International Journal On Engineering Technology and Sciences – IJETS™
ISSN (P): 2349-3968, ISSN (O): 2349-3976
Volume 1 Issue 6, October 2014

Privacy and Security Ensured Database Rights Management Scheme

Ms. S.Kokila
II Year M.E(CSE)
Shree Venkateshwara Hi-Tech Engg College, Gobi
kokiyellowever@gmail.com

Dr. T. Senthil Prakash
Professor & HOD
Shree Venkateshwara Hi-Tech Engg College, Gobi
jtyesp@yahoo.co.in

Ms. P.Maheswari
II Year M.E(CSE)
Shree Venkateshwara Hi-Tech Engg College, Gobi
mahe11191@gmail.com

ABSTRACT- *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. 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. 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 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.*

Keywords: *k-anonymity, l-diversity, Imprecision bounds, Access control*

1 INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation.

Privacy preserving data mining is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy preserving data mining is twofold. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem.

2 RELATED WORK

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

constraints that the privacy protection mechanism is required to meet along with the privacy requirements.

The challenges of privacy-aware access control are similar to the problem of workload-aware anonymization. In our analysis of the related work, we focus on query-aware anonymization. For the state of the art in k-anonymity techniques and algorithms, we refer the reader to a recent survey paper [3]. Workload-aware anonymization is first studied by LeFevre et al. [5] They have proposed the Selection Mondrian algorithm [4], which is a modification to the greedy multidimensional partitioning algorithm Mondrian. In their algorithm, based on the given query-workload, the greedy splitting heuristic minimizes the sum of imprecision for all queries. Iwuchukwu and Naughton have proposed an Rþ-tree based anonymization algorithm. The authors illustrate by experiments that anonymized data using biased Rþ-tree based on the given query workload is more accurate for those queries than for an unbiased algorithm. Ghinita et al. have proposed algorithms based on space filling curves for k-anonymity and l-diversity [10]. They also introduce the problem of accuracy-constrained anonymization for a given bound of acceptable information loss for each equivalence class [8]. Similarly, Xiao et al. [9] propose to add noise to queries according to the size of the queries in a given workload to satisfy differential privacy. Bounds for query imprecision have not been considered. The existing literature on workload-aware anonymization has a focus to minimize the overall imprecision for a given set of queries. Anonymization with imprecision constraints for individual queries has not been studied before. We follow the imprecision definition of LeFevre et al.and introduce the constraint of imprecision bound for each query in a given query workload.

3 PRIVACY-PRESERVING ACCESS CONTROL MODEL FOR RELATIONAL DATA

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

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

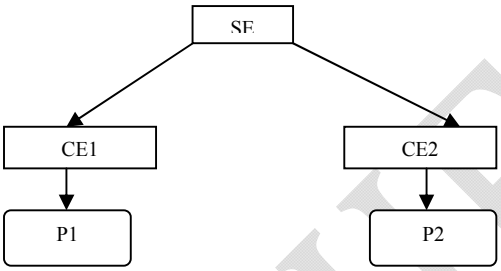
Example 1 (Motivating Scenario). Syndromic surveillance systems are used at the state and federal levels to detect and monitor threats to public health [7]. The department of health in a state collects the emergency symptoms, etc from county hospitals daily. Generally, each daily update consists of a 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 that allows the roles to access the tuples under the authorized predicate, e.g., Role CE1 can access tuples under Permission P1. 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.

The contributions of the paper are as follows. First, we formulate the accuracy and privacy constraints as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB) and give hardness results. Second, we introduce the concept of accuracy-constrained privacy-preserving access control for relational data. Third, we propose heuristics to approximate the solution of the k-PIB problem and conduct empirical evaluation.

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



Permission	Authorized Query Predicate
P1	Location =Country1^ Age=15-65 ^ Syndrome= Influenza
P2	Location =Country2^ Age=15-65 ^ Syndrome= Influenza

Fig. No: 3.1. Access Control Policy.

4 PROBLEM STATEMENT

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 problems are identified from the existing system.

- Static data based access control model
- Cell level access control is not supported
- Imprecision bound estimation is not optimized
- Fixed access control policy model

5 DATA PARTITIONING FOR PRIVACY PRESERVATION

In this section, three algorithms based on greedy heuristics are proposed. All three algorithms are based on kd-tree construction. Starting with the whole tuple space the nodes in the kd-tree are recursively divided till the partition size is between k and 2k. The leaf nodes of the kd-tree are the output partitions that are mapped to equivalence classes [1]. Heuristic 1 and 2 have time complexity of $O(d|Q|^2 n^2)$. 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 [2]. 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.

