ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

A New Approach For Collaborative Data Publishing Using Slicing And M-Privacy

Sathish.R[1], Silambarashi.G[2], Saranya.P[3], Santhosh Kumar.B[4]

Student, Department of Computer Science and Engineering, SNS college of technology, Coimbatore[1, 2, 3]

Assistant Professor, Department of Computer Science and Engineering, SNS college of technology, Coimbatore[4]

Abstract: We introduce the notion of m-privacy, which guarantees that the anonymized data satisfies a given privacy constraint against any group of up to m colluding data providers. Second, we present heuristic algorithms exploiting the equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy given a set of records. We present a data provider-aware anonymization algorithm with adaptive m- privacy checking strategies to ensure high utility and m-privacy of anonymized data with efficiency.We introduce a novel data anonymization technique called slicing to improve the current state of the art. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization.

Keywords: Monotonicity, Anonymized, Tuples, Bucketization

I. INTRODUCTION

DATA mining is widely used by researchers for science and business purposes. Data collected from individuals are important for decision making or pattern recognition. Privacy-preserving processes have been developed to sanitize private information from the samples while keeping their utility. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying

Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes.

II. SYSTEM ANALYSIS

Exiting System

The collaborative data publishing problem for anonymizing horizontally partitioned data at multiple data providers a new type of "insider attack" by colluding data providers who may use their own data records (a subset of the overall data) in addition to the external background knowledge to infer the data records contributed by other data providers. First, the notion of m-privacy, which guarantees that the anonymized data satisfies a given privacy constraint against any group of up to m colluding data providers. Second, heuristic algorithms exploiting the equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy given a set of records. Finally, a data provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonymized data with efficiency. Experiments on real-life datasets suggest that our approach achieves better or comparable utility and efficiency than existing and baseline algorithms while providing m-privacy guarantee.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Proposed System

We introduce a novel data anonymization technique called slicing to improve the current state of the art. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying.

III. DESCRIPTION OF PAPER

Overview of the Paper

The m-privacy verification algorithm, we can now use it in anonymization of a horizontally distributed dataset to achieve m-privacy. In this section, we will present a baseline algorithm, and then our approach that utilizes a data provider-aware algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy for anonymized data. The algorithm first generates all possible splitting points, π, for QI attributes and data providers . In addition to the multidimensional QI domain space, we consider the data provider or data source of each record as an additional attribute of each record, denoted as A0 .Introducing this additional attribute in our multi-dimensional space adds a new dimension for partitioning. This leads to more splits resulting a more precise view of the data and have a direct impact on the anonymized data utility. To find the potential split point along this dimension, we can impose a total order on the providers. We monitor that this multi set-based generalization is the same to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but split the association between them within one potential split point. We observe that while one-attribute-per-column slicing preserves attribute distributional information, it does destroy attribute correlation, for the reason that each attribute is in its own column. In slicing, one groups associated attributes together in one column and save their correlation. For instance, in the sliced table shown in Table correlations between Age and Sex and correlations between Zip code and Disease are conserved. In fact, the sliced table encodes the same amount of information as the original data with observe to correlations between attributes in the same column.

IV. MODULE DESCRIPTION

Original Data

We conduct the extensive workload experiments. Our results are confirm that slicing preserves much better data utility than the generalization. In these involving the sensitive attribute, slicing is also more effectual than bucketization. In some classification experiments, the novel technique i.e. slicing shows better performance than the original

Generalization And Suppression

A value is replaced by a less specific, more general value that is faithful to the original. In Figure the original ZIP codes {02138, 02139} can be generalized to 0213*,thereby stripping the rightmost digit and semantically indicating a larger geographical area.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

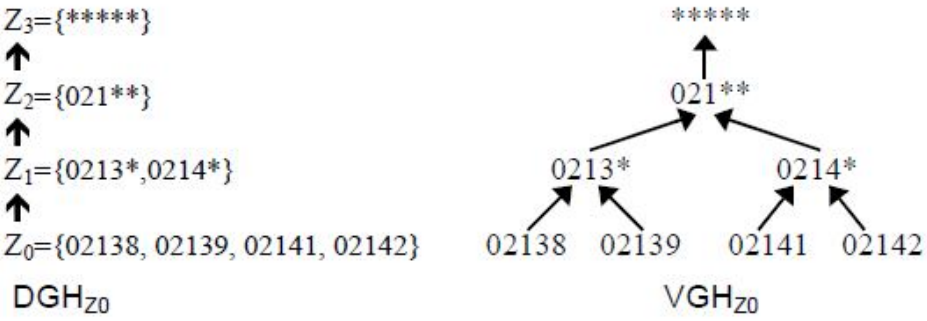
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014



In a classical relational database system, domains are used to describe the set of values that attributes assume. For example, there might be a ZIP domain, a number domain and a string domain. I extend this notion of a domain to make it easier to describe how to
generalize the values of an attribute. In the original database, where every value is as specific as possible, every attribute is considered to be in a ground domain.

