IPHV8I1021X

International Journal Of Advanced Research and Innovation -Vol.8, Issue .I

ISSN Online: 2319 – 9253
Print: 2319 – 9245

DATABASE CONTROL & PRIVACY: A FRAMEWORK FOR FAST DATA ANNONYMIZATION

#1MUSTYALA ROOPA, M.Tech Student,
#2A.SRINISH REDDY, Assistant Professor,
Dept of CSE,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, T.S., INDIA.

Abstract: Data privacy issues are increasingly becoming important for many applications. Protective individual privacy is a crucial downside. However, sensitive data will still be ill-used by approved users to compromise the privacy of shoppers. Traditionally, research in the database community in the area of data security can be broadly classified into access control research and data privacy research. Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users. Privacy Protection Mechanism (PPM) uses suppression and generalization of relational data to anonymize and satisfy privacy needs. Recent research studied the problem of publishing data in databases without revealing the sensitive information, moving to the privacy preserving paradigms of k-anonymity and L-diversity. While satisfying the privacy requirement, k-anonymity or l-diversity, the access control policies define selection predicates available to rolls. The PPM needs to satisfy an additional constraint namely the Imprecision Bound for each selection predicate. The literature survey might provide techniques for workload-aware anonymization for selection predicates, as the problem of satisfying the accuracy constraints for multiple roles has not been studied before. The purpose of the present project is to propose heuristics for anonymization algorithms and to show the viability of the proposed approach for empirically satisfying the imprecision bounds for more permission.

Key words: Access control, Privacy, k-anonymity, Precision, Imprecision, l-diversity.

I.INTRODUCITON

Organizations collect and analyze consumer data to improve their services. Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to the users. However, sensitive information can still be misused by authorized users compromising the privacy of consumers.

The concept of privacy-preservation for sensitive data can require the enforcement of privacy policies or the protection against identity disclosure by satisfying some privacy requirements.

The anonymity techniques can be used with an access control mechanism [1] to ensure both security and privacy of the sensitive information. The privacy is achieved at the cost of accuracy and imprecision is introduced in the authorized information under an access control policy. An integrated framework of achieving both privacy and security is proposed though the integration of Access Control Mechanism with Privacy Preservation [1] Technique to prevent the authorized user from misusing the sensitive information. The enforcement of privacy policies or the protection against identity disclosure satisfying some privacy requirements are the pre-requisites for privacy-preservation of sensitive data. Even after removal of identifying attributes, the sensitive information is susceptible to liking attacks by the authorized users. So the present investigation is proposed to study the area of micro data publishing and privacy definitions such as k-anonymity [2], l-diversity [3] and variance diversity.

The privacy requirements with minimal distortion of micro data can be satisfied by using suppression and generalization of anonymization algorithms. In a way to ensure security and privacy of sensitive information, the anonymity techniques can be used. To define a threshold on the amount of imprecision that can be tolerated for each permission, the concept of imprecision bound is to be used. A role based access [4][5] control is assumed in a way to focus on a static relational table that is anonymized only once.

In existing system [1] the heuristics proposed in this paper for accuracy constrained privacy-preserving access control are also relevant in the context of workload-aware anonymization. The framework is a combination of access control and privacy protection mechanisms. The concept of privacy-preservation for sensitive data requires the enforcement of privacy policies or the protection against identity disclosure by satisfying some privacy requirements by investigating privacy-preservation from the anonymity aspect.

The sensitive information, even after the removal of identifying attributes, is still susceptible to linking attacks by the authorized users. But it has some disadvantages such as – User doesn't have efficient privacy and accurate constraints. System fails to retrieve data in customized way. It minimizes the imprecision aggregate for all queries. The imprecision added to each permission/query in the anonymized micro data is not known, thus, not satisfying accuracy constraints for individual permissions in a policy/workload. System doesn't provide security for data which motivated me to work on this.

An accuracy-constrained privacy-preserving access control mechanism, illustrated in Fig.[1] (Arrows represent the direction of information flow), is proposed. The privacy protection mechanism ensures that the privacy and accuracygoals are met before the sensitive data is available to the access control mechanism. The permissions in the access control policy are based on selection predicates on the QI attributes. The policy administrator defines the permissions along with the imprecision bound for each permission/query, user-to-role assignments, and role-to permission assignments [6].The imprecision bound information is not shared with the users because knowing the imprecision bound can result in violating the privacy requirement. The privacy protection mechanism is required to meet the privacy requirement along with the imprecision bound for each permission.