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

Input : T,K,Q and BQ_j

Output : P

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

Algorithm 1: TDH1

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 Qj Pi value from the imprecision bound BQj of each query, for a Partition, say Pi, that is being added to the output. For example, if a partition of size k has imprecision 5 and 10 for Queries Q_1 and Q2 with imprecision bound 100 and 200, then the bounds are changed to 95 and 190, respectively. The best results are achieved if the kd-tree traversal is depth-first (preorder). Preorder traversal for the kd-tree ensures that a given partition is recursively split till the leaf node is reached. Then, the query bounds are updated. Initially, this approach favors queries with smaller bounds. As more partitions are added to the output, all the queries are treated fairly. During the query bound update, if the imprecision bound for any query gets violated, then that query is put on low priority by replacing the query bound by the query size. The intuition behind this decision is that whatever future partition splits TDH2 makes, the query bound for this query cannot be satisfied. Hence, the focus should be on the remaining queries.

Input : T,K,Q and BQj

Output : P

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

Algorithm 2: TDH2

The algorithm for TDH2 is listed in Algorithm 2. There are two differences compared to TDH1. First, the kd-tree traversal for loop in Lines 2-

International Journal On Engineering Technology and Sciences – IJETS™
ISSN (P): 2349-3968, ISSN (O): 2349-3976
Volume 1 Issue 6, October 2014

14 is preorder. Second, in Line 14, the query bounds are updated as the partitions are being added to the output (P). The time complexity of TDH2 is $O(d|Q|^2 n^2)$, which is the same as that of TDH1. In Section 5.3, we propose changes to TDH2 that reduce the time complexity at the cost of increased query imprecision.

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 (greater than 10 million tuples). In the Top-Down Heuristic 3 algorithm (TDH3, for short), 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. TDH3 algorithm is listed in Algorithm 3. In Line 4 of Algorithm 3, we use only one query for the candidate cut. In Line 6, the partition size ratio condition needs to be satisfied for a feasible cut. If a feasible query cut is not found, then the partition is split along the median as in Line 11.

