International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 263-269
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

AN EFFICIENT FRAMEWORK FOR CLUSTERING HIGH DIMENSIONAL DATA BASED ON SEMANTIC INDEXING

A V L N Sujith
Assistant Professor in CSE,
Vignan Institute of Information Technology,
Vizag

Dr. B Lalitha
Assistant Professor in CSE
JNTUA College of Engineering
Kalikiri

Abstract: In the new era of data processing data mash-up mechanism is emerged as an efficient technique for composing and processing data from several data providers. Software mash-ups are usually associated with software integration achieved at the User-Interface layer. In recent Web applications, this facilitates the usage of Web browser as an additional platform for running interactive Web applications composed form different existing Web sites, applications, services, and information sources Mash-ups are a new kind of interactive service application, built out of the composition of two or more existing service APIs and data sources. The mash up data from multiple sources often contains many data attributes. When enforcing existing reputed privacy models, the high-dimensional data would assist from the problem known as the annoyance of high dimensionality, which resulting an ineffective data for further information analysis. In this paper, we introduced a semantic indexing high dimensional data in regard to achieve anonymity during mash-up to provide the high dimensional privacy to the data and its sources.

Index terms: *Privacy protection, anonymity, data mash up, data integration, service-oriented architecture, high dimensionality, k-anonymity, indentified data*

I. INTRODUCTION

Mash-up service is a web technology that combines in order from multiple sources into a single web application. An in sequence service request could be a common count statistic task or a sophisticated data mining task such as categorization analysis. Upon receiving a service request, the data mash-up web application (mash-up coordinator) dynamically determines the data providers, collects information from them through their web service interface, and then integrates the collected information to fulfill the service request. A data mash-up purpose can help ordinary users search new knowledge; it might also be misused by adversaries to reveal susceptible in sequence that was not accessible before the mash-up. Hence it is obvious to conduct research that proposes novel architecture to achieve security in private data mash-ups.

1.1 The Challenge

Do we really think that our high dimensional, sensitive data and information are completely secured against all the attacks while mashing up the high dimensional data sent over on the unsecured Internet? Will this information still maintain 100% confidentiality and integrity at the reception? Can we be sure completely? These are some question marks which always give a birth to unreliability when we send very bulky and delicate information via an unsecured network.

As Internet has become a part of our daily lives; its harms are as high as its uses. When some data is sent over on the Internet, it must be made secured as Internet has proven itself as the most unsecured network. Sensitive information and data is always the core and important part for any enterprise, organization or an individual working around us, they require and prefer to send their data via secured networks. Such information requires a good level of security while is sent over an unsecured communication channel.

Not only the Internet provides unsecure and unreliable transmission, data can be insecure at the user interface level and in the data storage also. Therefore, the security provision is required at the three levels at the user level, data storage and the communication channel.

II. RELATED WORK

Information integration has been an energetic area of database research [15], [16]. This literature typically assumes that all information in each database can be freely shared [17].Secure multiparty computation (SMC) [18], [19], [20], on the other hand, allows sharing of the computed result (e.g., a classifier), but completely prohibits sharing of data. An example is the secure multiparty computation of classifiers [2], [2], [13]. In contrast, the privacy-preserving data mash-up problem studied in this paper allows data providers to share data, not only the data mining results. In many applications, data sharing gives greater flexibility than result sharing because the data recipients can perform their required analysis and data exploration [8].

