Secure Association Rule Mining for Distributed Level Hierarchy in Web

Gulshan Shrivastava,

Department of Computer Science & Engineering,
Ambedkar Institute of Technology,
Geeta Colony, Delhi, India

Dr. Vishal Bhatnagar

Department of Computer Science & Engineering,
Ambedkar Institute of Technology,
Geeta Colony, Delhi, India

***Abstract*—Data mining technology can analyze massive data and it play very important role in many domains, if it used improperly it can also cause some new problem of information security. Thus several privacy preserving techniques for association rule mining have also been proposed in the past few years. Various algorithms have been developed for centralized data, while others refer to distributed data scenario. Distributed data Scenarios can also be classified as heterogeneous distributed data and homogenous distributed data and we identify that distributed data could be partitioned as horizontal partition (a.k.a. homogeneous distribution) and vertical partition (a.k.a. heterogeneous distribution). In this paper, we propose an algorithm for secure association rule mining for vertical partition.**

Keywords- Data Mining, Association Rule Mining, Privacy Preserving, Web Log, Vertical Partition

I. INTRODUCTION

Data Mining refers to extraction or mining knowledge from huge amount of data. Mining encompasses various algorithms such as clustering, classification, association rule mining and sequence detection. Association rule mining is one of the best researched techniques of data mining that was first introduced in [2]. Its main goal is to frequent patterns, extract interesting correlations and association structures among itemsets in the transactional database or other data repositories. Basically association rule mining is association rule for satisfy minimal support and minimal confidence.

Internet is becoming immeasurably popular all across the globe. Web Mining is mainly use for discover and analyze the World Wide Web (Web) information. In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization's database. The type of collected data differs based on its source location. It also has extreme mutation both in its content (e.g. text, image, audio, symbolic) and Meta information, which might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are:

1. Unlabeled;
2. Distributed;
3. Heterogeneous (mixed media);
4. Semi structured;
5. Time varying;
6. High dimensional.

However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web [1]. Like any other technology, web mining may be misused. However, we have two common approaches: secure multiparty computation and data obscuration. This field is expected to flourish.

In contrast to the centralized model, the Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the mining results from all the data across these distributed sources. Since the primary (if not

only) focus is on efficiency, most of the algorithms developed to date do not take security consideration into account.

Centralized model will often fail to give globally valid results. Issues that cause a disparity between local and global results include:

1) Need for handling context sensitive and imprecise queries;
2) Need for summarization and deduction;
3) Need for personalization and learning.

The rest of this paper is arranged as follows: Section 2 gives an overview about the Background and related work in the area of privacy preserving association rule mining on distributed heterogeneous database. Section 3 the detail of the problem definition. Section 4 proposed algorithm for the for object level hierarchy in web for computing the distributed association rule mining to preserve the privacy of users. Section 5 experimental evolution of our propose algorithm with an example. Finally, some conclusion and prospect are put forward in Section 6.

II. BACKGROUND AND RELATED WORK

Lots of cryptography-based approaches have been developed in the context of privacy preserving data mining algorithms, to solve problems of the following nature i.e. two or more parties want to do a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. The issue here is how to conduct such a computation while preserving the privacy of the inputs. This problem is referred to as the Secure Multiparty Computation (SMC) problem. We have lots of approaches which have been developed by using the solution framework of secure multiparty computation. The data in all of the cases that this solution is adopted is distributed among two or more sites. Such as [4] [12]:

1) Vertically Partitioned Distributed Data Secure Association Rule Mining
2) Horizontally Partitioned Distributed Data Secure Association Rule Mining
3) Vertically Partitioned Distributed Data Secure Decision Tree Induction
4) Horizontally Partitioned Distributed Data Secure Decision Tree Induction
5) Privacy Preserving Clustering

There have been some cryptography based algorithms as well. Lindell and Pinkas [10] first introduced a secure multi-party computation technique for classification using the ID3 algorithm, over horizontally partitioned data. Lin and Clifton [11] propose a secure way for clustering using the EM algorithm [8] over horizontally partitioned data. Kantarcioglu and Vaidya [9] present architecture for privacy preserving mining of client information. Agrawal et al. [6] present a technique for computing set intersection, union, and equi-joins for two parties. Clifton et al. provide a good overview of tools for privacy preserving distributed data mining [7].

