Safeguarding Solitude for Outsourced Data Mining

Naveen N [1]
Research Scholar
Department of Computer Science & Engineering
VTU-RRC, Belagavi

Dr. K. Thippeswamy [2]
Professor
Department of Computer Science & Engineering
VTU-PG Centre Mysore

Abstract – The intense popularity of data mining strategies has amplified its demand across the information-oriented domain and has elevated its level manifold. Techniques such as the Data-Mining-as-a-service (DMaaS) prototype has been focused upon, which mainly emphasizes upon Privacy and security issues. Herein the Data owner prevents sharing of secured data with rest of the data owners or servers. A cloud-assisted secrecy secured data mining elucidation has been recommended for data off shoring that involves minimum peculation of crude data.

Index Terms - DMaaS, secrecy secured, semi-trusted, encryption

I. INTRODUCTION

There has been a significant rise in the digital data production which is being appropriately recorded according to [1]. Cloud computing has become the prime factor behind the upliftment of data mining and management since it offers the advantage of computation and flexible storage. Generally, it has become a common practice by the data owners to allocate the task of data mining task to the cloud service provider-server which in turn greatly minimizes the cost of storage, computation as well as management. With the help of Data mining, analysis of knowledge discovery can be effectively achieved in voluminous amount of data. Furthermore, it also helps in drawing out interesting patterns and knowledge from such data. Definitely, there lies ceaseless demand of the data mining techniques in almost all domains. The ARM technique helps in analysis of raw data for fetching significant patterns in the day-to-day business as mentioned by [2],[3]. This being referred to as market basket analysis by [5], [6]. In order to withdraw precise information or liaison rules from abundant data, there is requirement of hardware support, systematic data mining methods, human skills along with IT competency. Amazon S3, Microsoft Azure, and Google AppEngine are certain cloud service providers.

These CSPs offer services that are cost-effective as well as user-friendly. Moreover, the DMaaS prototype inherit various secrecy and sanctuary cases for economically evaluating voluminous data. The present research has pointed out that the Cloud/server are partially trustworthy. That is, the owner's secured data becomes accessible without the permission of the data owners after the data gets released to the server. Breaching of this particular security issue is known as the corporate privacy.

The sections presented ahead cover up assessment of certain privacy preserving techniques along with various popular and essential outsourced data mining approaches. Section II presents prior work associated with the study. Section III offers solution pertaining to the privacy preservation while outsourcing data mining. Section IV covers up the overall study analysis and the conclusion is covered up in Section V.

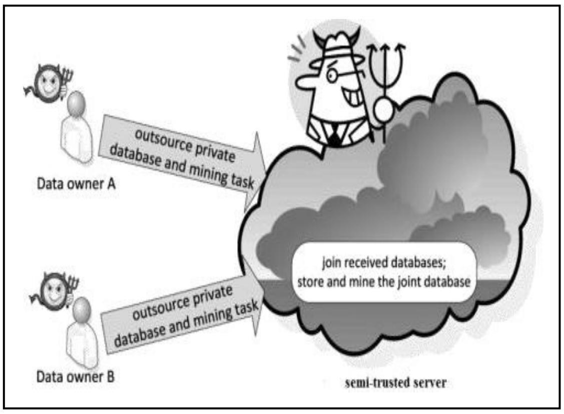
Data owner A
outsource private database and mining task
Data owner B
outsource private database and mining task
join received databases; store and mine the joint database
semi-trusted server

Figure 1: System model of Outsourced Data Mining

II. RELATED WORK

The concept of Privacy preservation has been of utmost significance over the last 20 years and there has been recommendation of enormous strategies concerning the same as pointed out by [4], [5], [6]. Lately, Privacy preservation in offshoring data mining frameworks has become a popular topic of study as witnessed by [2], [7]. The proposed infusion is of two types:

The first method involves transformation of raw data by inserting certain arbitrarily noise in such a way that the sensitive information remains hidden and at the same time the statistical data attributes remains preserved. Though the data mining precision gets impacted with such a method. The second method is adopted purely in distributed environment wherein multiple parties willingly share their private data in order to achieve the relevant output. Though such methods don't hold optimal for outsourcing.

