MULTI LEVEL PRIVACY PRESERVING IN MEDICAL DATA PUBLISHING

Neera Sali K* and Diya Thomas**

*****Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology, Kochi, India, Email: neerasali@gmail.com
** Department of Computer Science and Engineering, Rajagiri School of Engineering and Technology, Kochi, India, Email:diyaabs13@gmail.com

ABSTRACT: Privacy preserving data publishing(PPDF)is an emerging technology in the datamining field which performs datamining operations in a secured manner to preserve the confidential/sensitive information. Preserving privacy while publishing medical information has become an important challenge in this area due to its high confidentiality. While publishing medical data the PPDF scheme should maximize the data utility at the same time should have a minimum data disclosure risk. This paper concerned with privacy of medical data while publishing the patient information for research or analysis purposes. K-anonymity and l-diversity are the most popular techniques used for preserving privacy. These techniques does not consider the semantic relationship between the data values so they are prone to similarity attack .In this paper, we present a privacy-preserving data publishing framework for publishing large datasets with the goals of providing different levels of utility to the users based on their access privileges. The proposed system overcome the similarity attack by applying a privacy preservation approach which uses a key attribute masking technique and an anonymization process.The results showed that the semantic anonymization increases the privacy level with effective data utility.

KEYWORDS: Privacy preserving data publishing; k-anonymity; Categorical data; L-diversity.

INTRODUCTION

Past several years has witnessed fast development of information technology, which results in increasing ability to store, retrieve, and mine large number of personal electronic records. The storage and sharing of information has brought some substantial benefits to people. Information such as crime records, health data, and credit card information can be used to help companies, governments, health service organizations do specific research or make decisions, which will in turn benefit the society.Privacy has always been a great concern for patients and medical service providers while publishing medical related information's. Due to the Advances in the Technology and the government's push for Electronic Health Record (EHR) systems a large amount of medical data is collected and stored electronically. This data is a rich source for research and needs to be made available for mining, while at the same time patient privacy needs to be preserved in an highly efficient manner. The management of medical data is heavily regulated by the Health Insurance Portability and Accountability Act (HIPAA)[10] in the United States. This strong level of oversight and inherent characteristics of medical data make Privacy Preserving Medical Data Mining an important field of Privacy Preserving Data Mining (PPDM).

The goal of privacy preserving medical data publishing is to ensure that confidential patient data is not disclosed. Methods for preventing unauthorized disclosure of information include: restricting access, restricting the data, and restricting the output. Restricting the access by locking down the data is a relatively simple solution to the privacy problem, but it completely eliminates the utility of the data. It is critical that useful medical information be shared across research institutions. Restricting the data involves removing attributes or modifying the dataset with some form of generalization or anonymization of values. Restricting the output involves transforming the results of user queries while leaving the data unchanged. The restricted data approach allows for much more widespread sharing and distribution of the data.The proposed system uses series of anonymization and masking technique for restricting the data.

This paper focused on preserving privacy of the medical data by providing different levels of privilege for different users of the same data. Most of the existing work on privacy-preserving data publication targets at releasing safe versions of the dataset to provide accurate results for researchers. Such data releases assume that all users of the data share the same access privilege levels. .In this paper we are using different levels of access to the same data. For instance,in the patient information table, one may need to protect privacy and utility at different protection levels depending on the access privilege of the users. While some users (e.g.,less privileged data researchers) may be allowed to obtain only limited information without the patients confidential information. some others (e.g.,Doctors) may have access to some confidential information along with non sensitive data . In the same way, some less privileged users (like patients) may have more access than the researchers but lesser than doctors.

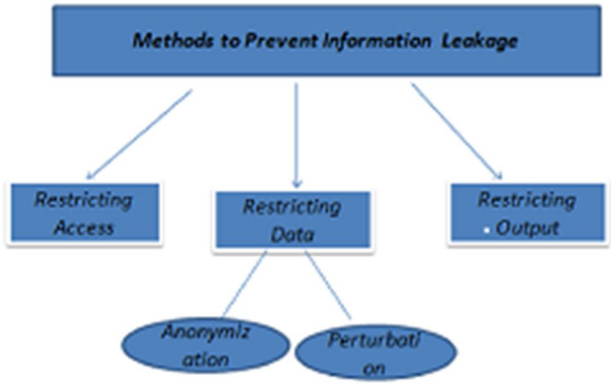

Figure. 1. Information Prevention

The data privacy scheme applied in this paper is a combination of masking and anonymization on which the masking scheme we have used is the popular Secure Hash Algorithm(SHA-1). anonymization is performed by dividing the entire dataset in to sensitive and nonsensitive partitions and applying semantic anonymization based on generalization and suppression on the sensitive attributes of the table.
This Paper is organized as follows: The related works to anonymization methods and their related techniques are discussed in section II. In section III,the proposed model is discussed.. Finally in section IV conclusion is provided