Kun Liu et al [5] discussed following problem that suppose there are N organizations O_1, $O_2 \ldots O_N$; each organization O_i has a private transaction database DB_i. A third party data miner wants to learn certain statistical properties of the union of these $\bigcup_{i=1}^{N} DB_i$. These organizations are Comfortable with this, but they are reluctant to disclose their raw data. How could the data miner perform data analysis without compromising the privacy of the data? In this scenario, the data is usually distorted and its new representation is released; anybody has arbitrary access to the published data. The authors [13] provide randomized multiplicative data perturbation technique to solve some of the problems of additive random perturbation.

Recently there has been a great deal of work in privacy preservation in social networks. The work that is most closely related to the setting of sparse multidimensional data that we study here is the one that models the social networks as graphs. Publishing information from a social network would lead to the revelation of a graph that connects each individual to his friends or other persons he socially interacts. [16] Describes a series of different attack models that rely on identifying sub-graphs in the published social network graph.

Our basic privacy guaranty comes from [17], but we follow a completely different path with respect to the data transformation and the type of data utility we are preserving.

III. PROBLEM DEFINITION

We consider the heterogeneous database scenario considered in [3], a vertical partitioning of the database between two parties P and Q. Now we need to discuss two key point of association rule i.e. support and confidence. So let $I = \{I_1, I_2, \ldots\ldots, I_n\}$ is a set of n distinct attributes, T is transaction that contain a set of items

such that $T \subseteq I$. An Association rule is implication of the form $P \Rightarrow Q$, where $P \subset I$, $Q \subset I$, and $P \cap Q = \phi$. So support of an association rule is defined as the fraction of records that contain $P \cup Q$ to total number of record in the database and confidence is defined as fraction of the number of transaction that contain $P \cup Q$ to total number of records that contain P that is measure of strength of the association rule[15][14].

In this paper, we focus on the preservation of privacy in vertical partitioned. Common source for such data are credit card log, web log etc. consider example a dataset P which contain web logs. If an attacker has background knowledge that associates queries to known user then the publication of P might lead to privacy breaches. For example, assume that attacker *Ravi* knows that the user *Ashu* was interested on *train ticket* to *Shimla*, so he have the background knowledge consisting of terms *Shimla* and *train ticket*. If P is published without any modification then the attacker can trace all record that contain both term *Shimla* and *train ticket*. If only one record exists, then he can easily get that this is the record of *Ashu*.

To counter such privacy leaks, we propose a technique work for many type of heterogeneous data but our work is significantly motivated by the need to publish web logs due to a large amount of information it contain. The records inside these groups are partitioned vertically to chunks of similar sub-records. The aim of vertical partitioning is twofold; on one hand, it aims at preserving combinations that are already frequent enough as not to be identifying. On the other hand, it aims at breaking up infrequent combinations so that the records that originally contain them cannot be identified.

Attacker Scenario: We can identify user through tracing web log records whose records have unique combination e.g. if an attacker has the information of a user A who has searched for D and G and if there is only single record that contains D and G in the dataset, the attacker is absolutely indisputable that this record is related to the user A. We assume that the attacker:

- Have the information about a user seems to the published dataset.
- Has background knowledge that equate to queries posed by any user.
- Does not have background knowledge sufficient to guess negative knowledge about a user. For example, if term A appears 1000 times in the data, we assume that the attacker does not have enough background knowledge that links A to 1000 difference persons, so they can rule out the possibility that A is associated to any other person.
- Is interested in identifying a record or a part of a record that is associated with a specific user.

