RESEARCH ARTICLE OPEN ACCESS

Privacy Preservation of Data in Data Mining

Prachi Kohale*, Sheetal Girase**
*(Department of Inforamation Technology, MIT, PUNE PUNE University)
** (Department of Information Technology,MIT PUNE, PUNE University)

ABSTRACT
For data privacy against an un-trusted party, Anonymization is a widely used technique capable of preserving attribute values and supporting data mining algorithms. The technique deals with Anonymization methods for users in a domain-driven data mining outsourcing.
Several issues emerge when anonymization is applied in a real world outsourcing scenario. The majority of methods have focused on the traditional data mining approach; therefore they do not implement domain knowledge nor optimize data for domain-driven usage. So, existing techniques are mostly non-interactive in nature, providing little control to users.
To successfully obtain optimal data privacy and actionable patterns in a real world setting, these concerns need to be addressed. The technique is based on anonymization framework for users in a domain-driven data mining outsourcing scenario. The framework involves several components designed to anonymize data while preserving meaningful or actionable patterns that can be discovered after mining.
In the medical field, for enhancing the quality and importance of healthcare, data mining plays an important role. For example classification analysis on patients data to determine their probability of having heart disease, so, appropriate actions can be taken, thus enhancing treatment, saving time and cost. It is helpful to anonymize data while preserving meaningful of actionable patterns that can be found after mining.
Keywords: DDDM, GT, BT

I. INTRODUCTION

In the real world data mining is constraint based as opposed to traditional data mining which is data driven trial and error process. In traditional data mining knowledge discovery usually performed without domain intelligence, thus affecting model or actionability for real business needs [1]. So there exists gap between objective and business goals as well as outputs and business expectations. Domain driven data mining aims to bridge the gap between these by involving domain experts and knowledge to obtain actionable patterns or models applicable to real world requirements.[7] Data anonymization in domain driven data mining uses methods like k anonymity and recent dynamic anonymization methods.

II. ANONYMIZATION

The fig.1 shows anonymization in domain driven data mining outsourcing. It involves several components designed to anonymize data while preserving meaningful and actionable patterns that can be discovered after mining [8]. In contrast to traditional data mining this framework integrates the knowledge to retain values after anonymization. Domain driven data mining DDDM is to make system deliver business friendly and decision making rules and actions that are of solid technical significance as well [2]. Its having effective involvement of domain intelligence, network intelligence, Data intelligence, human intelligence, social intelligence[9].

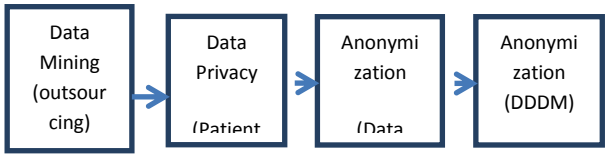

fig 1 anonymization in domain driven data mining outsourcing.

III. ANGEL TECHNIQUE

Generalization is well known method for privacy preservation of data. It has several drawbacks such as heavy information loss, difficulty of supporting marginal publication. [3]To overcome this Angel, a new anonymization technique that is effective as generalization in privacy protection but able to retain significantly more information in microdata. It is applicable to any principle l-diversity, t-closeness and to overcome the drawback of several methods. Compared to traditional generalization it ensures same privacy guarantee preserve significantly more information in microdata.

Table 1 ANGEL publication 1

Name	Age	Sex	Disease
Alan	21	M	pneumonia
Bob	23	M	pneumonia
Carrie	38	F	bronchitis
Daisy	40	F	bronchitis
Eddy	41	M	pneumonia
Frank	43	M	pneumonia
Gloria	58	F	bronchitis
Helena	60	F	bronchitis

(a) The microdata

Age	Sex	Disease
[21,40]	*	pneumonia
[21,40]	*	pneumonia
[21,40]	*	bronchitis
[21,40]	*	bronchitis
[41,60]	*	pneumonia
[41,60]	*	pneumonia
[41,60]	*	bronchitis
[41,60]	*	bronchitis