Samarati and Sweeney [20] propose the notion of K-anonymity. Datafly system [4] and -Argus system [9] use generalization to achieve K-anonymity. Preserving classification information in K-anonymous data is studied in [8] and [10]. Mohammed et al. [13] extended the work to address the problem of high-dimensional anonymization for the healthcare sector using LKC-privacy. All these works consider a single data source; therefore, data mash-up is not an issue. Joining all private databases from multiple sources and applying a single table anonymization method fails to guarantee privacy if a QIDspans across multiple private tables. Recently, Mohammed et al. [14] propose an algorithm to address the horizontal integration problem, while our paper addresses the vertical integration problem. Jiang and Clifton [6], [7] propose a cryptographic approach and Mohammed et al. [5] propose a top-down specialization algorithm to securely integrate two vertically partitioned distributed data tables to aK-anonymous table, and further consider the participation of malicious parties in [4]. Trojeret al. [3] present a service-oriented architecture for achieving K-anonymity in the privacy-preserving data mash-up scenario. Our paper is different from these previous works [6], [7], [5], [4], [3] in two aspects. First, this LKC-privacy model provides a stronger privacy guarantee than K-anonymity because K-anonymity does not address the privacy attacks caused by attribute linkages. Second, this method can better preserve information utility in high-dimensional mash-up data. High dimensionality is a critical obstacle for achieving effective data mash-up because the integrated data from multiple parties usually contain many attributes.

Enforcing traditional K-anonymity on high-dimensional data will result in significant information loss. Our privacy model resolves the problem of high dimensionality. This claim is also supported by our untried results.

Yang et al. [1] extend a cryptographic approach to learn categorization rules from a large number of data providers while susceptible attributes are protected. The difficulty can be viewed as a horizontally separation data table in which every transaction is owned by a dissimilar data provider. The output of their technique is a classifier, but the output of our technique is an anonymous mash-up data that supports general information analysis or classification analysis.

Jurczyk and Xiong [2], [11] present a privacy-preserving distributed data publishing for horizontally partitioned databases. The mash-up model considered in this paper can be observation as a vertically partitioned information table, which is very different from the model studied in [20], [12], and [1].

Jackson and Wang present a secure communication mechanism that enables cross-domain network requests and client-side communication with the goal of protecting the mash-up controller from malicious code through web services. In compare, this paper aims to preserve the privacy and information effectiveness of the mash-up data. In this regard Benjamin C.M. Fung et al [10] projected a Service-Oriented Architecture for High-Dimensional Private Data Mash-up, which is meant for privacy preserving in data mash-up. They considered the privacy threats caused by data mash-up and suggest a service oriented architecture and a privacy preserving information mash-up algorithm to securely incorporate person specific sensitive data from dissimilar data providers, wherein the integrated in sequence still retains the essential information for supporting common information exploration or a definite data mining task.

In regard to the motivation gained from the service oriented architecture for privacy preserving private data mash-up, here we propose an approach to index semantic observations for privacy preserving mash-up. Providing support to semantic web applications with dynamic data updates. The architecture provides a semantically integrated information space for updated and stored data drawn from heterogeneous autonomous data sources.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 263-269
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

III. SEMANTIC INDEXING BASED CLUSTERING FRAMEWORK FOR HIGH DIMENSIONAL DATA

3.1 The Framework Of Proposed Methodology

Our proposed model "Data anonymity by semantic indexing over High Dimensional Mash-up" consists of two layers; Mash-up layer and Transformation Layer to index and mix. The transformation layer is responsible to chunks the data and indexes to rationalize these chunks during mash-up. Each layer plays its own important role to achieve data anonymity during mash-up process, so the proposed model works efficiently and effectively with high performance and accuracy.

Transformation layer processes the data and converts them into the form various groups under unsupervised learning process and then indexes the data chunks by their semantic relevancy, which further uses to recognize the chunks with high and low risk to mash-up. Mash-up Layer achieves anonymity over mashing up the data chunks along with reducing the high dimensionality to low.

In the model explored here makes its distinction from its original in number of ways, firstly, the use of semantic indexing to recognize the sensitivity of data chunks association. And the most important scalable approach introduced here is achieving anonymity over multidimensional data chunks. This improves the efficiency of the system drastically.

3.2 Detailed design:

The processes explored at transformation Layer carried out at each data source, hence it is specific to each data source, but the activities related mash-up layer is carried out during the process of data mashing, hence it is specific to mash-up service source.

