An Efficient Algorithm for Differentially Private Data Release

Hanooja T, D.Ravi, *T*.K.P Rajagopal

***Abstract*— Data mining is an emerging technology which helps companies to focus on essential information about a customer's behavior. Sensitive data disclosure is a main problem addressed during publishing data. Sharing private data like electronic health records, financial transaction record poses a threat to individual privacy. Anonymizing data sets using Generalization is a widely used technique for preserving privacy. The proposed Top down specialization algorithm provides protection for sensitive information. Data sets are generalized in a top-down manner until k-anonymity problems are solved. The proposed system achieves differential privacy and privacy is achieved at the vertical level. The proposed algorithm can effectively preserve information for a data mining task.**

***Index Terms*— Generalization, k-anonymity, sensitivity, specialization.**

I. INTRODUCTION

A data owner wants to release a person-specific data table to another party or the public for the purpose of classification without scarifying the privacy of the individuals in the released data. Traditional data processing applications or on-hand database management tools becomes difficult to process for a large collection of data sets. The emergence of new paradigms such as cloud computing increases the amount of data distributed between multiple entities. These distributed data can be integrated to enable better data analysis for making better decisions and providing high-quality services the trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data. Cloud computing, a disruptive trend at present, poses a significant impact on current IT industry and research communities .Cloud computing provides massive computation power and storage capacity via utilizing a large number of commodity computers together, enabling users to deploy applications cost effectively without heavy infrastructure investment. However, numerous potential customers are still hesitant to take advantage of cloud due to privacy and security concerns. 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 organization. Data privacy issues need to be addressed urgently before data sets are shared on cloud. Data anonymyzation refers to as hiding sensitive data for owners of data records. Large-scale data sets are generalized using two phase top-down specialization for data anonymyzation.

As an example see the table below. If this table is to be anonymized with Anonymization Level (AL) set to 2 and the set of Quasi Identifiers as QI = {AGE, SEX, ZIP, PHONE}. The quasi-identifiers are identified by the organization according to their rules and regulations.

Table 1. Initial Table

NAME	AGE	SEX	ZIP	PHONE	DISEASE
Ali	20	M	190014	9419	Bronchitis
Bale	30	M	190001	9592	Lung Cancer
Calvin	40	M	192231	9823	STI
Doris	50	F	190001	8988	Skin Allergy
Elle	75	F	190002	8088	Skin allergy

The NAME attribute here is "Sensitive", so we would like to "suppress" this attribute before anonymizing the above table. After, suppression the table will look like as below:

Table 2. Table after Suppression

AGE	SEX	ZIP	PHONE	DISEASE
20	M	190014	9419	Bronchitis
30	M	190001	9592	Lung Cancer
40	M	192231	9823	STI
50	F	190001	8988	Skin Allergy
75	F	190002	8088	Skin Allergy

Anonymizing the data through Top-Down Specialization, each attribute value will be initialized to the root of Taxonomy Tree and will look like as below.

ISSN: 2278 – 1323

Table 3. Table after TDS Specialization

AGE	SEX	ZIP	PHONE	DISEASE
[0-100]	ANY	******	****	Bronchitis
[0-100]	ANY	******	****	Lung Cancer
[0-100]	ANY	******	****	STI
[0-100]	ANY	******	****	Skin Allergy
[0-100]	ANY	******	****	Skin Allergy

The data in the above table is highly privacy preserved, but the data utility is very low. The data is highly anonymized. We make a note here that Data Anonymization is not only the single goal that we are trying to achieve through Anonymization. We also make sure that data utility is high enough to make the information useful for mining.

The Top-Down Specialization Algorithm will iteratively specialize the attribute values till the k-anonymity is violated.

The given table after anonymizing it for k=2 will look like:

Table 4. Table after Anonymization

AGE	SEX	ZIP	PHONE	DISEASE
[0-50]	M	1900**	9***	Bronchitis
[26-50]	M	190001	9***	Lung Cancer
[26-50]	M	19*****	9***	STI
[26-50]	F	190001	8***	Skin Allergy
[51-100]	F	19000*	8***	Skin Allergy

II. RELATED WORK

Data mining addresses many problems like disclosing sensitive information while mining for useful information. Many papers are studied for solving this problem.

