Chapter 84
Privacy Preservation in Information System

D. P. Acharjya
VIT University, India

Geetha Mary A.
VIT University, India

ABSTRACT

The information technology revolution has brought drastic change in the way data is collected or generated for decision mining. The accumulated data has no relevance unless it provides certain useful information pertaining to the interest of an organization. The real challenge lies in converting high dimensional data into knowledge and to use this knowledge for the development of the organization. On the other hand, hiding an organization's sensitive information is a major concern. Much research has been carried out in this direction. This chapter discusses various privacy preservation techniques that can be employed in an information system to safeguard the sensitive information of an organization. This chapter also highlights sensitive fuzzy association rules that can be generated from an information system. The authors provide illustrations wherever necessary to give a clear idea of the concepts developed.

INTRODUCTION

Across the world, quintillions and trillions of bytes of data are produced and stored. The majority of the data are generated from social networks, email, blogs, sms etc. But most of these are unstructured and becomes hectic to generate knowledge from it. Business intelligence software's focuses on business point of view and it is not of much help. It is because the data collected from multiple sources may not be relevant for an organization. It leads to big data analysis which reacts to all the customer queries with an analysis. These analysis used for customer satisfaction like products a customer may like or a place a customer most likely to visit etc. Many of the institutions after generating the data go for publishing the data in World Wide Web for research purposes. According to Health Insurance Portability and Accountability Act, signed by President of United States, security rule updated on 2006, specifically states about technical safe guard of patient details. Patient safety rule of the Patient Safety and Quality

DOI: 10.4018/978-1-5225-7113-1.ch084

Improvement Act of 2005 specifies about confidentiality of data but allows minimum disclosure. During March and April of 2011, 8.3 and 10 million people were affected by security breaches. Almost 3 months once, privacy attacks happen and personal health information gets stolen. Nowadays patient details like transcription information are outsourced from USA to many of the countries like India, while doing so patient details need to be safeguarded. Therefore there is a need of privacy algorithms in order to adhere the privacy laws enforced at each country like USA, Italy, and India. Since a lot of projects are outsourced from USA, the data given outside the country should undergo privacy methods. In India, there is no specific act of privacy for patients regarding their health record disclosure, but the individual could file a petition in human rights law for privacy disclosure (Shrikant, 2010).

While publishing the data, institutions take enough measures to publish the data in a format that doesn't lead to identification of an individual person which in turn reveals the sensitive information of a person such as salary drawn by an employee or diseases a person is suffering from. Though personal identifiers like name, social security number, employee id, voters id, hospital id are removed from the published data, an individual can be identified if the hacker has some back ground information or any other supplementary materials by using data linkage techniques (Peter, 2012). The above said problems can be avoided by using privacy preserving data mining techniques. Usually data is either distorted or generalized in privacy preserving data mining, but the main decisive factor is the level to which the data is to be distorted or generalized so that there is no extensive change in exactness of data or the knowledge developed from it.

INFORMATION SYSTEM

The basic objective of inductive learning and data mining is to learn the knowledge for classification. However, in real world problems, we may not be faced with simply classification. One such problem is the ordering of objects. On the contrary, we are interested in hiding sensitive associations that is present in an information system. Before we discuss, various privacy preservation techniques to hide sensitive associations, one must know about an information system. An information system contains a finite set of objects typically represented by their values on a finite set of attributes. Such information system may be conveniently described in a tabular form in which each row represents an object whereas each column represents an attribute. Each cell of the information system contains an attribute value. Now, we define formally an information system as below.

An information system is defined as a quadruple $I = (U, A, V_a, f_a)$ where U is a finite nonempty set of objects called the universe, A is a finite nonempty set of attributes, V_a is a nonempty set of values for $a \in A$, $f_a : U \rightarrow V_a$ is an information function. For example, consider a sample information system as presented in Table 1 in which $U = \{o_1, o_2, o_3, o_4, o_5\}$ represents a nonempty finite set of objects; and A = {Humidity, Windy, Temperature} be a finite set of attributes. The information system presented in Table 1 is a qualitative system, where all the attribute values are discrete and categorical (qualitative).

In the information system shown in Table 2, $U = \{o_1, o_2, o_3, o_4, o_5\}$ represents a set of patients and A = {Temperature, Blood Pressure, Cholesterol} represents a finite set of attributes. This information system is a quantitative system, since all the attribute values are non categorical.