RELATED WORKS

This section reviews some of the previous work in this field Where privacy have become an important issue and considerable progress has been made with data anonymization. Most recent Studies Focused on devising anonymization algorithms for data publishing. One of the most popular anonymization method is the K-anonymization which is proposed by Samarati and sweeney[1].For K-anonymity the domain of each quasi-identifier attribute is partitioned into intervals and replace the values in the attribute with the values belonging to using a concept tree.The records are grouped by the same intervals of the quasi-identifiers if the sizes of all groups are at least k.However K-anonymity doesn't focus on sensitive information and it was not enough to protect the data which include linking attack.
The paper [2] proposes a novel, more flexible generalization scheme. The experimental results of their study indicate that their approaches produce k – anonymization with less generalization compared to previous approaches. They conclude that a bottom-up approach for k – anonymization is preferable for small number of quasi identifying attributes. Eventhough this method prevents the linkage attacks, it suffers from homogeneity and background knowledge attacks.
In[3]the authors proposed a task independent anonymization technique which preserve information privacy and utility of the data. Their algorithm is applied on the original data table to transform only the sensitive raw data before applying any mining methods. In most of the privacy preservation generalization methods, loss of information is due to transformation of QI attributes and sensitive attributes. They demand both privacy and no information loss by only transforming part of the QI and sensitive attributes and also algorithm handle any number of sensitive attributes. The complexity of the algorithm is based on the table size.
The paper proposed by Yan and Peng[4] used a modified L-diversity model in order to address the privacy of medical data. The k-anonymity model is fragile to linking attacks so for solving this problem l-diversity methods are devised. Here their paper proposed a modified entropy l-diversity model in which more detailed attacking conditions and characteristic of medical information are taken in to consideration. Approaches which address specific problems are also developed using anonymization method.

The k anonymity based method is illustrated in [5] is used to search for optimal feature set partitioning and [6] for cluster analysis. And [7] proposes a data reconstruction approach to achieve k-anonymity protection in predictive data mining. In this approach the potentially identifying attributes are first mapped using aggregation for numeric data and swapping for nominal data. A genetic algorithm technique is then applied to the masked data to find a good subset of it. This subset is then replicated to form the released dataset that satisfies the k-anonymity constraint. Another anonymization technique known, as Condensation is a statistical approach which constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters [9]. This method is called as condensation because of its approach of using condensed statistics of the clusters to generate pseudo data. It constructs groups of non-homogeneous size from the whole data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level. Here, pseudo data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data. This approach also can be effectively used for the problem of classification .The pseudodata provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data. Since the aggregate behaviour of the data is preserved, it becomes useful for a variety of data mining problems
The most popular privacy preservation techniques like K-anonymity and l-diversity etc.. are prone to several types of security threats. In this paper, we are presenting a new privacy-preserving data publishing framework for publishing large datasets with the goals of providing different levels of utility to the users based on their access privileges. The privacy preservation approach uses SHA-1 masking technique and a Semantic anonymization process to overcome the similarity attack.

PROPOSED METHOD

The main objective of this work is to provide privacy to the medical data by masking ,categorical and numerical attribute anonymization and adding different levels of privileges to different roles on same data.The goal of the work is to eliminate privacy breach and increase the data utility of a released data table.This is achieved by multilevel privacy preserving data publishing in which we applied a SHA-1 masking method and semantic anonymization on the l-diversity table for the sensitive attributes and other non sensitive attributes are directly published.