A. The transformation Layer

i. The Data structure

The data at source of services is considered to be structured data and can be expected to be stored in the form of records or XML elements.

ii. Preparing data chunks

The data chunks are said to be the preprocessed partitions of the structured data of each source. The text mining preprocessing steps such as tokenizing, stop word removal and stemming will be done. The resultant set of words of each record or element considered to be as a chunk

iii. Grouping chunks

Supervised learning approach can be used to group these chunks, any of the text mining classification models can be used in this regard.

iv. Indexing Chunks by semantic relevancy

Originally each group annotated by the conditions that are semantically associated to the each chunk of that group. As an example, semantic annotation to a group can be "recognition by zip code", "details of addresses", "details of diseases". The annotation process can be as (i) terms with high coverage of chunks of the same group will be recognized, (ii) The position similarity score of that terms will be measured, (iii) then the terms with high frequency of coverage and position similarity score will be used to annotate the group, (iv) finally the groups with similar annotations will be indexed as semantically relevant.

B. The Mash-up Layer

i. As fig 1 indicates, this layer receives requests from services about the data require to publish or use and collects that from divergent sources.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 263-269
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

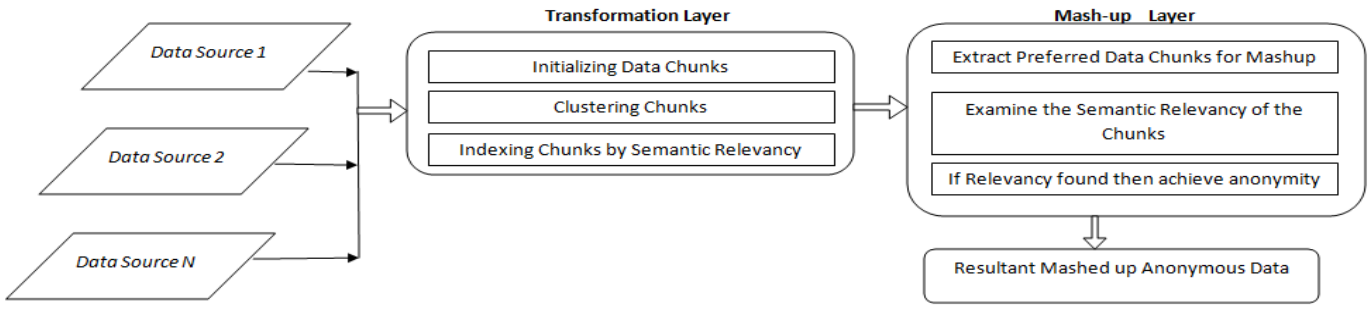

Fig 1: Architecture of the semantic indexing over high dimensional data for mashing up with anonymity

ii. If similarity score found to be greater than the opted sensitivity threshold, then it initiates the process of achieving anonymity as explored in following section.

3.3 Achieving anonymity at high dimension data chunks

In regard to the task of achieving anonymity, we devised an anonymization approach called l-suppression and k-anonymity (lsk-anonymity) which motivated by kactus[7] that is compatible to LKC-Privacy [6], in addition kactus is able to suppressing the dimensionality in the way of achieving scalable anonymity over multi dimensional data chunks. The exploration of the kaCtus[7] follows.

Lsk-Anonymity consists of two main phases: In the first phase, a tree that representing the relations is induced from the original data set; in the second, an algorithm lsk-anonymity developed in this study to anonymize the data set.

Phase 1: Deriving the Tree representation of relations

In this phase, an assessment tree inducer (denoted by ATI) to produce a decision tree denoted by AT is used. The tree can be consequent using various inducers. We focus on top-down univariate inducers which are well thought-out the most popular assessment tree inducers and consist of the well-known algorithms C4.5. Top-down inducers are greedy by quality and build the decision tree in a top-down recursive approach (also known as divide and conquer). Univariate means that the interior nodes are split according to the significance of a single aspect. For detailed exploration of building classification tree is available in [7].