There has been proposal of diverse secrecy-securing association rule mining methods. Herein, the data owner acts as the master and the rest behave as slaves. The slaves are responsible to insert certain fake data to their private datasets and then forward it to the master. In addition, a set of actual transaction's IDs are also being sent to the semi-trusted master. The master then mines the association rules from the joint database that holds the fake transactions. It effectively works towards fetching of sensitive information from the datasets even though there exist significant noise for minimizing the practicality of data. There exists possibility of variation in data mining precision. [9], [10] has adopted the asymmetric homomorphic encryption for computing supports of item sets. On the other hand [11], [12], [13] incorporates a secure scalar product protocol or a secret sharing technique pertaining to such computations. These techniques ensure precise depiction of element set to outright the data owners. There is no significant support through the recommended techniques by [10] which makes the computation complex.

III. PROPOSED SOLUTION

Solution for outsourced data in association rule mining involves the pre-processing and mining process. Preprocessing measure works at the client position.

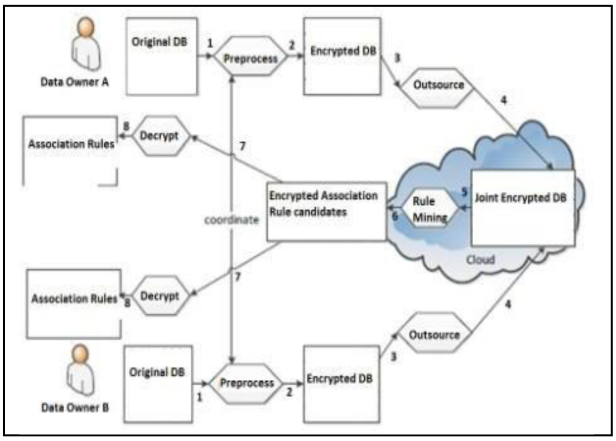

Figure 2: Proposed System Architecture of Privacy-preserving association rule mining

Initially, every data owner anonymizes a single item through the MD5 algorithm. Record of each item is enlisted with the matching message digest value. Using the RC4 symmetric encryption algorithm, there is encryption of all the customer IDs. After implementation of anonymization and encryption on original database, every data owner outsources his encrypted database to the third party (server).

Table 1: Original Database (Before Pre-processing)

Customer ID	Item ID
10001	bread, butter
10002	milk, bread
10003	milk
10004	milk, butter, bread
10005	beer
10006	bread, butter, beer
10007	milk, butter
10008	milk

Table 2: Pre-processed Database for Outsourcing

Customer ID	Item ID
Mg9BBRnC	d131dd02c5e6eec4, 693d9a0698af95c
MqNBDhmA	ae6dacd436c919c6, d131dd02c5e6eec4
NwxNJAnQ	ae6dacd436c919c6
MA5BDhZC	ae6dacd436c919c6, 693d9a0698af95c, d131dd02c5e6eec4
Zp8EDhmA	4004583eb8fb7f8
Nk7xYAnQ	d131dd02c5e6eec4, 693d9a0698af95c, 4004583eb8fb7f8
Bq9SSRnF	ae6dacd436c919c6, 693d9a0698af95c
Mx3BDhUL	ae6dacd436c919c6

Entire Computation activities in Mining takes place at the server end. Entire set of individual encrypted databases are merged into a joint database. Processing of the joint database is carried at the server end via distributed file system namely, Hadoop. HDFS and a frame-work analyses and transforms voluminous data sets by employing the Map Reduce framework.

