ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Empirical Evaluation of the Heuristics for Partitioning in Relational Data

Ireddy Rajender
M.Tech Student,
Department of CSE,
Global Group of Institutions,
Batasingaram, Ranga Reddy (Dist).

Mr. S Dilli Babu, M. Tech
Assistant Professor,
Department of CSE,
Global Group of Institutions,
Batasingaram, Ranga Reddy (Dist).

Mr. M V Narayana, M. Tech, Ph.D
Associate Professor & HOD,
Department of CSE,
Global Group of Institutions,
Batasingaram, Ranga Reddy (Dist).

Abstract:

While Sensitive information is being shared, an authorized user has to compromise with his privacy leading to identity disclosure. But Privacy Protection Mechanism (PPM) with its suppression and generalization of relational data anonymizes and satisfies privacy requirements using k-anonymity and l-diversity, against identity and attribute disclosure. Thus access control mechanism helps in protecting sensitive information from unauthorized users. Usually privacy is achieved at the cost of precision of authorized information. The present project focuses on an accuracy-constrained privacy-preserving access control framework. 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.

Index Terms:

Access control, privacy, k-anonymity, query evaluation.

INTRODUCTION:

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.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

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 accuracy goals 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.

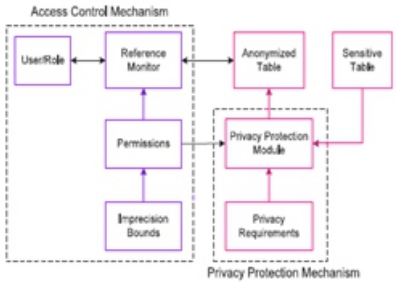


Fig:-access control mechanism.

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.

LITERATURE REVIEW:

Various papers were referred for the present research regarding access control mechanism, privacy preserving, k-anonymity, and for workload aware anonymity concepts. In this, the Private Data Anonymization was proposed or, kAnonymity Meets Differential Privacy was discussed by Li et al [7], they defined the privacy requirement in terms of kanonymity that after sampling, k-anonymity offers similar privacy guarantees as those of differential privacy. In this paper the privacy requirement were defined in terms of k-anonymity [2] 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 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, queryaware anonymization was focused. The problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class was proposed [8]. Similarly, Xiao et al. [9] proposed to add noise to queries according to the size of the queries in a given workload to satisfy differential privacy. However, 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. However, anonymization with imprecision constraints for individual queries has not been studied before.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

For the present study, the imprecision definition of Lefebvre et al. is followed[10] and the constraint of imprecision bound for each query in a given query workload is introduced. In which they concluded the problem of measuring the quality of anonymized data. The most direct way of measuring quality is with respect to the purpose for which the data will be used. For this reason, a suite of techniques for incorporating a family of tasks (comprised of queries, classification, and regression models) were introduced directly into the anonymization procedure.

The interaction between the access control mechanisms and the privacy protection mechanisms was discussed by Chaudhuri et al. [11]. Access control with privacy mechanisms in which they concluded with the sketch of an architecture for a hybrid system that enhances an authorization policy with the abstraction of noisy views that encapsulate previously proposed privacy mechanisms. Accessing data through a set of views is natural for users of database systems and thus the noisy views abstraction represents a natural progression of the concept of authorization views. It was also stated how noisy views based on differentially private algorithms could be implemented. A key advantage of the proposed hybrid system is its flexibility. It can support queries that refer to both the base tables and the differentially private views thus resulting in a system that is more powerful than using access control techniques or differential privacy techniques in isolation. While combining authorizations and differentially private views in this manner seems ad-hoc, it is shown to be a principled way to integrate differential privacy primitives with privacy guarantees [11].

The definition of differential privacy was used [12] whereby random noise is added to original query which results to satisfy privacy constraints. However, the accuracy constraints for permissions were not considered. But the present study defines the privacy requirement in terms of k-anonymity. Workload-aware anonymization is first studied by LeFevre et al. [10]. They have proposed the Selection Mondrian algorithm, which is a modification to the greedy multidimensional partitioning algorithm Mondrian [13]. In their algorithm, based on the given query-workload, the greedy splitting heuristic minimizes the sum of imprecision for all queries. The present study has considered the problem of measuring the quality of anonymized data. It is our position that the most direct way of measuring quality is with respect to the purpose for which the data will be used.

