Efficient Practices of Privacy Preservation in Relational Databases

Anusmitha.A[1], Renjana Ramachandran[2]

Abstract—**Data confidentiality, Availability and Integrity are essential part of database security. The access control mechanisms and privacy preservation mechanisms protect the data from unauthorized or third party user. When there is a lack in privacy preserving mechanism (PPM) and data is shared with others, the authorized user may need to compromise the privacy of data. The privacy preservation can be achieved through anonymization techniques like generalization or suppression. Along with privacy the precision of the authorized data is important. The aim of the work is to provide better security and minimum level of precision to the retrieved data, for that in this paper an accuracy constrained privacy preserving access control mechanism is implemented with additional constraint on each selection predicate called imprecision bounds. The accuracy constraints are satisfied for multiple roles also. Along with that a workload aware anonymization concept is used for selection predicates. The privacy preservation method introduced includes a multilevel anonymization in order to provide better results. The experimental results shows that the proposed system with multilevel anonymization works better in terms of precision, privacy for more permissions than current state of art.**

***Index Terms*-Access Control, Anonymization, Precision, Privacy preservation, Query evaluation**

I. Introduction

Every organization keeps a set of databases to store their information and there may be several situations to share those information with others. As we are living in the information age there are large sources of data around us. To improve the services the organizations collect and analyze the data. The Confidentiality, Integrity and Availability are termed as the [CIA-triad] designed to enable the information security within the organization. They are considered to be the essential components of the security. To ensure that only the authorized information are available only to the authorized users and access control mechanism is implemented in the databases. However there may happen the misuse of sensitive information by the authorized users to compromise the privacy of the customers. For the enhancement of the protection against the identity disclosure and enforcing the privacy policies, the concept of privacy preservation of sensitive data introduced by satisfying some privacy requirements [2].

The paper deals with the privacy preservation in the anonymity aspects. Sensitive information is the essential part of every database and even if we are implementing the privacy protection mechanisms [2] there may have the chance if linkage attacks [4] by the authorized users even after the removal of identifying attributes. This problem got discussed in the area of microdata publishing [3] and privacy definitions like k-anonymity [2], l-diversity [5] etc

Imprecision is a problem in information retrieval. A concept of imprecision bound is introduced in order to solve the problem of imprecision where a minimal level of tolerance or a threshold is defined for each permission [1]. The imprecision added to each permission/query and the aggregate imprecision for all queries got minimized in the workload aware anonymization techniques [5],[6]. The privacy of xml data are also discussed in [15], and also about the spatial database [16].

The topic of satisfying accuracy constraints for the individual permissions in the workload aware anonymization had not discussed yet. The accuracy constrained privacy preserving access control mechanism introduced is relevant in the workload aware anonymizations [1]. The topic of continuous data publishing anonymization [4] also introduced. A static relational table is used in this paper where the table is anonymized only once. A role based access control is also discussed where a concept like accuracy constraints for permissions applied to any privacy preserving security policy.

Normally the anonymization techniques are used to ensure the privacy and security for the data. In many of the works the anonymization technique like generalization is used[1]. The access control mechanisms are essential in order to preserve the confidentiality of the data by providing authentications whereas the privacy preservation is also important because it prevents the microdata or the sensitive information not to disclose with a third party user.

To improve the efficiency of the security methods, a privacy preservation module and a accuracy preserving module is combined [1]. The proposed system deals with a multilevel anonymization technique. Here instead of using single level anonymization like generalization or suppression, a combined form of anonymization technique introduced, which include generalization and suppression together.

The rest of this paper proceeds as follows. In section 2 preliminaries are discussed. The section 3 discusses the previous works. The problem definition and privacy preserving access control framework is discussed in section 4. Section 5 discusses the proposed system with multilevel

Manuscript received Feb, 2013.

Anusmitha.A, *Department of Computer Science and Engineering, Mangalam College of Engineering Ettumanoor. India.*
Renjana Ramachandran, *Department of Computer Science and Engineering, Mangalam College of Engineering, Ettumanoor, India.*

anonymization. Section 6 discusses the experimental results. The paper concludes in section 7.