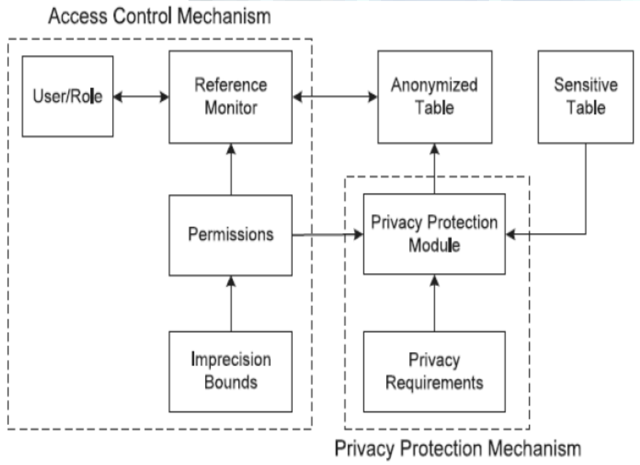

Fig 1: Accuracy-constrained privacy-preserving access control mechanism.

To overcome the disadvantages of existing system the heuristics proposed in this paper for accuracy constrained privacy-preserving access control are also relevant in the context of workload-aware anonymization. The framework is a combination of access control and privacy protection mechanisms. The access control mechanism allows only authorized query predicates on sensitive data. The privacy preserving module anonymizes the data to meet privacy requirements and imprecision constraints on predicates set by the access control mechanism and the advantages of proposed system are - formulate the accuracy and privacy constraints. Concept of accuracy-constrained privacy-preserving access control for relational data was studied and the solution of the k-PIB problem was approximated and empirical evaluation was conducted.

II.RELATED WORK

Access control mechanisms for databases allow queries only on the authorized part of the database. Predicate based fine-grained access control has further been proposed, where user authorization is limited to pre-defined predicates. Enforcement of access control and privacy policies has been studied. However, studying the interaction between the access control mechanisms and the privacy protection mechanisms has been missing. Recently, Chaudhuri et al. have studied access control with privacy mechanisms. They use the definition of differential privacy whereby random noise is added to original query results to satisfy privacy constraints. They have not considered the accuracy constraints for permissions. We define the privacy requirement in terms of k-anonymity. It has been shown by Li et al. [6] that after sampling, k-anonymity offers similar privacy guarantees as those of differential privacy. The proposed accuracy-constrained privacy preserving access control framework allows the access control administrator to specify imprecision constraints that the privacy protection mechanism is required to meet along with the privacy requirements.

The challenges of privacy-aware access control are similar to the problem of workload-aware anonymization. In our analysis of the related work, we focus on query-aware anonymization. For the state of the art in k-anonymity techniques and algorithms, we refer the reader to a recent survey paper [3]. Workload-aware anonymization is first studied by LeFevre et al. [5] They have proposed the Selection Mondrian algorithm [4], which is a modification to the greedy multidimensional partitioning algorithm Mondrian. In their algorithm, based on the given query-workload, the greedy splitting heuristic minimizes the sum of imprecision for all queries. Iwuchukwu and Naughton have proposed an Rþ-tree based anonymization algorithm.

The authors illustrate by experiments that anonymized data using biased Rþ-tree based on the given query workload is more accurate for those queries than for an unbiased algorithm. Ghinita et al. have proposed algorithms based on space filling curves for k-anonymity and l-diversity [10]. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class [8]. Similarly, Xiao et al. [9] propose to add noise to queries according to the size of the queries in a given workload to satisfy differential privacy. 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. Anonymization with imprecision constraints for individual queries has not been studied before. We follow the imprecision definition of LeFevre et al.and introduce the constraint of imprecision bound for each query in a given query workload.

III. PRIVACY-PRESERVING ACCESS CONTROL MODEL FOR RELATIONAL DATA

Organizations collect and analyze consumer data to improve their services. Access Control Mechanisms (ACM) is used to ensure that only authorized information is available to users. Sensitive information can still be misused by authorized users to compromise the privacy of consumers. The concept of privacy-preservation for sensitive data can require the enforcement of privacy policies or the protection against identity disclosure by satisfying some privacy requirements. In this paper, we investigate privacy-preservation from the anonymity aspect. The sensitive information, even after the removal of identifying attributes, is still susceptible to linking attacks by the authorized users. This problem has been studied extensively in the area of micro data publishing [3] and privacy definitions, e.g., k-anonymity, l-diversity and variance diversity. Anonymization algorithms use suppression and generalization of records to satisfy privacy requirements with minimal distortion of micro data.

The anonymity techniques can be used with an access control mechanism to ensure both security and privacy of the sensitive information. The privacy is achieved at the cost of accuracy and imprecision is introduced in the authorized information under an access control policy.