Phase 2: Anonymity Process

In this phase, we use the categorization tree that was fashioned in the first phase to produce the anonymous information set. We think that the categorization tree complies with the subsequent properties:

1. The categorization tree is univariate, i.e., every internal node in the tree refers to precisely one attribute.

2. All internal nodes refer to a quasi-identifier attributes. This is true because the decision tree was trained over the projection of the quasi-identifier set $(\pi_{Q\cup y}(S))$.

3. Assuming a top-down inducer, the characteristics are sorted (from left to right) according to their consequence for predicting the class (where the rightmost transmits to the least significant attribute).

4. Complete Coverage: Each occurrence is associated with precisely one path from root to leaf.

In the subsequently phase, we utilize these properties for the k-anonymity procedure. Given a tree AT and node x, we describe the following functions and procedures. Because these purposes are straightforward, they are worn here without provide pseudo code.

Algorithm K-anonymity (O, Qi, AT, k)

Input: O (Original dataset) , Qi (The quasi-identifier set) ,

k (Anonymity threshold

Output: O' (Anonymous dataset)

1. $AT \leftarrow ATI(\pi Qi \cup TO)$ /* phase 1 */
2. Return Anonymize (OD, Qi, AT, K)

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 263-269
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

/* phase 2 */

Algorithm 1. Anonymize (OD, Qi, AT, K)

- **Input:** OD (Orginal dataset to be anonymized), Qi (Quasi-Identifier set), AT (Classification tree), K (Anonymity Threshold)
- **Output:** O' Anonymity dataset

3. $O \leftarrow OD$ /* original dataset */
4. $O \leftarrow \emptyset$
5. WHILE height (root (AT) > 0)
6. P ← node is AT whose height=1
7. $nui \leftarrow 0$ /* number of uncomplying instances */
8. $sp \leftarrow \phi$ /* candidate instances to be suppressed */
9. $S_{extra} \leftarrow \phi$ /* extra complying instances */
10. For each $v \in children(p)DO$
11. IF $|\sigma ant(v)(S)| \geq K$ Then
12. $SV \leftarrow randomSelect$ $(|\sigma ant(v)(S)|) - k, ant(v)(S)$)
13. $S_{extra} \leftarrow S_{extra} \cup SV$
14. $SP \leftarrow SP \cup (\sigma ant(v)(S) - SV)$
15. ELSE
16. $nui \leftarrow nui + |(\sigma ant(v)(S)|$
17. End IF
18. End FOR
19. IF $nui < K$ THEN
20. $required \leftarrow K - nui$
21. IF $|S_{extra}| \geq$ required THEN
22. $S_{notrequired} \leftarrow randomselect$ $(|S_{extra}| - required, S_{extra})$
23. ELSE
24. $S_{notrequired} \leftarrow S_{extra}$
25. END IF
26. $SP \leftarrow SP \cup S_{notrequired}$
27. End IF
28. $suppressComplyingChildern$ (S', SP, Q, P)
29. $S \leftarrow S - SP$
30. Prune (CT,P)
31. END WHILE
32. IF $|S| \geq K$ THEN $s' \leftarrow s' \cup \sup press(s, Q, \phi)$

Algorithm 2. Suppress $(R, Q, \Pr ed)$

Input: R (Dataset), Q (The quasi-identifier set), v (Predicates)

Output: R' (Suppressed dataset)

33. $R' \leftarrow R$
34. for each $a \in Q$ do
35. If a does not appear in the antecedent in v
36. $\delta a \rightarrow '\iota'(R')$
37. ENF IF
38. END FOR
39. Return R'

Algorithm3.

