International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 1, January 2015, PP 17-21
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

A Survey on Privacy Preservation for Anonymzing Data

M. Saranya
P.G Scholar CSE
M.I.E.T Engineering College
Tamilnadu, India

R. Senthamil Selvi
Associate Professor CSE
M.I.E.T Engineering College
Tamilnadu, India

Abstract: *Private data such as electronic health records and banking transactions must be shared within the cloud environment to analysis or mine data for research purposes. In big data applications, data privacy is one of the most concerned issue because processing large-scale privacy-sensitive data sets often requires computation power provided by public cloud services. Introducing a technique called Data Anonymization, where the privacy of an individual can be preserved while aggregate information is shared for mining purposes.A key learning from Data Anonymization involves just deleting fields in the database.In this paper, a survey for existing techniques and to analyze the strength and weakness of these approaches is discussed.*

Keywords: *Data Anonymization, K-anonymous,L-diversity, Privacy preservation, Top Down Specialization*

1. Introduction

Privacy is one of the most concerned issues in cloud computing.Personal data like financial transaction records and electronic health records are extremely sensitive although that can be analyzed and mined by research organization. Data privacy issues need to be addressed before data sets are shared on cloud for analysis purpose. Data anonymization refers to as hiding sensitive data for owners of data records [1].

Large-scale data sets are generalized using two phase top-down specialization for data anonymization. This process split into two phases. In the first phase,original data sets are partitioned into a group of smaller data sets, and these data sets are anonymized in parallel,producing intermediate results. In the second phase, the intermediate results are integrated into one, and further anonymized to achieve consistent k-anonymous [6] data sets. In such cases, data sets are anonymized rather than encrypted to ensure both data utility and privacy preserving.Instead of encryption, anonymization of data can be done as a means of preserving privacy [2].

Data anonymization is widely used method for Privacy Preserving of data in non-interactive data publishing scenario Data anonymization refers to the hiding the identity or sensitive data for owners data record. The privacy of individual can be effectively preserved while some aggregate information is shared for data analysis and mining.Several models ofsecurity can improve Data Anonymization include k-anonymity and l-diversity.

2. Related Work

We briefly review recent research on data privacy preservation and privacy protection in Map Reduce and cloud computing environments. LeFevre et al. [2] uses the scalable decision trees and sampling techniques to address the scalability problem of anonymization algorithms. Iwuchukwu et al. [13] build a spatial index over data sets to achieve high efficiency, proposed an R-tree index-based approach.

However, the these approaches, fails to work in the Top- Down Specialization (TDS) approach. Fung et al. [15] proposed the Centralized TDS approach that produces anonymous data sets without the data exploration problem. A data structure Taxonomy Indexed PartitionS (TIPS) is used to improve the efficiency of TDS. But since it is centralized, leads to its inadequacy in handling large-scale data sets [1].

Now considering the distributed algorithms which are also proposed to preserve privacy. Jurczyk et al. [12] and Mohammed et al. [15] proposed distributed algorithms to anonymize horizontally partitioned data sets retained by multiple holders. Jiang et al. [14] and Mohammed et al. [3] proposed distributed algorithms to anonymize vertically partitioned data from different data sources without disclosing privacy information from one party to another. But these distributed algorithms mainly aim at secure integrating and anonymizing multiple data sources but not the scalability issue.

3. Privacy Preserving Approaches in Data Publishing

3.1. K-Anonymity

K-anonymity is a property possessed by certain anonymized data.Given person-specific field-structured data; produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals.

3.1.1. k-anonymization Methods

Suppression: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'

Generalization: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30'.

Table1. *Anonymized Table*

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related

3.2. L-Diversity

L-diversity is a form of group based anonymization that is used to preserve .The l-diversity model is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least k other records in the data.

The l-diversity model handles some of the weaknesses in the k-anonymity model where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity.

3.3. T-Closeness

Given the existence of attacks where sensitive attributes may be inferred based upon the distribution of values for l-diverse data, the t-closeness method was created to further l-diversity by additionally maintaining the distribution of sensitive fields.

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.

3.4. Generalization Approach

By creatively applying Map Reduce on cloud to Bottom Up Generalization (BUG) for data anonymization and deliberately design a group of innovative Map Reduce jobs to concretely accomplish the generalizations in a highly scalable way. Secondly, introduce a scalable Advanced BUG approach, which performs generalization on different partitioned data set and the resulting intermediate anonymizations are merged to find final anonymization which is used to anoymize the original data set. Results show that our approach can significantly improve the scalability and efficiency of BUG for data anonymization over existing approaches.

3.5. Top-Down Specialization

Generally, TDS is an iterative process starting from the topmost domain values in the taxonomy trees of attributes. Each round of iteration consists of three steps [4]:

- Finding The Best Specialization
- Performing Specialization
- Updating Values of The Search Metric For The Next Round

Such a process is repeated until k-anonymity is violated, to expose the maximum data utility. The goodness of a specialization is measured by a search metric.