II. Preliminaries

Given a relation T={A1, A2, A3……..An}, where Ai is an attribute, T* is the anonymized version of the relation T. We assume that T is a static relational table. The attributes can be of the following types:

- *Identifier*. Attributes, e.g., name and social security that can uniquely identify an individual. These attributes are completely removed from the anonymized relation.
- *Quasi-identifier (QI)*. Attributes, e.g., gender, zip code, birth date that can potentially identify an individual based on other information available to an adversary. QI attributes are generalized to satisfy the anonymity requirements.
- *Sensitive attribute*. Attributes, e.g., disease or salary, that if associated to a unique individual will cause a privacy breach.

A. Access control mechanism for relational databases

To define tuple-level permissions fine-grained access control like Oracle VPD [8] and SQL [9] are introduced in relational databases. Truman model [10] is introduced for evaluating user queries. A user query is modified by the access control mechanism and only authorized tuples are returned in this model. Column level access control allows queries to execute on the authorized column of the In relational data[8], [11] column level access control mechanisms allow queries to execute on authorized column, by replacing the unauthorized cell values by NULL values [12] cell level access control for relational data is achieved. For defining permissions on objects based on roles in an organization a Role-based Access Control (RBAC) was introduced. An RBAC policy configuration includes a set of Users(U), a set of Roles (R), and a set of Permissions (P). We assume that the selection predicates on the QI attributes define a permission for the relational RBAC model, [11] . UA is a user-to-role (U _ R) assignment relation and PA is a role to- permission (R _ P) assignment relation.

B. Privacy definitions

Here, privacy definitions related to anonymity are introduced.

Definition 1 (Equivalence Class (EC)). An equivalence class is a set of tuples having the same QI attribute values.
Definition 2 (k-anonymity Property). A table T* satisfies the k-anonymity property if each equivalence class has k or more tuples [2].

III. Previous works

Access control mechanisms for databases are an important concept that allows queries only on the authorized part of the database [8], [10]. Later a user authorization is limited to pre-defined predicates in a Predicate based fine-grained access mechanism [11]. Many techniques introduced for the enforcement of access control and privacy policies, they got discussed in[11]. The interaction between the access control mechanisms and the privacy protection mechanisms was missing in those studies.

Recently, Chaudhuri et al. have studied access control with privacy mechanisms [12]. Random noise was added to original query in differential privacy and the results which satisfy privacy constraints. But they do not considered the accuracy constraints for permissions. Li et al. [5] defined privacy in terms of K-anonymity where after sampling; k-anonymity offers similar privacy guarantees as those of differential privacy.

The accuracy-constrained privacy preserving access control framework [1] allows the access control administrator to specify imprecision constraints that the privacy protection mechanism is required to meet along with the privacy requirements. Both privacy-aware access control and problem of workload-aware anonymization are similar.

In our analysis of the related work, we focus on query-aware anonymization and a multilevel privacy ensurance which include the combination of the two anonymization techniques generalization and suppression.

We refer the recent survey paper [3] for k-anonymity techniques and algorithms. LeFevre et al. [5] in his work the workload aware anonymization techniques discussed for the first time, they proposed an algorithm named Selection Mondrian algorithm, it is a modification to the greedy multidimensional partitioning algorithm Mondrian [10]. In their algorithm, the greedy splitting heuristic minimizes the sum of imprecision for all queries on the basis of given query workload. a R+tree based anonymization algorithm was introduced by Iwuchukwu and Naughton in [7].

The anonymized data using biased R+tree based on the given query workload is more accurate for queries. Based on space filling curves for k-anonymity and l-diversity Ghinita et al. have proposed several algorithms [13]. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class [13].

Similarly, Xiao et al. [14] propose to add noise to queries according to the size of the queries in a given workload to satisfy differential privacy. In any of these works bounds for query imprecision have not been considered. The existing literature on workload-aware anonymization has a focus to minimize the overall imprecision for a given set of queries, but the anonymization with imprecision constraints for individual queries has not been discussed before. We follow the imprecision definition of LeFevre et al. [6] and introduce the constraint of imprecision bound for each query in a given query workload.

