Private Data Management in Collaborative Environments*

Larry Korba, Ronggong Song, George Yee, Andrew S. Patrick, Scott Buffett, Yunli Wang, and Liqiang Geng

Institute for Information Technology, National Research Council of Canada
Building M-50, Montreal Road, Ottawa, Ontario K1A 0R6
{Larry.Korba,Ronggong.Song,
George.Yee,Andrew.Patrick,Scott.Buffett,Yunli.Wang,
Liqiang.Geng}@nrc-cnrc.gc.ca
http://iit-iti.nrc-cnrc.gc.ca

Abstract. Organizations are under increasing pressures to manage all of the personal data concerning their customers and employees in a responsible manner. With the advancement of information and communication technologies, improved collaboration, and the pressures of marketing, it is very difficult to locate personal data is, let alone manage its use. In this paper, we outline the challenges of managing personally identifiable information in a collaborative environment, and describe a software prototype we call SNAP (Social Networking Applied to Privacy). SNAP uses automated workflow discovery and analysis, in combination with various text mining techniques, to support automated enterprise management of personally identifiable information.

Keywords: Privacy, compliance, workflow, social network analysis.

1 Introduction

The quantity of personal data that organizations must manage is increasing at a phenomenal rate. The main reason is the dramatic increase in the exploitation of communication and network technologies for collaboration, marketing, and sales. Other contributing factors include competitive pressures, as well as inexpensive computers and mass storage. While the amount of personal data is increasing, the prevalence of computer and network-based collaborations has made it very difficult for organizations to know where all the private data is stored and exactly how it is being used. This can occur despite attempts at controlling access to the data through centralization.

In the reality of the collaborative environments of today, the prospect of assuring privacy compliance, as may be required by legislation, regulations, or best practices, has become almost impossible. Our approach towards a solution is to combine several technologies in a manner that would allow organizations to understand and manage the life cycle of private data. The different technologies include: private data

* National Research Council Paper Number 49356.

Y. Luo (Ed.): CDVE 2007, LNCS 4674, pp. 88–96, 2007.
© Springer-Verlag Berlin Heidelberg 2007

discovery, social network (workflow) analysis, knowledge visualization, and effective human-computer interaction operating within a policy enforcement framework. Private data discovery involves text and data mining techniques to determine the location and use of personally identifiable information (PII) in different contexts across an organization. Social network analysis produces an understanding of workflow activities related to PII, providing measures for assessment of compliance with privacy policies, and a means for performing forensics analyses on the activities related to private data. In this paper we detail the challenges organizations have with privacy compliance, describe our progress in the development of a prototype for an automated privacy compliance system, outline early results, and detail the further challenges we are exploring.

2 Problem Description

The challenges for businesses today in the handling private data can be understood by referring to Figure 1. Key factors that put pressures on increasing the amount of private data collected include:

- Cheap Storage. Storage costs continue to drop so there are few impediments to collecting growing amounts of data of all sorts.
- Expanded Services. As new services are put in place, often more PII in different forms is collected and stored in order to assess product or service quality, and to facilitate follow-on sales.
- Marketing and Competition. Marketing pressures stemming from the desire to improve existing products and services or from competitive market conditions often lead to the collection and retention of more PII, e.g. e-service personalization where the consumer's contextual product selections are tracked for service improvement.
- Increasing Client Numbers. As an organization provides more products and services, and they become popular, more clients are garnered, leading to the collection and retention of larger amounts of PII.
- Computers Everywhere. Within organizations, desktop, portable, and handheld computers are being deployed at a high rate, due to their convenience and decreasing costs. The computers enable staff to share the work of creating and delivering services and products, which can lead to distributed, local storage of PII.

The pressures to decrease the amount of PII stored and managed within an organization include the following:

- Risk Management. Loss of customer PII is injurious not only to the customer but also to the organization. Data breaches can lead to lost sales due to a decrease in client trust. These risks can be reduced by minimizing PII collection and retention.
- Regulations and Policies. An organization may operate in a regulated sector (healthcare, banking, legal services, gaming, etc.) where there are specific, mandatory requirements for PII handling. In addition, an organization may have its own policies to manage its business and to evoke a stronger level of client trust.
- Legislation Enforcement. Beyond regulatory requirements, some jurisdictions, such as the European Union, may have legislation in place specifying how different types of PII must be handled.