Privacy Preserving Data Mining Model for Anonymizing Data Streams

R. Rajeswari
PG Scholar
Department of CSE
University College of Engineering
(BIT Campus)-Trichy.

Mrs. R. Kavitha
Assistant Professor
Department of CSE
University College of Engineering
(BIT Campus)-Trichy.

Abstract— The Access control mechanism avoids the unauthorized access of sensitive information. It protects the user information from the unauthorized access. The privacy protection mechanism is a much important concern in the case of sharing the sensitive information. The privacy protection mechanism provides better privacy for the sensitive information which is to be shared. The generally used privacy protection mechanism uses the generalization and suppression of the sensitive data. It prevents the privacy disclosure of the sensitive data. The privacy protection mechanism avoids the identity and attributes disclosure. The privacy is achieved by the high accuracy and consistency of the user information, ie., the precision of the user information. In this paper, it proposes a privacy persevered access control mechanism for data streams. For the privacy protection mechanism it uses the combination of both the k-anonymity method and fragmentation method. The k-anonymity method uses the suppression and generalization.

Keywords— Access control, Privacy, k-anonymity

I. INTRODUCTION

Data mining is the process of extracting the useful information from the database. It is possible to efficiently extract or mine knowledge from large amounts of vertically partitioned data within quantifiable security restrictions. In other words the data mining is the process of discovering the interesting knowledge from large amounts of data stored either in databases, data warehouses or other information repositories. Knowledge Discovery in Databases (KDD) is the process of extracting knowledge from large quantities of data. The KDD process assumes that all the data is easily accessible through centralized access mechanisms such as federated databases and virtual warehouses. Moreover, advances in information technology and the ubiquity of networked computers have made personal information much more available. Privacy advocates have been challenging attempts to bring more and more information into integrated collections. Database security is the important requirements of the database. Database security is a very broad area that addresses many issues, like legal and ethical issues regarding the right to access certain information. Some information may be stored to be private and cannot be accessed legally by unauthorized persons. The sensitive data is accessible to authorized users only. The database security provides the security for the sensitive information from the unauthorized access.

The database security is based on the access control mechanism and the privacy protection mechanism. The Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users The sensitive information which is the user is not authorized to access will not accessed by the user. The authorized user can only access the authorized data. In a multiuser database system, the Database Management System (DBMS) must provide techniques to enable certain users or user groups to access selected portions of a database without gaining access to the rest of the database. Its importance comes in a large organization where numerous workers working. There must be some important data which are not published to all the workers. There uses the access controls mechanism for providing the access to the secured data to the particular authorized user only. For example, sensitive information such as employee salaries or performance reviews should be kept confidential from most of the database systems users. A DBMS typically includes a database security and authorization subsystem that is responsible for ensuring the security of portions of a database against unauthorized access. Privacy is the one of the most important concern of human life. It gives more importance to protect the privacy of the personal life. In the case of database, there will be huge amount of data to be kept privately. These data may contain sensitive information about the persons, confidential information about some organizations and so on. These data has to be protected by using some methods. It is the privacy protection mechanism (PPM). The general method is to transform the original data into some anonymous form to prevent from accessing its record owners sensitive information. There are numerous methods to provide the privacy for the sensitive data. The anonymization method is one of the important privacy protection mechanisms. The anonymization process will transforms the sensitive information to some anonymized form. K-anonymity, l-diversity, etc., are some of the anonymization method. For a given query from unauthorized user, it will provide the anonymized data through the privacy preserving techniques. In this paper it deals about the privacy protected access control mechanism. It will provide the security for the sensitive information. For an example, in the case of hospital management system there should be a number of patients. Some of the patients may have the disease which has to be isolated and so on. While publishing the patient's data to the state medical board for

disease surveillance system, they should anonymize the personal data of the patient. For this purpose it can use the proposed method for the secured accesscontrol and privacy protection mechanism.