IV. Problem Definition and Privacy Preserving Access Control Framework

A. The k-PIB problem

The optimal k-anonymity problem is discussed in this section. This problem has been shown to be NP-complete for suppression [3] and generalization [4]. The hardness result for k-PIB follows the construction of LeFevre et al. [9] that shows the hardness of k-anonymous multi-dimensional with the smallest average equivalence class size. We show that finding k-anonymous partitioning that violates imprecision bounds for minimum number of queries is also NP-hard. A multiset of tuples is transformed into an equivalent set of distinct (tuple, count) pairs. The cardinality of Query *Qi* is the sum of count values of tuples falling inside the query hyper-rectangle. The constant *qv* defines an upper bound for the number of queries that can violate the bounds.

B. The Privacy preserving access control framework

The section describes about how the privacy preserving access control framework [1] works. It is illustrated using Fig.1 (The arrows represents the direction of information flow). Here an access control mechanism as well as a privacy preservation system is introduced together to improve security of the data. The privacy protection mechanism ensures that the privacy and accuracy of the data before the data available to the access control module. The permissions of the access control policy are based on the selection predicates.

The original data or tuple values in a relationship are replaced with the anonymized data normally the generalized data. Here a generalization concept is used for the anonymizations. Here the relaxed and strict access control are discussed, they enforce mechanisms over anonymized data. The access control by reference monitor can be of two type like

1. *Relaxed*. Use overlap semantics to allow success to all partitions that are overlapping the permissions.
2. *Strict*. Use enclosed semantics to allow access only to those partitions that are fully enclosed by the permissions.

V. Proposed System With Multilevel Anonymization

Privacy preservation and access control mechanism are important concept in every information sharing system. As the privacy preserving access control mechanism discussed in [1], it ensures the privacy and access control frameworks in one system. It provides efficient protection of information

Before getting the original data to the access control module the sensitive information passes through the privacy protection module and get anonymized, so that privacy become efficient when compared to the other mechanisms in[2].

Anonymization techniques replace the original data with some other text or symbols which cannot be easily identifiable by the users. In the privacy preserving access control mechanism [1], the anonymization technique used was generalization method. In generalization the values are replaced with a range of value (For eg. Let the age of bob 20, it becomes <10-30> range after generalization).

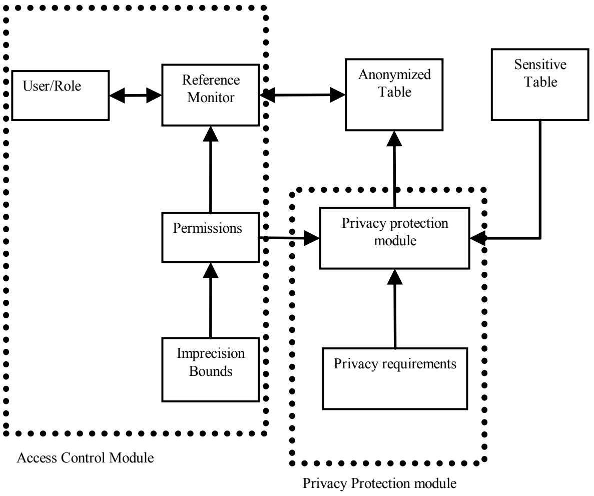


Fig 1 Accuracy-constrained privacy-preserving access control mechanism.

Here the anonymization is applied only once to the data for the security enhancement. The proposed system in which a multilevel anonymization is introduced in order to enhance the security more. In this system, a suppression technique also implemented along with generalization, where the data will be replaced with any symbols or letters (eg. Let the zip code of alice be 812345, after suppression the value can be represented as 8123**, 81****,******, etc).

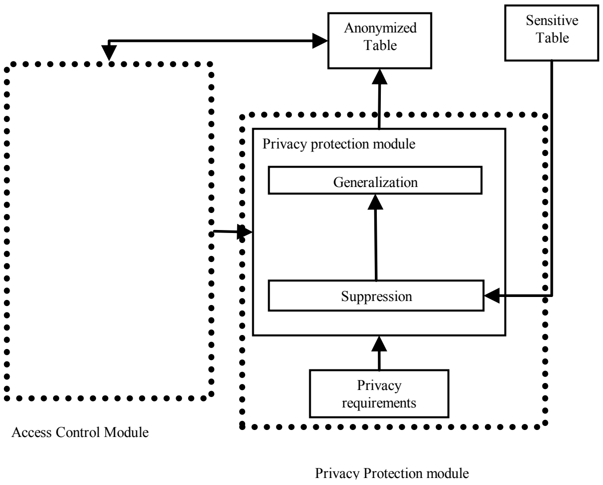


Fig 2. Accuracy-constrained privacy-preserving access control mechanism with multilevel

