Generalization Technique for Privacy Preserving of Medical information

Asmaa Hatem Rashid and Norizan Binti Mohd Yasin

***Abstract*—the development in health information technology have increased the need share data and knowledge especially medical information, researchers, public health organizations, others interested in medical data. Some medical records are often added and deleted in the practical applications. The leakage of privacy information caused by re-publishing datasets with multiple sensitive attributes becomes more likely than any other publication styles This paper attempts to fill the above gaps, presents a framework for de-identifying health information .we used the generalization technique and K-Anonymization model to protecting privacy of patient data. After the protecting privacy applicable prepare the data for analysis and extract the knowledge that supports decision-making, we offer a range of initial assessments show the effectiveness-of our-approach.**

***Index Terms*—Health information technology, protect privacy, Share knowledge, sensitive information, generalization technique, K-Anonymization, supports decision-making.**

I. Introduction

In recent years, increased use of the Internet and its applications in various aspects of life led to the more and more people to pay their attention to the publishing of personal data .Thus, privacy information caused by re-publishing data with multiple sensitive attributes becomes more likely than any other publication styles.

According to research paper [1], 87% of American had reported characteristic that likely made them unique based only on (zip code, date of birth, sex). In order to protect privacy information from such as linking attack. Sweeney present *k*-anonymization model. The process of *k*-anonymization model such as a table start removing all explicit identifier .the core idea of k-anonymization model is that each record in a table is indistinguishable from at least *k*-1 other record respect to the pre-determined quasi-identifier .

Since k-anonymization model is a sample and effective, it has been extensively studied and enhanced as a viable definition of privacy in data publishing. Conclusion about K-Anonymization model, a sample model based on a set on Techniques such as Generalization, Suppression and others. It convert private data to public data including the data benefits can used it at different processing. In this paper, we present one set of methods that would allow health information to be used and disclosed under existing legal frameworks is de-identification. According to El Emam *et al.*, [2], De-identification refers to a set of methods that can be applied to data to ensure that the probability of assigning a correct identity to a record in the data is very low".

As mentioned earlier the importance of the subject and really motivated to work in this area and encourage the work and research in the area of data confidentiality and privacy protection in various areas. Our research focus on privacy preserving at scientific research special in medical data [3]. Fig. 1. shows the privacy types.

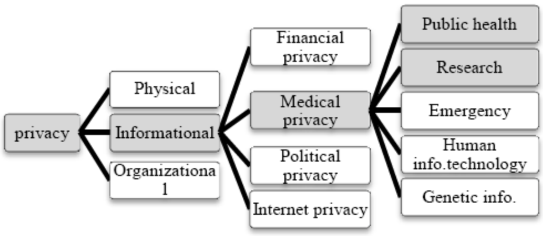

Fig. 1. Types of privacy.

II. Related Work

In the past few years, the issue of improving the control and sharing of data in knowledge management has attracted substantial interest among individual users and service providers such as research centers, companies, and governments. This growing interest confirms the importance of the subject and the sensitivity of the work and research involved. Most reviewers and researchers agree on the significant problems present in the control and sharing of data in knowledge management. The problems posed are under more than one research area. These include data confidentiality and privacy protection, with the review of suggested solutions to the problems involving confidential data for cryptographic information and hidden data [4], among others. The second area involves data mining and data-mining algorithms to ensure privacy [1], such as generalization and suppression techniques [5]. Other research areas of data management and control of sharing of data ensure integration especially in medical information systems [6] and knowledge base systems, which will be the focus of the present study. In most fields, sharing of data needs the control and management of such data to ensure system integration [4], such as patient data, without revealing any sensitive information that can identify a patient. There are several studies that focus on the management of data in medical applications to ensure system integration. However, this can result in information misuse. Nevertheless, there are many algorithms and methods that facilitate management of shared data using techniques such as removing sensitive characters from the information system. Such algorithms are used to prevent unauthorized access to the original data for illicit purposes.

Manuscript received July 20, 2013; revised September 11, 2013.

Asmaa Hatem Rashid and Norizan Binti Mohd Yasin are with University of Malaya, Faculty of Computer Science and Information Technology (e-mail: asmaarashid@siswa.um.edu.my)

DOI: 10.7763/IJET.2014.V6.709

In the present research, the main problem is the identification of an algorithm that provides control and management of shared data. Updating current data can be useful for future purposes such as analysis and knowledge management to support decisions in different fields of medical applications; however, the updated data should still represent real cases.

