Data Mining in the Telecommunications Industry

Gary M. 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,

& Li, 1999). These systems almost always utilize call detail and contract data, but also often use other data about the customer (credit score, complaint history, etc.) in order to improve performance. Churn prediction is fundamentally a very difficult problem and, consequently, systems for predicting churn have been only moderately effective—only demonstrating the ability to identify some of the customers most likely to churn (Masand et al., 1999).

Telecommunications Fraud Detection

Fraud is very serious problem for telecommunication companies, resulting in billions of dollars of lost revenue each year. Fraud can be divided into two categories: subscription fraud and superimposition fraud (Fawcett and Provost, 2002). *Subscription fraud* occurs when a customer opens an account with the intention of never paying the account and *superimposition fraud* occurs when a perpetrator gains illicit access to the account of a legitimate customer. In this latter case, the fraudulent behavior will often occur in parallel with legitimate customer behavior (i.e., is superimposed on it). Superimposition fraud has been a much more significant problem for telecommunication companies than subscription fraud. Ideally, both subscription fraud and superimposition fraud should be detected immediately and the associated customer account deactivated or suspended. However, because it is often difficult to distinguish between legitimate and illicit use with limited data, it is not always feasible to detect fraud as soon as it begins. This problem is compounded by the fact that there are substantial costs associated with investigating fraud, as well as costs if usage is mistakenly classified as fraudulent (e.g., an annoyed customer).

The most common technique for identifying superimposition fraud is to compare the customer's current calling behavior with a profile of his past usage, using deviation detection and anomaly detection techniques. The profile must be able to be quickly updated because of the volume of call detail records and the need to identify fraud in a timely manner. Cortes and Pregibon (2001) generated a signature from a data stream of call-detail records to concisely describe the calling behavior of customers and then they used anomaly detection to "measure the unusualness of a new call relative to a particular account." Because new behavior does not necessarily imply fraud, this basic approach was augmented by comparing the new calling behavior to profiles of generic fraud—and fraud is only signaled if the behavior matches one of these profiles. Customer level data can also aid in identifying fraud. For example, price plan and credit rating information can be incorporated into the fraud analysis (Rosset, Murad, Neumann, Idan, & Pinkas, 1999). More recent work using signatures has employed dynamic clustering as well as deviation detection to detect fraud (Alves et al., 2006). In this work, each signature was placed within a cluster and a change in cluster membership was viewed as a potential indicator of fraud.

There are some methods for identifying fraud that do not involve comparing new behavior against a profile of old behavior. Perpetrators of fraud rarely work alone. For example, perpetrators of fraud often act as brokers and sell illicit service to others—and the illegal buyers will often use different accounts to call the same phone number again and again. Cortes and Pregibon (2001) exploited this behavior by recognizing that certain phone numbers are repeatedly called from compromised accounts and that calls to these numbers are a strong indicator that the current account may be compromised. A final method for detecting fraud exploits human pattern recognition skills. Cox, Eick & Wills (1997) built a suite of tools for visualizing data that was tailored to show calling activity in such a way that unusual patterns are easily detected by users. These tools were then used to identify international calling fraud.

Telecommunication Network Fault Isolation and Prediction

Monitoring and maintaining telecommunication networks is an important task. As these networks became increasingly complex, expert systems were developed to handle the alarms generated by the network elements (Weiss, Ros & Singhal, 1998). However, because these systems are expensive to develop and keep current, data mining applications have been developed to identify and predict network faults. Fault identification can be quite difficult because a single fault may result in a cascade of alarms—many of which are not associated with the root cause of the problem. Thus an important part of fault identification is alarm correlation, which enables multiple alarms to be recognized as being related to a single fault.

The Telecommunication Alarm Sequence Analyzer (TASA) is a data mining tool that aids with fault identification by looking for frequently occurring temporal patterns of alarms (Klemettinen, Mannila & Toivonen, 1999). Patterns detected by this tool were then used to help construct a rule-based alarm correlation system. Another effort, used to predict telecommunication switch failures, employed a genetic algorithm to mine historical alarm logs looking for predictive sequential and temporal patterns (Weiss & Hirsh, 1998). One limitation with the approaches just described is that they ignore the structural information about the underlying network. The quality of the mined sequences can be improved if topological proximity constraints are considered in the data mining process (Devitt, Duffin and Moloney, 2005) or if substructures in the telecommunication data can be identified and exploited to allow simpler, more useful, patterns to be learned (Baritchi, Cook, & Lawrence, 2000). Another approach is to use Bayesian Belief Networks to identify faults, since they can reason about causes and effects (Sterritt, Adamson, Shapcott & Curran, 2000).

