Chapter 1.15
Data Mining in the Telecommunications Industry

Gary Weiss
Fordham University, USA

INTRODUCTION

The telecommunications industry was one of the first to adopt data mining technology. This is most likely because telecommunication companies routinely generate and store enormous amounts of high-quality data, have a very large customer base, and operate in a rapidly changing and highly competitive environment. Telecommunication companies utilize data mining to improve their marketing efforts, identify fraud, and better manage their telecommunication networks. However, these companies also face a number of data mining challenges due to the enormous size of their data sets, the sequential and temporal aspects of their data, and the need to predict very rare events—such as customer fraud and network failures—in real-time.

The popularity of data mining in the telecommunications industry can be viewed as an extension of the use of expert systems in the telecommunications industry (Liebowitz, 1988). These systems were developed to address the complexity associated with maintaining a huge network infrastructure and the need to maximize network reliability while minimizing labor costs. The problem with these expert systems is that they are expensive to develop because it is both difficult and time-consuming to elicit the requisite domain knowledge from experts. Data mining can be viewed as a means of automatically generating some of this knowledge directly from the data.

DOI: 10.4018/978-1-60566-010-3.ch076
BACKGROUND

The data mining applications for any industry depend on two factors: the data that are available and the business problems facing the industry. This section provides background information about the data maintained by telecommunications companies. The challenges associated with mining telecommunication data are also described in this section.

Telecommunication companies maintain data about the phone calls that traverse their networks in the form of call detail records, which contain descriptive information for each phone call. In 2001, AT&T long distance customers generated over 300 million call detail records per day (Cortes & Pregibon, 2001) and, because call detail records are kept online for several months, this meant that billions of call detail records were readily available for data mining. Call detail data is useful for marketing and fraud detection applications.

Telecommunication companies also maintain extensive customer information, such as billing information, as well as information obtained from outside parties, such as credit score information. This information can be quite useful and often is combined with telecommunication-specific data to improve the results of data mining. For example, while call detail data can be used to identify suspicious calling patterns, a customer’s credit score is often incorporated into the analysis before determining the likelihood that fraud is actually taking place.

Telecommunications companies also generate and store an extensive amount of data related to the operation of their networks. This is because the network elements in these large telecommunication networks have some self-diagnostic capabilities that permit them to generate both status and alarm messages. These streams of messages can be mined in order to support network management functions, namely fault isolation and prediction.

The telecommunication industry faces a number of data mining challenges. According to a Winter Corporation survey (2003), the three largest databases all belong to telecommunication companies, with France Telecom, AT&T, and SBC having databases with 29, 26, and 25 Terabytes, respectively. Thus, the scalability of data mining methods is a key concern. A second issue is that telecommunication data is often in the form of transactions/events and is not at the proper semantic level for data mining. For example, one typically wants to mine call detail data at the customer (i.e., phone-line) level but the raw data represents individual phone calls. Thus it is often necessary to aggregate data to the appropriate semantic level (Sasisekharan, Seshadri & Weiss, 1996) before mining the data. An alternative is to utilize a data mining method that can operate on the transactional data directly and extract sequential or temporal patterns (Klemettinen, Mannila & Toivonen, 1999; Weiss & Hirsh, 1998).

Another issue arises because much of the telecommunications data is generated in real-time and many telecommunication applications, such as fraud identification and network fault detection, need to operate in real-time. Because of its efforts to address this issue, the telecommunications industry has been a leader in the research area of mining data streams (Aggarwal, 2007). One way to handle data streams is to maintain a signature of the data, which is a summary description of the data that can be updated quickly and incrementally. Cortes and Pregibon (2001) developed signature-based methods and applied them to data streams of call detail records. A final issue with telecommunication data and the associated applications involves rarity. For example, both telecommunication fraud and network equipment failures are relatively rare. Predicting and identifying rare events has been shown to be quite difficult for many data mining algorithms (Weiss, 2004) and therefore this issue must be handled carefully in order to ensure reasonably good results.