III. Research Methodology and Discussion

The current study will address this above problem through analyzing and evaluating the following sub problem: There is no model that can identify the number of quasi-identifier characters in such a way that the shared data are managed and a new version of such data is always usable. There is a lack of trust among medical system providers in sharing data and managing knowledge. Aside from the lack of a centralized database to keep the collected data, the problem of case indexing is still left unresolved due to the inability to update data in such a way that it can be used for further analysis and studies. The lack of high-quality updated data and the possibility of errors that adversely affect the results of studies depend on the crucial task of updating the data. As such, the present study attempts to fill in these gaps. There is a need to build models or design algorithms that manage the sharing of data to avoid misuse. The goal is to bring authenticity to the data system. Guided by recent studies from the years 2005 to 2011 on the control and sharing of data in knowledge management [6], the current work notes that most reviewers and researchers have focused on ensuring the privacy of sensitive data. .following Fig. 2, explore the privacy preserving data mining algorithms [7].

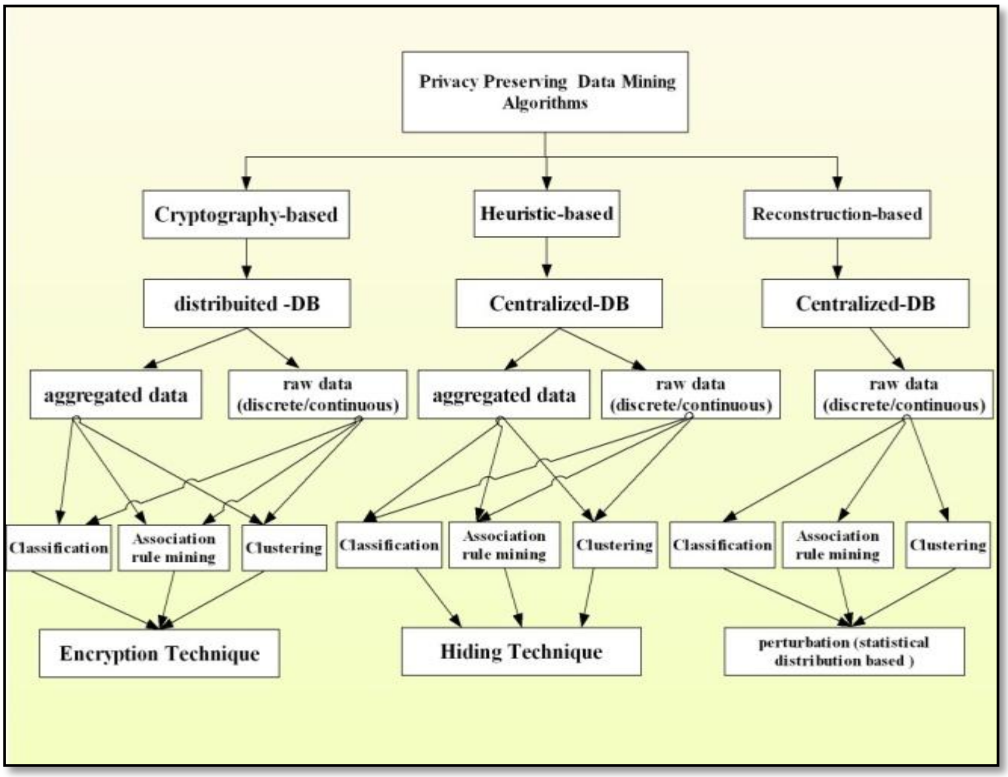

Fig. 2. A taxonomy of the developed PPDM algorithms[7].

In other words, great concern has been directed on the control of data and it's sharing to make it available to their owners. Some reviewers and researchers have even suggested the use of covert techniques which isolate data such as encryption technology. Different ways of protecting data have been dealt with in recent research. The methods previously introduced include information on how to spread and use data in research, decision making, scientific analyses, and other purposes [5]-[6]. First, the concern is how to control data sharing and management and avoid the risk of publishing data that may lead to revealing the real data. Second, there is lack of unity among the collected data, and their sources vary as they are collected from various points such as governments, hospitals, companies, and so on. Third, the data collected may contain errors. How data are processed and formatted before access requires a high level of analysis techniques to extract and determine knowledge and relationships hidden. To identify the relationships among different data and their influence on the results, they must be accurate and correct, as one type of data relies on the results of the analysis. Examples are the reasons for the spread of a particular disease in a particular area in the medical field, the losses incurred by a company after a change in business strategy, and the low standards of living in a society.