For this reason, a suite of techniques were developed for incorporating a family of tasks (comprised of queries, classification, and regression models) directly into the anonymization procedure. An extensive empirical study indicates that this typically leads to high-quality data. Further, the quality of the data with respect to a particular workload is not necessarily correlated with simple general-purpose measures that have been proposed in the previous literature. In the second half of the article, the problem of scalability is introduced. Two techniques were developed that allow our anonymization algorithms to be applied to datasets much larger than main memory.

The first technique is based on ideas from scalable decision trees [Gehrke et al. 1998], and the second is based on sampling. An experimental evaluation and analytical study indicate that these techniques work very well in practice. Iwuchukwu and Naughton have proposed an R+-tree based anonymization algorithm [14]. The authors illustrated by experiments that anonymized data using biased R+-tree based on the given query workload is more accurate for those queries than for an unbiased algorithm. Further Ghinita et al. have proposed algorithms based on space filling curves for k-anonymity and l-diversity [3]. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class [15].

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. However, 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. However, anonymization with imprecision constraints for individual queries has not been studied before. We follow the imprecision definition of LeFevre et al. [10] and introduced the constraint of imprecision bound for each query in a given query workload.

EXISTING SYSTEM:

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

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

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. Existing workload aware anonymization techniques 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 (e.g., increasing the value of k or l) results in additional imprecision for queries.

Dis-Advantages:

1.Their is no privacy for users
2.The sensitive information,even after the removal of identifying attributes,is still susceptible to linking attacks by the authorized users.

PROPOSED 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 anonymization for continuous data publishing has been studied in literature. 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. However, the concept of accuracy constraints for permissions can be applied to any privacy-preserving security policy, e.g., discretionary access control.

Advantages:

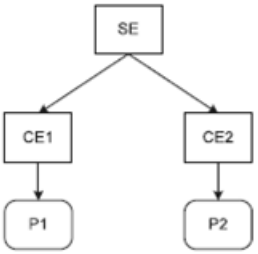
1.accuracy constrained privacy preserving access.

2.It's maintain data's in secure manner.

IMPLEMENTATION:

a)Access control policy:

Role	Designation
SE	State Epidemiologist
CE1	County 1 Epidemiologist
CE2	County 2 Epidemiologist



Permission	Authorized Query Predicate
P1	Location = County 1 ∧ Age = 15-65 ∧ Syndrome = Influenza
P2	Location = County 2 ∧ Age = 15-65 ∧ Syndrome = Influenza

Syndromic surveillance systems are used at the state and federal levels to detect and monitor threats to public health. The department of health in a state collects the emergency department data (age, gender, location, time of arrival, symptoms, etc.) from county hospitals daily. Generally, each daily update consists of a static instance that is classified into syndrome categories by the department of health. Then, the surveillance data is anonymized and shared with departments of health at each county. An access control policy is given in Fig. 1 that allows the roles to access the tuples under the authorized predicate, e.g., Role CE1 can access tuples under PermissionP1. The epidemiologists at the state and county level suggest community containment measures ,e.g., isolation or quarantine according to the number of persons infected in case of a flu outbreak. According to the population density in a county, an epidemiologist can advise isolation if the number of persons reported with influenza are greater than 1,000 and quarantine if that number is greater than 3,000 in a single day. The anonymization adds imprecision to the query results and the imprecision bound for each query ensures that the results are within the tolerance required. If the imprecision bounds are not satisfied then unnecessary false alarms are generated due to the high rate of false positives.

b)Anonymity:

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	Diarrhea
4	25	15	Fever
5	22	28	Flu
6	32	35	Fever
7	38	32	Flu
8	35	25	Diarrhea

(a) Sensitive table

	QI_1	QI_2	S_1
ID	Age	Zip	Disease
1	0-20	10-30	Flu
2	0-20	10-30	Fever
3	20-30	10-30	Diarrhea
4	20-30	10-30	Fever
5	20-30	10-30	Flu
6	30-40	20-40	Fever
7	30-40	20-40	Flu
8	30-40	20-40	Diarrhea

(b) 2-anonymous Table

anonymity is prone to homogeneity attacks when the sensitive value for all the tuples in an equivalence class is the same. To counter this shortcoming, l-diversity has been proposed and requires that each equivalence Access control policy. class of T_ contain at least l distinct values of the sensitive attribute. For sensitive numeric attributes, an l-diverse equivalence class can still leak information if the numeric values are close to each other. For such cases, variance diversity has been proposed that requires the variance of each equivalence class to be greater than a given variance diversity parameter. The table in Fig. 2a does not satisfy k-anonymity because knowing the age