Basic Notation

Let T [K,Q1,Q2,....,Qp,S] be a table. For example, T is a medical dataset. Let Q1, Q2..., QP denote the quasi-identifier specified by the application. Let S denote the sensitive attribute. A sensitive attribute is an attribute whose value for some particular individual must be kept secret from people who have no direct access to the original. data. Let K denote the key attributes of T which is to be removed before releasing a table. t[X] denote the value of attribute X for tuple t. Let T be the initial table and T' be the released micro data table. T' consists of a set of tuples over an attribute set. The attributes for k-anonymity table are classified into three categories namely quasi identifiers, Key attribute and Sensitive attributes.
Definition 1: Key Attribute
An attribute denoted by 'K' consists of values which is the most unique value for to identify the individual from dataset 'S'. Key attributes are used to identify a record, such as Name and patient-id, since our objective is to prevent sensitive information from being linked to specific respondents.
Definition 2: Quasi-identifiers
A set of non-sensitive attributes {Q1,Q2, …,Qp} of a table is called a quasi-identifier, if these attributes can be linked with external data to uniquely identify (can be called as candidate key) at least one individual in the general population. Quasi-identifier (QI) attributes are those, such as age and zip code, that in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in the micro data belong. Unlike identifier attributes, QI attributes cannot be removed from the micro data, because any attribute is potentially a QI attribute.
Definition 3: Sensitive attribute
A set 'A' consists of values which the user selects as most sensitive attribute from dataset 'S'. These attributes is what the researchers need, so they are always released directly. Sensitive attributes such as diagnostic report, salary and account number

System Architecture

The system architecture is shown in figure that will clearly describe each module used in this anonymization process.

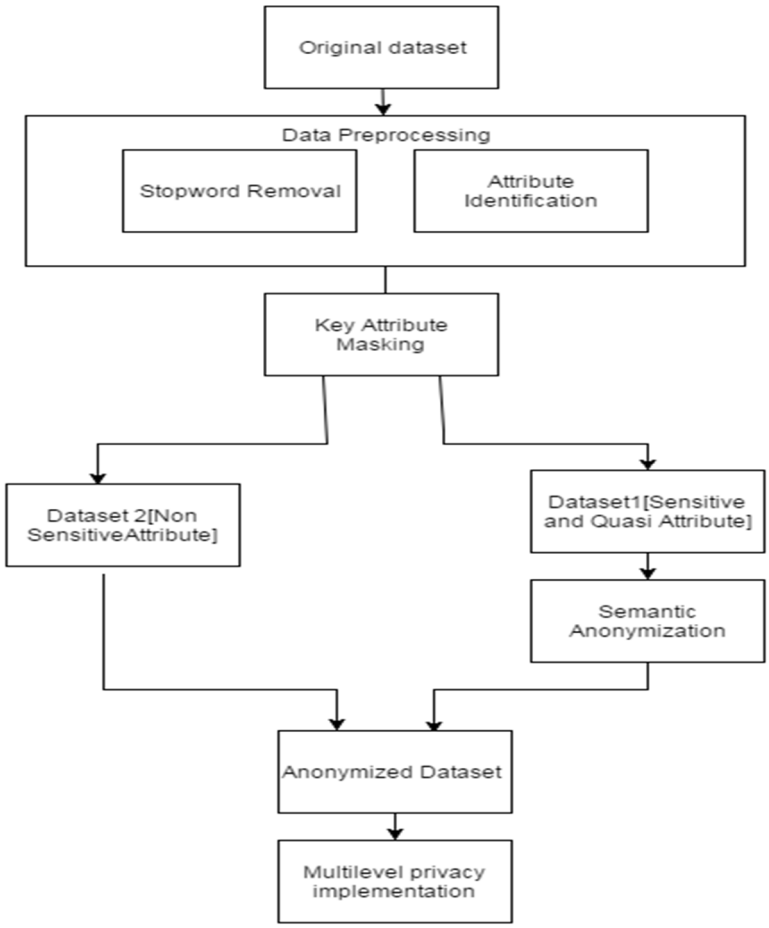

Figure. 2. System Architecture

Data Pre-processing

The Medical dataset is collected from different hospitals in this area that includes the patient information's like name, patient-id, disease, address, age, etc. In the pre-processing stage the missing attributes , datavalues in the datasets and stopwords are removed and dataset is converted to a text format for processing. The first stage in privacy preservation is the identification of different attributes in the dataset. The attributes in the data table comes under 4 categories Key attribute, Quasi Attribute, Sensitive and non-sensitive attributes that is already discussed in the previous section. The key identifier in the dataset is identified by using a duplicate check that will be performed in each column and the attribute without duplicate datavalues will be considered as the key attribute.

