Hierarchical Model for Preserving Privacy in Vertically Partitioned Distributed Databases

Rayudu Srinivas
Professor
SSAIST
Surampalem

A. Sri Rama Chandra Murthy
SSAIST
Surampalem

ABSTRACT

Data sharing is an important task in the world of globalization. Organizations, government offices and hospitals release data to the public for the purpose of analysis and research. While publishing data to public, privacy is to be provided to the sensitive information. Privacy preservation aims at limiting the risk of leaking published data to a particular individual. To provide privacy copious privacy techniques are available, sophisticated privacy preserving techniques are required for distributed data.
In distributed databases the data is distributed over the nodes. The data distribution may be horizontally partitioned or vertically partitioned. To provide privacy to the horizontally partitioned data various techniques proposed in [7,8] and for vertically partitioned data the method proposed in [15] if the data stored at two sites. Here we are proposed a novel method to provide privacy to the data if the data is vertically partitioned and distributed over sites. The major advantage of proposed method is to reduce transmission load and to provide privacy using multidimensional distributed method that is similar to the most popular method Mondrian multidimensional k-anonymity[5].

Keywords

Keywords: Distributed data, vertically partitioned data, Privacy, anonymous communication, Mix, Hierarchical Model, PPDM, anonymity

1. INTRODUCTION

An abundant information regarding individuals referred to as micro data that is published for research and analysis purpose. Three types of micro data attributes are related to privacy preservation. These are 1) sensitive attributes 2) quasi identifiers and 3) identifiers. Numerous methods were proposed to provide privacy to data. One of the approaches to provide privacy to the sensitive data is by removing identifiers from the data before publishing, But in some applications set of attributes are collectively used to identify the individual. In order to minimize record leakage, various techniques such as k-anonymity, suppression, generalization perturbations, permutation and swapping of certain values are used to meet privacy principles. Privacy preserving data publishing for a single database has been extensively used in recent years. To provide privacy for a single database, various methods were proposed. Among this techniques Mondrian multi dimensional k-anonymity became more popular.

The data sets may either be horizontally partitioned or vertically partitioned. In horizontal partitioned dataset, the individual records spread out across multiple entities, each of which has the same set of attributes. In vertical partitioned data set, the individual entities may have different attributes of the same record.

In societies throughout history, anonymity has always been a pervasive, dichotomous issue. Some believe anonymity is very essential in protecting privacy and freedom of expression while other believes that anonymity is superfluous and only encourages the propagation of dubious dogma as well as abusive, illegal activity

Anonymity and privacy are increasingly important issues. Individuals and organizations may prefer a certain degree of anonymity in three key activities those are Web-browsing, message-sending, and file-sharing over ubiquitous distributed environments [9]

Privacy plays a key role in distributed environment but providing privacy to the data in such environment is hard. To provide privacy to the distributed databases numbers of approaches are in use. A naïve approach is proposed in this paper. We assume that the data is vertically partitioned and stored at various sites; at least one common attribute has to be used among various attributes for performing join operations.

While data is transmitting, the information stored at individual nodes should not be revealed to other nodes. To achieve this objective here we are proposing novel approach. We assume that the nodes are semi trusted. All the sites participating in the query follows the instruction given by the leader. Leader is one among the nodes which is participated in query processing. The leader initially sends an encryption key to all the nodes. When the nodes in the model receives key, each individual site encrypts their data except common attributes used for join and sends this encrypted data to next level in the model. Each node in the model receives records from lower level and performs join operation then the result is forwarded to next level. The leader node is at the top level and receives data from its decedents and performs join operation and then decrypts the data. Once data is decrypted, data can be used for multidimensional k-anonymity. After performing Mondrian k-anonymity the result can be published to public.
The structure of the model is given in fig1.

