Volume 4, Issue 4, April 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Privacy for Data Publishing in Health Care Domain

Varsha Meshram*
Department of M.Tech CSE
R.T.M. Nagpur University, India

T. Raju
Asst. Prof. Department of M.Tech CSE
R.T.M. Nagpur University, India

Abstract-In the Healthcare Domain there is an increasing need for sharing data that contain personal information from distributed data base for Nationwide Health Information Network. To share information among hospital and other providers and supports appropriate use of Health information beyond direct patient care with privacy protection. In this paper enlist privacy checking strategy and algorithm which exploiting the privacy constraints and adaptive ordering techniques for efficiently checking set of records. There are many algorithm used to provide the privacy for maintaining the database we present an alternative, so that we can publish the data with privacy. For more efficiency of database we portioned the database in horizontally mannered so the thousands of records presents in database should be present in proper manner. While publishing the data we make that data anonymized so that the recipient will not be able to see some sensitive information. For making the data anonymized we use the concept of trusted third party which will helpful for avoiding the potential utility of the database.

Key Words– Data privacy, publishing data, distributed databases.

I. OVERVIEW

Now a day, the exchange of information between people all over the world is vast. In the health care domain also it's needed to share the data for research purpose. We used the concept of data publishing for data sharing among the users who uses the health care database. The concept of data publishing is used for research and expands on the 'why, when and how' of its collection and processing, leaving an account of the analysis and conclusions to a conventional article. A data publication should include metadata describing the data in detail such as who created the data, the description of the type of data, the versioning of the data, and most importantly where the data can be accessed (if it can be accessed at all).The main purpose of a data publication is to provide adequate information about the data so that it can be reused by another researcher in the future, as well as provide a way to attribute data to its respective creator.

Knowing who creates data provides an added layer of transparency, as researchers will have to be held accountable for how they collect and present their data. Ideally, a data publication would be linked with its associated journal article to provide more information about the research. While publishing the data we provide the privacy by making the data anonymised. Data anonymization is the process of destroying tracks or electronic trail, on the data that would lead an eavesdropper to its origins. An electronic trail is the information that is left behind when someone sends data over a network Companies use this concept for the privacy in order to track user data. There are two main settings are used for anonymization. One approach is for each provider to anonymize the data independently which results in potential loss of integrated data utility. A more desirable approach is collaborative data publishing which anonymizes data from all providers as if they would come from one source using either a trusted third-party (TTP) or Secure Multi-party Computation (SMC) protocols to do computations. We use the second approach to make the data anonymized.

II. RELATED WORK

A. TYPE OF ATTACKS.

As when data is collected into one centralised database there are probability increases that any attackers can attack on that database. We define some new type of attacks.

ATTACKS BY EXTERNAL DATA RECIPIENT

A data recipient, e.g. R0, may be an attacker and trying to gather additional information about the records using publish data (P*) and some background knowledge (Bk) such as freely available external data. Most literature on privacy preserving data publishing in a single provider setting considers only such attacks [2]. Many of them adopt a weak or relaxed adversarial or Bayes-optimal Privacy notion [9] to protect against specific types of attacks by assuming limited Background knowledge. For example, k-anonymity [11] prevents identity disclosure attacks by requiring each equivalence group, records with the same quasi-identifier values, to contain at least k records. Representative Constraints that prevent attribute disclosure attacks Include l-diversity, which requires each equivalence group to contain at least l "well-represented" sensitive values [9], and t-closeness [11], which requires the distribution of a sensitive attribute in any equivalence group to be close to its distribution in the whole population. In contrast, differential privacy [1], [3] publishes statistical data or computational results of data and gives unconditional privacy guarantees independent of attacker's background knowledge.

ATTACKS BY DATA PROVIDERS USING INTERMEDIATE RESULT

We assume the data providers are not completely honest [7], [8], commonly used in distributed calculation situation. They can attempt to infer additional information about data coming from other providers by analyzing the data received during the anonymization. A trusted third party (TTP) or Secure Multi-Party Computation (SMC) protocols can be used to guarantee there is no disclosure of intermediate information during the anonymization. However, either TTP or SMC do not protect against data providers to infer additional information about other records using the anonymized data and their own data (discussed below). Since the problem is orthogonal to whether a TTP or SMC is used for implementing the algorithm, without loss of generality, we have assumed that all providers use a TTP for anonymization and note that an SMC alternative can be implemented.

