Chapter 34
Mutual Correlation-Based Anonymization for Privacy Preserving Medical Data Publishing

Ashoka Kukkuvada
Bapuji Institute of Engineering and Technology, India

Poornima Basavaraju
Bapuji Institute of Engineering and Technology, India

ABSTRACT

Currently the industry is focused on managing, retrieving, and securing massive amounts of data. Hence, privacy preservation is a significant concern for those organizations that publish/share personal data for vernacular analysis. In this chapter, the authors presented an innovative approach that makes use of information gain of the quasi attributes with respect to sensitive attributes for anonymizing the data, which gives the fruitfulness of an attribute in classifying the data elements, which is a two-way correlation among attributes. The authors show that the proposed approach preserves better data utility and has lesser complexity than former methods.

1. INTRODUCTION

The advancements in the field of information technology has improved our standard of living. With the lightening growth in computing, networking and database technologies results into collection and integration of tremendous amount of digital data. Data Mining involves the process of deriving functional, interesting and previously concealed information from the collection of large data bases. Present industry is focused on retrieving, managing and securing huge amount of data. For the purpose of business analytics or because of government policies, this data need to be shared/published among various organizations. For example, The US government open data and the data of 105 departments of government of India is published in the open data portals (U.S. Government's open data, (n.d.); Open

DOI: 10.4018/978-1-5225-7113-1.ch034

Government Data, (n.d.). Also, sharing of healthcare data helps in computer assisted clinical decision support. For example, Red Cross Blood Transfusion Service (BTS) is an organization that provides services that includes collecting and examine the blood from the donors and dispense the blood to various public hospitals. Government Health Agency in United States of America systematically collects patient's information from public hospitals that contains patient specific healthcare data. This patient specific healthcare data is shared with Red Cross Blood Transfusion Service (BTS) for the purpose of auditing and data analysis which can improve the estimated future blood consumption at different hospitals and also makes recommendations on the blood usage medical cases. Here the patient's privacy must be protected while sharing data between Government Health Agency and the Red Cross BTS. Figure-1 depicts the various stakeholders in the Red Cross BTS system. The blood is collected from the donors and after examination it will be distributed to various public hospitals. The hospitals transfuse the blood to the needed patients, also the hospitals are responsible for maintaining the patient health records and the blood transfusion information like name of the doctor in charge, type of illness, reason and amount of blood transfusion etc. Periodically public hospitals have to put forward, blood usage data along with individual patient‘s surgery data to Government Health Agency. The Government Health Agencies in turn submit this data to the Red Cross BTS for the purpose of auditing and analysis. The intention of this auditing and analysis is to enhance the subsequent blood consumption in several hospitals and to make suggestions on the imminent medical cases. Here, patient's privacy must be protected while sharing the data between hospitals and the Red Cross BTS.

Data publishing exists in other domains also. For example, the popular online movies rental service provider-Netflix, published a data set that consists of movie ratings of 500,000 members, to enhance the perfection of movie recommendations depending on personal preferences (Bennett & Lanning, 2007); AOL-a web portal and online service provider based in New York, published the query logs of 650,000 users, but deleted immediately for privacy matters.

Typically the data will be gathered from different locations in different format, and compiled into the format that is suitable to store in Data Warehouse. In this scenario the Data Warehouse is the data receiver who collect data from multiple data publishers. The data publisher, generally an independent

Figure 1. Scenario of Red Cross BTS system