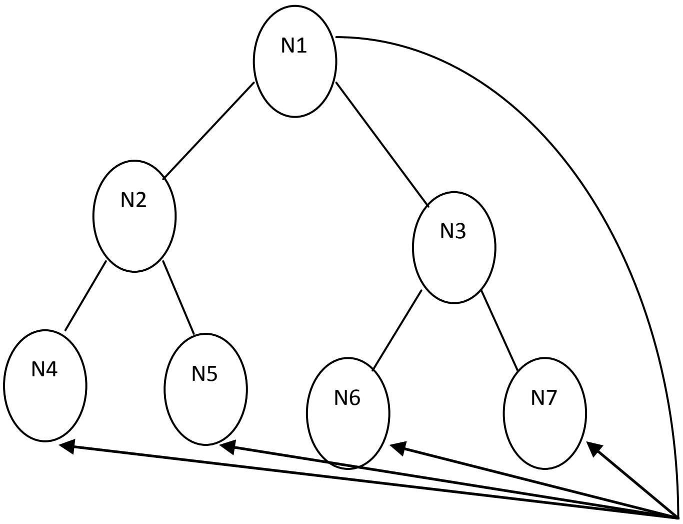

Fig 1.

2. METHODOLOGY

Binary trees consist of at most two children. A complete Binary tree structure can be used for connecting the nodes of a network. We assume that the nodes are connected randomly in the tree structure. When a node in the network receives a query from user, broadcast the query to the other nodes in the network.

The leader site receives the query from user and this query is broadcasted to all the other nodes in the network. Each node in the network receives the query and the results of this query are used for join operation. The leader node initiates the join operation by encrypting its data except attribute which is used for join operation. The same method is followed by all the nodes in the network. Each node in the network receives the data from its decedents present in the hierarchical model and performs the join operation based on the join attribute. The join attribute is common for all the nodes in the network. Once the join operation is performed then the result is forwarded to its ancestor node. The leader node is the one which finally receives data from all other nodes in the network. To decrypt the data decryption key is used. Once the data is decrypted it removes the identifiers from the data and it apply multidimensional k-anonymity for providing privacy to sensitive information.

The leader node sends a global metadata table (**GMDT**) to all leaf nodes by filling values of the attributes with the available data. Example is shown in table 1. Categorical attribute values are converted to integer by assigning proper values to each possible attribute value.

S.No	Attribute name	Lower bound	Upper bound
1	Age	21	65
2	Salary	10000	1000000
3	Address	1	2
4	…….	……..	………
…..	…….	……..	………
N	…….	……..	………

Table 1

When leaf nodes receive this GMDT they update the attribute range values and forward this GMDT to next level. If any node receives more than one table i.e if it is having two child nodes then update the field values based on the table entries and its attribute values and forward this updated table. Leader node on receiving these tables, performs updation on the GMDT tables and uses this updated table to perform multidimensional k-anonymity.

Algorithm at leader site

1. Wait until query has been received, once query is received it sends this query to all the other nodes in the network
2. The leader node itself executes this query and from the result it removes all the identifiers.
3. Identifier removed records are used for encryption except join attribute all other attributes were encrypted and this result is used to prepare the set and these sets are randomly forwarded to the leaves in the hierarchical model and wait for receiving result from its decedents
4. Once it receives the result it decrypts the result by using decryption key to decrypt the data .Once the data is decrypted it removes the identifiers from the data and it apply multidimensional k-anonymity for providing privacy to sensitive information.
5. The result obtained from step 4 is published to user.

Algorithm at non leader site

1. Remains in waiting state until the query has been received from leader node.
2. The node itself executes this query and from the result it removes all the identifiers.
3. Identifier removed records are used for encryption except join attribute every other attribute is encrypted and this result is used to perform join operation with records received from decedents in the hierarchical model.
4. Once it receives records from decedents apply join operation and this result is forwarded in the network.

To deliver the result to user, the data must be collected at one site. To collect data at one site first leader transmits its data to other nodes. The records available at leader node are encrypted first except the primary key which is used for join operation. These encrypted records are divided into number of groups the distribution of groups depends on the number of leaves. The data groups are transmitted to leaves randomly. The Sample model shown in fig 2

In Fig 2. N1 is the leader node which initiates the activity. At the beginning, node N1 broadcasts the query it received and executes the same query on its database. From the result on applying encryption method identifiers are removed except for join attribute this encrypted result partitioned into sets these sets are randomly send to leaves N4 to N7. The nodes in the network on receiving the query performs identifiers removal operation and performs encryption on attributes except for join attribute and wait for the result from decedents. All leaf nodes N4, N5, N6 and N7 receives records from N1 and perform join operation with the records it contains and forward these records to next level.