In the first level anonymization the suppressed information of original table is used and in the second level of anonymization a generalized value is used. So that there it can assure a multilevel security for the data. This also provides a minimum level of precision to the data, along with that the sensitive information will get protected. The proposed system depicted in fig. 2

VI. Experimental Evaluation

The system is implemented in dreamweaver8 with JSP support. Tomcat server is used as application server and mySQL as backend database.

The section describes about the experimental evaluation done in a medical dataset. The privacy preserving access control model and the multilevel access control model are used to show the experimental results. The fig. 3 shows the experimental result.

The fig 3 shows that the number of tuples retrieved by the query given is increased with the increase in the predicates of the query. The blue line shows the number of tuples retrieved when the number of predicate is one, two, and three with three predicates in privacy preserving access control module. The red line shows the number of tuples retrieved by the queries with multiple predicates in multilevel access control model. It clearly shows that the proposed method performs better than the previous methods in terms if number of tuples retrieved.

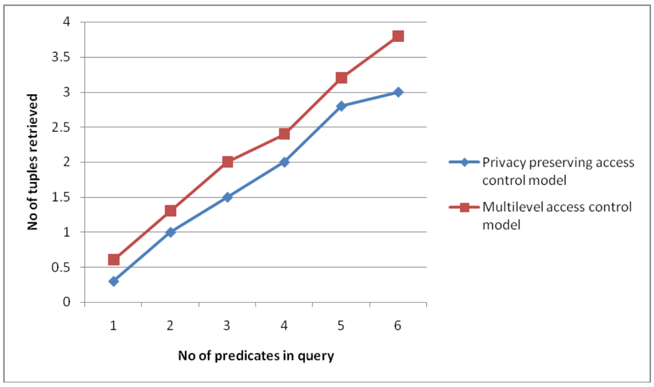


Fig. 3. Experimental results on number of query predicates and number of tuples retrieved

The fig 4 shows the pie chart that depict the relation between the filter count (number of predicates) and result count(number of tuples retrieved).

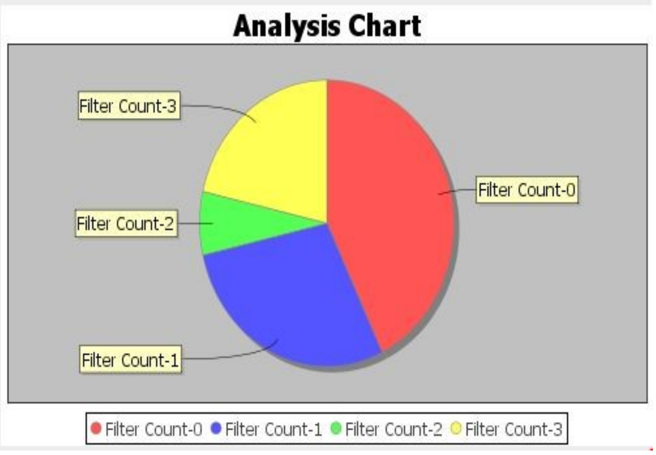
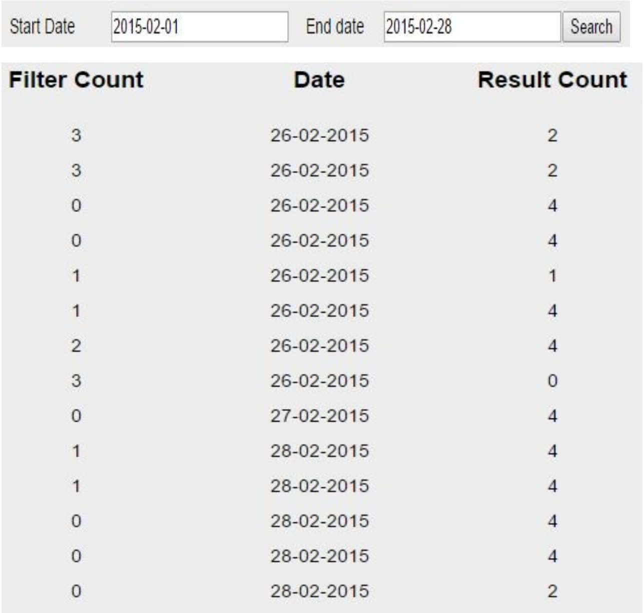


Fig 4. Analysis chart on the basis of filter count and result counts

The pie chart in the fig 4 is drawn with the history of search in the month of February 2015. The Fig 5 shows the details with which the chart was drawn, which is the history of search with different number of predicates in the month of February 2015.



