Developing a Blueprint for Preserving Privacy of Electronic Health Records using Categorical Attributes

T.Kowshiga¹, T.Saranya², T.Jayasudha³, Prof.M.Sowmiya⁴ and Prof.S.Balamurugan⁵
Department of IT, Kalaignar Karunanidhi Institute of Technology, Coimbatore, TamilNadu, India¹,²,³,⁴,⁵

ABSTRACT: Cloud computing offers unique opportunities for supporting long-term record preservation. MyPHRMachines, a patient owned health record system prototype based on remote virtual machines hosted in the cloud. MyPHRMachines is particularly promising for countries with a very heterogeneous architecture of systems across hospitals and other care institutions. In the view of developer PHRs should be portable. PHR systems typically offer functionality to share, visualize and analyze PHR data. Secure lifelong management of patient medical records since data are stored in the cloud and do not have to be carried around by patients. We also present method for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set.


I. INTRODUCTION

The publishing of data to third parties is also as important assure as masking of data. Because the hackers can be traced with good amount of evidence, the leakage of data is detected. To tackle this problem an image is attached with the masked data and then it is distributed to the agents. Using steganography the masked data is shared to agent. The attached image contains the key which will give alert message to the distributor while agent distributed to any other third parties.

If the distributor sees enough evidence that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings. In this paper we develop a model for assessing the guilt of agents. We also present method for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set. Such objects do not correspond to real entities but appear. If it turns out an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty.

II. LITERATURE SURVEY

k-Anonymity concept used for solve the tension between data utility and respondent privacy for individual data protection. The generalization and suppression approaches proposed in literature to achieve K-Anonymity is not equally suited for all types of attributes:

- Generalization/suppression is one of the few possibilities for nominal categorical attributes.
- It is one possibility for ordinal categorical attributes.
- It is completely unsuitable for continuous attributes, as it causes them to lose their numerical meaning.

The fundamental rights of patient to have their privacy protected by health care organizations. This information used to identify particular individual is not used to reveal sensitive patient such diagnoses, etc. If the degree of anonymity of a disseminated data set could be measured. Privacy protection in disseminated databases can be facilitated by use of special ambiguities algorithm. Generalization involves replacing a value with less specific but semantically consistent value. Suppression involves not replacing a value at all. In this paper authors plead on the need of knowledge intensive
tools is data privacy more especially we discuss the role of knowledge related tools in data protection and in disclosure risk assessment. Statistical Disclosure Control(SDC) family of method for micro data masking the micro data so that they can be released while preserving the privacy. Aggregate original database records into small groups prior to publication. Each group should contain k-records.k->constant value. Recently micro aggregation achieve k-Anonymity in addition. Optimal micro aggregation computed in polynomial time for univariate data. This paper present new data oriented heuristics which improve on the trade off between computational complexity and information loss and are thus usable for large data sets. Microaggregation is the well known Microdata protection method, ensuring confidentiality. Authors propose and use for new approach like text documents. This method relies on word net framework that provide full semantic relationship taxonomy between words. Authors aim to ensure confidentiality of text document, at the same time preserve general meaning by applying some measures to evaluate the quality of the protection method relying on information loss. Inference control in data base also known as SDC. This is an important application in several areas such as official statistics, health statistics, e-commerce, etc. Hence it refers to data modification, challenge for SDC is to achieve protection with minimum loss of accuracy database we discuss several information loss and disclosure risk measures and analyse several ways of combining them to assess the performance of the various method. In US, the Health Insurance Portability and Accountability Act(HIPAA) protects the confidentiality of patient data and approval of internal review Board to use data for research but these requirements can be waived if data is de-identified. The De-identification of narrative text documents often realized and require significant resources. In this method based performed better with PHI is rarely mentioned in clinical text but are more difficult to generalize.

