ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Enhanced Security for Two Party Confidential Data

G. Sumathi[*1], M.Indumathi[2]

[*1] Assistant Professor, Dept Of Computer Science And Engineering, [2]Muthayammal Engineering College, Namakkal, India

Abstract

Data mining is the process of extracting hidden predictive information from large databases to support decision making process. There may be a chance of revealing sensitive information stored in dataware house during extraction of hidden details. To protect this consequence a method is proposed to securely integrate person-specific sensitive data from two data providers, whereby the integrated data still retains the essential information for supporting data mining tasks. Most real-life scenarios are in need for simultaneous data sharing and privacy preservation of person sensitive data. Enhanced security for two party confidential data adopts differential privacy, provides a rigorous privacy model and makes no assumption about an adversary's background knowledge. Two party algorithm is used in existing system for private data release for vertically-partitioned data between two parties in the semi-honest adversary model. A differentially-private mechanism is proposed to ensure that the probability of any output is equally likely from all nearly identical input data sets and thus guarantees that all outputs are insensitive to any individual's data. Updating sensitivity scores is done to set scores for the sensitive details. Sha-2 algorithm is used for the purpose of user authentication to avoid masquerade attacks.

Keywords: Differential privacy, Two party algorithm, Sensitive score updation, Sha-2 algorithm.

Introduction

There is a great demand for database due to the advancement in information exchange, data storage and retrieval. Each organizations/institutions hold their own database future references for , e.g., student records by colleges/schools, customers sopping details by super market, loan and financial details by banks, and etc..

Due to the enhancement in computer science new paradigms such as cloud computing created a great demand for integrating data between different database holders. These integrated data enable better data analysis for making better decisions and providing high-quality services. For example, data can be integrated to improve scientific research, extended services for users. Data integration should be done in such a way that no sensitive details of an particulars is revealed *i.e* no individual records should be at risk due to participating in dataset. And another thing to be considered At the same time, new knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration. An algorithm is proposed to securely integrate person-specific sensitive data from two data providers, whereby the integrated data still retains the essential information for supporting data mining tasks.

Related Theory

K-ANONYMOUS DATA MINING[2]- Hillol Kargupta And Souptik.

Privacy preserving data mining techniques clearly depend on the definition of privacy, which captures what information is sensitive in the original data and should therefore be protected from either direct or indirect (via inference) disclosure. K-Anonymous presents a specific aspect of privacy that has been receiving considerable attention recently, and that is captured by the notion of k-anonymity. K-anonymity is a property that models the protection of released data against possible re-identification of the respondents to which the data refer. Intuitively, k-anonymity states that each release of data must be such that every combination of values of released attributes that are also externally available and therefore exploitable for linking can be indistinctly matched to at least k respondents. K-anonymous data mining has been recently introduced as an approach to ensuring privacy preservation when releasing data mining results.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

K-ANONYMITY

K-anonymity is a property that captures the protection of released data against possible re-identification of the respondents to whom the released data refer. Consider a private table PT, where data have been deidentified by removing explicit identifiers (e.g., SSN and Name). However, values of other released attributes, such as ZIP, Date of birth, Marital status, and Sex can also appear in some external tables jointly with the individual respondents' identities. If some combinations of values for these attributes are such that their occurrence is unique or rare, then parties observing the data can determine the identity of the respondent to which the data refer or reduce the uncertainty over a limited set of respondents. k-anonymity demands that every tuple in the private table being released be indistinguishably related to no fewer than k respondents. Two main techniques have been proposed for enforcing k-anonymity on a private table:

- Generalization
- Suppression

Generalization consists in replacing attribute values with a generalized version of them. Generalization is based on a domain generalization hierarchy and a corresponding value generalization hierarchy on the values in the domains. Typically, the domain generalization hierarchy is a total order and the corresponding value generalization hierarchy a tree, where the parent/child relationship represents the direct generalization/specialization relationship. Generalization can be applied at the level of single cell (substituting the cell value with a generalized version of it) or at the level of attribute (generalizing all the cells in the corresponding column).