Noman Mohammed, Dima Alhadidi, Benjamin C.M. Fung [1] acquires privacy by vertically partitioning the data. N. Mohammed ,B.C.M Fung, and M. Debbabi [2] solves a privacy preserving data sharing problem and can integrate two data sets securely. However,. Anonymization problem occur due to intuitiveness.

N. Mohammed ,B.C.M Fung, P.C.K Hung, and C. Lee[4] prevent linkage attacks by means of anonymity operations such as generalization and suppression .However, it is difficult to simulate the background knowledge of attackers.

W. Jiang and C. Clifton [5] enable two parties to integrate their data satisfying the k-anonymity privacy model. R. Agrawal, A. Evfimievski, and R. Srikant [6] proved that these protocols disclose minimal information apart from the query result .However, this protocol leaks some information about which tuples joined, based on the distribution of duplicates.

A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan [3] prove a lower bound on accuracy of approximating the Hamming distance between two vectors-the classical problem in two-party computation-with two-sided guarantees of differential privacy. However, there may be leakage of data within the system

III. EXISTING SYSTEM

Data anonymyzation has been extensively studied and widely adopted for data privacy preservation in non interactive data publishing and sharing scenarios. Data anonymyzation refers to hiding identity and/or sensitive data for owners of data records. Then, the privacy of an individual can be effectively preserved while certain aggregate information is exposed to data users for diverse analysis and mining. A variety of anonymyzation algorithms with different anonymyzation operations have been proposed. However, the scale of data sets that need anonymizing in some cloud applications increases tremendously in accordance with the cloud computing .Data sets have become so large that anonymizing such data sets is becoming a considerable challenge for traditional anonymyzation algorithms. The researchers have begun to investigate the scalability problem of large-scale data anonymyzation.

The centralized data mining model assumes that all the data required by any data mining algorithm is either available at or can be sent to a central site. A simple approach to data mining over multiple sources that will not share data is to run existing data mining tools at each site independently and combine the results. However, this will often fail to give globally valid results. Issues that cause a disparity between local and global results include: Values for a single entity may be split across sources. In existing system, data partitioning is only achieved at the horizontal level. When combining data from different sources, it could potentially reveal person-specific sensitive information. Due to the use of single party protocol, existing systems are time consuming. Dumped file is accessed by an unauthorized user or application, private information within the program may be leaked.

IV. PROPOSED SYSTEM

I propose Top down Specialization (TDS) approach for data anonymyzation. To make full use of the parallel capability, specializations required in an anonymyzation process are split into two phases. In the first one, 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 one, the intermediate results are integrated into one, and further anonymized to achieve consistent k-anonymous data sets. Here TDS approach is used

ISSN: 2278 – 1323

to accomplish the concrete computation in both phases. First, TDS is applied for data anonymyzation, then do specialization in a highly scalable fashion. Second, I propose a two-phase TDS approach to gain high scalability via allowing specializations to be conducted on multiple data partitions in parallel during the first phase.

V. METHODOLOGY ADOPTED

A. *Suppression*

Initially sensitive data are identified and suppress those attributes before anonymyzation. Suppression is the process of hiding the sensitive information from the data set. After applying suppression, the highly sensitive information is removed from the table. The sensitive score for each attribute is calculated and the attribute which is having highest score is selected for suppression. After applying suppression, highly sensitive data become secured for a meanwhile.

B. *Data Partition*

The data partition is performed on the data sets. Large data sets are collected and split it into small data sets. Then provides the random no for each data sets .Partitioning is the process of determining which reducer instance will receive which intermediate keys and values. Each mapper must determine for all of its output (key, value) pairs which reducer will receive them. The partitioner's are able to partition data independently they should never need to exchange information with one another to determine the partition for a particular key. It is necessary that for any key, regardless of which mapper instance generated it, the destination partition is the same. If the key cat is generated in two separate (key, value) pairs, they must both be reduced together. It is also important for performance reasons that the mappers be able to partition data independently they should never need to exchange information with one another to determine the partition for a particular key.

C. Generalization

Generalization is the process of giving a random value to a set of data. A range of value is given to each attribute, so that it is difficult to identify a single person or attribute from a group. A range is specified for highly sensitive data. Highly sensitive data are somewhat secured after applying generalization

D. Anonymyzation

Data Anonymyzation is the process of either encrypting or removing personally identifiable information from data sets, so that the people whom the data describe remain anonymous. Anonymyzation of data can mitigate privacy and security concerns and comply with legal requirements. Anonymization is not invulnerable countermeasures that compromise current Anonymyzation techniques can expose protected information in released datasets. After getting the individual data sets it applies the anonymyzation. The anonymyzation means hide or remove the sensitive field in data sets. Then it gets the intermediate result for the small data sets. The intermediate results are used for the specialization process. Data anonymyzation algorithm converts clear text data into a nonhuman readable and an irreversible form.