Patient record data are highly sensitive so their secondary use raises both ethical and data protection issues. Disclosure of patient data could cause serious difficulties so individual damaging for patient and clinicians. In this paper grid based medical data repository accessing risk and suggest a new model for Statistical Disclosure Control(SDC) of patient data. It provides enormous opportunities for data mining tasks. K-Anonymity which are vulnerable to privacy attacks based on background knowledge. Set-value data could be efficiently released under differential privacy with guaranteed utility help of taxonomy trees. Top down partitioning algorithm to generate a differentially privacy release scale with input data size. Protect of personal data in statistical data base has become major concern before they released to public use they applied statistical data bases. Microaggregation for SDC is to protect micro data that is record on individual compares. Micro data into groups at least K-records replace the record in each group. DBA(Density Based Algorithm) it form descending order of their densities in reverse order and compare with latest microaggregation methods.

III. BASIC PRIMITIVES AND TERMINOLOGIES

Micro Data

The customers or patient’s data’s are collected for this process. In this, we consider micro data such as census data and medical data. Typically, microdata is stored in a table, and each record corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories:

- **Identifier**: Identifiers are attributes that clearly identify individuals. Examples include Social Security Number and Name.
- **Quasi-Identifier**: Quasi-identifiers are attributes whose values when taken together can potentially identify an individual. Examples include Zip-code, Birthdate, and Gender.
- **Sensitive Attribute**: Sensitive attributes are attributes whose values should not be associated with an individual by the adversary. Examples include Disease and Salary.

Data Privacy

This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that i-closeness substantially limits the amount of useful information that can be extracted from the released data. This limits the amount of sensitive information about individuals while preserves features and patterns about large groups.
View Agents List
The Admin can view all registered agents using this module. It contains the agent’s full details for reference. It contains like Agent ID, Agent Name, Contact Number, Mail ID etc.,

Attach Fake Object: Fake objects are objects generated by the distributor that are not in set T. It contains the secret file and saved location and secret keys.

Steganography (Secret File Sharing) Steganography is an alternative to encryption for keeping data or correspondence confidential. The Secret is embedding with in image. After that the key generated for secure sharing.

View Distribution List The Admin also known as distributed agent’s data’s. This module contains the details about already distributed data’s in agent wise.

Data Leak Report
View Leaked Agent
The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. It displays the Agent ID, Agent Name etc.,

Image Extraction When the receiver gets the image, he will use the same random number generator.

AGENT
Agent Registration The registration module contains the agents personal details like Agent ID, Agent Name, Contact Number, Mail ID etc., and The Agent choose click points at the time of registration. It’s very sensitive data’s transaction.

Agent Login
Image Authentication Image authentication has been proposed as a user-friendly alternative to password generation and authentication.

Receive Data: Random allocation has also performance, since as the number of agents increases, the probability that at least two agents receive many common objects becomes higher. The every agent has successfully login into their application he can view the received sensitive file.

IV. WORKING METHODOLOGY

In this paper we describe understanding the textual data requires exploitation and integration of clinical resources. In past several approaches for assessing word similarity by exploiting different knowledge source have been proposed. These measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data.
Fig. 2 Flowchart for Agent-Based Leakage Detection

Fig. 3 Classification of Data by Distributor
Distributor can send the data to different agent with hide of original data. The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. It displays the Agent ID, Agent Name etc., the objects are designed to look like real objects.

V. CONCLUSION AND FUTURE WORK

This paper detailed about various methods prevailing in literature for protecting privacy of anonymized medical data. Ontology Based measure to compute semantic similarity in Biomedicine is studied. Ordinal, continuous and heterogeneous K-Anonymity through Microaggregation are dealt in detail. Protecting patient privacy by quantifiable control of disclosure in disseminated databases and achieving k-Anonymity privacy protection using generalization and suppression are discussed in detail. Efficient Multivariate data-Oriented Micro aggregation of Categorical data for confidential documents is examined. Differential Privacy for Automatic De-Identification of textual documents in the electronic health records and Statistical Disclosure control for patient records in biomedical information System is considered. Density-based microaggregation for statistical disclosure control and anonymization of Set-Valued Data via Top-Down, Local Generalization are also aggregated in brief.

In this paper we developed a model for assessing the guilt of agents. We also presented method for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set.

REFERENCES


