International Journal of Technology and Engineering System (IJTES)
Vol 8. No.1 – Jan-March 2016 Pp. 90-94

available at : www.ijcns.com
ISSN: 0976-1345

ANONYMIZATION TECHNIQUE FOR PRIVACY PROTECTION SHARING OF SENSITIVE DATA STREAM

G.Nivedha,
PG Scholar, Department of CSE,
E.G.S Pillay Engineering college,
Nagapattinam, India
nivethaudhai@gmail.com

Mr.M.Chinnadurai,
Associate professor, CSE
E.G.S Pillay Engineering College,
Nagapattinam, India.
mchinna81@gmail.com

ABSTRACT

The two mechanism have been proposed for data stream namely Access control mechanisms and Privacy Protection Mechanisms (PPM). The access control for a data stream allows restrict to authorized users using role based access to tuples satisfying an sliding-window query. It prevents the unauthorized access of user data. The PPM provides better privacy for the tricky information which is to be shared. The PPM meets privacy requirements, e.g., k-anonymity or l-diversity technique by generalization of stream data. Imprecision introduced by generalization can be reduced by delaying the proclaim of stream data. In order to attain better efficiency, the suspension in sharing the stream tuples can lead to false-negatives if the tuples are occupied by PPM while the query predicate is evaluated. Controller of an access control policy defines the fault constrained for each query. The objection for PPM is to upgrade the delay in publishing of stream data so that the defect bound for the maximum number of queries is satisfied. It proposes the fidelity-bounded access control for privacy-protection data streams problem, present the heuristics method for partitioning the data in an anonymization process,

Keywords*: L-diversity, K-anonymity, Suppression, privacy.*

1.INTRODUCTION

Data mining is the process of extracting data or information from large databases and also popularly known as Knowledge Discovery in Databases, refers to the nontrivial derivation of constant, already hidden and probably useful information from knowledge in databases. Although data mining and knowledge discovery in databases are periodically deal with synonyms, data mining is indeed factor of the knowledge discovery process.

In principle, data mining is not unique to one type data which can be any kind of information repository applicable to the level. Data

mining is being put into use for many types such as relational databases, data warehouses, transactional databases, World Wide Web, spatial databases, multimedia databases, time-series databases and textual databases. The rapid increase in database management systems has also supply to new extensive gathering of all sorts of information. Today, there is a need to handle more information such as business transactions and scientific data, to satellite pictures, text reports and military intelligence.

Information retrieval is the practice of recovering information relevant to the resource that can be used for later retrieve purposes. These needs are self-regulating declaration of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data.

With the massive amount of data stored in files, databases, and other repositories, it is progressively significant, to establish powerful means for analysis and perhaps interpretation of such data and for the separation of exotic knowledge that could help in decision-making.

II. REQUIREMENTS

The main objective of this project privacy protection mechanism applies generalization to the stream data such that the privacy requirement and imprecision bound for the maximum number of sliding-window queries is satisfied. Anonymization technique is used for privacy preserving of sharing data.

A release of data is said to have the *k*-anonymity property if the information for every person contained in the release cannot be differed from at least k-1 individuals whose information also appear in the release.In the condition of *k*-anonymization problems, a database contains table with *n* rows and *m* columns. Each row of the table represents a record relating to a particular member of a population and the entries in the various rows need not be unique. The values in the different columns are the values of attributes associated with the members of the population. The attributes are valid to an adversary are called "quasi-identifiers". Each "quasi-identifier" tuple appear in at least k records for a dataset with k anonymity.

There are two common methods for achieving *k*-anonymity for some value of *k.*

1. **Suppression**: In this method, some values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'.After applying this method it can replace all the values in the 'Name' and all the values in the 'Religion' indication have been replaced by the symbol'*'.
2. **Generalization**: In this approach, particular values of attributes are replaced by with a range category. For example, the value '15' of the attribute 'Age' may be restored by ' $\leq$ 30', the value '37' by '30 < Age $\leq$ 40' , etc.

L-diversity is a form of group based anonymization that is used to protectprivacy in data sets by minimizing the granularity of a data representation. This reduction is a trade off that ends in some loss of efficiency of data management or mining algorithms in order to rise some confidentiality. The l-diversity model is an expansion of the k-anonymity model which cut down the granularity of data representation adopting performance along with generalization and suppression alike any given record maps onto at least k other report in the data. The l-diversity model handles a bit of the weaknesses in the k-anonymity model where preserved identities to the level of k-individuals is not similar to protect the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group perform homogeneity.