II. RELATED WORK

In the related work, first the literature related to access control on data streams is reviewed and then research related to privacy preserving publishing of data streams is discussed. To the best our knowledge both the precision- bounded access control and privacy together for data streams has not been investigated before. Nehme et al. propose security punctuation-based access control framework for data streams [1], [9]. Security punctuation is a predicate that defines access to stream data and is created by the user generating stream data. The security punctuation tuples are then interleaved in the data stream. The subjects are assigned roles on the server and can execute authorized queries on the incoming data stream. The server allows the roles access to stream tuples according to the embedded security punctuation. Role-based access control for data streams have been proposed by Carminati et al. [2], [10]. In their framework, there are two types of temporal constraints. First is the interval constraint during which the role can access stream data. Second is the window constraint that limits access to the data stream for each role according to the authorized view defined by the sliding-window query predicate. They consider two types of privileges over the authorized data that is read privilege for selection and projection operations and aggregate privilege for Min, Max, Count, Avg, and Sum operations. In the current paper, we follow the access control specification of Nehme et al. and Carminati et al. but further consider the privacy preservation along with the precision-bounded access control. Cao et al. have proposed CASTLE for continuously anonymizing data streams [3].They extend the definition of k-anonymity for data streams and propose a clustering algorithm that publishes anonymized clusters before a given maximum delay deadline. The measure used to assess the quality of published clusters is the information loss metric [11] that does not consider the information loss due to delay in publishing. To overcome this shortcoming, Zhou et al. proposed a delay-based anonymization quality measure that increases the information loss as the publishing delay increases [4]. They propose a randomized algorithm based on the R-tree. The data stream tuples are added to the active R-tree and the leaf nodes of the tree due at each time instance are published. The due time for each leaf node is evaluated randomly based on the information loss. They further use the distribution density of the data stream to improve the algorithm. Both Cao et al. and Zhou et al. suppress the time-stamp attribute in the anonymized stream. However, the time-stamp attribute is required to evaluate any predicate sliding-window query over the anonymized stream. Dwork et al. have proposed differential privacy for data streams considering a single aggregate query [12]. Cao et al. further extend the model to sliding-window queries over data streams [13]. Differential privacy is achieved by adding random noise to original query results and offers better privacy guarantees than generalization, however syntactic anonymization techniques (e.g., generalization) provide better precision [14]. In the current paper our focus is on generalization and we further explore precision bounds for sliding-win privacy-preserving data streams. Access control and privacy techniques have been investigated for static relational data. K.LeFevre et al. [5], [7] and Iwuchukwu et al. [8] have proposed workload-aware anonymization for micro data publishing. Work has been done on micro data anonymization with accuracy and privacy constraints [6], [15], [16]. We have proposed the concept of imprecision bounds for accuracy-constrained access control on relational data [6]. However, the access control on data streams presents different challenges because of the temporal constraints defined by sliding window query predicates. In the data stream literature, access control and privacy-preserving publishing have been considered in isolation. However, we propose a unified precision-bounded access control framework for privacy- preserving data streams.

III. METHODOLOGY

There are lots of methods for providing the privacy for the sensitive information stored in the database and there are different access control methods for accessing the secured information stored in a database. In my project it deals with the introduction of both the access control mechanism and the privacy protection mechanism together for protecting the sensitive information. Here it uses the anonymity method for the privacy protection.

A. Suppression

In suppression-based anonymization, the database can be classified into two subsets are suppressed attributes and non-suppressed attributes. When the tuple T is k- anonymous, then for every tuple t is a subset of the tuple T. In the database the corresponding value is replaced by * (indicating suppressions of the original values). Suppression is used to reduce the content of the database or to minimize the size of the database.

B. Generalization

For generalization-based anonymization, each attribute value can be mapped to a more general value. The main step in most generalization based k-anonymity is to replace a specific value with a more general value. When the tuple T is k-anonymous, they can delete duplicate tuples. After the suppression and generalization, nobody can access the original database.

IV. ARCHITECTURE DIAGRAM

The architecture defines the entire privacy protection mechanism. The proposed system uses secure reversible Accuracy-Constrained Privacy-Preserving Access Control for relational database. The proposed

method provides data publication in a privacy preserved method. The framework of the proposed method is a combination of access control and privacy protection mechanisms. The access control mechanism allows only authorized queries predicates on sensitive data. The privacy preserving module anonymized the data to meet privacy requirements. Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users. However, there is chance of sensitive information can still be misused by authorized users for their use. The confidential data can also be misused. The concept of privacy-preservation for sensitive data requires the enforcement of privacy of the secured sensitive data and privacy policies or the protection against identity disclosure by satisfying some privacy requirements. In the proposed method, it investigate privacy-preservation from the anonymity aspect. The sensitive information, even after the removal of identifying attributes, is still in danger to linking attacks by the authorized users. Here it uses the data fragmentation and the anonymization method for the purpose of the privacy protection mechanism. Anonymization algorithms use suppression and generalization of records to satisfy privacy requirements. Anonymization is the process of making the data anonymized, i.e., the sensitive data is made privacy protected. In this proposed method it uses the k-anonymity method and the data fragmentation method for the privacy protection. The important term used here is the imprecision bound.

