ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Accuracy Confined Access Control (Acac) For Privacy Preserving Data Streams

S.Kishore Verma[*1], A.Rajesh[*2], S.Anuradha[*3] and J.S.Adeline Johnsana[*4]
[*1]Computer Science and Engineering,SCSVMV University,Kanchipuram,India.
[*2, *3] Computer Science and Engineering,C.Abdul Hakeem College of Engineering andTechnology,Melvisharam,India.
[*4]Computer Science and Engineering,St.Peters University,Avadi,India .
***Corresponding author: E-Mail: kishore.saj3@gmail.com**

ABSTRACT

Privacy preserving data mining is a rising field of research which is concentrating on the two conflicting objectives (i) Utility and (ii) Privacy. In recent years many access control mechanisms and privacy protection mechanism have been proposed for the ceaseless data streams. Access control mechanisms for a data stream permits access to approve sliding window by just taking into account the consent utilized on each roles (role-based access control).Privacy protection mechanisms accomplishes privacy prerequisites by generalization of the stream data. Generalization produces imprecision results which can be reduced by delaying the publishing of stream data which in turn prompts to false-negatives. The challenge is to optimize the increase in imprecision due to the overlap of the rectangle bounds. To enforce accuracy confined access control for privacy preserving data stream an indexing mechanism of R*-tree is formulated that effectively reduces the imprecision and produces the optimized results.

KEY WORDS: Privacy, access control, data stream

1. INTRODUCTION

Transactional data is processed by Data Stream Management Systems (DSMS) e.g., web traffic, health monitoring and sensor networks. Access control mechanisms for data streams guarantee that just the approved parts of the stream are accessible to each user or role. Objects secured by access control mechanism are the queries of the data stream. If the delicate data in the approved perspective of the data streams is not security ensured then the protection of a individual can be traded off even in the vicinity of access control. Moreover, prior to distributing these authorization perspectives to the stakeholders privacy of patients must be secured. Most popular privacy preserving techniques such as k-anonymity and l-diversity have also been employed for securing ceaseless data streams. To the best of our insight precision bounded access control along with security has not been researched before for data streams.

Temporal queries are evaluated by using time stamp information. Records in the data stream can be generalized to preserve privacy requirements. This generalization introduces imprecision which can be reduced by delaying the publishing of stream data. It introduces false-negatives in the case that tuples are not made accessible in the access control mechanism at the time of querying.

The focus of this paper is to minimize the increase in imprecision due to duplication caused by the overlaps of the rectangle bounds in R+-tree (Sellis, 1987). To overcome the difficulty caused by the overlaps in R+-tree a new anonymization technique that uses R*-tree (1990) has been proposed. It has the following optimization criteria such as

(i)The area covered by a directory rectangle is minimized.
(ii)The overlap between directory rectangles is minimized which in turn reduces the duplication. This reduction will also reduce the increase in imprecision.
(iii)The margin of the rectangle is minimized.
(iv)Storage utilization is maximized since it avoids duplication.

Related Work: In this section, first we discuss literature related to the access control then the research related to the privacy preserving for the data streams is discussed. Nehme propose security punctuation-based access control framework for data streams (Nehme, 2008). Here the security restriction policies are streamed together called as the security punctuation and are interleaved in the data streams rather than storing permissions in the server. The roles are assigned permissions based on this security punctuation. To prevent unauthorized access, a general framework based on Role based access control (RBAC) is proposed by carminati (2010). It exploits a query rewriting mechanism. It rewrites the user queries according to the specified access control policies. It also provides two constraints. First, is the interval constraint in which the role can access the stream data and second, is the window constraint in which the access is limited by the sliding-window query predicate for each role. An accuracy-constrained privacy-preserving access control framework is proposed by Pervaiz (2014). 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). This framework also gives hardness results for the k-PIB problem and presents heuristics for partitioning the data to the satisfy the privacy constraints and the imprecision bounds. This framework concentrates on static data rather than stream data.

