International Journal of Engineering Research
Volume No.5, Issue No.9, pp :775-777
ISSN:2319-6890)(online),2347-5013(print)
1 September 2016

Efficient Approach for Anonymizing Tree Structured Dataset Using Improved Greedy Search Algorithm

Ruchira Warekar [1] , Savitri patil [2]
Department of Computer Engineering, GHRCEM, Wagholi, Pune, India
ruchirawarekar@gmail.com[1] , savithri3010@gmail.com[2]

Abstract: ***As increasing companies, the collection of storing personal information is needed, arises problem for maintaining security. Data anonymization techniques use to provide security for personnel information, as like generalization; bucketization doesn't ensure the privacy preservation of data collection. In this paper we describe privacy preservation methods for data integration.***

Keywords: Privacy, anonymity, security, integrity.

Introduction

Increasing population and users of technology results in collection of personal data in companies and organization. Data anonymization technique can provide security privacy to managing of personal information without losing personal identity. But these techniques imposed some issues described in existing work. When more powerful guaranty is applied it strictly asks for utility requirements which cannot meet.

In this paper, we are trying to overcome these limitations of greedy search algorithm by using automatic improved greedy algorithm. This will provide more accuracy and use automatic methods to find out correct results. We are presenting k(m;n)-anonymity privacy guarantee which addresses value and structure using improved and automatic greedy algorithm.

Existing system

Existing system addressing the complication of anonymizing tree structured data in the presence of structural knowledge. They propose k(m;n) anonymity privacy assurance. It present an anonymization algorithm which is able to generate k(m;n)-anonymous datasets, by employing value generalization and a novel data transformation, which we term structural disassociation.

Problem statement

In companies and organizations gathering of personal statistics is must. This is the serious problem of maintaining the privacy of this personal information. Data anonymization techniques proposed recently in order to deliver security to personal data of users. Below four are recent research problems in this domain to accomplish: a) in many cases there are utility requirements that cannot be met when more powerful assurance are applied, b) there is often inability to characterize attributes as sensitive or non sensitive, c) the privacy security law in most countries usually focuses on identity, and d) recent methods apply the greedy algorithm for establishing anonymization large scale tree structured dataset.

Data Anonymization

It is the process of de-identifying sensitive data while preserving its format and data type. Technically, data masking refers to a technique that replaces the data with a special character whereas data anonymization constitutes hiding of data and this would imply replacement of the original data value with a value preserving the format and type. It reduces the risk of identity disclosure whereas the data remains still realistic

k-anonymity

K-anonymity requires each tuple in the published table to be indistinguishable from at least k-1 other tuples. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given record cannot be distinguished from at least (k-1) other records. k-anonymity cannot provide a safeguard against attributes disclosure.

l-diversity

It is used to overcome drawback of k-anonymity and tries to put constraints on minimum number of distinct values seen within an equivalence class for any sensitive attributes.

Classification

Classification is used to predict the value of a certain attributes for future records, where the values are not known. This is usually done with a decision tree, where each node in the tree filters the records it receives into two or more distinct partitions based on their value for an attribute. This partition can then be split into more partitions until a termination requirement is met.

I. METHODOLOGY

A. Proposed idea

To overcome form the existing system problems proposed system provides techniques to provide privacy security to the organizational personal information of users. Where we are applying data anonymization techniques to hide the personal information from anonymous. It provide privacy guaranty without losing person's identity. By applying improved automatic greedy search techniques proposed system publish the data from the data sets.

B. Architecture

Goal of the development of this system is to provide security to the personal data of the organisational user from the anonymous data serchers which can help to provide secured and anonymous data channel for organisational users.

ISSN:2319-6890)(online),2347-5013(print)
1 September 2016

It increses the privacy policy of the proposed system and also data generation speed as well.

The proposed system takes the input as a releaseble datasets then it applies anonymization methos like generalisation to view the data in structuralal tree data set. Finally we are applying improved greedy cut serch algorith with l-diversity for automation and currect result generation.

Improved greedy cut serch algorithm with l-diversity proveds more efficient way for imformation retraction without compromoising anonymous identity.

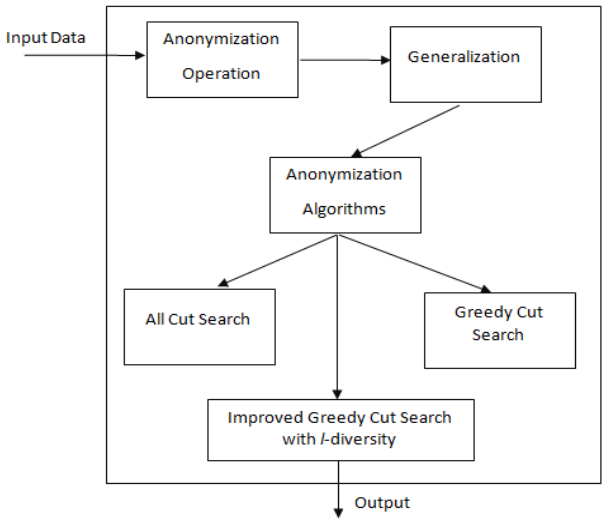

Figure 1: System architecture

C. WORKING

Anonymization operation with using k-anonymity dataset gets anonymized. Generalization operation generalized the dataset in normal form. Greedy search algorithm is used for automatic data mining.