(b) 2-diverse generalization

Suppose that we want to publish the microdata of Table 1, conforming to 2-diversity. ANGEL first divides the table into batches:
Batch 1: {Alan, Carrie},
Batch 2: {Bob, Daisy},
Batch 3: {Eddy, Gloria},
Batch 4: {Frank, Helena}.
Observe that each batch obeys 2-diversity: it contains one pneumonia- and one bronchitis-tuple. ANGEL creates a batch table (BT), as in Table 3.1a, summarizing the Disease-statistics of each batch. For example, the first row of Table 3.1a states that exactly one tuple in Batch 1 carries pneumonia. Then, ANGEL creates another partitioning into buckets (which do not have to be 2-diverse).
Bucket 1: {Alan, Bob},
Bucket 2: {Carrie, Daisy},
Bucket 3: {Eddy, Frank},
Bucket 4: {Gloria, Helena}

Table 2 Microdata

Name	Age	Sex	Zip	Disease
Allan	21	M	10k	Pneumonia
Bob	23	M	58k	Flu
Carrie	58	F	12k	Bronchitis
Daisy	60	F	60k	Pneumonia
Eddy	70	M	78k	Flu
Frank	72	M	80k	Bronchitis

Table 3 2-diverse generalization

Age	Sex	Zipcode	Disease
[21, 23]	M	[10k, 58k]	Pneumonia
[21, 23]	M	[10k, 58k]	Flu
[58, 60]	F	[12k, 60k]	Bronchitis
[58, 60]	F	[12k, 60k]	Pneumonia
[70, 72]	M	[78k, 80k]	Flu
[70, 72]	M	[78k, 80k]	Bronchitis

Zipcode	Disease
[10k, 12k]	Pneumonia
[10k, 12k]	Bronchitis
[58k, 60k]	Flu
[58k, 60k]	Pneumonia
[78k, 80k]	Flu
[78k, 80k]	Bronchitis

Finally, ANGEL generalizes the tuples of each bucket into the same form, producing a generalized table (GT). Table 3 demonstrates the GT. Note that GT does not include the Disease attribute, but stores, for each tuple of the microdata, the ID of the batch containing it. For instance, the first tuple of Table has a Batch-ID 1, because its owner Alan belongs to Batch 1. Tables 3.a and 3.b are the final relations released by ANGEL[10].

IV. CONCLUSION

Angelization is new anonymization technic for privacy preserving publication which is applicable to any monotonic anonymization principle. It produces anonymized relation that achieve privacy guarantee as conventional generalization but permits much more accurate reconstruction of original data distribution. It offers simple and rigorous solution which was difficult issue.

V. ACKNOWLEDGEMENTS

I would like to take opportunity to express my humble gratitude to my guide Prof. Shital P. Girase, Asst. Prof. Maharashtra Academy of Engineering & Educational Research's, Maharashtra Institute of Technology, Pune, who gave me assistance with their experienced knowledge and whatever required at all stages of my project.

References

[1] Longbing Cao "*DDDM challenges and prospectus* " *june 2010,755-769*
[2] Longbing Cao, Philip S. Yu, Chengqi Zhang, Yanchang Zhao "Domain Driven Data Mining"
[3] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. "*Achieving anonymity via clustering*". In PODS, 2006.
[4] Josep Domingo-Ferrer and Vicenc¸ Torra "*A Critique of k-Anonymity and Some of Its Enhancements*". *3 International* Conference on Availability, Reliability and Security
[5] V.Narmada "*An enhanced security algorithm for distributed databases in privacy preserving data ases*" *(IJAEST) INTERNATIONAL JOURNAL OF*

ADVANCED ENGINEERING SCIENCES AND TECHNOLOGIES".

[6] R. Agrawal and R. Srikant. "*Privacy-preserving data mining*".

[7] J. Han and M. Kamber Morgan Kaufmann "*Data Mining: Concepts and Techniques*".2000.

[8] Brian, C.S. Loh and Patrick, H.H. Then l. "*Ontology-Enhanced Interactive Anonymization in Domain-Driven Data Mining Outsourcing*" *2010 Second International* Symposium on Data, Privacy, and E-Commerce

[9] CHARU C. AGGARWAL,PHILIP S. YU "*PRIVACY-PRESERVING DATA MINING:MODELS AND ALGORITHM*" IBM T. J. Watson Research Center, Hawthorne, NY 10532

[10] Yufei Tao1 Hekang Chen2 Xiaokui " *Enhancing the Utility of Generalization for Privacy Preserving data publishing*"