IJRRECS/January 2014/Volume-2/Issue-1/1905-1910 **ISSN 2321-5461**

INTERNATIONAL JOURNAL OF REVIEWS ON RECENT ELECTRONICS AND COMPUTER SCIENCE

AN APPROACH TOWARDS SAFE MAINTENANCE OF DATA PUBLISHING

A.Nagaraja[1], A.Pravin[2]

[1]M.E Student, Dept of CSE, Sathayabama University, Chennai, T.N, India

[2]Associate Professor, Dept of CSE, Sathayabama University, Chennai, T.N, India

ABSTRACT:

An environment that assists large-scale data mining and data analysis is the collection of digital information. The accessibility of high quality data and helpful information sharing depends on data mining and the main obstacle to the advancement of technology is the lack of trust in data mining. The techniques of privacy-preserving are often required to reduce the possibility of identifying sensitive information about individuals, when a data set is released to other parties for data analysis. A novel technique of data anonymization known as slicing was introduced to get better the existing state of art. Slicing protects confidentiality because it breaks the associations between unconnected attributes, which are infrequent and thus identifying.

Keywords: Data Analysis, Privacy-preserving, Slicing, Anonymization.

1. INTRODUCTION:

Encryption intends to prevent an unofficial party from accessing the data, but enable an authorized party to have full access to the data. In privacy preserving data publishing, it is the authorized party who may also play the role of the adversary with the goal of inferring sensitive information from the data received. Thus, encryption may not be directly applicable to some privacy preserving data publishing problems. Many techniques are proposed for protecting individual privacy and sensitive information in order to avoid the obstacles. To confine the types of publishable information in addition to agreements on the applying as

well as storage of sensitive information the current privacy protection practice primarily depends up on the policies and guidelines [4]. The constraint of this approach is that the data is distorted extremely or it requires a trust level that is impractically high in many data-sharing scenarios. The privacy concern related to the input of data mining methods is addressed by the Confidence bounding which is the first contribution, but the output of data mining methods could also cause confidentiality threats [8]. Encryption is another commonly employed technique for confidentiality protection. Even though an encrypted record communicates to a real life patient, the encryption hides the semantics required for acting on the corresponding patient. Even though it can be used to assume perceptive properties about record holders the output is an aggregate pattern, it is not intended to identify a record holder [1]. Preservation of the data intended for a required data analysis in addition to limit the usefulness of unnecessary inferences that may possibly be resulting from the data release. The accessibility of high quality data and eventual information sharing depends on data mining [12]. Intended for publishing information in an additional hostile atmosphere a task of the utmost importance is to develop methods as well as tools with the intention that the published data remains almost functional while confidentiality of individual is preserved.

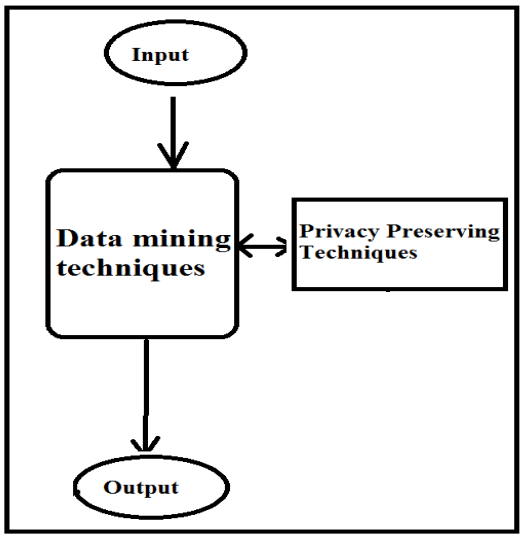

Fig 1: An overview of Privacy preserving data publishing

2. METHODOLOGY:

In some data publishing scenarios, it is important that each published record corresponds to an existing individual in real life. Privacy preserving in data distribution has become one of the most significant research topics in the field of data safety and it has turn out to be a severe unease in publication of personal data in recent years. The public has an increased sense of privacy invasion due to the increased level of

