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.

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.

MAIN FOCUS

Numerous data mining applications have been deployed in the telecommunications industry. However, most applications fall into one of the following three categories: marketing, fraud detection, and network fault isolation and prediction.

Telecommunications Marketing

Telecommunication companies maintain an enormous amount of information about their customers and, due to an extremely competitive environment, have great motivation for exploiting this information. For these reasons the telecommunications industry has been a leader in the use of data mining to identify customers, retain customers, and maximize the profit obtained from each customer. Perhaps the most famous use of data mining to acquire new telecommunications customers was MCI's Friends and Family program. This program, long since retired, began after marketing researchers identified many small but well connected subgraphs in the graphs of calling activity (Han, Altman, Kumar, Mannila & Pregibon, 2002). By offering reduced rates to customers in one's calling circle, this marketing strategy enabled the company to use their own customers as salesmen. This work can be considered an early use of social-network analysis and link mining (Getoor & Diehl, 2005). A more recent example uses the interactions between consumers to identify those customers likely to adopt new telecommunication services (Hill, Provost & Volinsky, 2006). A more traditional approach involves generating customer profiles (i.e., signatures) from call detail records and then mining these profiles for marketing purposes. This approach has been used to identify whether a phone line is being used for voice or fax (Kaplan, Strauss & Szegedy, 1999) and to classify a phone line as belonging to a either business or residential customer (Cortes & Pregibon, 1998).

Over the past few years, the emphasis of marketing applications in the telecommunications industry has shifted from identifying new customers to measuring customer value and then taking steps to retain the most profitable customers. This shift has occurred because it is much more expensive to acquire new telecommunication customers than retain existing ones. Thus it is useful to know the *total lifetime value* of a customer, which is the total net income a company can expect from that customer over time. A variety of data mining methods are being used to model customer lifetime value for telecommunication customers (Rosset, Neumann, Eick & Vatnik, 2003; Freeman & Melli, 2006).

A key component of modeling a telecommunication customer's value is estimating how long they will remain with their current carrier. This problem is of interest in its own right since if a company can predict when a customer is likely to leave, it can take proactive steps to retain the customer. The process of a customer leaving a company is referred to as *churn*, and *churn analysis* involves building a model of customer attrition. Customer churn is a huge issue in the telecommunication industry where, until recently, telecommunication companies routinely offered large cash incentives for customers to switch carriers. Numerous systems and methods have been developed to predict customer churn (Wei & Chin, 2002; Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000; Mani, Drew, Betz & Datta, 1999; Masand, Datta, Mani,