Economics, Bengbu, China "Survey on Privacy Preserving Data Mining" .
[5] Shashank,(PG Student II year MCA) S.K.Saravan, G.Rekha(Assistant Professor) Department of Computer Applications, Valliammai Engineering College, SRM Nagar, Kattankulathur. "Information Security in Big Data Using Encryption and Decryption Techniques".