SuppressComplyingChildern (S', S, Q, P)

Input: S' (anonotmous dataset), S (original dataset), Q (Quasi-identifier), P (Parent node)

40. FOR each $v \in childern(p)$ DO
41. IF $(|\sigma ant(v)(S)|) \geq k$ THEN
42. $sv \leftarrow \sigma ant(v)(S)$
43. $s' \leftarrow s' \cup \sup press(sv, Q, ant(v))$
44. END IF

End FOR

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 8, ISSN No. 2455-2143, Pages 263-269
Published Online June - July 2016 in IJEAST (http://www.ijeast.com)

IV. EXPERIMENTS AND PERFORMANCE ANALYSIS

The proposed semantic indexing over high dimensional data (SIO-HDD) for mashing up with anonymity has been tested in distributed environment build by using RMI under java run time environment. The considered distributed environment is a multi party service structure such that each party capable to provide data. The proposed semantic indexing performed at each party and the dimension suppression and achieving anonymity is performed at mash-up service provider.The objective is to evaluate the benefit of data integration without losing its privacy by SIO-HDD. The experiments conducted to verify the scalability in view of anonymity scope, breakability, Informative without losing generality. The same compared with PHDMashup [6] model that considered as motivation to our SIO-HDD. The fig 3 and 4 indicates the advantage of SIO-HDD over PHDMashup, which is due the semantic indexing and KACTUS. The performance analyzed by comparing the true positives of the anonymity breaching and understandability. The scope of true positives observed under breaching the anonymity can be explored by fig 2. The understandability of anonymized data scaled in fig 3.

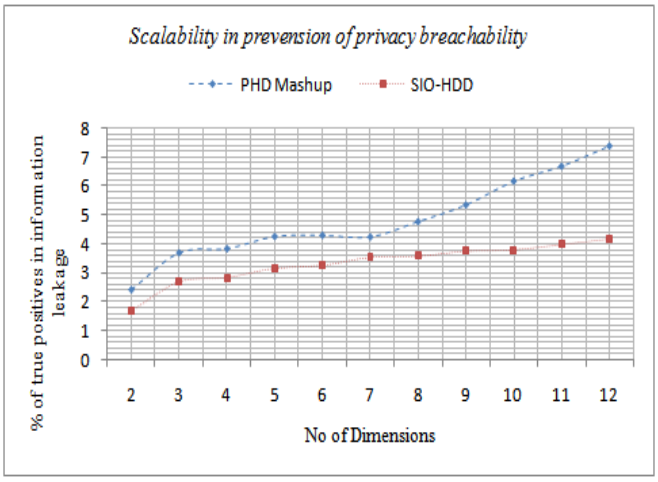

Fig 2: Privacy preserving Scalability of SIO-HDD over PHDMashup

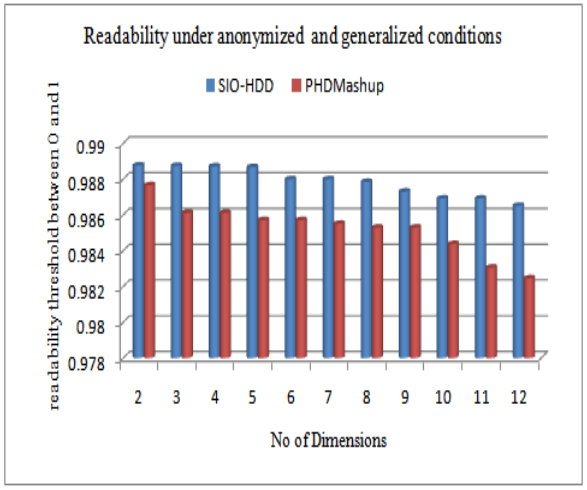

Fig 3: Readability of the mashed up under anonymized and generalized data by SIO-HDD and PHDMashup

V. CONCLUSION

Here we proposed and devised a semantic indexing and anonymization approach to achieve stability and scalability towards achieving anonymity with high readability of high dimensional mash up data. In this regard here we devised an ontology based semantic indexing approach and dimensionality suppression and anonymity (lsk-anonymity) algorithm, which is influenced by kACTUS [7]. The performance analysis indicating that the proposed SIO-HDD is scalable and stable over PHDMashup. In future this approach can be enhanced to achieve anonymity with novel perturbation techniques.