Map Reduce is the data processing framework enabling the applications to store files in a distributed environment such as the HDFS. It involves two functions namely, a Mapper, and a Reducer which are executed on the nodes in the cluster. Initially, the joint encrypted database is read as input by the Map Reduce code from the HDFS storage. Each item is then processed as a separate key for computing the frequent 1-itemset. Thereafter, pair of items from the frequent 1-itemset helps in generating the frequent 2-itemsets. The process is repeated till every number of repetitions are identified for the required number of item sets. Following is the framework for Map Reduce:

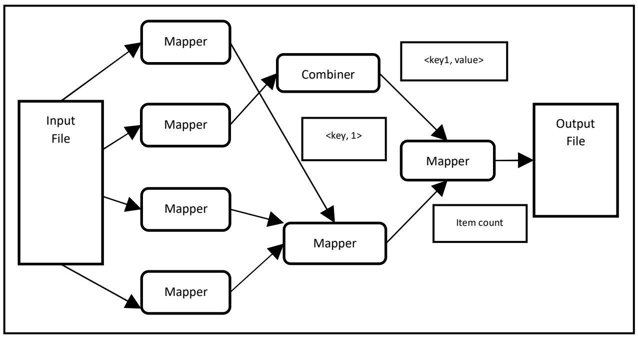

Figure 3: Map Reduce Framework

1. **RC4 Encryption:** Represents one of the symmetric key algorithms where data stream will be XORed with the generated key sequence. 40 and 128-bit keys have been utilized. This algorithm is in various commercial software packages like the Lotus Notes and Oracle Secure SQL. RC4 Encryption is nearly 10 times faster than DES.

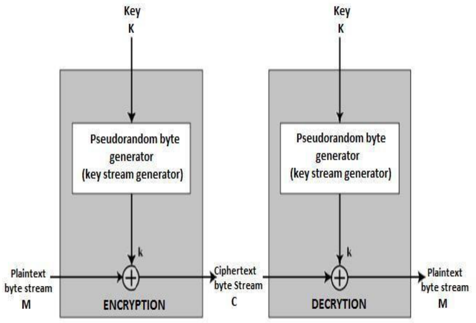
Key
K
Pseudorandom byte generator (key stream generator)
k
Plaintext byte stream M
ENCRYPTION
Ciphertext byte Stream C
Key
K
Pseudorandom byte generator (key stream generator)
k
DECRYTION
Plaintext byte stream M

Figure 4: RC4 Encryption

2. **MD5 Algorithm:** Represents one-way cryptographic function which takes a message of arbitrary length as input and the output is in the form of a fixed- length (128 bit) digest value that is employed for authentication of the original message.

3. **Hadoop Distributed File system**: The Hadoop Distributed File System (HDFS) is designed for storing very large data sets reliably. In a big cluster, thousands of servers are able for both hosting directly attached storage and executing user applications. Hadoop cluster scales storage capacity, computation capacity, and I/O bandwidth up by simply adding more commodity servers and distributing computation and storage among many servers.

IV. ANALYSIS

Here there is analysis of the prevailing approaches, their assumptions, results and restrictions along with the analysis of the proposed solution.

Table 3: Comparison of previously proposed solutions

Approach	Assumption	Result	Limitation
Master-slave based	Master node is semi-trusted	To minimize the usability of data all slave nodes insert some fictitious transactions to the dataset	Not reliable; because only master node is responsible for all the computations
Substitution Cipherbased	Third party Server is semi-trusted	All the raw data is encrypted prior to outsourcing	Frequency-analysis attack is possible
K-anonymity based	Third party Server and other data owners are semi-trusted	Exact support of items are hidden by inserting some fictitious transactions tothe dataset	Fake transactions may affect data mining precision.

Privacy Analysis -Against Server's Attacks: MD5 algorithm is adopted for combating the attack from semi trusted servers. This one-way function is crypto-graphically highly secure that generates arbitrary 128-bit digests values. In the transaction database, each item gets converted into their matching digest values without disclosing the information pertaining to the raw data. Information pertaining to the actual items is contained solely with the data owners though no information can be generated at the server end that matches a particular digest value.