III. PROPOSED SYSTEM

- Introduce the abstraction of precision-bounded access control for privacy-preserving data Streams. Formulate the Precision-bounded Access Control for

privacy- protection data streams (PACE) problem and give hardness results along with probabilistic analysis for query bound violation.

- Propose a heuristic for an approximate solution of the PACE problem and conduct empirical evaluation.

- Query evaluation semantics and give definitions for the imprecision, imprecision bound and average query bound violation (AQV).

- An access control policy administrator can use this analysis to revise the imprecision bounds for the queries if the probability of satisfying the bound for a large number of queries at any time instance is very low.

- The Privacy Protection Mechanism (PPM) ensures that the privacy and precision goals are met before the sensitive stream data is made available to the access control mechanism.

- The access control policy administrator defines sliding-window queries that define the authorized view of the data stream for each role.

- The PPM uses generalization of stream data tuples to anonymize and satisfies the given privacy requirement.

ADVANTAGES

- The generalized time-stamp value for specific Equivalence Class (EC) must be involved in the anonymized stream.

- The time-stamp value for a person in a relational stream data can allow to find the associated sensitive value

IV. ARCHITECTURE

INTRODUCTION

System design is the process of explaining the design,architecture, components, modules, and data for a system to satisfy specified requirements. It may be the application of systems theory to promote the outcome. There is some overlap with the disciplines of analysing the system, systems architecture and systems engineering. If the broader topic of product improvement fusion the perspective of retailing, layout, and manufacturing into a single access to product development, then layout is the action of taking the retailing intelligence and establishing the method of the product to be constructed. System design is therefore the process of describing and advancing systems to satisfy stated requirements of the user.

ALGORITHM

The top down heuristic algorithm is introduced for obtaining superior conclusion than the actual top down selection Mondrian algorithm. Consider a partition that extend along a query. If the median also decline inside the query then even after splitting the partition, the defect for that query will not change as both the new partitions still overlap the query. In TDSM, the partitions are division along the median, so it helps the median cut. In the heuristic method, the splitting of the partition is observed along the query cut. By using the query cut method it will choose the dimension where, the imprecision should be least possible for all the queries. If there present multiple queries that extend along a partition, then it has to choose the query for the query cut.

For this purpose the queries having the imprecision greater than zero will taken and sorted. The queries having the defect bound is small will taken because the queries with less imprecision bound. If the queries does not allows the query cut then it will uses the median cut. The feasible cut is the partition resulting from the split should perform the privacy requirements.

MODULE

MICRO-DATA CREATION MODULE / DATA COLLECTION MODULE

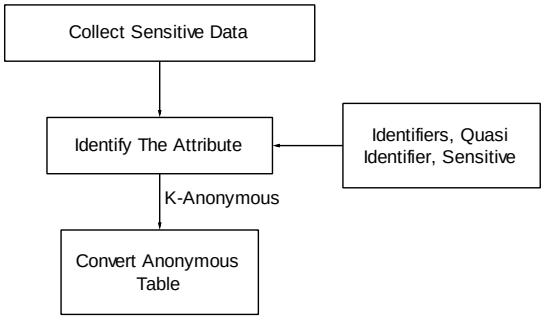
Collect patient micro data through any hospital or internet and import to our application. These data place in Identifiers, Quasi Identifiers, Sensitive Information; this is a normal table ex: Voter Table, Patient Table.

ANONYMIZATION CREATION MODULE

Medical records from a made up hospital located in upstate any country. Note that the table contains no uniquely analyzing attributes like name, social security number, etc.

Divide the attributes into two groups namely the *sensitive* attributes (consisting only of medical condition) and the *non-sensitive* attributes (zip code, age, and nationality). An attribute is noted sensitive if an adversary must not be owned to design the value of that attribute for any peculiar in the dataset.

Attributes not marked sensitive are non-sensitive. Furthermore, let the collection of attributes {zip code, age, nationality} be the quasi-identifier for this dataset. A 4-anonymous table derived from the (here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range [14850−14859] and "age=3*"means the age is in the range [30 − 39]). Note that in the 4-anonymous table, each tuple has the similar values for the quasi-identifier as at least three other tuples in the table.



DATA PARTITION MODULE

The Expected False- Positives for a leaf-node Partition P (EFPP , for short) is defined as the sum of false-positives for all queries resulting from Partition P, provided the partition is published at the current time instance. Sliding-window queries that add false-negatives are evaluated in the next time instance.

