ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

Enhanced Slicing Technique for Improving Accuracy in Crowdsourcing Database

T.Malathi[1], S. Nandagopal[2]

PG Scholar, Department of Computer Science and Engineering, Nandha College of Technology, Erode, Tamilnadu, India [1]

Professor & Head, Department of Information Technology, Nandha College of Technology, Erode, Tamilnadu, India[2]

Abstract – In recent years, privacy preserving has seen rapid growth which leads to an increase in the capability to store and retrieve personal dataset without revealing sensitive information about the individuals. Different techniques have been proposed to improve accuracy in crowdsourcing database. Anonymization techniques such as, generalization and bucketization, are designed for improving accuracy in privacy preserving method. But the malicious workers can hack the private information of the user and misuse it. Recent work has been shown that k-anonymity for generalization losses considerable amount of information especially for higher dimensionality data. l-diversity for bucketization does not able to prevent membership disclosure. In this paper we introduce a novel technique called overlapped slicing, which partitions the data in both horizontal and vertical manner. Slicing preserves better data utility than generalization and bucketization techniques. As an extension we proposed a technique called overlapped slicing, in which an attribute is divided into more than one column. The release in each column consists of more attribute correlations. Important advantage of this work is to handle high-dimensional data and also preserves better privacy than the previous techniques.

Index terms – crowdsourcing, k-anonymity, l-diversity, generalization, bucketization.

I. INTRODUCTION

Privacy is one of the most important properties that an information system must satisfy, with rapid development of Internet technology privacy preserving [2] data publication became one of the most important research topics and also a serious concern in publication of personal data in recent years. Data mining is the process of extracting knowledge from large amount of database popularly known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of previously unknown and potentially useful information from databases.

Crowdsourcing database [13] is process of getting work from a group of people based on the user requirement. Each query from the user is allotted to the workers [9] and results related to the concern event are returned to database. Operators process each relevant answer from the workers and publish the records for further processing in the database. Some of the current crowdsourcing platforms are Amazon AMT and Crowd flower [1] [9] which adopts a labor as a service model. Each database may consist of not only the records but also personal details such as salary, educational qualification, working experience etc., which can be revealed by an adversary who can retrieve individual's sensitive information. The main part of operators is to collect the answers related to the queries [5] which are published in earlier. For example consider human resource agents receive thousands of applications from both the users and companies. Here the users submit the curriculum vitae and companies with their job positions. These records are processed by an operator and queries with an equivalent answers are returned to the users.

In order to provide security for each individual data user's formal protection generalization: model k-anonymity [14] used. A data release which cannot be distinguished from at least k-1 data individual. In order to preserve the privacy the attributes each record in the k-anonymity database consists of three types of attributes, identifiers, quasi identifiers and sensitive attributes [3]:

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

1. Identifiers (ID): Identifiers are attributes that clearly identify individuals.
   Examples: Social Security Number and Name.

2. Quasi-identifiers (QI): Quasi-identifiers are attributes that the values are taken together can identify an individual.
   Examples: Zip-code, Date of Birth, and Gender.
   An attacker may already know the QI values of some individuals in the data. The knowledge can be either from personal contact information or from other publicly available databases.

3. Sensitive Attributes (SAs): values should not be associated with an individual by an attacker.
   Examples: Disease, salary.

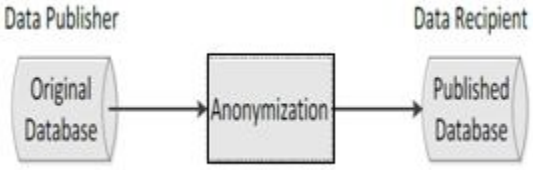

Fig.1.privacy preserving data model

In both generalization and bucketization one first removes the identifiers and partitions the tuples into buckets. The generalization mechanism produces a release candidate by generalizing (coarsening) some attribute values in the original table. The need of anonymization is privacy preserving data model is represented in Fig.1.The basic idea is that, after generalizing some attribute values, some records would become identical when projected on the set of quasi-identifier (QI) attributes (e.g., age, gender, zip code).Each group of records that have identical QI attribute values is called an equivalence class. Bucketization mechanisms do not have a clear separation between quasi identifiers and sensitive attributes. Several anonymization techniques are used such as generalization for k-anonymity [6] and l- diversity for bucketization [4].

II. EXISTING SYSTEM

Anonymization Methods:

Two different anonymization methods [3] are generalization and bucketization. The main advantage of bucketization over generalization is that bucketization does not generalize the QI attributes.

Generalization

Generalization [2][3] is the process of replacing the information with semantically consistent value. It replaces quasi-identifier values with values that are less-specific but semantically consistent. Due to the high-dimensionality of the quasi-identifier, with different possible items in thousands of order, that any generalization method will cause high information loss and also rendering the data in useless. In order to improve generalization in efficient manner, records in the same bucket must be similar to each other so that during generalization the records would not lose too much information.