RANDOMIZATION METHODS FOR PRIVACY PRESERVING DATA MINING[3] - CHARU C. AGGARWAL, PHILIP.

In the randomization method, noise is added to the data in order to mask the values of the records. The noise added is sufficiently large so that the individual values of the records can no longer be recovered. However, the probability distribution of the aggregate data can be recovered and subsequently used for privacy-preservation purposes.

The earliest work on randomization describes that, it has been used in order to eliminate evasive answer bias. In it has been shown how the reconstructed distributions may be used for data mining. The specific problem which has been discussed in is that of classification, though the approach can be easily extended to a variety of other problems such as association rule mining.

Problem Definition

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. Data mining at individual sites will be unable to detect cross-site correlations.
- The same item may be duplicated at different sites, and will be over weighted in the results.
- Data at a single site is likely to be from a homogeneous population, hiding geographic or demographic distinctions between that population and others.

Algorithms have been proposed for distributed data mining. Distributed classification has also been addressed. A meta-learning approach has been developed that uses classifiers trained at different sites to develop a global classifier. This could protect the individual entities, but it remains to be shown that the individual classifiers do not disclose private information. Recent work has addressed classification using Bayesian Networks in vertically partitioned data, and situations where the distribution is itself interesting with respect to what is learned.

Proposed System

Differential privacy is a recent privacy definition that provides a strong privacy guarantee. It guarantees that an adversary learns nothing more about an individual, regardless of whether her record is present or absent in the data. Even when data collectors try to protect the privacy of their customers by releasing anonymized or aggregated data, this data often reveals much more information than intended.

To reliably prevent such privacy violations, need to replace the current ad-hoc solutions with a principled data release mechanism that offers strong, provable privacy guarantees. Recent research on differential privacy has brought us a big step closer to achieving this goal. Differential privacy allows us to reason formally about what an adversary could learn from released data, while avoiding the need for many assumptions (e.g. about what an adversary might

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

already know), the failure of which have been the cause of privacy violations in the past. However, despite its great promise, differential privacy is still rarely used in practice. Proving that a given computation can be performed in a differentially private way requires substantial manual effort by experts in the field, which prevents it from scaling in practice.

Differential privacy offers a way to answer queries about sensitive information while providing strong, provable privacy guarantees, ensuring that the presence or absence of a single individual in the database has a negligible statistical effect on the query's result. Proving that a given query has this property involves establishing a bound on the query's sensitivity how much its result can change when a single record is added or removed. Differential privacy offers strong statistical privacy guarantees for certain types of queries, even in worst-case scenarios. Intuitively, this is accomplished by

a) admitting only queries whose result does not depend too much on the data of any single individual, and

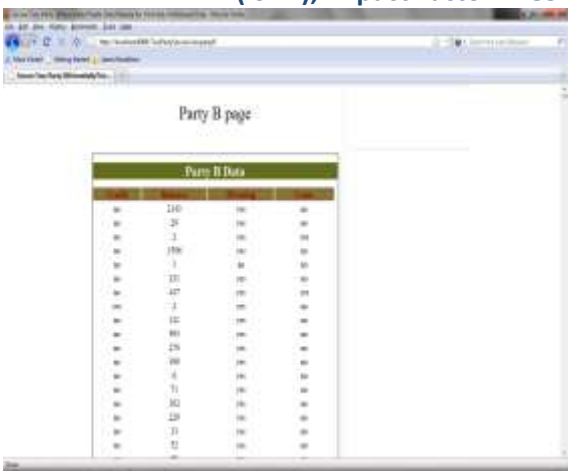
b) adding some random noise to the result of each query.