Input : T, K, Q and BQ_j

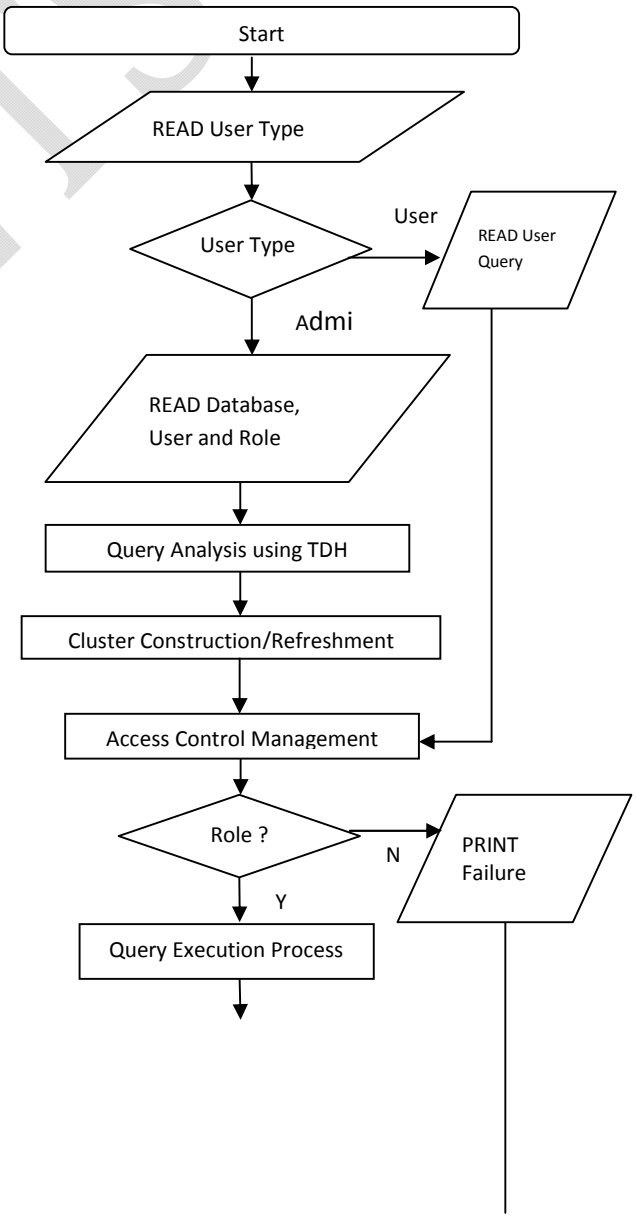
Output : P

1 Initialize set of candidate partitions ($CP \leftarrow T$)
2 for ($CP_i \in CP$) do //Depth first preorder traversal
3 Find the set of queries QO that overlap CP_i Such that $ic_{CPi}^{QOj} > 0$
4 Select query from QO with smallest BQ_j
5 Create query cuts in each dimension
6 Reject cuts with skewed partitions
7 Select dimension and cut having least overall Imprecision for all querie in Q
8 if (feasible cut found) then
9 Create new partitions and add to CP
10 else
11 Split CP_i recursively along median till anonymity requirement is satisfied
12 Compact new partitions and add to P
13 Update BQj according to ic_{pi}^{Qj}, $\forall$ Qj $\in$ Q
14 return (P)

Algorithm 3: TDH3

6 DATABASE RIGHTS MANAGEMENT WITH INCREMENTAL MINING MECHANISM

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.



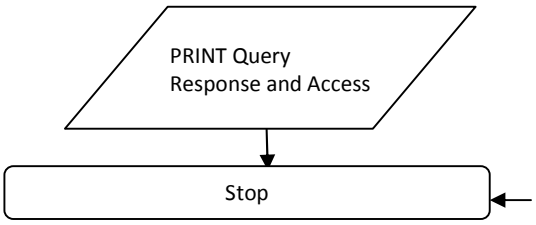


Fig. No: 6.1. Privacy Preserved Database Rights Management Scheme

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.

6.1. 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.

6.2 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.

6.3 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.

6.4 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.

6.5 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.

6.6 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.

7 CONCLUSION

Access control mechanism for relational data is constructed with the privacy preservation based model. 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] Noman Mohammed and Mourad Debbabi, "Secure Two-Party Differentially Private Data Release for Vertically Partitioned Data" IEEE Transactions On Dependable And Secure Computing, Vol. 11, No. 1, January/February 2014

[2] Russell Paulet, Md. Xun Yi and Elisa Bertino, "Privacy-Preserving and Content-Protecting Location Based Queries" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014

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

[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] N. Li, W. Qardaji, and D. Su, "Provably Private Data Anonymization: Or, k-Anonymity Meets

Differential Privacy," Arxiv preprint arXiv:1101.2604, 2011.

[7] J. Buehler, A. Sonricker 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] G. Ghinita, P. Karras and N. Mamoulis, "A Framework for Efficient Data Anonymization Under Privacy and Accuracy Constraints," ACM Trans. Database Systems, vol. 34, no. 2, article 9, 2009.

[9] X. Xiao, G. Bender, M. Hay, and J. Gehrke, "Ireduct: Differential Privacy with Reduced Relative Errors," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2011.

[10] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast Data Anonymization with Low Information Loss," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 758-769, 2007.

AUTHORS BIOGRAPHY

Ms.S.Kokila pursuing M.E(CSE) degree in Shree Venkateshwara Hi-Tech Engineering College, Erode, India in 2014 and B.E(CSE) degree from Velalar college of Engineering and Technology, Erode, India in 2012. She has published 1 National conferences, 4 workshops. She is a Member of Computer Society of India(CSI). Her research interests include Data mining, Databases.

Dr.T.Senthil Prakash received the Ph.D. degree from the PRIST University, Thanjavur, India in 2013 and M.E(CSE) degree from Vinayaka Mission's University, Salem , India in 2007 and M.Phil.,MCA.,B.Sc(CS) degrees from Bharathiyar University, Coimbatore India, in 2000,2003 and 2006 respectively, all in Computer Science and Engineering. He is a Member in ISTE New Delhi, India, IAENG, Hong Kong, IACSIT, Singapore SDIWC, USA. He has the experience in Teaching of 10+Years and in Industry 2 Years. Now He is currently working as a Professor and Head of the Department of Computer Science and Engineering in Shree Venkateshwara Hi-Tech Engineering College, Gobi, Tamil Nadu, and India. His research interests include Data Mining, Data Bases, Artificial Intelligence, Software Engineering etc.,He has published several papers in 17 International Journals, 43 International and National Conferences.

Ms.P.Maheswari pursuing M.E(CSE) degree in Shree Venkateshwara Hi-Tech Engineering College, Erode, India in 2014 and B.E(CSE) degree from Kathir college of Engineering, Coimbatore, India in 2013. She has published 3 National conferences, 6 workshops. She is a Member of Computer Society of India(CSI). Her research interests include Data mining, Databases.