Three categories of generalization:

1. Global recoding- Values are generalized to the same level of the hierarchy. In other words all values of an attribute come from the same domain level in the hierarchy. It has the property that multiple occurrences of the same value are always replaced by the same generalized value.

2. Regional recoding- It is also known as multidimensional recoding which partitions the domain space into noninterest regions and data points in the same region are represented by the region they are in. it allows different values of an attribute to be generalized to different levels.

3. Local recoding- Does not have the above constraints and allows different occurrences of the same value to be generalized differently. It allows the same value to be generalized to different values in different records. Here there are different values mappings can be chosen across different anonymized groups.

Limitation of Generalization

Our intention is not to eliminate generalization and there is no doubt that generalization is an important technique, partly proved by the fact that it has received much attention in the literature. An alternative approach has advantages, since it can retain a larger amount of

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

data characteristics .Two main problems of generalization are:

1. Fails on high-dimensional data due to the curse of dimensionality
2. Too much information loss due to uniform-distribution.

Bucketization

Bucketization [11][14], is the process of partitioning the tuples in the table into buckets and then to separate the sensitive attribute from the non-sensitive attributes by randomly permuting the sensitive attribute values within each bucket. Sanitized data set consists of the buckets with permuted sensitive attributes. Partitioning the tuples into buckets and within each bucket, here we apply an independent random permutation to each column. Then the resulting bucket is published. Bucketization preserves better data utility than generalization.

Limitations of Bucketization

1. Does not prevent membership disclosure.
2. Requires a clear separation between QIs and SAs.
3. Breaks the attribute correlations between the QIs and the SAs by separating the SA from the QI attributes.

At first bucketization partitions tuples in the table into buckets and then separate quasi identifiers with sensitive attribute by randomly permuting the sensitive attribute values in each bucket in the table. The anonymized data set consist of a set of buckets with permuted sensitive attribute values in given record. It is important to note that bucketization has been used for anonymizing high-dimensional data. The given approach assumes a clear separation between QIs and SAs, because the exact values of all QIs are released and membership information is disclosed in the data set.

Privacy Preserving Techniques

There are two types of privacy preserving techniques used in existing method. They are k-anonymity and l-diversity.

K-anonymity

Anonymity is used to prevent identification of individual records in the given data set. The database is said to be K-anonymous where attributes are generalized until each row is identical with at least k-1 rows. K-Anonymity guarantees that the data released is accurate. The two different techniques used by k-anonymity [14] method are, generalization and suppression. To protect respondent's identity when releasing data, data operators often remove explicit identifiers like names, social security numbers etc. One of the interesting aspects of k-anonymity is its association with protection techniques that preserve originality of the data in each record. Basic approach towards privacy protection in data mining has to perturb the data before it is mined.

The guarantee given by k-anonymity [11] is that no information can be linked to groups of less than k individuals. In Generalization for k-anonymity which losses considerable amount of information, for higher-dimensionality data set. K-anonymity model for multiple sensitive attributes consist of three kinds of information disclosure:

1. Identity Disclosure: An individual who can link to a particular record in the published data set is known as identity disclosure.
2. Attribute Disclosure: When the sensitive information regarding particular individual revealed is known as attribute disclosure.
3. Membership Disclosure: The information regarding individual belongs from data set is presented or not revealed is a membership disclosure.

Anonymity refers to a state where data does not show its identity. A dataset which satisfies k-anonymity if every record in the dataset is not distinguished from at least k-1 other records with respect to every set of quasi-identifier attributes is known as k-anonymity dataset.

Limitations

1. Does not be able to hide whether a given individual is in the database.
2. Reveals individuals' sensitive attributes.
3. The attack based on background knowledge is not prevented.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

4. Cannot be applied to high-dimensional data without data loss.
5. Different methods are required for a dataset which is anonymized and published more than once.

Attacks on k-anonymity are homogeneity attack and background knowledge attack.
Homogeneity Attack: Sensitive information in the dataset may be revealed based on the known information, if the non sensitive information of an individual is revealed to an adversary. The method of information revealing is known as positive disclosure.

Background Knowledge Attack: If the user has some external information that can be linked to the released data which helps in neglecting some of the sensitive attributes, which may reveal personal information of an individual. The method of information revealing is known as negative disclosure.

l-diversity

This prevents the association of an individual record with sensitive attribute value. ℓ-diversity [4][8] is a distribution of a sensitive attribute in each equivalence class which has at least l "well represented" values to protect against attribute disclosure. The distribution of target values within a group is referred to as "ℓ-diversity".