We use the concept of imprecision bound for each permission to define a threshold on the amount of imprecision that can be tolerated. Existing workload aware anonymization techniques [5] minimize the imprecision aggregate for all queries and the imprecision added to each permission/query in the anonymized micro data is not known. Making the privacy requirement more stringent results in additional imprecision for queries. The problem of satisfying accuracy constraints for individual permissions in a policy/workload has not been studied before. The heuristics proposed in this paper for accuracy-constrained privacy-preserving access control are also relevant in the context of workload aware anonymization. The anonymization for continuous data publishing has been studied in literature [3]. In this paper the focus is on a static relational table that is anonymized only once. To exemplify our approach, role-based access control is assumed. The concept of accuracy constraints for permissions can be applied to any privacy-preserving security policy, e.g., discretionary access control.

IV.DATA PARTITIONING FOR PRIVACY PRESERVATION

In this section, three algorithms based on greedy heuristics are proposed. All three algorithms are based on kd-tree construction. Starting with the whole tuple space the nodes in the kd-tree are recursively divided till the partition size is between k and 2k. The leaf nodes of the kd-tree are the output partitions that are mapped to equivalence classes [1]. Heuristic 1 and 2 have time complexity of O(d 2 *Q* n 2). Heuristic 3 is a modification over Heuristic 2 to have O(d|Q|nl gn) complexity, which is same as that of TDSM. The proposed query cut can also be used to split partitions using bottom- up (Rþtree) techniques.

4.1 Top-Down Heuristic 1 (TDH1)

In TDSM, the partitions are split along the median. Consider a partition that overlaps a query. If the median also falls inside the query then even after splitting the partition, the imprecision for that query will not change as both the new partitions still overlap the query as illustrated. In this heuristic, we propose to split the partition along the query cut and then choose the dimension along which the imprecision is minimum for all queries [2]. If multiple queries overlap a partition, then the query tobe used for the cut needs to be selected. The queries having imprecision greater than zero for the partition are sorted based on the imprecision bound and the query with minimum imprecision bound is selected.

The intuition behind this decision is that the queries with smaller bounds have lower tolerance for error and such a partition split ensures the decrease in imprecision for the query with the smallest imprecision bound. If no feasible cut satisfying the privacy requirement is found, then the next query in the sorted list is used to check for partition split. If none of the queries allow partition split, then that partition is split along the median and the resulting partitions are added to the output after compaction.

The TDH1 algorithm is listed in Algorithm 1. In the first line, the whole tuple space is added to the set of candidate partitions. In the Lines 3-4, the query overlapping the candidate partition with least imprecision bound and imprecision greater than zero is selected. The while loop in Lines 5-8 checks for a feasible split of the partition along query intervals. If a feasible cut is found, then the resulting partitions are added to CP. Otherwise, the candidate partition is checked for median cut in Line 12. A feasible cut means that each partition resulting from split should satisfy the privacy requirement. The traversal of the kd-tree for partitions to consider in Set CP can be depth-first or breadth-first. The order of traversal for TDH1 does not matter.

Input: T,K,Q and BQ *j*
Output: P

```
1 Initialize set of candidate partitions(CP □T)
2 for (CP i □CP) do
3 Find the set of queries QO that overlap CP i
such that ic jiQO
CP > 0
4 sort queries QO in increasing order of BQj
5 while (feasible cut is not found) do
6 Select query from QO
7 Create query cuts in each dimension
8 Select dimension and cut having least overall imprecision
for all queries in Q
9 if (feasible cut found) then
10 Create new partitions and add to CP
11 else
12 Split CP i recursively along median till anonymity
requirement is satisfied
13 Compact new partitions and add to P
14 return (P)
```

Algorithm 1: TDH1

4.2 Top-Down Heuristic 2 (TDH2)

In the Top-Down Heuristic 2 algorithm, the query bounds are updated as the partitions are added to the output. This update is carried out by subtracting the ic Qj Pi value from the imprecision bound BQj of each query, for a Partition, say Pi, that is being added to the output. For example, if a partition of size k has imprecision 5 and 10 for Queries Q1 and Q2 with imprecision bound 100 and 200, then the bounds are changed to 95 and 190, respectively. The best results are achieved if the kdtree traversal is depth-first (preorder). Preorder traversal for the kd-tree ensures that a given partition is recursively split till the leaf node is reached. Then, the query bounds are updated. Initially, this approach favors queries with smaller bounds. As more partitions are added to the output, all the queries are treated fairly. During the query bound update, if the imprecision bound for any query gets violated, then that query is put on low priority by replacing the query bound by the query size. The intuition behind this decision is that whatever future partition splits TDH2 makes, the query bound for this query cannot be satisfied. Hence, the focus should be on the remaining queries.

