ISSN 23499842(Online), Volume 1, Special Issue 2(ICITET 15), March 2015

International Journal of Innovative Trends and Emerging Technologies

DYNAMIC DATA AND POLICY BASED ACCESS CONTROL SCHEME FOR RELATIONAL DATA

Mrs.N.M.INDUMATHI[1], Mrs.M.MOHANASUNDARI [2],
[1]PG Scholar, [2]Assistant Professor(Sr.Gr)
[1,2]Department of Computer Science and Engineering
[1,2]Velalar College of Engineering and Technology,Erode-12
induksrce@gmail.com

Abstract

Privacy preserved data mining methods are used to protect the sensitive attributes in knowledge discovery process. Privacy preservation is used to protect private data values. Anonymity is considered in the privacy preservation process. Clustering method is used to group up the records based on the relevancy. Distance or similarity measures are used to estimate the transaction relationship. Census data and medical data are referred as micro data.

User permissions are managed with dynamic data and policy management mechanism with privacy. Privacy Protection Mechanism (PPM) uses suppression and generalization of relational data to anonymize and satisfy privacy needs. Accuracy-constrained privacy-preserving access control framework is used to manage access control in relational database. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k-anonymity or l-diversity. Imprecision bound constraint is assigned for each selection predicate. Role-based Access Control (RBAC) allows defining permissions on objects based on roles in an organization. Top Down Selection Mondrian (TDSM) algorithm is used for query workload-based anonymization. Query cuts are selected with minimum bounds in Top-Down Heuristic 1 algorithm (TDH1). The query bounds are updated as the partitions are added to the output in Top-Down Heuristic 2 algorithm (TDH2). The cost of reduced precision in the query results is used in Top-Down Heuristic 3 algorithm (TDH3). Repartitioning algorithm is used to reduce the total imprecision for the queries.

The policy based access control mechanism is enhanced to support dynamic data management technique. Data insert, delete and update operations are connected with the partition management mechanism. Cell level access control is provided with differential privacy method. Dynamic role management model is integrated with the access control policy mechanism for query predicates.

1. Introduction

Privacy is today a key issue in information technology and has received increasing attention from consumers, companies, researchers and legislators. Legislative acts, such as Health Insurance Portability and Accountability Act (HIPAA) for healthcare and Gramm Leach Bliley Act (GLBA) for financial institutions, require enterprises to protect the privacy of their customers. Although enterprises have adopted various strategies to protect customer privacy and to communicate their privacy policies to customers, such as publishing a privacy policy on websites possibly based on P3P, or incorporating privacy seal programs, in these approaches there are not systematic mechanisms that describe how consumer personal data is actually handled after it is collected. Privacy protection can only be achieved by enforcing privacy policies within an enterprise's online and offline data processing systems. Otherwise, enterprises' actual practices might intentionally or unintentionally violate the privacy policies published at their websites.

Conventional access models, such as Mandatory Access Control (MAC), Discretionary Access Control (DAC) and Role Based Access Control (RBAC), are not designed to enforce privacy policies and barely meet privacy protection requirements, particularly, purpose binding, conditions and obligations. The significance of purposes, conditions and obligations originates from OECD Guidelines on the Protection of Privacy and Transformer Flows of Personal Data, current privacy laws in the United States and public privacy policies of some well know organizations. The OECD guidelines are, to the best of our knowledge, the most well known set of private information protection principles, on which many other guidelines, data-protection laws and public privacy policies are based. Purposes are directly applied in the OECD *Data Quality Principle*, *Purpose Specification Principle* and *Use Limitation Principle*. Purposes are also widely used for specifying privacy rules in legislative acts and actual public policies. HIPPA rules clearly state purposes. The majority of public privacy documents posted at well known sites also specify purposes.

Obligations, that is, actions to be performed after an action has been executed on data objects, are necessary for some cases. For example, the OECD *Accountability Principle* states that "A data controller should be accountable for complying with measures which give effect to the principles stated above". A common approach to implement this principle in operating systems or DBMS is to log each data access as an event. Executing logging actions could be an obligation for the majority of privacy policies. Conditions, that is, prerequisites to be met before any action can be executed, are critical in some cases. One of these cases is related to children information. One of the most important rules in COPPA is the so called *Verifiable Parental Consent* (VPC): before collecting, using or disclosing personal information from a child, an operator must obtain verifiable parental consent from the child's parent. The VPC is a condition that must be satisfied before collecting and accessing personal information related to children under thirteen.