Data Masking

After Data preprocessing and finding unique Attribute the next stage is the masking of key attribute.In This work we are using the Secure Hash Algorithm(SHA-1) for data masking.SHA-1 is cryptographic hash function designed by the United States National Security Agency.SHA-1 produces a 160-bit hash value known as a message digest.This work assumes patient-id as the key attribute and the SHA-1 algorithm generate a 160 bit hash value for each patient-id.The identifier which will identify an individual is removed from the dataset.The output of this stage contains the masked key attribute value that is hidden for all levels of users.The SHA-1 algorithm mainly consists of 6 Stages mainly

Task 1. Appending Padding Bits. The original message is "padded" (extended) so that its length (in bits) is congruent to 448, modulo 512. The original message is always padded with one bit "1" first. Then zero or more bits "0" are padded to bring the length of the message up to 64 bits fewer than a multiple of 512.

Task 2. Appending Length. 64 bits are appended to the end of the padded message to indicate the length of the original message in bytes. The length of the original message in bytes is converted to its binary format of 64 bits. If overflow happens, only the low-order 64 bits are used.Break the 64-bit length into 2 words (32 bits each).

Task 3. Preparing Processing Functions. SHA1 requires 80 processing functions defined as:

f(t;B,C,D) = (B AND C) OR ((NOT B) AND D) (0 <= t <= 19)

f(t;B,C,D) = B XOR C XOR D (20 <= t <= 39)

f(t;B,C,D) = (B AND C) OR (B AND D) OR (C AND D) (40 <= t <= 59)

f(t;B,C,D) = B XOR C XOR D (60 <= t <= 79)
Task 4. Preparing Processing Constants. SHA1 requires 80 processing constant words defined as:
K(t) = 0x5A827999 (0 <= t <= 19)
K(t) = 0x6ED9EBA1 (20 <= t <= 39)
K(t) = 0x8F1BBCDC (40 <= t <= 59)
K(t) = 0xCA62C1D6 (60 <= t <= 79)
Task 5. Initializing Buffers. SHA1 algorithm requires 5 word buffers with the following initial values:
H0 = 0x67452301 H1 = 0xEFCDAB89 H2 = 0x98BADCFE H3 = 0x10325476 H4 = 0xC3D2E1F0
Task 6. SHA-1 Process Message in 512-bit Blocks. This is the main task of SHA1 algorithm, which loops through the padded and appended message in blocks of 512 bits each. For each input block, a number of operations are performed in each rounds.After the masking process the entire data set is divided in to 2 partitions namely sensitive and non-sensitive attributes and further anonymization is done on the sensitive partition.

Data Anonymization
The Semantic Anonymization process consists of 2 main stages ,effective semantic rule determination and Anonymization based on that rules.The process applied on a anonymized table resulted from the L-diversity process .l-diversity model provides an extension to k-anonymity and requires that each equivalence class also contain atleast l well separated distinct values for a sensitive attribute to avoid the homogeneous sensitive information revealed for the group. In this process approach starts by checking the anonymized table for finding the semantic similarity based on the rules from the repository.If there is an extracted piece of information that impacts privacy then store that rule as an effective rule to use in the anonymization process
Anonymization process is based on the effective rules obtained from the previous stage. Anonymizer assign suitable anonymization action (generalization) to each values depends on the rules.In the generalization the anonymizer find a general representation of the QI values in the EC that is susceptible to privacy attack when compared to other ECs in the table.This process is performed by using a merge method and if there is no EC that can be merged with the EC that is susceptible to privacy attack then suppression technique is used. The Semantic Extraction Algorithm can be concluded as follows