A cluster based scheme called Castle (2011) is proposed by Cao for Anonymizing continuous data streams. It is used to K-anonymize streams on the fly by using clustering algorithm and at the same time ensures the freshness of anonymized data by satisfying the delay constraints. Here the information loss does not consider the loss due to the delay. Zhou.at.al proposed a delay based anonymization to overcome this shortcoming (Zhou, 2009) in which the information loss will increase as the delay increase. They proposed continuous anonymization which is based on the R-tree in which the incoming tuples are added to the R-tree and the leaf nodes are published which is due at the particular time instance. Differential privacy is a recent privacy technique which is adapted to the problem of statistical disclosure control. Dwork (2010) have proposed differential privacy for data streams considering a single aggregate query. Issues are identified in the areas of k-anonymity and l-diversity of the syntactic anonymity. Differential privacy which is based on adding noise to the analysis outcome has been promoted to answer to privacy preserving data mining Clifton (2013). It looks at the issues and criticisms involved in both these approaches. These two approaches are sometimes perceived as competing approaches, and that one can be used instead of the other. This conception is wrong. We explained that the syntactic models are designed for privacy-preserving data publishing (PPDP) while differential privacy is typically applicable for privacy-preserving data mining (PPDM). Hence, one approach cannot replace the other, and they both have a place alongside the other. The concept of Precision bounded access control is proposed without the overlap of bounding rectangle by Pervaiz, (2015) In this paper, we propose an accuracy confined access control by considering overlapping of bounding rectangle for privacy preserving data streams.

Background: Given a data stream R[i]={ID,TS,A_1,A_2,…..,A_d, SA},Where ID is an identity attribute, TS is the time-stamp attribute that represents the advent time of the record, A_j is a Quasi-identifier (QI)attribute, SA is a sensitive attribute, d represents the number of QI attribute, and i is the current time stamp. R[i] represents all the data stream records that have advented up to the time instance i. The identity attribute (e.g., social security number) can remarkably recognize a person in a data stream. QI properties (e.g., address, age) can be utilized with the background information to recognize an individual even in the event that identity attribute has been suppressed. If the SA is connected with a one of a kind individual, it results in security infringement.

Privacy definitions: K-anonymity for streaming data has been proposed by Cao (2011) and Zhou (2009). They both have suppressed the TS attribute in the anonymized stream to assure privacy. To evaluate streaming data for sliding window queries TS attribute is required. The generalized time-stamp value should be incorporated for each equivalence class in the anonymized stream. The TS attribute is a QI attribute and knowing the TS value of an individual can allow to find its associated Sensitive attribute SA and can bring about infringement of security of that individual.

Definition 1 (Equivalence Class). An equivalence class is a set of records having the same Quasi-identifier attribute and time-stamp value.

Definition 2 (Stream k_s-anonymity Property (Zhou, 2009): A data stream $R^p[i]$ satisfies the k_s-anonymity if each published equivalence class has k or more records and if t_1.ID=t_2.ID then EC (t_1)≠EC(t_2) for any $t_1,t_2 \in R[i]$

Here $R^p[i]$ is the anonymized view of the stream data which is published up to the time instance i. In $R^p[i]$ the ID is suppressed, the QI and the TS values are generalized and the sensitive attribute is published. The time stamp attribute gives the advent time of a record $t \in R[i]$.The publishing delay is equal to t.PUB-t.TS where t.PUB is the time when the record is published. Records with the same identifiers must be in different equivalence classes.

Definition 3 (Delay Constraint): The maximum delay before which a record should be published is given by the delay constraint δ

The delay constraint can be set according to the requirement of the data stream application with respect to the availability of the anonymized records. $R^h[i]$ is the set of records that are put on hold and yet to be anonymized at time instance i.

Stream query model: The Sliding window query is characterized by two parameters (i) Range, which is characterized by the size of the querying window and (ii) Slide, which is characterized by the progression of the window movement. If the slide of the window is lesser than range then the overlap of sliding window occurs. Apart from that if the range and slide are equal then the non-overlapping of windows occur and it is called as tumbling window. Sliding window query can be record-count sliding window or time-sliding window. In this paper time-sliding window query is focused.

Role based access control:

Definition 4 (RBAC Policy): An RBAC policy is a record(U,R,P,UA,PA,RH),where U represents Users, R represents Roles, P represents Permissions, RH is a Role Hierarchy, UA represents user-to-role assignment and RA is role-to-permission assignment.

Accuracy Confined Access Control (Acac) For Privacy Preserving Data Streams

Predicate assessment and imprecision: Predicate Sliding-window query is assessed for a data stream R[i],by including all the stream records that fulfill the query predicate. For assessment over an anonymized data stream. $R^p[i]$ we include all the records in equivalence classes that overlap the query predicate range. This semantic of query evaluation is called as overlap semantics. Another possible semantics for query evaluation is to include all records in the equivalence classes that are fully enclosed inside the query predicate range these semantics is called as the enforced semantics.

Definition 5 False-Positive record: A record is a false-positive when it does not fulfill the sliding-window query predicate at the time instance of query evaluation but is included in the query result as the equivalence class in Rp[i] that contains the record overlaps the query predicate.