E. T*op down Specialization*

In Top-Down Specialization, all the attribute values are initialized to the root value of the hierarchy tree. The specialization is carried out iteratively over the attribute values, until the k- anonymity is violated. The specialization is performed by replacing the parent attribute value by its child value in Taxonomy Tree.

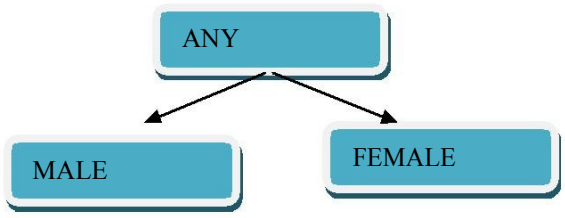

Fig 1. The Taxonomy Tree for the attribute SEX

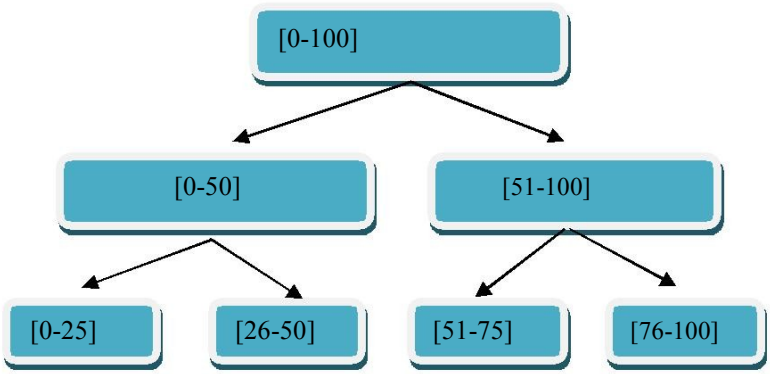

Fig 2. Taxonomy Tree for continuous attribute AGE

Records consist of a large number of attributes. This attributes are arranged in a tree like structure. **.** Initially, Cut_i contains only the top most value for its attribute. Then specialization is applied to the data sets. After gets the intermediate result those results are merged into one. Again applies the anonymyzation on the merged data its called specialization.

At each iteration find the highest score. When Top-Down specialization is applied to the taxonomy tree structure first it finds the best specialization, and then performs specialization again and finally update values for the next round. Values for the each specialization are analyzed. The highest value for specialization is regard as the best specialization. The values are updated until k-anonymity problems are solved.

ISSN: 2278 – 1323

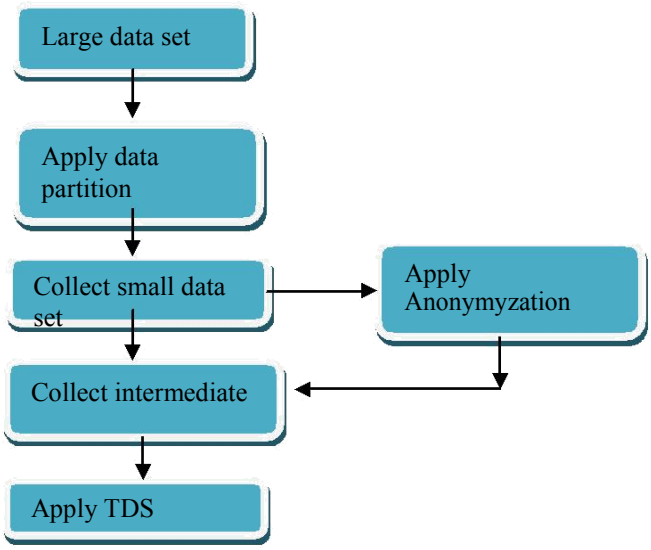

Fig 3. Data Flow Diagram

Top down Specialization Algorithm

1. Initialize every value in T to the top most value.
2. Initialize Cut_i to include the top most value.
3. **while** some $x \in UCut_i$ is valid do
4. Find the Best specialization of the highest Score in $UCut_i$.
5. Perform the Best specialization on T and update $UCut_i$.
6. Update score(x) and validity for $x \in UCut_i$.
7. **end while**
8. **return** Generalized T and $UCut_i$.