TABLE 2: The Generalized Table

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*	Dyspepsia
[54-64]	*	4730*	gastritis

Bucketized Data:

In that we show the efficiency of slicing in membership disclosure protection. For this purpose, we calculate the number of fake tuples in the sliced data. Also we compare the number of matching buckets for original tuples and that for fake tuples. Our experimental results illustrate that bucketization does not prevent membership disclosure as almost every tuple is distinctively identifiable in the bucketized data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

TABLE 3: The Bucketized Table

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

Multiset-Based Generalization Data

We monitor that this multi set-based generalization is the same to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but split the association between them withinone bucket.

TABLE 4: Multiset-Based Generalization

Age	Sex	Zip code	Disease
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	Dysp.
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	Flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	Flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	Bron.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	Flu
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	Dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	Dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	Gast.

One-Attribute-Per-Column Slicing Data:

We observe that while one-attribute-per-column slicing preserves attribute distributional information, it does destroy attribute correlation, for the reason that each attribute is in its own column. In slicing, one groups associated attributes together in one column and save their correlation. For instance, in the sliced table shown in Table correlations between Age and Sex and correlations between Zip code and Disease are conserved. In fact, the sliced table encodes the same amount of information as the original data with observe to correlations between attributes in the same column.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

TABLE 5: One-Attribute-Per-Column Slicing

Age	Sex	Zip code	Disease
22	F	47906	Flu
22	M	47905	Flu
33	F	47906	Dysp.
52	F	47905	Bron.
54	M	47302	Dysp.
60	F	47304	Gast.
60	M	47302	Dysp.
64	M	47304	Flu

Sliced Data

Another important advantage of slicing is its capability to handle high-dimensional data. By dividing attributes into columns, slicing condense the measurements of the data. Each of which column of the table can be viewed as a sub-table with a lesser dimensionality. Slicing is also not similar from the approach of publishing multiple independent sub-tables in that these sub-tables are associated by the buckets in slicing

TABLE 6: The Sliced Table

(Age,Sex)	(Zipcode,Disease)
(22,M)	(47905,Flu)
(22,F)	(47906,dysp.)
(33,F)	(47905,bron.)
(52,F)	(47906,flu)
(54,M)	(47304,gast.)
(60,M)	(47302,Flu)
(60,M)	(47302,dysp.)
(64,F)	(47304,dysp.)

V. CONCULTION AND FUTURE WORK

A novel data anonymization technique called slicing is used to improve the current state of the art. Slicing partitions the data set both vertically and horizontally. Multi-dimensional space adds a new dimension for partitioning. There is no future work in this paper if any method is introduced for privacy preserving then this paper should be implemented in that new method.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

REFERENCES

[1] DATA MINING Concepts , Tasks and Techniques Author-S N Sivanandam , S Sumathi

[2] http://www.slideshare.net/akannshat/data-mining-15329899

[3] Closeness: A New Privacy Measure for Data PublishingNinghui Li, Tiancheng Li, and Suresh VenkatasubramanianIEEE TRANSACTIONS ON KNOWLEDGE & DATA ENGINEERING, VOL. 22, NO. 7, JULY 2011

[4] B.santhosh kumar ,"PRIVACY PRESERVING DATA PUBLISHING FOR MULTIPLE SENSITIVE ATTRIBUTES ,"

[5] Slawomir Goryczka, Li Xiong, and Benjamin C. M. Fung, "m-Privacy for Collaborative Data Publishing,"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2013

[6] G. Cormode, D. Srivastava, N. Li, and T. Li, "Minimizing minimalityand maximizing utility: alyzing method-based attacks on anonymized data,"Sept. 2010

[7] Pui K. Fong and Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, FEBRUARY 2012

[8] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang, "Enabling Multilevel Trust in Privacy Preserving Data Mining," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,SEPTEMBER2012

[9] Tiancheng Li, Ninghui Li,Slicing: A New Approach for Privacy Preserving Data Publishing IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3,MARCH 2012

BIBLIOGRAPHY

Guide:
B.Santhosh Kumar
Department of Computer Science and Engineering
SNS college,of technology
Coimbatore

SATHISH.R
Department of Computer Science and Engineering
SNS college,of technology
Coimbatore
Anna University(2011-2014)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

SILAMBARASHI.G
Department of Computer Science and Engineering
SNS college,of technology
Anna University(2010-2014)
Coimbatore

SARANYA.P
Department of Computer Science and Engineering
SNS college,of technology
Anna University(2010-2014)
Coimbatore