Existing access control technology can be used as a starting point for managing personal identifiable information in a trustworthy fashion. A language used for privacy policies must be the same as or integrated with the language used for access control policies, because both types of policy usually control access to the same resources and should not conflict with one another [3]. Hence, we propose a family of Privacy-aware Role Based Access Control (PRBAC) models that naturally extend classical RBAC models to support privacy policies.

We believe that an RBAC-based solution to the problem of privacy aware access control may have a great potential. It could be easily deployed in systems already adopting RBAC and would thus allow one to seamlessly introduce access control policies specialized for privacy enforcement. The goal of the work reported in this paper is to extend the RBAC model in order to support privacy-aware access control. In our model, referred to as PRBAC, privacy policies are expressed as permission assignments (PA); these permissions differ from permissions in classical RBAC because of the presence of additional components, representing privacy related information. We also develop conflict analysis algorithms to detect conflicts among PA, thus avoiding the problems that EPAL rules have because of its sequential semantics.

2. Related Work

Data privacy has been an active research topic in the statistics, database and security communities for the last three decades. The proposed methods can be roughly categorized according to two main scenarios:

- Interactive versus noninteractive. In an interactive framework, a data miner can pose queries through a private mechanism and a database owner answers these queries in response. In a noninteractive framework, a database owner first anonymizes the raw data and then releases the anonymized version for data analysis. Once the data are published, the data owner has no further control over the published data. This approach is also known as privacy preserving data publishing (PPDP).
- Single versus multiparty. Data may be owned by a single party or by multiple parties. In the distributed scenario, data owners want to achieve the same tasks as single parties on their integrated data without sharing their data with others.

Our proposed algorithm addresses the distributed and noninteractive scenario. Below, we briefly review the most relevant research works. Single-party scenario. We have already discussed different privacy models. Here, we provide an overview of some relevant anonymization algorithms. Many algorithms have been proposed to preserve privacy, but only a few have considered the goal for classification analysis. Iyengar has presented the anonymity problem for classification and proposed a genetic algorithmic solution. Bayardo and Agrawal have also addressed the classification problem using the same classification metric. Fung et al. have proposed a top-down specialization (TDS) approach to generalize a datatable. LeFevre et al. have proposed another anonymization technique for classification using multidimensional recoding. More discussion about the partition-based approach can be found in the survey of Fung et al. [7]. Differential privacy has recently received considerable attention as a substitute for partition-based privacy models for PPDP. So far most of the research on differential privacy concentrates on the interactive setting with the goal of reducing the magnitude of the added noise, releasing certain data mining results [8], [9], or determining the feasibility and infeasibility results of differentially-private mechanisms [6]. Research proposals [1] that address the problem of noninteractive data release only consider the single-party scenario. Therefore, these techniques do not satisfy the privacy requirement of our data integration application for the financial industry. A general overview of various research works on differential privacy can be found in the survey of Dwork [12].

Distributed interactive approach. This approach is also referred to as privacy preserving distributed data mining (PPDDM). In PPDDM, multiple data owners want to compute a function based on their inputs without sharing their data with others. This function can be as simple as a count query or as complex as a data mining task such as classification, clustering and so on. For example, multiple hospitals may want to build a data mining model for predicting disease based on patients’ medical history without sharing their data with each other. In recent years, different protocols have been proposed for different data mining tasks including association rule mining, clustering and classification.

None of these methods provide any privacy guarantee on the computed output. On the other hand, Dwork et al. and Narayan and Haeberlen [10] have proposed interactive algorithms to compute differentially private count queries from both horizontally and vertically partitioned data, respectively. When compared to an interactive approach, a non interactive approach gives greater flexibility because data recipients can perform their required analysis and data exploration, such as mining patterns in a specific group of records, visualizing the transactions containing a specific pattern, or trying different modeling methods and parameters. Distributed non interactive approach. This approach allows anonymizing data from different sources for data release without exposing the sensitive information.

