Computers and Electrical Engineering 45 (2015) 134–140

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

An approach for prevention of privacy breach and information leakage in sensitive data mining ☆

M. Prakash [a,*], G. Singaravel [b]

[a] Department of Computer Science and Engineering, K.S.R. College of Engineering (Affiliated to Anna University, Chennai), Tiruchengode, Tamilnadu, India
[b] Department of Information Technology, K.S.R. College of Engineering (Affiliated to Anna University, Chennai), Tiruchengode, Tamilnadu, India

ARTICLE INFO

Article history:
Received 14 August 2014
Received in revised form 19 January 2015
Accepted 19 January 2015
Available online 25 February 2015

Keywords:
Anonymization
Data mining
Privacy
Privacy preserving
Privacy preserving techniques
Sensitive data publishing

ABSTRACT

Government agencies and many non-governmental organizations often need to publish sensitive data that contain information about individuals. The sensitive data or private data is an important source of information for the agencies like government and non-governmental organization for research and allocation of public funds, medical research and trend analysis. The important problem here is publishing data without revealing the sensitive information of individuals. This sensitive or private information of any individual is essential to several data repositories like medical data, census data, voter registration data, social network data and customer data. In this paper a personalized anonymization approach is proposed which preserves the privacy while the sensitive data is published. The main contributions of this paper are three folds: (i) the definition of the data collection and publication process, (ii) the privacy framework model and (iii) personalized anonymization approach. The experimental analysis is presented at the end; it shows this approach performs better over the distinct *l*-diversity measure, probabilistic *l*-diversity measure and *k*-anonymity with *t*-closeness measure.

1. Introduction

The data collected by public organizations and private organizations are increasing every day and stored in electronic repository. The data being collected includes private or sensitive data. More and more data mining techniques are becoming now a days and this can be used for assisting decision making process. The data mining techniques are used to extract the hidden knowledge from huge data collections in the form of trends, models and patterns. While doing the data mining process the personal data of any individual need to be protected from privacy concerns [1–4]. Privacy means the individual's personal information known as sensitive data has to be protected while publishing the data. In Schoeman [5] point of view three possible privacy definitions are proposed.

- Privacy is the right of a person. The person can decide which personal information to be communicated or published to others.
- Privacy is restricted access to an individual and to any or all the features associated with the person.
- Privacy is that the management over access to data or information relating to oneself.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. H. Abdul Shabeer.
* Corresponding author.

http://dx.doi.org/10.1016/j.compeleceng.2015.01.016
0045-7906/

In the above definitions the knowledge inferred is the Controlled Information Release, means that the personal data need to be hide while publishing for research and other purposes [6]. The Microdata or sensitive data is the individual's private data like salary, age, medical information, etc. A Microdata set can be viewed as a file with n records, where each and every record contains m attributes [7]. The attributes can be classified as follows.

1.1. Identifier

The attribute that unambiguously identify the individual is the identifier. Examples are name, social security number, passport number, etc.

1.2. Quasi-identifier

The attributes that identify the individual with some degree of ambiguity is the quasi-identifier. Examples are age, gender, address, telephone number, etc.

1.3. Confidential outcome attributes

The attributes that contains the sensitive information of an individual is the confidential outcome attributes. Examples are salary, health condition, religion, etc.

1.4. Non-confidential outcome attributes

The attributes that do not fall in any categories above are the non confidential outcome attributes.

The paper is organized as follows: The definition of the data collection and publication process is given in Section 2. In Section 3, the privacy framework model is described and the personalized anonymization approach is discussed. Finally in Section 4 the experimental analysis is presented with example. The efficiency factor is compared with distinct l-diversity measure, probabilistic l-diversity measure and k-anonymity with t-closeness measure.

2. Data collection and publication

The data collection and data publishing phases are described in Fig. 1. Here, in the data collection phase, the data are collected by the data publisher from the record owners. In the data publishing phase, the collected data are released by the data publisher to a researcher or to the public or to the data miner, called the data recipient. For example, a hospital collects the data from their patients and publishes the collected records to the external agency or medical centre for further research. In Fig. 1 John, Peter, Raj, Mary and Christy are the patients, who are the data owners. The hospital is the data publisher who collects the data from the data owners (patients like John, Peter) and publishes to the data recipient

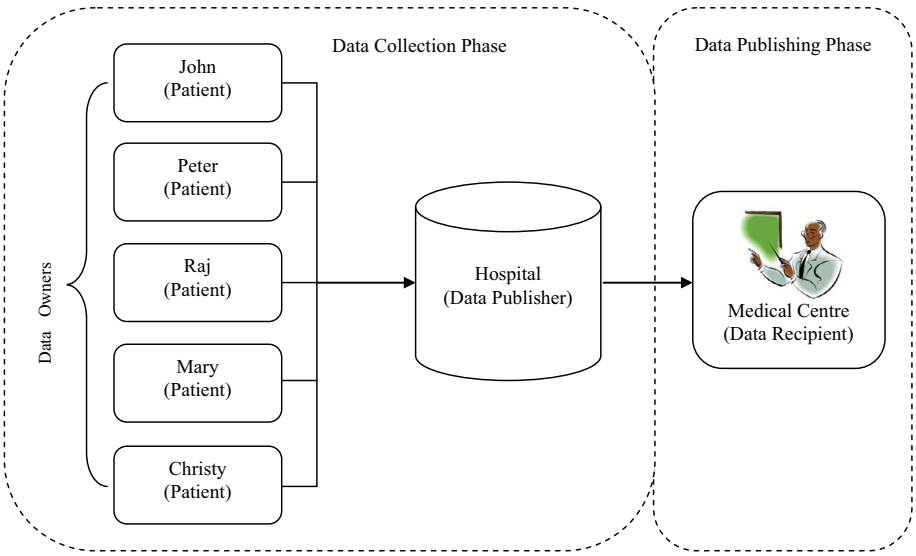

Fig. 1. Data collection and data publishing phase.