From the limitations of k-anonymity, ℓ-diversity for privacy preserving provides that data publisher who does not know what kind of knowledge is possessed by an attacker. The important concept of ℓ-diversity is based on requirement is that values of the sensitive attributes are well-represented in each group. In l-diversity group of k different records share a particular quasi-identifier so that an adversary cannot identify the individual based on the QI attribute [6]. The distribution of target values within a group is referred to be known as ℓ-diversity. This principle represents an important step beyond k-anonymity in protecting against attribute disclosure.

Limitations

1. May be difficult to achieve.
2. Insufficient to prevent the attribute disclosure.
3. Does not consider overall distribution of sensitive values.
4. Semantic meanings of sensitive values are not considered.
5. Not be able to prevent probabilistic attack.

Some of the attacks by which limitation occurs are:
Skewness Attack: When the given overall distribution is skewed satisfying the l-diversity does not prevent attribute disclosure. Similarity Attack: When the sensitive values in a QI group are distinct but semantically similar, an adversary can able to learn information.

III. MOTIVATION OF SLICING

It has been shown that generalization for k-anonymity [3] [4] losses considerable amount of information in the data set, especially for higher dimensionality data. This is due to the following reasons:

1. Generalization for k-anonymity suffers from the curse of dimensionality. For generalization to be efficient, records in the same bucket must be close to each other so that generalizing the records would not lose too much information
2. Generalization significantly reduces the data utility of each generalized data.
3. Because of each attribute is generalized separately correlations between different attributes are lost.

While bucketization has better data utility than generalization, it has several limitations. Bucketization does not able to prevent membership disclosure in data set. Because bucketization publishes the quasi identifiers values in their original forms, so attacker can find out whether an individual has a record in the published data or not.

In this paper, we present a new data anonymization technique called slicing [7][8] to improve the current state of the art. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is grouping of attributes into columns based on the correlations among the attributes. Each column within the original table contains a subset of attributes which are highly correlated with each other. Horizontal partitioning is grouping of tuples into buckets. At final step, values in

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

each column are randomly permutated within each bucket to break the linking between different columns.

The basic idea of slicing [12] is to break the association cross columns, it preserves the association within each column. It reduces the dimensionality of the data and also preserves better utility than generalization and bucketization. Slicing preserves better utility because it groups highly correlated attributes together, and preserves the correlations between attributes. Slicing protects privacy because it breaks the associations between each uncorrelated attributes in the column, which are infrequent and thus identifying. The key intuition that slicing provides privacy protection is that the slicing process ensures that for any tuple, there are some multiple matching buckets. Given a tuple: t =<v1, v2. . . vc> where c is the number of columns and vi is the value for the ith column. A bucket is a matching for t if and only if for each i (1<= i<= c), vi appears at least once in the ith column of the bucket. It also contains, a matching bucket can be due to containing other tuples each of which contains some but not all vi's.

IV. PROPOSED METHOD

Overlapped Slicing

In this section, we present a new enhanced slicing technique called overlapped slicing for privacy-preserving data publishing.
Slicing has several advantages:

1. It preserves better data utility than generalization.
2. It preserves more attribute correlations with the SAs than bucketization.
3. It can also handle high-dimensional data and data without a clear separation of QIs and SAs.

The algorithm partitions attributes into columns, applies column generalization, and partitions tuples into the buckets. Attributes which are highly correlated are in the same column preserves the correlations between such attributes. Here each association between uncorrelated attribute is broken. It provides better privacy as the associations between such attributes are less frequent and potentially identifying.

The architectural diagram of the proposed work is shown in Fig.2.Generally in privacy preservation do not provide efficient security for each data in its original form. The privacy protection is not effective due to the presence of the adversary's background knowledge in real life application. Data in its original set contains sensitive information about each individual. When the data sets are published it will cause violation in privacy. So this kind of approach may lead to insufficient protection. The main idea of slicing is to break the contribution of cross columns and also conserve within each column. Many algorithms like bucketization and generalization have been tried to preserve better privacy but they exhibit attribute disclosure. To overcome the problem slicing technique is used. Lessens the dimensionality of data and conserves better utility than the generalization and bucketization.

Architecture design

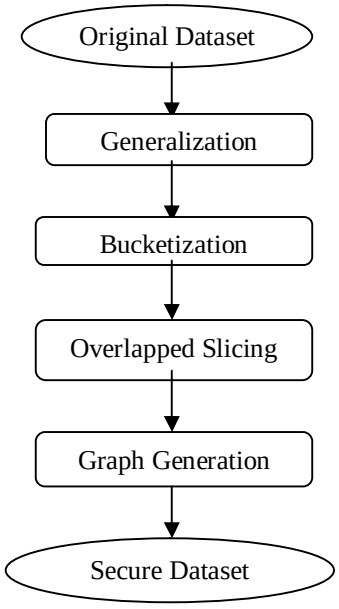

Fig.2.Architectural Diagram

Overlapped Slicing partitions the data set both vertically and horizontally. Vertical partitioning can be done by grouping attributes into columns based on the correlations among the attributes. Horizontal partitioning can be done by grouping tuples into buckets. Slicing method consists of three phases. Which are explained in the following:

1. Attribute Partitioning
2. Column Generalization

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

3. Tuple Partitioning

1. Attribute Partitioning

Algorithm partitions attributes so that highly correlated attributes are in the same column. It is good for both data utility and data privacy. In terms of data utility, grouping highly correlated attributes preserves the correlations among those attributes. In privacy preserving method, association of uncorrelated attributes presents higher identification risks than association of highly correlated attributes due to the associations of uncorrelated attribute values which are less frequent and also more identifiable.