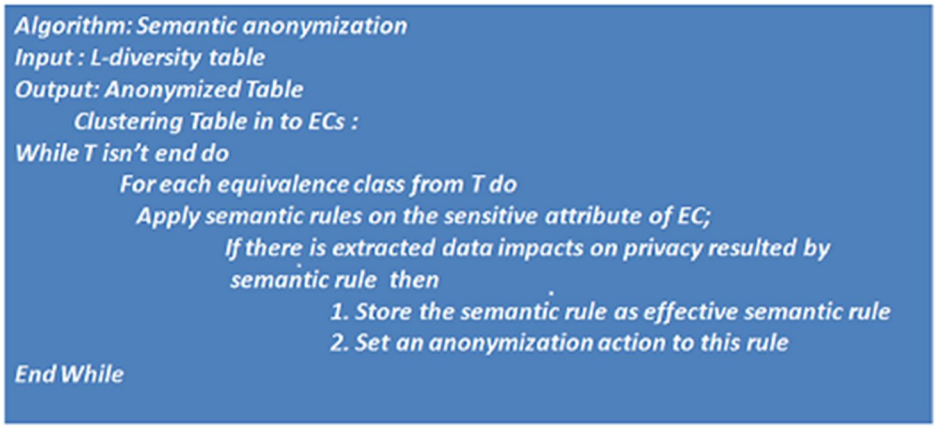

Figure. 3. Semantic Anonymization Algorithm

Multi-level privilege Addition
The privacy-preserving data publishing framework provide different levels of utility to the users based on their access privileges.In contrast to existing anonymization techniques that provide information only at a single safe level, here in our approach we are providing data to different users depending on their Privileges that will increase the systems security. Users are provided with their keys for accessing the data depending on their roles.Here we are applying a key-based data access mechanism.
Privacy levels: We consider 3 levels of privacy/utility
Level 1 - Medical analysts: These users obtain only non sensitive information about the data
Level 2 - Researchers: level 2 users have access to Quasi attributes along with non-sensitive attributes.
Level 3 - Doctors -All access users: Users obtain original dataset without the key It represents the highest level of access to the dataset

CONCLUSION

In this paper, we present a privacy-preserving data publishing framework for publishing large data sets with the goals of providing different levels of utility to the users based on their access privileges.The proposed system uses a privacy preservation approach which uses the SHA-1 masking technique and an anonymization process based on the semantic relationship between the data values of sensitive attribute to overcome the similarity attack.We expect that this system will produce an anonymized data set with high degree of utility and minimum disclosure risk and this will expected perform better than the traditional anonymization Techniques. To improve the performance of the system further the future work focus on developing a novel algorithm replacing SHA-1.The cloud implementation of this work will also provide us a greater scope to explore challenges in medical data publishing.

REFERENCES

[1] L. Sweeney, "*k-Anonymity: A Model for Protecting Privacy*" ZWInternational Journal on Uncertainty Fuzziness Knowledge based Systems, 2002.

[2] Tiancheng Li, Ninghui Li, "*Towards Optimal k-anonymization*", Data and Knowledge Engineering, 2008 Elsevier

[3] E. Poovammal and M. Ponnavaikko,"*Task Independent Privacy Preserving Data Mining on Medical Dataset*",\International Conference on Advances in Computing, Control and Telecommunication Technologies, 2009

[4] Yan ZHU and Lin PENG, "*Study on K-anonymity Models of Sharing Medical Information*", 1-4244-0885-7/07/ © 2007 IEEE

[5] Nissim Matatov, Lior Rokach, Oded Maimon, "*Privacy-preserving data mining: A feature set partitioning approach*"Information Sciences 180 (2010) 2696-2720

[6] Dan Zhu, Xiao-Bai Li, Shuning Wu, "*Identity disclosure protection: A data reconstruction approach for privacy preserving data mining*", Decision Support Systems 48 (2009) 133-140.

[7] Benjamin C. M. Fung, Ke Wang, Lingyu Wang, Patrick C.K. Hung, "*Privacy-preserving data publishing for cluster analysis*" ,Data and Knowledge Engineering 68 (2009) 552-575

[8] Yang Xu,Tinghuai Ma,Meili Tang and Wei Tian,"*A Survey of Privacy Preserving Data Publishing using Generalization and Suppression*"International Journal of Computer Applications,March 2014.

[9] Bhavana Abad (Khivsara) and Kinariwala S.A , "*A Novel approach for Privacy Preserving in Medical Data Mining using Sensitivity based anonymity*",International Journal of Computer Applications,March 2012.

[10] Zhiyuan Chen, Tamas Sandor , "*Privacy preserving data mining for medical data*", University of Maryland at Baltimore County Catonsville , MD, USA © 2011.

[11] P.Samarati, "*Protecting respondents identities in microdata release*", IEEE Transactions on Knowledge and Data Engineering, 13(6):10101027.2001.