Against Data Owner's Attacks: Even though the data owners join together to mine their datasets but still the data owners remain inquisitive for seeking information of remaining data owners. For resolving this issue, all the data owner encrypts the customer IDs with RC4 encryption method via their own private key. Hence, its not possible for the data owner to decrypt customer information of other data owners. Once the mining results are fetched from the server, it gets decrypted by the data owner for obtaining the original results.

V. CONCLUSION

The research proposes a privacy-preserving technique pertaining to the outsourcing data mining scenarios. With such techniques, multiple data owners can outsource their data mining tasks in a privacy-preserving way. The proposed technique helps in safeguarding the data owner's raw data from the third party and other data owners too. Both the performance as well as privacy-preservation is enhanced. This technique is most optimal for the multi-party scenario with no compromising on performance.

REFERENCES

[1] Vaidya, J. and C. Clifton. Privacy-preserving data mining: why, how, and when. IEEE Security and Privacy, 2(6):1927, 2004.

[2] Qiu, L., Y. Li, and X. Wu. An approach to outsourcing data mining tasks while protecting business intelligence and customer privacy. In Work-shops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), pages 551558, Hong Kong, China, December 1822, 2006.

[3] Molloy.L, N. Li, and T. Li, On the (in)security and (im)practicality of outsourcing precise association rule mining, in Proc. ICDM, Dec. 2009, pp. 872877.

[4] Agrawal, R., A. Evmievski, and R. Srikant. Information sharing across private databases. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Database, pages 8697, San Diego, California, 2003.

[5] Agrawal, S. and J. R. Haritsa. A framework for high-accuracy privacy preserving mining. In Proceedings of the 21th IEEE International Conference on Data Engineering (ICDE 2005), pages 193204, Tokyo, Japan, 2005.

[6] Kargupta, H., S. Datta, Q. Wang, and K. Sivakumar. Random data perturbation techniques and privacy-preserving data mining. Knowledge and Information Systems: an International Journal, 7(4):387414, 2005.

[7] Qiu, L., Y. Li, and X. Wu. Protecting business intelligence and customer privacy while outsourcing data mining tasks. Knowledge and Information Systems: an International Journal, November 16 2007.

[8] Rozenberg.B and E. Gudes, Association rules mining in vertically partitioned databases, Data Knowl. Eng., vol. 59, no. 2, pp. 378396, 2006.

[9] Zhan.J, S. Matwin, and L. Chang, Privacy-preserving collaborative association rule mining, in Proc. DBSEC, 2005, pp. 153165.

[10] Zhong.S, Privacy-preserving algorithms for distributed mining of frequent itemsets, Inf. Sci., vol. 177, no. 2, pp. 490503, 2007.

[11] Vaidya.J and C. Clifton, Secure set intersection cardinality with application to association rule mining, J. Comput. Secur., vol. 13, no. 4, pp. 593622, 2005.

[12] X. Ge, L. Yan, J. Zhu, and W. Shi, Privacy-preserving distributed association rule mining based on the secret sharing technique, in Proc. SEDM, Jun. 2010, pp. 345350.

[13] R. Kharat, M. Kumbhar, and P. Bhamre, Efficient privacy preserving distributed association rule mining protocol based on random number, in Intelligent Computing, Networking, and Informatics. Raipur, Chhattisgarh, India: Springer, 2014, pp. 827836

Authors Profile

Naveen N. Currently working as Assistant Professor in Department of Information Science & Engineering at Kalpataru Institute of Technology, Tiptur. Pursuing PhD degree in Computer Science and Engineering from VTU, Karnataka, India

Dr. K. Thippeswamy. Currently working as Professor in CSE at VTU Centre for PG Studies, Mysore, Karnataka, India. His research interest includes Data Mining & Cloud Computing.

.