STREAM QUERY MODULE

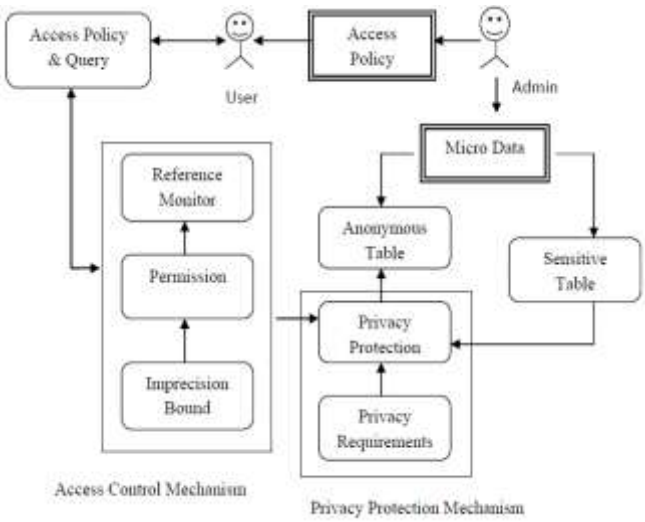
An access control policy is given in this module; that allows the roles to access the tuples under the authorized predicate, e.g., Role based CE1 can access tuples under Permission P1. The epidemiologists at the state and county level suggest community containment measures, According to the population 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. Predicate window queries have been proposed for streaming data management systems. The sliding window query is defined by two parameters: 1) Range that defines the size of query window and 2) Slide that defines the step by which the window moves. If the slide of the window is less than the range, then the query sliding windows overlap. Otherwise, if the slide is equal to the range, then the windows are non-overlapping and are also known as tumbling windows.

DATA STREAM MODULE

After successfully convert to anonymous table and also assign a role based access mechanism over internet application to publish the anonymous table to other research centre over the world. These data fully secured and publish to internet.

DATA ACCESSING MODULE

The research centre to access the anonymous data through our access keys or our SSN number over government server. These user already assign some roles in data access control mechanism module.



SYSTEM ARCHITECTURE

V. CONCLUSION

Precision-bounded access control for privacy preserving data streams has been proposed. The access control administrator defines the permitted view of the data stream along with the required precision. The privacy protection mechanism applies generalization to the stream data such that the privacy requirement is met and imprecision bound for the maximum number of sliding-window queries is satisfied. An algorithm has been proposed to minimize the total imprecision have been performed to compare the performance and plan to extend the access control enforcement to enclosed semantics and extend the differential privacy model for sliding-window queries over binary data streams.

REFERENCES

[1] Bache. K and Lichman. M, "UCI machine learning repository," School of Information and Computer Sciences, University of California, Irvine, 2013, http://archive.ics.uci.edu/ml.

[2] Buehler. J, Sonricker. A, Paladini. M, Soper. P, and Mostashari. F,"Syndromic surveillance practice in the united statesAdv. Disease Surveillance, vol. 6, no. 3,pp. 1–20, 2008.

[3] Cao. J, Carminati. B, Ferrari. E, and Tan. K, "Castle: Continuously anonymizing data streams," IEEE Trans. Dependable Secure Computing., vol. 8, no. 99, pp. 337–352, May/Jun. 2011.

[4] Carminati. B, Ferrari. E, Cao. J, and Tan. K, "A framework to enforce access control over data streams, "ACM Trans. Inf. System .Security, vol. 13, no. 3, p. 28, 2010.

[5] Golab. L and Ozsu .M , "Issues in data stream management," ACM Sigmod Rec., vol. 32, no. 2, pp. 5–1 2003 .

[6] LeFevre. K, DeWitt. D, and Ramakrishnan. R, "Workload-aware anonymization techniques for large-scale datasets," ACM Trans. Database Syst., vol. 33, no. 3, pp. 1–47, 2008.

[7] Machanavajjhala. A, Kifer. Gehrke. D,J, and Venkitasubramaniam. M, "l-diversity: Privacy beyond k-anonymity,"ACM Trans. Knowl. Discov. Data, vol. streams," IEEE Trans. Dependable Secure Comput.,vol. 8, no. 99, pp, May/Jun. 2011.

[8] Nehme R, Rundensteiner. E, and Bertino. E, " A security punctuation framework for enforcing access control on streaming data," in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 406–415.

[9] Ruggles. S, Alexander. J.T, Genadek. K, Goeken. R, Schroeder. M.B, and Sobek .M, "Integrated public use microdata series:Version 5.0 [machine-readable database]," Univ. Minnesota,Minneapolis, MN, USA, 2010.

[10] Pervaiz. Z, Aref. W.G, Ghafoor A, and Prabhu .N, "Accuracy constrained privacy-preserving access control mechanism forrelational data," IEEE Trans. Knowl. Data Eng., vol. 26, no. 4, pp. 795–807, Apr. 2014.