B. DATA PUBLISHING TECHNIQUES

We publish the data by make that data anonymized and then portion that data. Portioning is the concept to reduce the amount of data we uses the concept for the simplified data management and increase the performance. Basically there are two types of portioning is used. First is the Horizontal portioning and second one is the vertical portioning. In horizontal portioning we simply splitting a single large entity into several too many smaller entities. Vertical partitioning involves creating tables with fewer columns and using additional tables to store the remaining columns. Normalization also involves this splitting of columns across tables, but vertical partitioning goes beyond that and partitions columns even when already normalized. Different physical storage might be used to realize vertical partitioning as well; storing infrequently used or very wide columns on a different device, for example, is a method of vertical partitioning done explicitly or Implicitly, this type of partitioning is called "row splitting" (the row is split by its columns). A common form of vertical partitioning is to split dynamic data (slow to find) from static data (fast to find) in a table where the dynamic data is not used as often as the static.

C. PRUNING STRATEGIES

Pruning means to change the model by deleting the child nodes of a branch node. The pruned node is regarded as a leaf node. Leaf node cannot be pruned. Pruning is useful for to make the data generalize so that the size of data get optimize. Here we present two types of pruning strategies. In downward pruning if a league is not able to violate privacy, then all its sub league will not able to do so and hence do not need to be checked. In upward pruning if a league is able to violate privacy, then its entire sub league will not be able to do so and hence do not need to be checked.

For the fast pruning we use the Heuristic algorithm. The key idea of heuristic algorithm is to efficiently search the adversary space with effective pruning such that all adversaries need to be checked. This is achieved by two different pruning strategies. An adversary ordering and a set of search strategies that enable fast pruning.

III. PROBLEM DEFINITION

Privacy preserving data analysis and data publishing has received considerable attention in recent years as promising approaches for sharing data while preserving individual Privacy. When the data are distributed among multiple data providers or data owners, two main settings are used for anonymization. One approach is for each provider to anonymize the data independently which results in potential loss of integrated data utility.

Existing system uses the conception of differential privacy where background knowledge was embrace. Which consists of the problems that, the easily larceny data from the database may shows the diverse effects on research field. Again the un-availability of memory space is the big problem, as database get increases, not uses the concept of removing unwanted data from the database.

IV. OBJECTIVES

While publishing the data on social networking sites, those sites have proper privacy so that no one can attacks on sensitive data base.

- The lot of work in this research area is done but there are number of problems in existing systems. So the objectives to be recovered in the future may be,
- The problem of availability of memory space can occur.
- It is time consuming process.
- Proper privacy fitness provided.

To publish the data on the social networking sites we will make that data annonymized it and present it horizontally fashion for better utilization of database.

So there is a concept of data privacy which provides the proper privacy for database. With help of this we will be able to minimize attacks of data attackers. It will also help for fast and accurate search by maintaining the database.

V. CONCLUSION

We considered a new type of potential attackers in shared data publishing – a coalition of data providers. To prevent privacy disclosure by any adversary we provide proper privacy which is enough. We discussed some fundamental limitations of the problem of privacy-preservation in the presence of increased amounts of public information and

background knowledge. Finally, we discussed a number of diverse application domains for which privacy-preserving data mining methods are useful.

Our approach is to achieve better or comparable utility than existing algorithms while ensuring privacy efficiently. We are trying to define a proper privacy fitness score for different privacy constraints. We eliminate the unwanted data from the database so the size of database should be maintained.

REFERENCES

[1] C. Dwork, "Differential privacy: a survey of results," *in Proc. of the 5th Intl. Conf. on Theory and Applications of Models of Computation*, 2008, pp. 1–19

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent Developments," *ACM Comput. Surv.* vol. 42, pp. 14:1–14:53, June 2010.

[3] C. Dwork, "A firm foundation for private data analysis," *Commun*.ACM, vol. 54, pp. 86–95, January 2011..

[4] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-Dimensional healthcare data,"*ACMTransactions on Knowledge Discovery from Data (TKDD),* vol.4, no. 4, pp. 18:1–18:33, October 2010.

[5] W. Jiang and C. Clifton, "Privacy-preserving distributed k-anonymity," *in Data and Applications Security XIX, ser*. Lecture Notes in Computer Science, 2005, vol. 3654, pp. 924–924.

[6] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity, *VLDBJ*" vol. 15, no. 4, pp. 316–333, 2006

[7] O. Goldreich, *Foundations of Cryptography: Volume 2, Basic Applications*. Cambridge University Press, 2004.

[8] Y. Lindell and B. Pinkas, "Secure multiparty computation for privacy-preserving datamining," *TheJournal of Privacy and Confidentiality*, vol. 1, no. 1, pp. 59–98, 2009.

[9] Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *in ICDE*, 2006, p. 24.

[10] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzz*., vol.10, no. 5, pp. 557–570, 2002.

[11] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *In Proc. of IEEE 23rd Intl. Conf. on Data Engineering (ICDE)*, 2007.