The number of False-Positive records in the result of a predicate sliding-window query, say $Q_j[i]$, at any time instance i, is as follows:

$FP_{Qj[i\ ]} = |Q_j(R^p[i])|-|Q_j(R[i]) - R^h[i])|;$

Where $|Q_j(R^p[i])|= \sum_{EC(overlaps)Q_j} |EC|$

Definition 6 False-Negative record: A record is a false-negative when it satisfies the predicate sliding-window query at the time instance of query evaluation but is not included in the query result due to being put on hold.

The number of False-Negative records for a query, say $Q_j[i]$, evaluated at time instance i, is as follows:

$FN_{Qj[i\ ]}=|Q_j(R^h[i])|$

Definition 7 Sliding-Window Query Imprecision: The imprecision for query $Q_j[i]$ evaluated at time instance i is denoted by $imp_{Qj}[i]$ and is equal to the sum of false-positives and false-negatives for a predicate sliding-window query evaluated on an anonymized stream $R^p[i]$.

$imp_{Qj}[i]=FP_{Qj}[i] + FN_{Qj}[i]$

Definition 8 Query Imprecision Limit: The query imprecision limit, denoted by $LQ_j[i]$, is the total imprecision acceptable to the access control mechanism when the sliding-window query predicate $Q_j[i]$ is evaluated at time instance i.

Definition 9 Average Query-Limit Violation: The Average Query-limit Violation for a query Q_j is the average number of times the query imprecision limit is violated over a given time period. In other words,

$AQV_{Qj} = V_{Qj}/N_{Qj}$

where N_{Qj} is the number of steps Query Q_j takes till the current time instance and V_{Qj} is the number of times the imprecision bound is violated for these steps.

Definition 10 Expected False-Positives: The Expected False-Positives for a leaf node of Partition P at the current time instance is given as sum of the false-positives for all queries resulting from Partition P.

$EFP_P = \sum_{Q_j \epsilon Q} |P - Q_j|$

Definition 11 Expected False-Negatives: The Expected False-Negatives for a leaf node of Partition P is given as the sum of the false-negatives for all the queries which is to be executed at the next time instance from partition P if the partition is held by the PPM at the current time instance.

$EFN_P = \sum_{Q_j \epsilon Q} | Q_j(P)|$

Accuracy confined access control (ACAC)

Definition 12 ACAC Problem: Given a data stream R[i],a set of predicate sliding-window queries Q, and privacy parameter k_s, the Accuracy-Confined Access Control for privacy- preserving data streams (ACAC) problem is to generate an anonymized stream $R^p[i]$ such that the sum of the average query bound violation for all queries $q \epsilon Q$ is minimized.

Accuracy Confined access control for privacy preserving data streams is proposed as shown in Fig.1. The privacy protection guarantees that the privacy and precision goals are met before the sensitive stream data are made available to the access control mechanism. The access control administrator provides sliding-window query that defines the approved view of the data stream. The PPM satisfies the privacy requirements by using the generalization of the stream data for anonymization of records. The anonymization is carried by the data stream anonymization algorithm. Generalization produces imprecision which can be decreased by delaying the publishing of the stream data. These delay can lead to false-negatives if some of the records were held by the PPM. These imprecision examination is done by the Optimized Imprecision Minimization (OIM) algorithm. The administrator provides the imprecision bound for each query. PPM must make sure that the sum of false-negatives and false-positives is less than the imprecision bound during the instance of query evaluation. Reference monitor is used to set the overlap semantics or enclosed semantics. False-positives due to generalization in overlap semantics implies that access is provided to unauthorized records. False-negatives in enclosed semantics although denies access to approved data will not violate the access control policy. In this paper our task is to reduce the overlap by R*-tree so that the imprecision gets reduced so we focus on the semantics called as the overlap semantics.

