Chapter 7.10
Sanitization and Anonymization of Document Repositories

Yücel Saygin
Sabanci University, Turkey

Dilek Hakkani-Tür
AT&T Labs—Research, USA

Gökhan Tür
AT&T Labs—Research, USA

ABSTRACT

Information security and privacy in the context of the World Wide Web (WWW) are important issues that are still being investigated. However, most of the present research is dealing with access control and authentication-based trust. Especially with the popularity of WWW as one of the largest information sources, privacy of individuals is now as important as the security of information. In this chapter, our focus is text, which is probably the most frequently seen data type in the WWW. Our aim is to highlight the possible threats to privacy that exist due to the availability of document repositories and sophisticated tools to browse and analyze these documents. We first identify possible threats to privacy in document repositories. We then discuss a measure for privacy in documents with some possible solutions to avoid or, at least, alleviate these threats.

INTRODUCTION

Information has been published in various forms throughout the history, and sharing information has been one of the key aspects of development. The Internet revolution and World Wide Web (WWW) made publishing and accessing information much easier than it used to be. However, widespread data collection and publishing efforts on the WWW increased the privacy concerns since most of the gathered data contain private information. Privacy of individuals on the WWW may be jeopardized via search engines and browsers or sophisticated text mining tools that can dig through mountains of Web pages. Privacy concerns need to be addressed since they may hinder data collection efforts and reduce the number of publicly available databases that are extremely important for research purposes such as in machine learning, data mining, information extraction/retrieval, and natural language processing.

In this chapter, we consider the privacy issues that may originate from publishing data on the WWW. Since text is one of the most frequently and conveniently used medium in the WWW to convey information, our main focus will be text documents. We basically tackle the privacy problem in two phases. The first phase, referred to as *sanitization*, aims to protect the privacy of the contents of the text against possible threats. Sanitization basically deals with the automatic identification of named entities such as sensitive terms, phrases, proper names, and numeric values (e.g., credit card numbers) in a given text, and modification of them with the purpose of hiding private information. The second phase, called *anonymization*, makes sure that the classification tools cannot predict the owner or author of the text.

In the following sections, we first provide the taxonomy of possible threats. In addition to that, we propose a privacy metric for document databases based on the notion of *k*-anonymity together with a discussion of the methods that can be used for preserving privacy.

BACKGROUND AND RELATED WORK

Privacy and security issues were investigated in the database community in the context of statistical databases, where the users are limited to statistical queries. In statistical databases, privacy is protected by limiting the queries that can be issued by the user to non-confidential values, or statistical operations (Adam & Wortmann, 2004). Security leaks resulting from the intersection of multiple queries are investigated, and privacy is defined by the concept of *k*-anonymity. A database table provides *k*-anonymity if it cannot be used to unambiguously identify less than k entities (Samarati & Sweeney, 1998).

Currently, the data mining community is investigating how data could be mined without actually seeing the confidential values. This is called privacy preserving data mining which was introduced in Agrawal and Srikant (2000) for the case of classification model construction. Further research results have been published on various data mining models for preserving privacy (Evfimievski, Srikant, Agrawal, & Gehrke, 2002; Rizvi & Haritsa, 2002). Privacy preserving data mining on distributed data sources was another interesting research direction, which was addressed in Clifton and Kantarcioglu (2004) and Vaidya and Clifton (2004) for association rule mining and classification model construction. Another aspect of privacy issues in data mining is to protect the data against data mining algorithms. This result is due to the fact that data mining tools can be used to discover sensitive information. Hiding sensitive association rules by database sanitization is proposed in Saygin et al. (2001) and Verykios et al. (2004). Further research was conducted for data sanitization to protect the data against data mining tools (Oliveira & Zaïane, 2002, 2003). However, there is not much work about preserving privacy for natural language databases and its effects, except the studies of Ruch et al. (2000) and Sweeney (1996) who have worked on sanitization of medical reports on a limited domain.

On the other hand, information extraction (IE) has been studied extensively in the natural language processing community. IE is the task of extracting particular types of entities, relations, or events from natural language text or speech. The notion of what constitutes information extraction has been heavily influenced by the