VI. Conclusion

In this paper, a highly secured TDS approach is proposed. Data sets are partitioned and anonymized in parallel in the first phase, producing intermediate results. Then, the intermediate results are merged and further anonymized to produce consistent k-anonymous data sets in the second phase. Experimental results on real-world datasets have demonstrated that with our approach, the scalability and privacy of TDS are improved significantly over existing approach. The proposed algorithm is differentially private and secure under the security definition of the semi-honest adversary model.

The paper solves the problem of disclosing sensitive information using the TDS algorithm. Here the security is achieved both at the horizontal and vertical level. Differential privacy is the main thing achieved with this project. A differentially private mechanism ensures that the probability of any released data is equally likely from all nearly identical input data sets and, thus, guarantees that all outputs are insensitive to any individual's data.

Acknowledgment

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this M.Tech project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

I express my warm thanks to the Head of Department Dr.Subathra madam, and Dr.T.K.P Rajagopal sir for their support and guidance. I express my gratitude to my project guide Ravi sir and all the people who provide me with the facilities being required and conductive conditions for my project.

References

[1] Noman Mohammed, Dima Alhadidi, Benjamin C.M. Fung,, "Secure Two-Party Differentially Private Data Release for Vertically Partitioned Data,". Transactions on dependable and secure computing, vol. 11, no. 1, January 2014.

[2] N. Mohammed, B.C.M. Fung, and M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Participants,"Very Large Data Bases J., vol. 20, no. 4, pp. 567-588, Aug.2011.

[3] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. Vadhan, "The Limits of Two-Party Differential Privacy,"Proc. IEEE Symp. Foundations of Computer Science (FOCS '10), 2010.

[4] N. Mohammed, B.C.M. Fung, P.C.K. Hung, and C. Lee, "Anonymizing Healthcare Data: A Case Study on the Blood Transfusion Service,"Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '09), 2009.

[5] Paul Bunn and Rafail Ostrovsky, " Secure Two-Party k-means Clustering ," Very Large Data Bases J.,vol. 15, no. 4,pp. 316-333, Nov. 2006.

[6] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing Across Private Databases,"Proc. ACM Int'l Conf. Management of Data, 2003.

AUTHORS BIOGRAPHY

Hanooja T: She Completed UG from Government Engineering College, BartonHill , Thiruvananthapuram, India. She is currently doing PG in computer Science and Engineering from Kathir College of Engineering, Coimbatore, India. She is an active member of IEEE student branch. Attended various International, National Conferences and Attended various work shop based on Image processing and Data mining. Her research interests are Image processing, Data mining**.**

ISSN: 2278 – 1323

Ravi.D: He received the MCA degree a from the IGNOU, New Delhi in the year 2003, the M.Phil from the Bhrathidasan University, Trichy in the 2005. The M.E (CSE) in Anna University of Technology, Coimbatore in the year 2011. He has 11 years of Teaching Experience. His area of Interest is in Web Mining and Personalization. He published 04 National Conference and 04 International Conference. He is a member of ISTE, IACSIT and IAENG.

T.K.P. Rajagopal M.A., M.Phil., M.C.A., M.Phil., M.E.(CSE), M.Tech.(IT), Ph.D'., is working as an Associate Professor in the Department of Computer Science and Engineering at Kathir College of Engineering, Coimbatore, Tamailnadu, India. He has 15 years of teaching experience. His research areas are Network Security, Data Mining and Digital Image Processing. He has published 3 books, 9 research papers in international journals and 2 research papers in national journals, Presented 17 papers in International Conference, 19 papers in National level conferences and has attended 3 QIP, 5 FTP, 10 Seminars and 33 Workshops. He is a life member of various professional societies like ISTE, DOEACC, CSTA, CSI, ACS, IAS, IACSIT, IIRJC, SDIWC, ISI, IoN, UACEE, theIRED, IAENG Society of Artificial Intelligence, Computer Science, Data Mining, Software Engineering, and Wireless Networks, Hong Kong, Serving as a reviewer in IJITCE, WCSIT, CSC-IJCSS, CSC-IJCN, IJCSIS, Scientific & Technical Committee & Editorial Review Board on Engineering and Natural Sciences in WASET.

ISSN: 2278 – 1323