Jurczyk and Xiong have proposed an algorithm to securely integrate horizontally partitioned data from multiple data owners without disclosing data from one party to another. Mohammed et al. [4] have proposed a distributed algorithm to integrate horizontally partitioned high dimensional health care data. Unlike the distributed anonymization problem for vertically partitioned data studied in this paper, these methods propose algorithms for horizontally partitioned data. Jiang and Clifton have proposed the Distributed k Anonymity (DkA) framework to securely integrate two data tables while satisfying the k anonymity requirement. Mohammed et al. [2] have proposed an efficient anonymization algorithm to integrate data from multiple data owners. To the best of our knowledge, these are the only two methods generate an integrated anonymous table for vertically partitioned data. Both methods adopt k-anonymity or its extensions as the underlying privacy principle and therefore, both are vulnerable to the recently discovered privacy attacks [5].

3. Problem Formulation

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. Accuracy-constrained privacy-preserving access control framework is used to manage access control in relational database. The access control policies define selection predicates available to roles while the privacy requirement is to satisfy the k-anonymity or l-diversity. Imprecision bound constraint is assigned for each selection predicate. k-anonymous Partitioning with Imprecision Bounds (k-PIB) is used to estimate accuracy and privacy constraints. Role-based Access Control (RBAC) allows defining permissions on objects based on roles in an organization.

Top Down Selection Mondrian (TDSM) algorithm is used for query workload-based anonymization. The Top Down Selection Mondrian (TDSM) algorithm is constructed using greedy heuristics and kd-tree model. Query cuts are selected with minimum bounds in Top-Down Heuristic 1 algorithm (TDH1). The query bounds are updated as the partitions are added to the output in Top-Down Heuristic 2 algorithm (TDH2). The cost of reduced precision in the query results is used in Top-Down Heuristic 3 algorithm (TDH3). Repartitioning algorithm is used to reduce the total imprecision for the queries. The following drawbacks are identified from the existing system. They are static data based access control model, cell level access control is not supported, imprecision bound estimation is not optimized and fixed access control policy model.

4. Role-Based Access Control (RBAC) Techniques

Role-Based Access Control (RBAC) is a promising access control technology for the modern computing environment. In RBAC permissions are associated with roles and users are assigned to appropriate roles thereby acquiring the roles' permissions. This greatly simplifies management. Roles are created for various job functions in an organization and users are assigned roles based on responsibilities and qualifications. Users can be easily reassigned from one role to another. Roles can be granted new permissions as new applications come on line and permissions can be revoked from roles as needed. Role-role relationships can be established to lay out broad policy objectives.

RBAC is policy neutral and flexible. The policy en-forced is a consequence of the detailed configuration of various RBAC components. RBAC allows a wide range of policies to be implemented. Administration of RBAC must be carefully controlled to ensure the policy does not drift away from its original objectives. In large systems the number of roles can be in the hundreds or thousands, users can be in the tens or hundreds of thousands and permissions in the millions. Managing these roles and users and their interrelationships is a formidable task that cannot realistically be centralized in a small team of security administrators. Decentralizing the details of RBAC administration without losing central control over broad policy is a challenging goal for system designers and architects [11]. There is tension here between the desire for scalability through decentralization and maintenance of tight control.

Since the main advantage of RBAC is to facilitate administration, it is natural to ask how RBAC itself can be used to manage RBAC. The use of RBAC for managing RBAC will be an important factor in its long-term success. There are many components to RBAC. RBAC administration is therefore multi- faceted. In particular we can separate the issues of assigning users to roles, assigning permissions to roles and assigning roles to roles to define a role hierarchy. These activities are all required to bring users and permissions together. In many cases, they are best done by different administrators or administrative roles. Assigning permissions to roles is typically the province of application administrators. Thus a banking application can be implemented so credit and debit operations are assigned to a teller role, whereas approval of a loan is assigned to a managerial role. Assignment of actual individuals to the teller and managerial roles is a personnel management function. Assigning roles to roles has aspects of user-role and permission-role administration. More generally, role-role relationships establish broad policy.

An administrative model called ARBAC97 was recently introduced by Sandhu et al. ARBAC97 has three components: URA97 is concerned with user-role administration; PRA97 is concerned with permission-role administration and is a dual of URA97 and RRA97 deals with role-role administration.

5. Privacy Preserved Access Control Model

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. Heuristic 1 and 2 have time complexity of $O(d|Q|^2 n^2)$. 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 (Rþ-tree) techniques.

5.1. 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. 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.

5.2. 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 Q_j P_i value from the imprecision bound BQ_j of each query, for a Partition, say P_i, that is being added to the output. For example, if a partition of size k has imprecision 5 and 10 for Queries Q_1 and Q_2 with imprecision bound 100 and 200, then the bounds are changed to 95 and 190, respectively. The best results are achieved if the kd-tree traversal is depth-first. 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.