2. Column Generalization

Column generalization is required for identity or membership disclosure protection. If a column value is unique, then a tuple with in this unique column value can have only one matching bucket. It will not be efficient for privacy preserving, in the case of generalization and bucketization each tuple will belong to only one equivalence bucket.

3. Tuple Partitioning

The algorithm maintains two data structures: a queue of buckets Q and a set of sliced buckets SB. At first stage Q contains only one bucket which includes all tuples and sliced bucket with empty value. In each step the algorithm removes a bucket from Q and splits the bucket into two buckets .If the sliced table after splitting satisfies l-diversity technique, then algorithm provides two buckets at the end of the queue Q. Otherwise, we cannot be able to split the bucket further. Then algorithm puts the bucket into set of sliced buckets.

Slicing with Tuple grouping algorithm provides efficient random tuple grouping for micro data publishing. Each column contains sliced bucket (SB) that permutated random values for each partitioned data. It is also permutated the frequency of the value in each one of the diversity algorithm checks the diversity when the each sliced table.

V. CONCLUSION

This paper presents a new approach called overlapped Slicing a new approach for data anonymization. Overlapped slicing overcomes the limitations of generalization and bucketization and which also preserves better data utility while protecting against privacy threats. In this paper we describe about how to use overlapped slicing to prevent attribute disclosure and membership disclosure. The general methodology proposed by this work is before anonymization of each data one can analyze the characteristics of the data and use these characteristics in data anonymization technique. The main reason is that we can design better data anonymization techniques when we know that given data is in a better manner. This work will motivate several directions for future research. In this paper, we consider overlapped slicing, which reduplicates an attribute in more than one column. The release in each column consists of more attribute correlations.

REFERENCES

[1]Sai Wu, Xiaoli Wang, Shen Wang, Zhenjie Zhang and Anthony K.H. Tung, "K-Anonymity for crowdsourcing database" 2013.

[2]R. J. B. Jr. and R. Agrawal, "Data privacy through optimal k-anonymization", in ICDE, 2005.

[3]Sweeney.L, "Achieving k-anonymity for privacy protection using generalization and suppression," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, 2002.

[4]Machanavajjhala. A,Kifer.D, Venkitasubramaniam.M, Gehrke.J "l-Diversity: Privacy Beyond k-Anonymity," ACM Transactions Knowledge Discovery Data, volume 1, issue 1, March 2007.

[5]Feng.A, Franklin.M, Kossmann.D, Kraska.T, Madden.S, Ramesh.S, Wang.A, and Xin.R, "CrowdDB: Query Processing with the VLDB Crowd," PVLDB, volume 4, issue 12, pp. 1387-1390, 2011.

[6] Li. T and Li. N, "On the Tradeoff between Privacy and Utility in Data Publishing," In KDD, pp. 517-526, 2009.

[7] Mohanapriya.D, Dr.Meyyappan.T, "Slicing Technique for Privacy Preserving Data Publishing," International Journal of Computer Trends and Technology (IJCTT) volume 4, Issue 5, May 2013.

[8] Kiruthika.S and MohamedRaseen.M, "Suppression Slicing–using l-diversity," IJCA Proceedings on Amrita International Conference of Women in Computing, pp. 1-6, January 2013.

[9] Sorokin.A and Forsyth.D, "Utility data annotation with Amazon Mechanical Turk," in First IEEE Workshop on Internet Vision at CVPR, 2008.

[10] Marcus.A, Wu.E, Madden.S, and Miller.R.C, "Crowdsourced Databases: Query Processing with People," in CIDR, pp. 211-214, 2011.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

[11] K. LeFevre, R. Ramakrishnan, and D. J. DeWitt "Mondrian multidimensional k-anonymity", in ICDE, 2006.

[12]Tiancheng Li, Ninghui Li, Jian Zhang, Lan Molloy, "Slicing: A new approach to privacy preserving data publishing".120-150, 2012.

[13] X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu and M. Zhang, "Cdas: A crowdsourcing data analytics system," vol. 5, no. 10, pp. 1040–1051,2012.

[14] Sweeney.L, "k- Anonymity: A Model for Protecting Privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, volume 10, issue 5, pp.557-570, 2002.