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

and zip code of a person allows associating a disease to that person. The table in Fig. 2b is a 2-anonymous and 2-diverse version of table. The ID attribute is removed in the anonymized table and is shown only for identification of tuples. Here, for any combination of selection predicates on the zip code and age attributes, there are at least two tuples in each equivalence class

c)Anonymization with impresision Bounds:

we formulate the problem of k-anonymous Partitioning with Imprecision Bounds and present an accuracy-constrained privacy-preserving access control framework. Imprecise data means that some data are known only to the extent that the true values lie within prescribed bounds while other data are known only in terms of ordinal relations. Imprecise data envelopment analysis (IDEA) has been developed to measure the relative efficiency of decision-making units (DMUs) whose input and/or output data are imprecise. In this paper, we show two distinct strategies to arrive at an upper and lower bound of efficiency that the evaluated DMU can have within the given imprecise data. The optimistic strategy pursues the best score among various possible scores of efficiency and the conservative strategy seeks the worst score. In doing so, we do not limit our attention to the treatment of special forms of imprecise data only, as done in some of the studies associated with IDEA. We target how to deal with imprecise data in a more general form and, under this circumstance, we make it possible to grasp an upper and lower bound of efficiency.

d)Accuracy-Constrained Privacy-Preserving Access Control:

An accuracy-constrained privacy-preserving access control mechanism. (arrows represent the direction of information flow), is proposed. The privacy protection mechanism ensures that the privacy and accuracy goals 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 .The specification of the imprecision bound ensures that the authorized data has the desired level of accuracy. 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.

e)Top-Down Heuristic:

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. If multiple queries overlap a partition, then the query to be 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.

Algorithm 1: TDH1

Input : T, k, Q, and B_{Q_j}
Output: P

Initialize Set of Candidate Partitions($CP \leftarrow T$)
for ($CP_i \in CP$) **do**
  Find the set of queries QO that overlap CP_i such that $ic_{CP_i}^{QO_j} > 0$
  Sort queries QO in increasing order of B_{Q_j}
  while *(feasible cut is not found)* **do**
    Select query from QO
    Create query cuts in each dimension
    Select dimension and cut having least overall imprecision for all queries in Q
  if *(Feasible cut found)* **then**
    Create new partitions and add to CP
  else
    Split CP_i recursively along median till anonymity requirement is satisfied
    Compact new partitions and add to P
return (P)

CONCLUSION:

An accuracy-constrained privacy-preserving access control framework for relational data has been proposed. The planned additive approach of access management and privacy protection mechanisms in our system provides a lot

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

of security and information is retrieved during a custom-made approach which will build users to access during as lot of versatile approach. Any access management concentrates on anomaly users to avoid privacy problems security .The ACM allows solely licensed user predicates on sensitive information and PPM anonymizes the information to satisfy privacy necessities and inexactness constraints on predicates set by the access management mechanism. 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. This interaction is formulated as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB). Hardness results are given for the k-PIB problem and the heuristics for partitioning the data are presented to satisfy the privacy constraints and the imprecision bounds. In the current work, static access control and relational data model has been assumed. The proposed privacy-preserving access is extended to control incremental data and cell level access control.

REFERENCES:

[1] 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.

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

[3] 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.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubrama- niam, "L-Diversity: Privacy Beyond k-anonymity," ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

[5] 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.

[6] 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.

[7] J. Buehler, A. Sonricker, M. Paladini, P. Soper, and F. Mostashari, "Syndromic Surveillance Practice in the United States: Findings from a Survey of State, Territorial, and Selected Local Health Departments," Advances in Disease Surveillance, vol. 6, no. 3, pp. 1- 20, 2008.

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

[9] A. Rask, D. Rubin, and B. Neumann, "Implementing Row-and Cell-Level Security in Classified Databases Using SQL Server 2005," MS SQL Server Technical Center, 2005.

[10] 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.

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

[12] K. LeFevre, R. Agrawal, V. Ercegovac, R. Ramakrishnan, Y. Xu, and D. DeWitt, "Limiting Disclosure in Hippocratic Databases," Proc. 30th Int'l Conf. Very Large Data Bases, pp. 108-119, 2004.

[13] D. Ferraiolo, R. Sandhu, S. Gavrila, D. Kuhn, and R. Chandra- mouli, "Proposed NIST Standard for Role-Based Access Control," ACM Trans. Information and System Security, vol. 4, no. 3, pp. 224- 274, 2001.

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

[15] J. Friedman, J. Bentley, and R. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Mathe- matical Software, vol. 3, no. 3, pp. 209-226, 1977.