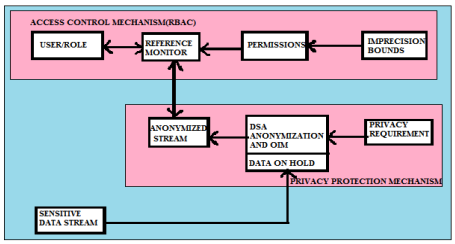


Figure.1. Accuracy Confined Access Control Framework

Algorithm for accuracy confined anonymization: Cao (2011) have proposed a clustering algorithm for anonymization of a data stream. Another approach proposed by Zhou (2009) uses an R-tree (Guttman, 1998) based algorithm to anonymize. The stream tuples are added to leaf nodes in an R-tree with a constraint that each node should have between k_s to $2k_s$ tuples. When a leaf node is published, that node is removed from the R-tree. The proposed heuristic listed in Algorithm 1 Data Stream Anonymization (DSA) can be applied to both techniques for a given predicate sliding-window query workload. We follow the approach suggested by Zhou. But use an R*-tree (1990) instead of an R-tree. R*-tree overcome the increase in imprecision that is caused by the overlap of the leaf nodes. The increase in imprecision is due to the repetition of the records in the leaf nodes of R+-tree caused by overlap. Total imprecision Minimization (TIM) is done without the repetition of the leaf nodes with non-overlapping rectangle by using R+-tree now we are considering the overlapping rectangles which reduces the duplication by different optimization criteria of R*-tree and proposing Algorithm 2 Optimized Imprecision Minimization (OIM). An R*-tree based index is maintained by the PPM in which the data stream records R[i] are added to the R*-tree at each time instance. Initially the R*-tree is empty (Iwuchukwu, 2007) and the ordered leaf nodes N are added to the partition of R*-tree in lines 1-8.If the leaf nodes N is greater than M which is the maximum number of entries in leaf node then forced reinsertion is set in line 9 since it is costlier. Only one reinsertion will be set in the same level otherwise split of the partition will take place. Due to this restructuring of nodes by forced reinsert, less number of split occurs and it also improves the storage utilization by avoiding duplication of leaf nodes. The anonymization range will be updated for quasi-identifier and the records to be held at the current time instance which is going to be added in the next time instance are given in lines 13-16

Algorithm 1. Data Stream Anonymization(DSA) algorithm

Input: a stream of records R[i], set of ordered leaf nodes N,parameter k
Output: Equivalence classes EC1,EC2..with respect to set of partitions S.
begin
1.Set the active R*-tree to an empty R*-tree
2.Initialization
3.While N not equal to empty.
4.P->empty partition
5.While $|P| \leq M$ (where M represents max.no.of entries in leaf node)
6.L->next leaf node in N
7.Add all records in L to P
8.N->N-L
9.If the total number of records in the remaining leaf nodes in N is less than M then remove those records from N and add them to P
else
For all the excess entries of the node N for the first call in the same level.
Compute the distance between the centers of their rectangles and the center of the bounding rectangle
Sort the entries in decreasing order
Remove the first entry and adjust the bounding rectangle.
else
remove the leaf node T,partition them into T1 and T2 and add them to P.
10.Update generalised quasi-identifier values for every record in P.
11.S->SUP
12.Continue step 3.
13.For each equivalence class EC that is due at the time instance i do
14.Examine and publish EC of the set of records to be held
15.remove and Re-insert all those records held into R*-tree
16.End for

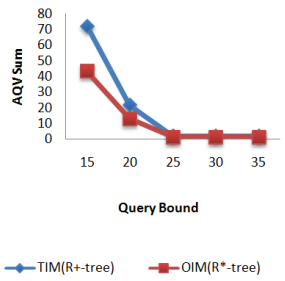
In algorithm 2 we made a call to the DSA algorithm so that the anonymization is carried out. A leaf node in R*-tree can have False-positive (if published) or False-negative(if held)toward sliding-window queries. The false-

positives represent the information loss obtained by generalization while false-negatives represent the information loss obtained by publishing delay. Therefore, we choose the option that contributes less imprecision for all the queries with respect to a partition. In other words, a leaf-node partition can be held in the active R*-tree until EFP_P is smaller than EFN_P . We also define w_{FP} and w_{FN} as weights where $0 < w_{FP}, w_{FN} \leq 1$. The weight assignment should be done according to requirements of the application. The leaf node is split if the size of new leaf nodes is greater than k_s. The leaf node is split along the median in the dimension having the least expected false-positives.The for loop in Line 5 checks all the leaf nodes of the active R*-tree. If the expected false-negatives by holding a leaf node are more than the false-positives as a result of publishing the node, then the node is published as an equivalence class.