IV. PROPOSED ALGORITHM

In vertical partition is applied to each cluster independently. Let us suppose a cluster C and let T^C be he set of term that is in C.

Step 1: Partition C into $C_1,\ldots C_n$ of n records chunks and C_T as term chunks.

Step 2: Divide T^C into n+1 subsets $T_1,\ldots..,T_n, T_T$.

Step 3: Now that subsets are Pair wise disjoint i.e.

$T_i \cup T_j = \phi$, $i \neq j$ and jointly exhaustive i.e. $\cup T_i = T^C$.

Step 4: Here in these step the subsets $T_1,\ldots.T_n$ are used for record chunks $C_1,\ldots..C_n$ while subset T_T is used to define term chunk C_T.

$C_i = \{\{T_i \cap r \mid \text{for every record } r \in C\}\}$ (1)

Where {{…...}} denotes collection of record where duplicate records are allowed in C_i.

Step 5: Finally the term chunk is defined as

$C_T = T_T$. (2)

V. EXPERIMENTAL EVOLUTION

In this section, we present the details of our solution, which is based on simple heuristics for the vertical partitioning phases. To vertically partition phase, we follow the Greedy method and greedy method is a method of choosing a subset of the dataset as the solution set that result in some profit.

TABLE 1. ORIGINAL WEB LOG DATASET

Session	Records Sequentially Visited in Session
S_1	$\{P_1, P_2, P_3, P_4, P_5\}$
S_2	$\{P_3, P_2, P_6, P_5, P_7, P_8\}$
S_3	$\{P_1, P_3, P_7, P_4, P_8\}$
S_4	$\{P_1, P_2, P_6\}$
S_5	$\{P_1, P_2, P_3, P_7\}$

In this example the process is illustrated table 1 that contain web log consisting of 5 records is presented. An attacker that knows few query term may easily identify a specific user. For instance, a user request to information about P_2 and P_8, an attack easily identify session S_2 as web history.

We have our data in cluster and we partition in 3 chunks (2 record chunk and 1 term chunk). In this cluster $T_1 = \{P_1, P_2, P_3\}$, $T_2 = \{P_7, P_8\}$ and $T_T = \{P_4, P_5, P_6\}$ Note that the record and term chunks are populated according to equation 1 and 2 respectively.

TABLE 2. RESULT DATASET

Record Chunks		Term Chunk
C_1	C_2	C_T
$\{P_1, P_2, P_3\}$		
$\{P_3, P_2\}$	$\{P_7, P_8\}$	P_4, P_5, P_6
$\{P_1, P_3\}$	$\{P_7, P_8\}$	
$\{P_1, P_2\}$	$\{P_8\}$	
$\{P_1, P_2, P_3\}$	$\{P_7\}$	

Now that the attacker that could identify session S_2 in the original dataset using term P_2 and P_8 can't do so in the result dataset since they don't appear in sub-record together. Additionally, the attacker is not able to identify even a sub-record of the original record S_2 since P_2 occurs in 4 sub-records of the first chunk and P_8 occurs in 3 sub-records of the second chunk.

VI. CONCLUSION AND FUTURE WORK

The major contributions of this paper are a privacy preserving association rule mining algorithm given a secure web mining. Our grand goal is to develop algorithm that can be done at vertical partitioned, while respecting their privacy policies. In this paper, we proposed an algorithm for privacy preserving in distributed level hierarchy in web.

In future we aim to improve our algorithm and implement it in real dataset. Additionally we plan to investigate how to quality of published dataset will be improved.

ACKNOWLEDGMENT

The authors would like to grateful for constructive suggestions and thoughtful comments from reviewers who improved the content of the paper.

REFERENCES

[1] Shrivastava, G., Sharma, K., Kumar, V., "Web Mining: Today and Tomorrow" In Proceedings of the IEEE 3rd International Conference on Electronics Computer Technology, 2011.

[2] Agrawal, R., Imielinski, T., and Swami, A. N., "Mining association rules between sets of items in large databases" In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216. 1993.