The main objective of the present research is to control management and sharing of data in the medical field, which mainly involves "patient data." Our main objective is to propose means to preserve information. The secondary objectives, which relate to the removal of sensitive data, are as follows: To evaluate and identify the parameters that negatively affect the management of shared patient data, thus determining the reasons behind the decrease in trust between private and health information communities, To evaluate and measure the efficiency of k-anonymization and generalization methods in privacy and misuse protection [2], To build a model that can help prevent shared patient data from being misused ,To test the information metric method using real medical information, To ensure high-quality information in every stage of the model. Some research questions on the control and sharing of data in knowledge management are as follows: How can data are kept unidentified? How can shared data be managed, ensuring that these benefit the target communities? What indexing methods should be followed to facilitate accurate and fast indexing of a case? How should the effect of perturbation on scientific analysis be measured, and what is an acceptable effect?

IV. The Proposed Model for Privacy Preserving of Medical Information

The proposed model by the present work consists of three stages. As explained in Fig. 3, the first stage is when the provider sends data from different databases into an expert database. At this stage, the problem is how to preserve the confidentiality of data sent to the main database. We assume that the connection between the data provider and the centralized database is characterized by trust [3]. The second stage is when the expert receives data in the database and recreates (re-processes) these before sending to the anonymizer engine that applies the k-anonymization and generalization technique. Thus, the second stage is designed for preparing and obtaining data. The third stage applies data mining algorithms such as analysis, which should identify the hidden relationships among various data and extract results supporting scientific research and decision making. The last stage is the publication of the results on the Web site. The interface used (published data) and the last version (results) should appear in a simple style to ensure understanding by the recipient [3].

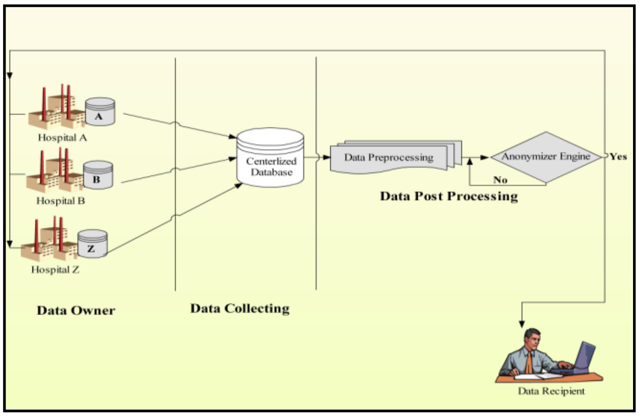

Fig. 3. The proposed model for privacy preserving of medical information

V. Conclusion

This work demonstrates the effort needed to set up a policy framework for the control and sharing of data in knowledge management in the medical field. Data sharing can help guide the nation’s adoption of health information technologies and improve the availability of health information and the quality of health care. The proposed control and sharing of data in knowledge management uses the k-anonymization model and generalization technique. The efficiency of these processes has been confirmed through the study and analysis of all processes involved and recent scientific research in the same domain. The control and sharing of data in knowledge management of medical information secure data between health care consumers and providers. The broad use of the proposed system has the potential to improve health care quality and prevent medical errors, thus increasing the efficiency of the care provided and reducing unnecessary health care costs. Moreover, the proposed system would increase administrative efficiency, expand access to affordable care, improve people’s health, and provide relevant data to support scientific research.

References

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, vol.10, no. 5, pp. 557-570, 2002.

[2] K. E. Emam, E. Jonker, and A. Fineberg. (2011). The Case for De-identifying Personal Health Information. [Online]. Available: http://www.researchgate.net/publication/228226957_

[3] S. Ray *et al.*, "Verification of data pattern for interactive privacy preservation model," in *Proc. the 2011 ACM Symposium on Applied Computing*, pp. 1716-1723, 2011.

[4] D. Jiang *et al.*, "Privacy-preserving dbscan on horizontally partitioned data," in *IT in Medicine and Education*, pp. 20-22, 2008.

[5] A. H. Rashid and A. F. Hegazy, "Protect privacy of medical informatics using k-anonymization model," in *Proc. Conference on 2010 The 7th International Informatics and Systems (INFOS)*, pp. 2-4, 2010

[6] K. El Emam, E. Jonker, and A. Fineberg, "The case for de-identifying personal health information," *Electronic Health Information Laboratory, Children's Hospital of Eastern Ontario Research Institute*, pp. 3, Canada, 2011.

[7] E. Bertino, I. N. Fovino, and L. P. Provenza, "A framework for evaluating privacy preserving data mining algorithms", *Data Mining and Knowledge Discovery*, 2005, vol. 11, no. 2, pp. 121-154.

Asmaa Hatem Rashid was born on February 25, 1983; she got his master degree in information science 2010, her joined into PhD program in University of Malaya\ department of Information Science - Faculty of Computer Science and Information Technology - Malaysia. she current research interests include Healthcare Information Systems (HISs) and privacy issues in this area regard to sharing healthcare data in research communities based on privacy preservation and keep data utility. Member in International Association of Computer Science and Information Technology (IASCIT).