Filter Count	Date	Result Count
3	26-02-2015	2
3	26-02-2015	2
0	26-02-2015	4
0	26-02-2015	4
1	26-02-2015	1
1	26-02-2015	4
2	26-02-2015	4
3	26-02-2015	0
0	27-02-2015	4
1	28-02-2015	4
1	28-02-2015	4
0	28-02-2015	4
0	28-02-2015	4
0	28-02-2015	2

Fig 5. History of search in February

VII. Conclusion

The paper discusses about the need of the security in the databases. The access control and privacy preserving modules are combined in order to provide better results. Here the sensitive information in original database will only be available to the access control modules after providing some privacy to data. Mainly anonymization techniques are used to enforce the privacy and mainly generalization techniques are implemented. In the proposed system a multilevel anonymization is introduced in order to improve the efficiency in privacy preserving access control mechanism. The proposed method works better in terms of tuple retrieval and privacy preservation and it is showed by the experimental results.

Acknowledgment

The first author would like to thanks all those people, who guided and supported. Without their valuable guidance and support, this task was not possible and also likes to thank colleagues for their discussions and suggestions.

REFERENCES

[1] Zahid Pervaiz, Walid G.Aref, Arif Gafoor, Nagabhushana Prabhu "Accuracy constrained privacy preserving access control mechanism for relational databases" IEEE Transaction on Knowledge Engineering, vol.26, No.4, April 2014, pp.795-807

[2] E. Bertino and R. Sandhu, "Database Security-Concepts, Approaches, and Challenges," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, Jan.-Mar. 2005.

[3] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov. 2001.

[4] B. Fung, K. Wang, R. Chen, and P. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, article 14, 2010..

[5] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond k-anonymity,"

ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

[6] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization Techniques for Large-Scale Datasets," ACM Trans. Database Systems, vol. 33, no. 3, pp. 1-47, 2008.

[7] T. Iwuchukwu and J. Naughton, "K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 746-757, 2007.

[8] K. Browder and M. Davidson, "The Virtual Private Database in oracle9ir2," Oracle TechnicalWhite Paper, vol. 500, 2002.

[9] S. Rizvi, A. Mendelzon, S. Sudarshan, and P. Roy, "Extending Query Rewriting Techniques for Fine-Grained Access Control," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 551-562, 2004.

[10] S. Chaudhuri, T. Dutta, and S. Sudarshan, "Fine Grained Authorization through Predicated Grants," Proc. IEEE 23rd Int'l Conf. Data Eng., pp. 1174-1183, 2007.

[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. Data Eng., pp. 25- 25, 2006.

[12] N. Li, W. Qardaji, and D. Su, "Provably Private Data Anonymization: Or, k-Anonymity Meets Differential Privacy," Arxiv preprint arXiv:1101.2604, 2011.

[13] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast Data Anonymization with Low Information Loss," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 758-769, 2007.

[14] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "A Framework for Efficient Data

[15] Anonymization Under Privacy and Accuracy Constraints," ACM Trans. Database Systems, vol. 34, no. 2, article 9, 2009.

[16] E. Bertino, S. Castano, and E. Ferrari "Securing XML Documents with Author-X", IEEE Internet Computing, vol. 5, no. 3, pp. 21-30, 2001.

[17] Neha Sehta et al "A Fine Grained Access Control Model for Relational Databases", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1) , 2012, 3183 – 3186, 3183.

Anusmitha.A, Department of Computer Science & Engineering, Mangalam college of Engineering, Ettumanoor, Kerala, India.

Renjana Ramachandran, Assistant Professor, Department of Computer Science and Engineering, Mangalam college of Engineering, Ettumanoor, Kerala, India