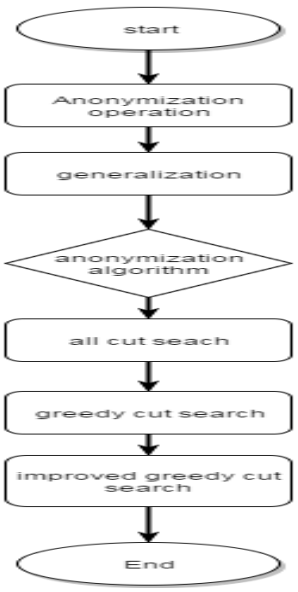

Figure 2: process of data anonymization

C. Algorithm

Algorithm: L-Diversity
Input: A releasable dataset *D*.
Output:
A releasable dataset $Tn*$, which ensures that each equivalence class has the same sensitive attribute values set before and after update and has minimal information loss.

1. Go to step 5 if the number of sensitive attribute values in $\Delta Tn-1$ is less than l.
2. $Tn* = Tn-1*$.
3. Merge the independent l-diverse equivalence classes generated from $\Delta Tn-1$ with $Tn*$.
4. Remove the corresponding records from$\Delta Tn-1$.
5. For each record r in $\Delta Tn-1$
6. Generate the candidate equivalence classes Cr in $Tn*$ according to its sensitive attribute value;
7. Insert the record r into a selected candidate equivalence class, which results the minimal information loss;
8. $\Delta Tn-1 = \Delta Tn-1 - r$.
9. For each equivalence class whose size is more than$2l-1$and each sensitive attribute value exists at least two times.
10. Divide the equivalence class if no inference channels are generated.
11. Return $Tn*$.

This l-diversity algorithm provides the automatic result generation. ℓ-diversity is used to overcome the limitation of k anonymity and tries to include constraints on minimum number of discrete values seen within any correspondence class for any sensitive attribute.
The l-diversity Principle: An equivalence class is said to have l-diversity if there are at least l "well -represented" values for the sensitive attribute. A table is said to have l -diversity if every equivalence class of the table has l-diversity. The given data set is said to be l-diversified if every correspondence classes in the table contains at least l well represented sensitive attribute values. l-diversity must assurance that the SA value of a particular person cannot be identified unless the opponent has enough background knowledge to eliminate l−1 SA values in the person's Equivalence Class.

III. RESULT ANANLYSIS

Age	Sex	Zip code	Disease
22	M	47906	Dyspepsia
22	F	47906	Flu
22	F	47905	Flu
33	F	47905	Bronchitis
54	M	47302	Flu
60	M	47302	Dyspepsia
64	F	47304	gastritis
54	M	47304	gastritis

Table 1: Original data

ISSN:2319-6890)(online),2347-5013(print)
1 September 2016

The above table shows original data of the DB, which contains sensitive attributes, quasi identifiers, non-sensitive data.

Quasi -identifier			Sensitive data
Age	Sex	Zip code	Disease
22	M	47906	Dyspepsia
22	F	47906	Flu
22	F	47905	Flu
33	F	47905	Bronchitis
54	M	47302	Flu
60	M	47302	Dyspepsia
64	F	47304	gastritis
54	M	47304	gastritis

Table 2: classified data

The above table 2 shows the classified table of the original data.

Age	Sex	Zip code	Disease
20-52	*	4790*	Dyspepsia
20-52	*	4790*	Flu
20-52	*	4790*	Flu
20-52	*	4790*	Bronchitis
54-64	*	4730*	Flu
54-64	*	4730*	Dyspepsia
54-64	*	4730*	gastritis
54-64	*	4730*	gastritis

Table 3: 2-anonymized dataset

SCOPE OF WORK

- Some methods focusing to protect identity or other properties of single node which resulted into non-scalable method.
- Some methods provide privacy but data loss resulted.
- Greedy algorithm cannot work for some anonymization problems correctly.
- Greedy algorithm required manual working to achieve the correct results.

System efficeincy

Existing system does not provide privacy guaranty for sensitive data from the data set. The proposed system can solve this limitation with providing security with automatic correctness.

IV. CONCLUSION AND FUTURE WORK

Increasing the data collection for the organization. So, maintaining this information will arises multiple security issues. Some methods focusing to protect identity or other properties of single node which resulted into non-scalable method.Greedy algorithm cannot work for some anonymization problems correctly.

ACKNOWLEDGMENT

I would wish to thank all the people who showered their kind support required for the whole analysis. The authors also want to thank the anonymous reviewers for their useful and constructive comments.

References

i.G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving Anonymity via Clustering. In PODS, 2006.

ii.G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation Algorithms for k-Anonymity. Journal of Privacy Technology, 2005

iii.R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k- Anonymization. In ICDE, pages 217–228, 2005.

iv.R. Chaytor and K. Wang. Small-domain randomization: Same privacy more utility. In VLDB, 2010.

v.J. Cao and P. Karras. Publishing microdata with a robust privacy guarantee. PVLDB, 5(11):1388–1399, 2012.

vi.J.Cheng, A.W.-c. Fu, and J. Liu. K-isomorphism: privacy preserving network publication against structural attacks. In SIGMOD, 2010.

vii.R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. PVLDB, 4(11):1087–1098, 2011

viii.C. Clifton and T. Tassa. On syntactic anonymity and differential privacy. In PRIVDB, 2013.

ix.G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In SIGKDD, pages 1253–1261, 2011.

x.R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. PVLDB, 4(11):1087–1098, 2011

Ruchira Warekar. Ruchira Warekar received her bachelor's degree in information technology from A. G. Patil institute of technology, Solapur university. She is currently working towards master degree at Savitribai phule university, pune. She is teaching assistant at department of computer engineering, Government polytechnic, Mumbai, India

Prof. Savitri patil. Savitri patil received her bachelors degree in computer science and engg. from hirasugar institute of technology, nidasoshi, visvesvarya technological university, Belgaum, india. She received M.Tech degree in computer science and Engg. From basaveshwar engineering college, bagalkot, visvesvarya technological university, Belgaum, India. She is working as assistant professor in GHRCEM college, Wagholi, Pune, India. Her research are as are biometrics, image processing and pattern recognition.