3.6. Map Reduce: A Large-Scale Data Processing Framework

To address the scalability problem of the Top-Down Specialization (TDS) approach for large scale data set used a widely adopted parallel data processing framework like Map Reduce. In first phase, the original datasets are partitioned into group of smaller datasets and these datasets are anonymized in parallel producing intermediate results. In second phase, these intermediate results are integrated into one and further anonymized to achieve consistent k-anonymous dataset.

Mapreduce is used to split up the large input data into chunks of more or equal size, spinning up a number of processing instances for the map phase apportioning data to each of the mappers, tracking the status of each mapper, routing the map results to the reduce phase and finally shutting down the mappers and the reducers when the work has been done. It is easy to scale up MapReduce to handle bigger jobs or to produce results in a shorter time by simply running the job on a larger cluster. When Mapreduce is not used the process fails in distribution system.

3.7. Two-Phase Top-Down Specialization (TPTDS)

A TPTDS approach in TDS is a highly scalable and efficient approach. The two phases of our approach are based on the two levels of parallelization provisioned by Map Reduce on cloud. Basically, Map Reduce on cloud has two levels of parallelization

- Job level
- Task level

Job level parallelization deals multiple MapReduce jobs that can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, MapReduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, for example, Amazon Elastic MapReduce service [5]. Task level parallelization refers to that multiple mapper/reducer tasks in a MapReduce job are executed simultaneously over data splits.By parallelizing multiple jobs on data partitions in the first phase to achieve high scalability,but the resultant anonymization levels are not identical. To obtain finally consistent anonymous data sets, the second phase is necessary to integrate the intermediate results and further anonymize entire data sets.

4. EVALUATION

Table2. *Different privacy preserving approaches*

TITLE	*AUTHOR*	*CONCEPT*	*STRENGTH*	*WEAKNESS*
"A Scalable Two-Phase Top-Down Specialization Approach For Data Anonymization Using Mapreduce On Cloud"[6]	Xuyun Zhang, Laurence T. Yang, Senior Member, Ieee, Chang Liu, And Jinjun Chen, Member, Ieee	A Highly Scalable Two-Phase TDS Approach Is Proposed Using Map Reduce On Cloud	Multiple Data Partition To Improve Scalability And Privacy	The Privacy Preservation In Cloud For Data Analysis, Sharing And Mining Is A Challenging Issue Inadequacy In Handling Large Datasets
Incognito: Efficient Full-Domain K-Anonymity" [8]	K. Lefevre, D.J. Dewitt, And R. Ramakrishnan, Proc.Acm Sigmod Conf. Management Of Data	Implementation Framework For Full Domain Generalization Using Multidimensional Data Model Together With Suite Of Algorithms	To Produce Minimal Full-Domain Generalizations Perform Up To An Order Of Magnitude Faster Than Previous Algorithms On Two Real-Life Databases	Performance Of Incognito Can Be Enhanced By Materializing Portions Of The Data Cube, Including Count Aggregates At Various Points In The Dimension Hierarchies

“Workload-Aware Anonymization Techniques For Large-Scale Data Sets”[7]	K. Lefevre, D.J. Dewitt, And R. Ramakrishnan, Acm Trans. Database Systems, Vol. 33, No. 3, Pp. 1-47, 2008	Anonymization Algorithms That Incorporate A Target Class Of Workloads, Consisting Of One Or More Data Mining Tasks As Well As Selection Predicates And The Datasets Much Larger Than Main Memory	High Efficiency And Quality Data Overcomes Problem Of Scalability	Problem Of Measuring The Quality Of Anonymized Data Fails To Work In The Top-Down Specialization (TDS) Approach
“Privacy-Preserving Data Publishing: A Survey Of Recent Developments”[1]	B.C.M. Fung, K. Wang, R. Chen, And P.S. Yu, Acm Computing Surveys, Vol. 42, No. 4, Pp. 1-53, 2010	Provides Methods And Tools For Publishing Useful Information While Preserving Data Privacy	PPDP Has Received A Great Deal Of Attention In The Database And Data Mining Research Communities	Degradation Of Data / Service Quality Loss Of Valuable Information Increased Costs
“K-Anonymization As Spatial Indexing: Toward Scalable And Incremental Anonymization,”[9]	T. Iwuchukwu And J.F. Naughton, Proc. 33rd Int’l Conf. Very Large Data Bases (Vldb ’07), Pp. 746-757,2007	K-Anonymizing A Data Set Is Similar To Building A Spatial Index Over The Data Set Using R-Tree Index Based Approach	Achieve High Efficiency And Quality Anonymization Multidimensional Generalization, High Accuracy	More Compaction Is Needed To Achieve High Quality Anonymization Different Indexing Algorithm Provide Different Issues
“Mapreduce: Simplified Data Processing On Large Clusters,”[10]	J. Dean And S. Ghemawat, Comm. Acm, Vol. 51, No. 1, Pp. 107-113,2008	Mapreduce Is A Programming Model Implementation For Processing And Generating Large Data Sets Performs Map() And Reduce()	Allows Us To Handle Lists Of Values That Are Too Large In Memory The Model Is Easy To Use	Mapreduce Is Not Suitable For A Short On-Line Transactions
“Sedic: Privacy-Aware Data Intensive Computing On Hybrid Clouds”[15]	Zhang K, Zhou X, Chen Y, Wang X, Ruan Y	To Protect Data Privacy During Map-Reduce Programming	Sensitive User Data Protection High Privacy Assurance Ease To Use Fully Preserve The Scalability	Scalability Problem Occurs
“Data & Knowledge Engineering”[16]	Jiuyong Li , Jixue Liu , Muzammil Baig, Raymond Chi-Wing Wong	Data Generalization In Anonymization Should Be Determined By The Classification Capability Of Data Rather Than The Privacy Requirement	More Accurate Classification Models And Is Faster Than A Benchmark Utility-Aware Data Anonymization Algorithm	Not As Much Faster Than Other Anonymization Methods