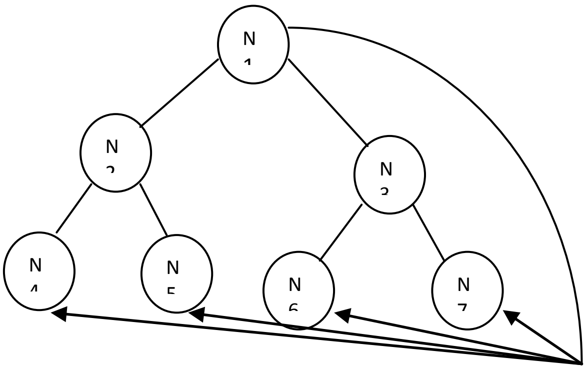

Fig2.

Node N2 receives data from nodes N4 and N5 and performs secured join operation on both received data and its self data. Similarly node N3 receives data from N6 and N7, performs secure join operation with its self data and forwards the data to its successor node. After performing join operation nodes N2 and N3 forward their data to the next level. The node N1 receives data from N2 and N3. At this stage node N1 has the complete result of the query; this result is published to the public after applying privacy preserving methods.

3. CONCLUSION

In some applications for providing privacy the data may be vertically partitioned and stored at various sites. In order to get best results for research we should collect these records to provide privacy to information and the site. We proposed the Hierarchal model to transfer vertically partitioned distributed data. Once leader node receives data from all the nodes which contains the result of the query, then the leader node decrypts the data and publishes the data to the end user.

4. REFERENCES

[1] Latanya Sweeney: “ K- anonymity : A model for protecting privacy” IJUFKS 10(5) :2002; 557-570

[2] K. LeFevre, D.J.DeWitt, and R. Rama Krishna. “Incognito: Efficient full domain k-anonymity. In SIGMOD conference, pages 49-60,2005.

[3] R.J. Bayardo and R.Agrawal. “ Data privacy through optimal k-anonymity”, In Proc. 21st Intnl.Conf.Data Engg(ICDE),pages217-228,USA,2005.

[4] Ashwin Machanavajjhala , Johannes Gehrke, Daniel Kifer : “l- diversity: Privacy beyond k-anonymity” In Proc. 22nd Intnl. Conf.Data Engg.(ICDE),pages 24,2006

[5] Kristen LeFevre, David J.DeWitt, Raghu Ramakrishnan: “Mondrian multidimensional k-anonymity” In ICDE, page 25,2006.

[6] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian:” t closeness : privacy Beyond k-anonymity and l-diversity” ,In ICDE, pages106-115,2007

[7] Pawel Jurczyk,Li Xiong : “Privacy preserving data publishing for horizontally partitioned databases” Technical Report TR-2008-013,EmoryUniversity,Math&CSDept,2008.

[8] R. Srinivas et al., Hierarchical Model for Preserving Privacy in Horizontally Partitioned Databases” IJETTCSVolume2 Issue1 jan-feb-2013

[9] Guan, Yong, Xinwen Fu, and Riccardo Bettati, "An Optimal Strategy for Anonymous Communication Protocols," Proceedings of the 22nd International Conference on Distributed Computing Systems, College Station, TX, 2002.

[10] Pawel Jurczyk,Li Xiong .”DObjects : enabling distributed data services for meta computing platform” In proc. Of the ICCS, 2008.

[11] Wongil choi, Joonsuk ryu, Won young kim, Ungmo kim . “Simple data transformation method for privacy preserving data republication” IEEE 2009 : 978-1-4244-428-7109.

[12] Ren Xiangmin, Yang Jinf, Zhang Jianpei, Wang Kecho: ” Research on CBK(L,K)- anonymity algorithm” IJACT volume 3: number-4 may 2011.

[13] R. Srinivas et al.,. “Preserving Privacy in Horizontally Partitioned Databases Using Hierarchical Model” *IOSR Journal of Engineering May. 2012, Vol. 2(5) pp: 1091-1094*

[14] R. Srinivas et al.,. “Effective Bandwidth Utilization using Trusted LPEs in Anonymous Communication” *International Journal of Computer Applications (0975 – 888) Volume 47– No.7, June 2012*

[15] Wei jian and Chris Clifton, Dept of CS, Purdu university” Privacy Preserving Distributed K-anonymity” CERIAS Tech Report:2005-13