Algorithm 2 Optimized Imprecision Minimization

Input: R[i], k_s, Q, and LQ_j
Output: EC1,EC2, . . .
Begin
1.Call DSA algorithm.
2. if there is a overlapping of two rectangle choose the rectangle which needs least enlargement and resolve the ties between them.
3. if (Size of new leaf nodes after splitting is > k_s) then
4.Split the leaf node;
5.for (all leaf nodes P in active R*-tree at time instant i) do
6.Update the imprecision cost of each leaf node;
7.if ((w_{FN} * EFN_P > EFP_P * w_{FP}) OR ((i – t_m.TS) ($^{\delta}$- 1)))
then
8.Publish the leaf node as EC and remove from active R*-tree;

Experimental results: The adult dataset from UC Irvine Machine learning repository (Bache, 2013) is considered having 30,163 records with attributes such as sex,age, race, marital-status, education, native country, work class, occupation, salary-class.For the k_s-anonymity experiment we use the age and salary-class as quasi-attributes. To model the data set we assume that 1000 tuples are received at each time instance. The maximum delay constraint $^{\delta}$ is set to five time units. It is assumed that the time interval between the two time instances is enough to update the R*-tree and the query imprecision at each time instance.100 queries are used as the workload. In this approach, two records are selected randomly from the record space and a query is formed by making a bounding box of these two tuples. The bounding box gives the predicates for the sliding-window query. The window size and step are also selected randomly from a fixed range. The range for the window size is 20-30 and for the step the range is 10-20. The imprecision bounds for all sliding-window queries are set based on the query size at the time of query evaluation. For the k_s-anonymity experiments, the value of k_s is fixed and the query imprecision bound is varied from 15 to 35 percent with increments of 5 and the sum of the average query-bound violation for all predicate sliding-window queries is evaluated. The results for k_s-anonymity are given in Fig.2 for the Adult dataset for k_s value of 3and Fig.3 for k_s value of 4.The OIM of R*-tree gives the better results when compared to the TIM (Precision, 2015) of R+-tree (Sellis, 1987) since sum of AQV produce the decrease in imprecision.



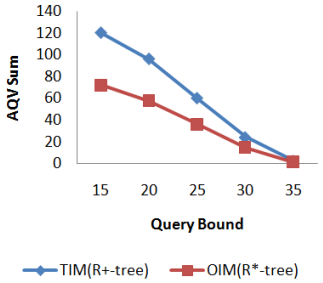


Figure.2.AQV sum when k_s=3 Figure.3.AQV sum when k_s=4

2. CONCLUSION

In this paper accuracy confined access control for privacy preserving data streams is proposed based on R*-tree in which the optimized reduction of imprecision is achieved. Though we get optimized results it seems to slightly costlier.Our future concentrates on the area of how we can reduce these cost and we plan to extend the differential privacy model for sliding –window queries over binary data streams to relational data streams.

REFERENCES

An R*-tree, An Efficient and robust access methods for points and rectangles Norbert Beckmann, Hans-peter kriegar, Ralf schneidar, Bernhard seegar, Praktische Informatik, Universitaet Berman, D-2800 Berman, 33,West Germany,1990.

Bache K and Lichman M, UCI machine learning repository, School of Information and Computer Sciences, University of California, Irvine, 2013.

Cao J, Carminati B, Ferrari E and Tan K, Castle, Continuously anonymizing data streams, IEEE Trans. Dependable Secure Comput, 8(99), 2011, 337–352.

Carminati B, Ferrari E, Cao J and Tan K, A framework to enforce access control over data streams, ACM Trans. Inf. Syst.Security, 13(3), 2010, 28.

Clifton C and Tassa T, On syntactic anonymity and differential privacy, in Proc. IEEE Int. Conf. Data Eng. Workshop Privacy- Preserving Data Publication Anal, 2013, 88–93.

Dwork C, Naor M, Pitassi T and Rothblum G.N, Differential privacy under continual observation, in Proc. 42nd ACM Symp.Theory Comput, 2010, 715–724.

Guttman A, R-trees, A dynamic index structure for spatial searching, in Readings in Database Systems, 3rd ed. Cambridge,MA, USA, MIT Press, 1998, 90–100.

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

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, 406–415.

Pervaiz Z, Aref W.G, Ghafoor A and Prabhu N, Accuracy constrained privacy-preserving access control mechanism for relational data, IEEE Trans. Knowl. Data Eng, 26(4), 2014, 795–807.

Precision-Bounded Access Control Using Sliding-Window Query Views For Privacy-Preserving Data Streams, Zahid Pervaiz, Arif Ghafoor, Fellow, IEEE and Walid G.Aref, Senior Member, IEEE, 2015

Sellis T, Roussopoulos N and Faloutsos C, The r+-tree, A dynamic index for multi-dimensional objects, in Proc. 13th Int.Conf. Very Large Data Bases, 1987, 507–518.

Zhou B, Han Y, Pei J, Jiang B, Tao Y and Jia Y, Continuous privacy preserving publishing of data streams, in Proc. 12th Int.Conf. Extending Database Technol, Adv. Database Technol, 2009, 648–659.