[3] Chen, R., Sivakumar, K., and Kargupta, H.,"Distributed Web mining using Bayesian networks from multiple data streams". In Proceedings of the IEEE International Conference on Data Mining. 2001.

[4] Verykios,S.V. Bertino,E. Fovino, I.N. Provenza,L.P. Saygin,Y. Theodoridis,Y. "State-of-the-art in Privacy Preserving Data Mining" ACM SIGMOD Vol. 33 Issue 1, March 2004.

[5] Liu, K. and Ryan, J., "Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining". IEEE Trans. on Knowledge and Data Eng. 18, 1 (Jan. 2006), 92-106.

[6] Agrawal, R., Evmievski, A. and Srikant, R., "Information sharing across private databases" In Proceedings of ACM SIGMOD International Conference on Management of Data, San Diego, CA, June 9-12 2003.

[7] Chris Clifton, Murat Kantarcioglu, Xiaodong Lin, Jaideep Vaidya, and Michael Zhu., "Tools for privacy preserving distributed data mining". SIGKDD Explorations, 4(2):28{34, January 2003

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm (with discussion)". Journal of the Royal Statistical Society, B 39:1{38, 1977.

[9] Murat Kantarcioglu and Jaideep Vaidya. "An architecture for privacy-preserving mining of client information." IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, volume 14, pages 37{42, Maebashi City, Japan, December 9 2002. Australian Computer Society.

[10] Yehuda Lindell and Benny Pinkas. "Privacy preserving data mining". In Advances in Cryptology CRYPTO 2000, pages 36, Springer-Verlag, August 20-24 2000.

[11] Xiaodong Lin, Chris Clifton, and Michael Zhu. "Privacy preserving clustering with distributed EM mixture modeling". Knowledge and Information Systems, 2004.

[12] H. Pang, X. Ding, and X. Xiao, "Embellishing text search queries to protect user privacy," *PVLDB*, vol. 3, no. 1, 2010.

[13] J. Cheng, A.W.-c. Fu, and J. Liu, "K-isomorphism: privacy preserving network publication against structural attacks," in *SIGMOD*, 2010.

[14] K. Liu and E. Terzi, "A framework for computing the privacy scores of users in online social networks," *Data Mining, IEEE International Conference on*, vol. 0, pp. 288–297, 2009.

[15] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," *TKDE*, vol. 16, no. 4, 2004.

[16] Backstrom, L., Dwork, C., and Kleinberg, J., "Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography," in *WWW*, 2007.

[17] M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-preserving Anonymization of Set-valued Data," *PVLDB*, vol. 1, no. 1, 2008.

AUTHORS PROFILE

Gulshan Shrivastava, has obtained a degree of MBA (IT & Finance) and Pursuing M.Tech. in Information Security from Ambedkar Institute of Technology, New Delhi after completing his B.Tech. & Polytechnic in Computer Science from Hindu Society. He has rich experience in teaching the classes of Graduate and Post-Graduate in India and Abroad. He is a Sun Certified Java Programmer. He has been continuously imparting corporate training to the experienced professionals of multinational IT giants in the area of Java Programming & Information Security. He has participated in many National & International Workshop and Technical Fest. He has contributed to numerous International journal & conference publications in various areas of Computer Science. His area of interest includes Java Programming, Website Designing, Data Mining and Information Security.

Dr. Vishal Bhatnagar, Associate-Professor (CSE), has obtained his Ph.d. degree in 2010 and has done M.Tech. (IT) from Punjab University in the year 2005 and completed his B.E. (CSE) from Nagpur University in the year 1999. He has more then 12 years of experience. His area of Interest is Database and Data Mining, Data Warehouse, and application of DWDM in business domain. He joined as an Assistant Professor (CSE) in the department of Computer Science and Engineering in Ambedkar Institute of Technology, Geeta Colony, Delhi. He is currently working as a Associate Professor and HOD (CSE Deptt.) in A.I.T., New Delhi.