security after revolutionary attacks [3]. The data publisher collects data from record owners in the data collection phase. The data publisher releases the collected data to a data miner or to the public in the data publishing phase, called the data recipient, who will then conduct data mining on the published data. The data publisher is trustworthy and record owners are willing to provide their personal information to the data publisher in the trusted model; however, to the data recipient the trust is not transitive [14]. Every data publishing scenario shown in fig1 has its own assumptions and requirements of the data publisher, the data recipients, and the data publishing purpose. A novel technique of data anonymization known as slicing was introduced to get better the existing state of art. Slicing protects confidentiality because it breaks the associations between unconnected attributes, which are infrequent and thus identifying. The data set is partitioned both vertically and horizontally by slicing [2] [9]. Preservation of the data intended for a required data analysis in addition to limit the usefulness of unnecessary inferences that may possibly be resulting from the data release. To preserve the association within each column the essential thought of slicing is to break the connection cross columns. Different columns within each bucket, the values within every column are at random permutated. Based on the associations between the attributes, the partitioning of vertical is completed by means of attributes grouping into columns. By grouping tuples into buckets horizontal partitioning is done. Finally, to break the linking between each column is randomly permutated. Slicing which does not have horizontal partitioning is the first marginal publication which can be viewed as a special case [7]. Therefore, correlations among attributes in different columns are lost in marginal publication. The data preserves better utility than generalization and bucketization and reduces the dimensionality. Generalization and bucketization are two popular anonymization techniques. For generalization various schemes of encoding have been introduced such as: regional recoding, universal recoding and local recoding [15]. Global recoding contains the assets of abundant occurrences of the similar value are constantly replaced through the similar comprehensive value. The main problems with generalization are: First is due to the curse of dimensionality it fails on

high-dimensional data. Second is, due to the uniform-distribution assumption it causes too much information loss. With the responsive attribute by means of arbitrarily permuting the sensitive attribute values in each bucket, bucketization initially partitions tuples within the table into buckets as well as subsequently divides the quasi identifiers. Bucketization has been intended for anonymizing information of high-dimension in particular [12]. To marginal publication slicing has some connections; both of them release correlations among a subset of attributes. Attribute correlations between different columns are preserved by horizontal partitioning. Overlapping vertical partitioning which is left as our future work is similar to Marginal publication [5]. The multidimensional recoding which is also called regional recoding partitions the space of domain into the regions of non-intersect in addition to data points within the identical region are represented by means of the region they are into. Attribute correlations between different columns are preserved by horizontal partitioning. Overlapping vertical partitioning which is left as our future work is similar to Marginal publication. To be generalized differently the local recoding does not have the above constraints in addition to permitting unusual incidents of the identical value [10]. Anonymizing classification is a primary difficulty in analysis of data. To access large collection of data a classifier is required. A threat may pose to an individual's privacy by releasing person-specific data. The first involvement is to proficiently identify a k-anonymous solution which preserves the classification structure [6]. By the considering the both information and privacy the search is done by featuring the information level in a manner of top-down and this refinement algorithmic framework is highly efficient and natural for handling different types of attributes [13]. This approach exploits by the noise and redundant structures in the data for achieving both a privacy goal and a classification goal.

3. RESULTS:

Slicing protects confidentiality because it breaks the associations between unconnected attributes, which are infrequent and thus identifying. By involving the sensitive attribute, slicing conserves improved utility of data compared to generalization and is more efficient than bucketization within workloads. The associations between columns values of a

bucket are randomly generated and this may lose data utility. While protecting against privacy threats the limitations of generalization and bucketization are overcome and preserves better utility by slicing.

4. CONCLUSION:

In recent years, with the fast improvement of Internet technology the privacy safeguarding of data has become one of the most important research topics and turn out to be a serious unease in publication of personal data. Many techniques are proposed for protecting individual privacy and sensitive information in order to avoid the obstacles. A novel technique of data anonymization known as slicing was introduced to get better the existing state of art. To preserve the association within each column the essential thought of slicing is to break the connection cross columns. By involving the sensitive attribute, slicing conserves improved utility of data compared to generalization and is more efficient than bucketization within workloads.

REFERENCES:

[1] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In ICDE, pages 126–135, 2007.

[2] I. Dinur and K. Nissim, "Revealing Information while Preserving Privacy," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Proc. Theory of Cryptography Conf. (TCC), pp. 265-284, 2006.

[4] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SULQ Framework," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.

[5] C. Dwork, "Differential Privacy," Proc. Int'l Colloquium Automata, Languages and Programming (ICALP), pp. 1-12, 2006.

[6] A. Inan, M. Kantarcioglu, and E. Bertino, "Using Anonymized Data for Classification," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 429-440, 2009.

[7] V. Rastogi, D. Suciu, and S. Hong. The boundary between privacy and utility in data publishing. In VLDB, pages 531–542, 2007.

[8] J.H. Friedman, J.L. Bentley, and R.A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic

Expected Time," ACM Trans. Math. Software, vol. 3, no. 3, pp. 209-226, 1977.

[9] Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.

[10] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.

[11] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In KDD, pages 754–759, 2006.

[12] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[13] G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[14] C. Dwork, "Differential Privacy: A Survey of Results," Proc. Fifth Int'l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008.

[15] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.