Chapter 8
Data Mining Challenges in the Context of Data Retention

Konrad Stark
University of Vienna, Austria

Michael Ilger
Vienna University of Technology & University of Vienna, Austria

Wilfried N. Gansterer
University of Vienna, Austria

ABSTRACT

Retaining electronic communication and internet traffic data imposes novel technical and organisational challenges for internet service providers as well as for government authorities. ISP companies are not only burdened by storing extraordinary amounts of data, but also must develop and adhere to data protection and data security policies in order to protect the data against unauthorised access or disclosure and against accidental destruction. The authors present distributed, horizontally partitioned data warehouse architecture for retaining data at each internet service provider separately. Moreover, they elaborate a data warehouse schema for storing e-mail data according to the European data retention directive which facilitate parameterised data retrieval. The authors show how their system allows for applying various types of data mining techniques to both internet access and communication data. Finally, they discuss issues related to data security, cost and performance, and reveal limitations of data retention systems.

INTRODUCTION

The EU Data Retention directive 2006/24/EC ("Data Retention Directive") of the European Parliament, published on 15.03.2006, requires the operators of publicly accessible electronic communication networks to store ("retain") certain data which is generated or processed in their networks to serve the investigation, detection, and prosecution of serious crime (European Parliament, 2007). National service providers are required to implement and maintain the technical means needed to store and provide this data to government authorities upon request. For various categories of electronic communication, including Internet access, Internet e-mail and Internet telephony, the directive defines

DOI: 10.4018/978-1-60566-906-9.ch008

which data has to be retained. Affected are traffic and location data (but not the contents of the communication) for a period of time between six months and two years.

In this chapter, we discuss the application of data mining methods in the context of implementing this EU Directive which has implications for both, public and private sectors. Retaining electronic communication and internet traffic data imposes novel technical and organisational challenges for internet service providers as well as for government authorities. These challenges not only relate to the collection and the management of the data to be retained, but also to the analysis of the data, for example, when having to respond to queries posted by government authorities.

Challenges for the ISP company: A data retention system has to respond to enquiries of competent authorities 'without undue delay' (Elizalde, 2006). That is, instead of storing e-mail communication in log files, data has to be stored in a structured way facilitating *efficient* data retrieval. Although the directive specifies the mandatory information to be stored for each e-mail communication, no technical guidelines are given about how the information may be stored to support parameterised queries.

Challenges for the government authorities: The retained data is distributed among various ISPs which is particularly complicating the analysis of e-mail data. For instance, if an e-mail is sent from person A to person B which are customers of two different providers PV1 and PV2, two separate enquiries are necessary to identify both individuals. If the common social relationships of A and B are surveyed, two result sets are delivered by PV1 and PV2. In order to combine the result sets and perform analyses, standardised data structures are essential.

From the legal point of view, the ISP company is the owner of the customer data and responsible for it. The company must not retain e-mail data externally. Hence, a central data retention system hosted by authorities is not allowed. Further, for competitive reasons companies are usually not interested in storing valuable customer-related data outside their control. An authority may formulate an enquiry for a person as a result of an order of the court. In this case ISPs do have to deliver the communication data for a person timely. Thus, a data retention system is needed allowing distributed, standardized and protected data storage for ISPs and a secure central enquiry interface for the authority. Therefore, we encourage using a distributed data warehouse system to meet all these requirements.

Definitions

'A data warehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of management decisions' (Kimbal 1996).

Data warehouses (DWH) are designed to facilitate so-called Online Analytic Processing (OLAP) of data. That is, data is analysed interactively based on hypotheses. DWH data is stored in proprietary schemas that are optimised for data analysis.

In the following, we propose a data retention system with standardised data structures, query interfaces, data linkage and data analysis tools. We elaborate a distributed data warehouse that is composed of local data warehouse nodes residing at the ISP companies. We design a data warehouse schema for retaining e-mail and internet access data which satisfies the following requirements:

- store mandatory data according to the EU data retention directive,
- support person, time, and location-related enquiries by appropriate dimensions, and
- store additional information useful for data analysis.

RELATED WORK

Over the past years information technology faced novel demands from data retention requirements (Stampfel et al., 2008; Stampfel, G., &