5.3. Top-Down Heuristic 3 (TDH3)

The time complexity of the TDH2 algorithm is $O(d|Q|^2 n^2)$, which is not scalable for large data sets. In the Top-Down Heuristic 3 algorithm (TDH3), we modify TDH2 so that the time complexity of $O(d|Q|n\lg n)$ can be achieved at the cost of reduced precision in the query results. Given a partition, TDH3 checks the query cuts only for the query having the lowest imprecision bound. Also, the second constraint is that the query cuts are feasible only in the case when the size ratio of the resulting partitions is not highly skewed. We use a skew ratio of 1:99 for TDH3 as a threshold. If a query cut results in one partition having a size greater than hundred times the other, then that cut is ignored.

6. k-Anonymity Process

Given a data set T, T[c][r] refers to the value of column c, row r of T. T[C] refers to the projection of set of columns C on T and T[.][r] refers to selection of row r on T. Although there are many ways to generalize a given data value, in this paper, we stick to generalizations according to domain generalization hierarchies (DGH). We also abuse notation and write $\Delta^{-1}(v^*)$ to indicate the children of v^* at the leaf nodes. For example, given DGH structures Δ_1(USA) = AM,

Δ_{2} (Canada) =*; $\Delta_{0.1}$ (<M, USA>) = <M, AM>,
Δ(USA) = {USA, AM,*},Δ^{-1}(AM)
= {USA, Canada, Peru, Brazil}.

Since k-anonymity does not enforce constraints on the sensitive attributes, sensitive information disclosure is still possible in a k-anonymization. This problem has been enforcing diversity on sensitive attributes within a given equivalence class. It should be noted that even extensions to k-anonymity have vulnerabilities in the case of external knowledge. As our focus in this paper is the look-ahead process, we do not present further detail. For the sake of simplicity, from now on we assume data sets contain only QI attributes unless noted otherwise.

7. Dynamic Data and Policy based Access Control Scheme

The privacy preserved access control framework is enhanced to provide incremental mining features. Data insert, delete and update operations are connected with the partition management mechanism. Cell level access control is provided with differential privacy method. Dynamic role management model is integrated with the access control policy mechanism for query predicates. The cluster based access control system is designed with incremental mining mechanism. The system also provides cell level access control mechanism. The system uses the differential privacy to protect cell level access. The system is divided into six major modules. They are data preprocess, role management, query level analysis, clustering process, incremental mining and data retrieval process.

Data preprocess module is designed to perform noise elimination process. User level access permissions are assigned role management process. Query and associated data ranges are analyzed in query level analysis module. Data partitioning is performed in clustering process module. Incremental mining module is designed to modify the database transactions. Data retrieval module is designed to fetch data using query values.

7.1. Data Preprocess

Data populate process is performed to transfer textual data into relational database. Meta data provides the information about the database transactions. Data cleaning process is initiated to correct noisy transactions. Missing values are updated using aggregation based data substitution mechanism.