VI. REFERENCES

[1] P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers,"Proc. 23rd Ann. IFIP WG 11.3 Working Conf. Data and Applications Security (DBSec),2014.

[2] P. Jurczyk and L. Xiong, "Privacy-Preserving Data Publishing for Horizontally Partitioned Databases,"Proc. 17th ACM Conf. Information and Knowledge Management,Oct. 2015.

[3] T. Trojer, B.C.M. Fung, and P.C.K. Hung, "Service-Oriented Architecture for Privacy-Preserving Data Mashup,"Proc. IEEE Seventh Int'l Conf. Web Services,pp. 767-774, July 2009.

[4] N. Mohammed, B.C.M. Fung, and M. Debbabi, "Anonymity Meets Game Theory: Secure Data Integration with Malicious Partici-pants,"Int'l J. Very Large Data Bases,vol. 20, pp. 567-588, 2014.

[5] N. Mohammed, B.C.M. Fung, K. Wang, and P.C.K. Hung, "Privacy-Preserving Data Mashup,"Proc. 12th Int'l Conf. Extending Database Technology (EDBT),pp. 228-239, Mar. 2015.

[6] Fung, B.C.M.; Trojer, T.; Hung, P. C K; Li Xiong; Al-Hussaeni, K.; Dssouli, R., "Service-Oriented Architecture for High-Dimensional Private Data Mashup," Services Computing, IEEE Transactions on , vol.5, no.3, pp.373,386, Third Quarter 2012; doi: 10.1109/TSC.2011.13

[7] Slava Kisilevich, Lior Rokach, Yuval Elovici, and Bracha Shapira. 2010. Efficient Multidimensional Suppression for K-Anonymity. IEEE Trans. on Knowl. and Data Eng. 22, 3 (March 2010), 334-347. DOI=10.1109/TKDE.2009.91 http://dx.doi.org/10.1109/TKDE.2009.91

[8] 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.

[9] V.S. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc. Eighth ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD), pp. 279-288, July 2002.

[10] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization,"Proc. 12th ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD),Aug. 2006.

[11]C.C. Aggarwal, "Onk-Anonymity and the Curse of Dimension-ality,"Proc. 31st Very Large Data Bases,pp. 901-909, 2015.

[12]B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments,"ACM Computing Surveys,vol. 42, no. 4, pp. 14:1-14:53, June 2010.

[13]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. 15th ACM SIGKDD Int'l Conf. Knowl-edge Discovery and Data Mining,pp. 1285-1294, June 2009.

[14] N. Mohammed, B.C.M. Fung, P.C.K. Hung, and C.K. Lee, "Centralized and Distributed Anonymization for High-Dimen-sional Healthcare Data," ACM Trans. Knowledge Discovery from Data,vol. 4, no. 4, pp. 18:1-18:33, Oct. 2010

[15] A. Jhingran, "Enterprise Information Mashups: Integrating In-formation, Simply," Proc. 32nd Int'l Conf. Very Large Data Bases, pp. 3-4, 2006.

[16] G. Wiederhold, "Intelligent Integration of Information,"Proc. ACM Int'l Conf. Management of Data (SIGMOD),pp. 434-437, 1993.

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

[18] O. Goldreich,Foundations of Cryptography: Vol. II Basic Applications. Cambridge Univ. Press, 2004.

[19] Y. Lindell and B. Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining,"J. Privacy and Confidentiality, vol. 1, no. 1, pp. 59-98, 2009.

[20] A.C. Yao, "Protocols for Secure Computations,"Proc. 23rd Ann. Symp. Foundations of CS,pp. 160-164, 1982.