REFERENCES

[1] B.C.M. Fung, K. Wang, R. Chen and P.S. Yu, “Privacy-Preserving Data Publishing: A Survey of Recent Developments,” ACM Comput. Surv., vol. 42, no. 4, pp.1-53, 2010.

[2] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, “Workload- Aware Anonymization Techniques for Large-Scale Datasets,” ACM Trans. Database Syst., vol. 33, no. 3, pp. 1-47, 2008.

[3] B. Fung, K. Wang, L. Wang and P.C.K. Hung, “Privacy- Preserving Data Publishing for Cluster Analysis,” Data Knowl.Eng., Vol.68, no.6, pp. 552-575, 2009.

[4] B.C.M. Fung, K. Wang, and P.S. Yu, “Anonymizing Classification Data for Privacy Preservation,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[5] Amazon Web Services, “Amazon Elastic Mapreduce,”http://aws.amazon.com/elasticmapreduce/, 2013.

[6] Xuyun Zhang, Laurence T. Yang, Senior Member, IEEE, Chang Liu, and Jinjun Chen, Member, IEEE “A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud”, vol. 25, no. 2, february 2014.

[7] K. LeFevre, D.J. DeWitt and R. Ramakrishnan,“Workload- Aware Anonymization Techniques for Large-Scale Datasets,” ACM Trans. Database Syst., vol.33, no. 3, pp. 1-47, 2008.

[8] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Incognito: Efficient Full-Domain K-Anonymity,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’05), pp. 49-60, 2005.

[9] T. IwuchukwuandJ.F. Naughton, “K-Anonymization as Spatial Indexing: Toward Scalable and Incremental Anonymization,” Proc. 33rdInt'lConf. VeryLarge DataBases (VLDB'07), pp.746-757, 2007

[10] J. Dean and S. Ghemawat, “Mapreduce: Simplified Data Processing on Large Clusters,” Comm. ACM, vol. 51, no. 1, pp. 107-113,2008.

[11] A.Machanavajjhala, J.Gehrke, and D.Kifer, et al, “ℓ-diversity: Privacy beyond k-anonymity”, In Proc. ofICDE, Apr.2006.

[12] Dean J, Ghemawat S. Mapreduce: a flexible data processing tool. Communications of the ACM 2010;53(1):72–77. DOI: 10.1145/1629175.1629198.

[13] P.JurczykandL.Xiong, “Distributed Anonymization:Achieving Privacy for Both Data Subjects and Data Providers, ”Data and Applications Security XXIII(DBSec'09),pp.191-207,2009.

[14] N. Mohammed, B.C. Fungand M. Debbabi, “Anonymity Meets Game Theory: Secure Data Integration with MaliciousParticipants,”VLDBJ.,vol.20,no.4,pp.567-588,20.

[15] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness:Privacy Beyond k-anonymity and l-Diversity”, In Proc.of ICDE, 2007, pp. 106-115.

[16] Jiuyong Li, Jixue Liu , Muzammil Baig , Raymond Chi-Wing Wong” Information based data anonymization for classification utility”

AUTHORS’ BIOGRAPHY

Saranya has received B.E degree in Computer Science and Engineering from Anna University in 2013. She is currently pursuing her M.E degree in Computer Science and Engineering under the same University.

Senthamil Selvi has been working as an associate professor under Anna University affiliated college and pursuing her Ph.D.