IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com

ITB11710

This chapter appears in the book, *Web and Information Security*
edited by Elena Ferrari and Bhavani Thuraisingham © 2006, Idea Group Inc.

Chapter XIV

Privacy-Preserving Data Mining on the Web: Foundations and Techniques

Stanley R. M. Oliveira, Embrapa Informática Agropecuária, Brazil

Osmar R. Zaïane, University of Alberta, Edmonton, Canada

Abstract

Privacy-preserving data mining (PPDM) is one of the newest trends in privacy and security research. It is driven by one of the major policy issues of the information era—the right to privacy. This chapter describes the foundations for further research in PPDM on the Web. In particular, we describe the problems we face in defining what information is private in data mining. We then describe the basis of PPDM including the historical roots, a discussion on how privacy can be violated in data mining, and the definition of privacy preservation in data mining based on users' personal information and information concerning their collective activities. Subsequently, we introduce a taxonomy of the existing PPDM techniques and a discussion on how these techniques are applicable to Web-based

applications. Finally, we suggest some privacy requirements that are related to industrial initiatives and point to some technical challenges as future research trends in PPDM on the Web.

Introduction

Analyzing what right to privacy means is fraught with problems, such as whether the exact definition of privacy constitutes a fundamental right and whether people are and should be concerned with it. Several definitions of privacy have been given, and they vary according to context, culture, and environment. For instance, in a seminal paper, Warren and Brandeis (1890) defined privacy as "the right to be alone". Later on, Westin (1967) defined privacy as "the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitude, and their behavior to others". Schoeman (1984) defined privacy as "the right to determine what (personal) information is communicated to others" or "the control an individual has over information about himself or herself". More recently, Garfinkel (2001) stated that "privacy is about self-possession, autonomy, and integrity". On the other hand, Rosenberg (2000) argues that privacy may not be a right after all but a taste: "If privacy is in the end a matter of individual taste, then seeking a moral foundation for it—beyond its role in making social institutions possible that we happen to prize—will be no more fruitful than seeking a moral foundation for the taste for truffles".

The above definitions suggest that, in general, privacy is viewed as a social and cultural concept. However, with the ubiquity of computers and the emergence of the Web, privacy has also become a digital problem. With the Web revolution and the emergence of data mining, privacy concerns have posed technical challenges fundamentally different from those that occurred before the information era. In the information technology era, privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others (Cockcroft & Clutterbuck, 2001).

In the context of data mining, the definition of privacy preservation is still unclear, and there is very little literature related to this topic. A notable exception is the work presented in Clifton, Kantarcioglu, and Vaidya (2002), in which PPDM is defined as "getting valid data mining results without learning the underlying data values". However, at this point, each existing PPDM