Input : T,K,Q and BQj
Output: P

```
1 Initialize set of candidate partitions (CP □ T)
2 for (CP i □CP ) do //Depth first preorder traversal
3 Find the set of queries QO that overlap CP I such that ic QOj
CPi > 0
4 Sort queries QO in increasing order of BO j
5 While (feasible cut is not found) do
6 Select query from QO
7 Create query cut each dimension
8 select dimension and cut having least Overall imprecision
for all queries in Q
9 if (Feasible cut found) then
10 Create new partitions and add to CP
11 else
12 Split CP i recursively along median till anonymity
requirement is satisfied
13 Compact new partitions and add to P
14 Update BQ j according to ic Qipi , □ Q j□Q
15 return (P)
```

Algorithm 2: TDH2

V.CONCLUSION

An accuracy-constrained privacy-preserving access control framework for relational data has been proposed. The framework is a combination of access control and privacy protection mechanisms. The access control mechanism allows only authorized query predicates on sensitive data. The privacy preserving module anonymizes the data to meet privacy requirements and imprecision constraints on predicates set by the access control mechanism. We formulate this interaction as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB). We give hardness results for the k-PIB problem and present heuristics for partitioning the data to the satisfy the privacy constraints and the imprecision bounds. In the current work, static access control and relational data model has been assumed. For future work, we plan to extend the proposed privacy-preserving access control to incremental data and cell level access control.

This paper presents a heuristics method for partitioning the data to satisfy the privacy constraints and the imprecision bounds. This proposed paper gives a secured access control mechanism and privacy protection mechanism for the relational data. In the current work, static access control and relational data model has been assumed. For future work, it plan to extend the proposed privacy-preserving access control to cell level access control and can use the l-diversity instead of k-anonymity method.

REFERENCES

[1] ZahidPervaiz, Walid G. Aref, ArifGhafoor, and NagabhushanaPrabhu, Accuracy-Constrained Privacy-Preserving Access Control Mechanismfor Relational Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, NO. 4, April 2014.
[2] Ms. S.Kokila, Dr. T. SenthilPrakash, and Ms. P.Maheswari, " Privacy and Security Ensured Database Rights Management Scheme", International Journal On engineering Technology and Sciences – IJETS™ ISSN (P): 2349-3968, ISSN (O): 2349-3976 Volume 1 Issue 6, October 2014.
[3] Yung-Wang Lin, Li-Cheng Yang, Luon-Chang Lin, and Yeong-Chin Chen, "Preserving Privacy in Outsourced Database", International Journal of Computer and Communication Engineering, Vol. 3, No. 5, September 2014.
[4] T.Sujitha, V.Saravanakumar, C.Saravanabhavan, "An Efficient Cryptographic approach For Preserving Privacy In Data Mining", International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013.
[5] ZahidPervaiz, ArifGhafoor, and Walid G. Aref , "Precision bounded access control for privacy preserving data stream", CERIAS Tech Report 2013-7.
[6] Alaa H Al-Hamami, and Suhad Abu Shehab, "An Approach for Preserving Privacy and Knowledge In Data Mining Applications", Journal of Emerging Trends in Computing and Information Sciences, Vol. 4, No. 1 Jan 2013.
[7] N.Punitha, R.Amsaveni, "Methods and Techniques to Protect the Privacy Information in Privacy Preservation Data Mining" IJCTA | NOV-DEC 2011.
[8] S. Chaudhuri, R. Kaushik, and R. Ramamurthy, "Database Access Control & Privacy: Is There a Common Ground?" Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR), pp. 96-103, 2011.
[9] Gabriel Ghinita, PanosKalnis and Yufei Tao," Anonymous Publication of Sensitive Transactional Data", IEEE Transactions on Knowledge and Data Engineering, vol. 23, Issue.2,pp.161-174,2011.
[10] N. Li, W. Qardaji, and D. Su, "Provably Private Data Anonymization: Or, k-Anonymity Meets Differential Privacy," Arxiv preprint arXiv:1101.2604, 2011.
[11] Shahidul Islam Khan, Dr. A. S. M. LatifulHoque, "A New Technique for Database Fragmentation in Distributed Systems", International Journal of Computer Applications (0975 – 8887) Volume 5– No.9, August 2010.
[12] Xin Jin, Nan Zhang, Gautam Das, "Algorithm-Safe Privacy-Preserving Data Publishing" EDBT 2010, March 22–26, 2010.
[13] 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.
[14] Gabriel Ghinita, PanagiotisKarras, And PanosKalnis," A Framework for Efficient Data Anonymization under Privacy and Accuracy Constraints", ACM Transactions on Database Systems, Vol. 34, No. 2, Article 9, Publication date: June 2009.
[15] LalanthikaVasudevan , S.E. DeepaSukanya, and N. Aarthi, "Privacy Preserving Data Mining Using Cryptographic Role Based Access", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 VolI IMECS 2008, 19-21 March, 2008.