Thus, if any individual record is picked and all of individual's data are released from the database before answering a query, the probability that the result is any given value remains almost the same. This limits the amount o

information that can be learned about a single individual by observing the result of the query.

Modules

Dataset collection



Two party algorithm implementation

Data between two parties where integrated by using shared identifier such as ssn, name, employee id. Integrated data is preprocessed ie.. removing all the explicit identifiers such as name,age,etc.. but there may be a existence of pseudo identifiers which may lead to link attack. Integrated data gets generalized to hide the sensitive details. Owner of the data generalizes the details by assuming some of the field as sensitive. Hence security is satisfied statistically. A method is proposed to provide dynamic security called differential privacy which does not assume about adversaries background knowledge.

User authentication

Sha algorithm is used for authentication purpose. User key is generated using sha algorithm and sent to the registered mail address. If he is an valid user key will be sent to the user else key will not be sent. Advantage by using sha, it was one way hash function it cannot be decrypted.

Sensitive score updating

User will request the needed data, this requested data will be sent to the admin having the data set. Administrator can decide which data can be provided to the user. Based on the user requests all the details of the user will be viewed by the admin such

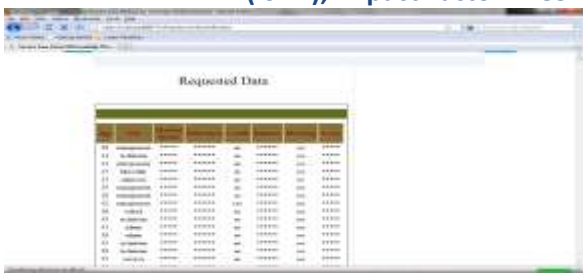
ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

as job, education. Admin use to decide which data can be provided to the third party based on their particular details.

For ex. If the user is an employee of income tax department, he is requesting salary detail of the user admin will provide the data by setting the sensitivity score less than 5. If adversaries needs any sensitive details then admin can set the sensitivity score greater than 5.

Sharing data to users

Requested data get shared to the users, if the user is authorized to view all the data. If admin does not want to provide all the data to the user noise is added to the final results.



Conclusion

Data integration between autonomous entities is conducted in such a way that no more information than necessary is revealed between the participating entities. At the same time, new knowledge that results from the integration process is not misused by adversaries to reveal sensitive information that was not available before the data integration. Two party private data release algorithm for vertically-partitioned data is used in existing system, where the data is presented to the user in generalized manner and it provides static security and the generalized data may reveal some sensitive information to the adversaries. Differential private data release mechanism is proposed it provides dynamic security to the original data by adding noise to the sensitive details. For user authentication sha algorithm is used to ensure data integrity. Hence proposed method provides similar data utility compared to the recently proposed algorithms and it also supports in effective decision making process. This can be further enhanced by implementing algorithm for more than two parties.

References

[1] *Noman Mohammed, Dima Alhadidi, Benjamin C. M. Fung, and Mourad Debbabi, "Secure Two-Party Differentially Private Data Release for Vertically-Partitioned Data", 2013.*

[2] *Hillol Kargupta And Souptik Data, Qi Wang And Krishnamoorthy Shivakumar,"Random data perturbation techniques and privacy preserving data mining", IEEE international conference on Data mining, 2003.*

[3] *Charu C. Aggarwal, Philip,"A Survey of Randomization Methods for Privacy-Preserving Data Mining", 2007.*

[4] *Josep Domingo-Ferrer, "Inference Control Methods for Privacy-Preserving Data Mining", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005).*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

[5] Noman Mohammed, Rui Chen, Benjamin C. M. Fung, Philip S. Yu, "Differentially Private Data Release for Data Mining", KDD'11, August 21–24, 2011, San Diego, California, USA. Copyright 2011 ACM.

[6] Roberto J. Bayardo, Rakesh Agrawal,"Data Privacy Through Optimal k-Anonymization", Proceedings of the 21st International Conference on Data Engineering (ICDE 2005).