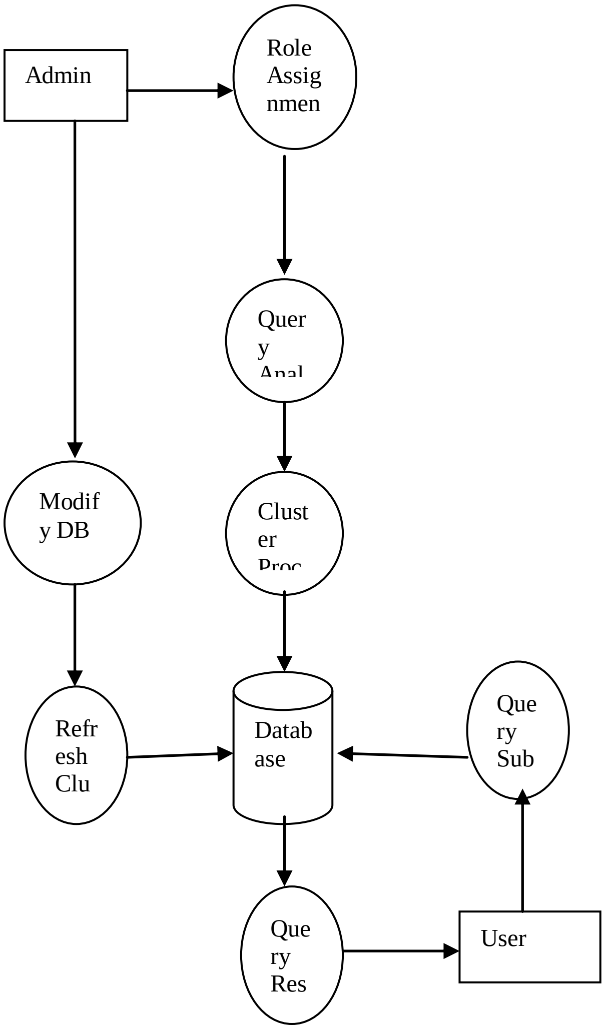

Fig. No: 7.1. Dynamic Data and Policy based Access Control Scheme

7.2. Role Management

User details and their access permissions are maintained in the role management process. Sensitive attributes selection is carried out to perform data anonymization process. Each user is assigned with different query values. The query values are used to manage the access permissions to the users.

7.3. Query Level Analysis

User query values are analyzed to estimate the data ranges. Data boundary for each query is estimated using Top-Down Heuristic 1 algorithm (TDH1). TDH2 algorithm is used to update the query bounds as initial partitions. Query results are verified with precision reduction level using TDH3 algorithm.

7.4. Clustering Process

Clustering process is applied to partition the transaction table with query results. TDH based partitioning algorithm is used to cluster the transaction data values. Data partitioning is performed on Anonymized data values. Data partitions are updated into the database.

7.5. Incremental Mining

Data insert, update and delete operations can be performed on the database tables. Tables are associated with the partitioned data values. Reclustering process is performed for the entire database transactions. Cluster refresh process is used to adjust the partitioned data values in incremental mining process.

7.6. Data Retrieval Process

Data retrieval process is carried out using user query values. User query and data retrieval rate are updated into the access logs. User data access is verified with imprecision bound levels. Cell level access control is provided in the query execution process.

8. Conclusion

Privacy preserved relational database access control model is upgraded with dynamic data policy management methods. Role Based Access Control (RBAC) scheme protects the sensitive data with minimum imprecision values. K-Anonymity model is integrated with minimum imprecision based data access control mechanism. Privacy preserved data access control mechanism is improved with incremental mining model and cell level access control. The system reduces the imprecision rate in query processing. Access control mechanism is adapted for incremental mining model. Time complexity is reduced in the system. The system provides the dynamic policy management mechanism.

REFERENCES

[1] Mohammed, Fung and P.S. Yu, "Differentially Private Data Release for Data Mining," Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining , 2011.

[2] N. Mohammed, B.C.M. Fung and M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants," Very Large Data Bases J., vol. 20, no. 4, pp. 567-588, Aug. 2011.

[3] A. H. Anderson. A comparison of two privacy policy languages: Epal and xacml. In *SWS '06: Proceedings of the 3rd ACM workshop on Secure web services,* pages 53–60, New York, NY, USA, 2006. ACM Press.

[4] N. Mohammed and C. Lee, "Centralized and Distributed Anonymization for High-Dimensional Healthcare Data," ACM Trans. Knowledge Discovery from Data, Oct. 2010.

[5] R.C.W. Wong, Y. Xu and P.S. Yu, "Can the Utility of Anonymized Data be used for Privacy Breaches?" ACM Trans. Knowledge Discovery from Data, vol. 5, no. 3, article 16, Aug. 2011.

[6] A. McGregor and S. Vadhan, "The Limits of Two-Party Differential Privacy," Proc. IEEE Symp. Foundations of Computer Science, 2010.

[7] B.C.M. Fung and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM , 2010.

[8] K. Chaudhuri and A. Sarwate, "Differentially Private Empirical Risk Minimization," J. Machine Learning Research, July 2011.

[9] K. Chaudhuri, A.D. Sarwate and K. Sinha, "Near-Optimal Differentially Private Principal Components," Proc. Conf. Neural Information Processing Systems, 2012.

[10] A. Narayan and A. Haeberlen, "DJoin: Differentially Private Join Queries over Distributed Databases," Proc. 10th USENIX Conf. Operating Systems Design and Implementation, 2012.

[11] Tamir Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases" IEEE Transactions on Knowledge and Data Engineering, April 2014

[12] C. Dwork, "A Firm Foundation for Private Data Analysis," Comm. ACM, 2011.