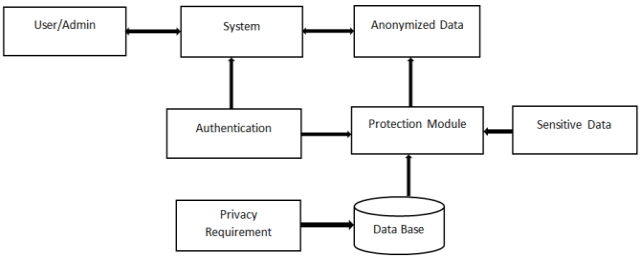

Fig. 1. Privacy Preserving Architecture.

V. CONCLUSION

In this paper a privacy preserving access control system for relational data has been proposed. The proposed system is a mix of access control and security assurance instruments. The entrance control component permits just the approved question predicates on delicate information. The access control administrator defines the permitted view of the data stream along with the required precision. The privacy protection mechanism applies generalization to the stream data such that the privacy requirement is met.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their significant and constructive critiques and suggestions, which improved the paper very much.

REFERENCES

[1] R. Nehme, E. Rundensteiner, and E. Bertino, “A security punctuation framework for enforcing access control on streaming data,” in IEEE 24th International Conference on Data Engineering, pp. 406–415, IEEE, 2008.

[2] B. Carminati, E. Ferrari, J. Cao, and K. Tan, “A framework to enforce access control over data streams,” ACM Transactions on Information and System Security (TISSEC), vol. 13, no. 3, p. 28, 2010. 14

[3] J. Cao, B. Carminati, E. Ferrari, and K. Tan, “Castle: Continuously anonymizing data streams,” IEEE Transactions on Dependable and Secure Computing, no. 99, pp. 1–1, 2008.

[4] B. Zhou, Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia, “Continuous privacy preserving publishing of data streams,” in Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 648–659, ACM, 2009.

[5] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Workloadaware anonymization techniques for large-scale datasets,” ACM Transactions on Database Systems (TODS), vol. 33, no. 3, pp. 1–47, 2008.

[6] Z. Pervaiz, W. Aref, A. Ghafoor, and N. Prabhu, “Accuracyconstrained Privacy-preserving Access Control Mechanism for Relational Data,” IEEE Transactions on Knowledge and Data Engineering, 2013.

[7] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in Proceedings of the 22nd International Conference on Data Engineering, pp. 25–25, IEEE, 2006.

[8] T. Iwuchukwu, Anonymization techniques for large and dynamic data sets. PhD thesis, The University of Wisconsin-Madison, 2008.

[9] R. Nehme, H. Lim, and E. Bertino, “Fence: Continuous access control enforcement in dynamic data stream environments,” in IEEE 26th International Conference on Data Engineering (ICDE), pp. 940–943, IEEE, 2010.

[10] J. Cao, B. Carminati, E. Ferrari, and K. Tan, “Acstream: Enforcing access control over data streams,” in IEEE 25th International Conference on Data Engineering, pp. 1495–1498, IEEE, 2009.

[11] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, “Fast data anonymization with low information loss,” in Proceedings of the 33rd international conference on Very large data bases, pp. 758–769, VLDB Endowment, 2007.

[12] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, “Differential privacy under continual observation,” in Proceedings of the 42nd ACM symposium on Theory of computing, pp. 715–724, ACM, 2010.

[13] J. Cao, Q. Xiao, G. Ghinita, N. Li, E. Bertino, and K.-L. Tan, “Efficient and accurate strategies for differentially-private sliding window queries,” in Proceedings of the 16th International Conference on Extending Database Technology, pp. 191–202, ACM, 2013.

[14] C. Clifton and T. Tassa, “On syntactic anonymity and differential privacy,” in ICDE Workshop on Privacy-Preserving Data Publication and Analysis (PRIVDB), 2013.

[15] S. Chaudhuri, R. Kaushik, and R. Ramamurthy, “Database access control & privacy: Is there a common ground?,” in Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR), pp. 96–103, 2011.

[16] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, “A framework for efficient data anonymization under privacy and accuracy constraints,” ACM Transactions on Database Systems (TODS), vol. 34, no. 2, p. 9, 2009.