FUTURE TRENDS

Data mining should play an important and increasing role in the telecommunications industry due to the large amounts of high quality data available, the competitive nature of the industry and the advances being made in data mining. In particular, advances in mining data streams, mining sequential and temporal data, and predicting/classifying rare events should benefit the telecommunications industry. As these and other advances are made, more reliance will be placed on the knowledge acquired through data mining and less on the knowledge acquired through the time-intensive process of eliciting domain knowledge from experts—although we expect human experts will continue to play an important role for some time to come.

Changes in the nature of the telecommunications industry will also lead to the development of new applications and the demise of some current applications. As an example, the main application of fraud detection in the telecommunications industry used to be in cellular cloning fraud, but this is no longer the case because the problem has been largely eliminated due to technological advances in the cell phone authentication process. It is difficult to predict what future changes will face the telecommunications industry, but as telecommunication companies start providing television service to the home and more sophisticated cell phone services become available (e.g., music, video, etc.), it is clear that new data mining applications, such as recommender systems, will be developed and deployed.

Unfortunately, there is also one troubling trend that has developed in recent years. This concerns the increasing belief that U.S. telecommunication companies are too readily sharing customer records with governmental agencies. This concern arose in 2006 due to revelations—made public in numerous newspaper and magazine articles—that telecommunications companies were turning over information on calling patterns to the National Security Agency (NSA) for purposes of data mining (Krikke, 2006). If this concern continues to grow unchecked, it could lead to restrictions that limit the use of data mining for legitimate purposes.

CONCLUSION

The telecommunications industry has been one of the early adopters of data mining and has deployed numerous data mining applications. The primary applications relate to marketing, fraud detection, and network monitoring. Data mining in the telecommunications industry faces several challenges, due to the size of the data sets, the sequential and temporal nature of the data, and the real-time requirements of many of the applications. New methods have been developed and existing methods have been enhanced to respond to these challenges. The competitive and changing nature of the industry, combined with the fact that the industry generates enormous amounts of data, ensures that data mining will play an important role in the future of the telecommunications industry.

REFERENCES

Aggarwal, C. (Ed.). (2007). *Data Streams: Models and Algorithms*. New York: Springer.

Alves, R., Ferreira, P., Belo, O., Lopes, J., Ribeiro, J., Cortesao, L., & Martins, F. (2006). Discovering telecom fraud situations through mining anomalous behavior patterns. *Proceedings of the ACM SIGKDD Workshop on Data Mining for Business Applications* (pp. 1-7). New York: ACM Press.

Baritchi, A., Cook, D., & Holder, L. (2000). Discovering structural patterns in telecommunications data. *Proceedings of the Thirteenth Annual Florida AI Research Symposium* (pp. 82-85).

Cortes, C., & Pregibon, D (1998). Giga-mining. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 174-178). New York, NY: AAAI Press.

Cortes, C., & Pregibon, D. (2001). Signature-based methods for data streams. *Data Mining and Knowledge Discovery*, 5(3), 167-182.

Cox, K., Eick, S., & Wills, G. (1997). Visual data mining: Recognizing telephone calling fraud. *Data Mining and Knowledge Discovery*, 1(2), 225-231.

Devitt, A., Duffin, J., & Moloney, R. (2005). Topographical proximity for mining network alarm data. *Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data* (pp. 179-184). New York: ACM Press.

Fawcett, T., & Provost, F. (2002). Fraud Detection. In W. Klosgen & J. Zytkow (Eds.), *Handbook of Data Mining and Knowledge Discovery* (pp. 726-731). New York: Oxford University Press.

Freeman, E., & Melli, G. (2006). Championing of an LTV model at LTC. *SIGKDD Explorations*, 8(1), 27-32.

Getoor, L., & Diehl, C.P. (2005). Link mining: A survey. *SIGKDD Explorations*, 7(2), 3-12.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2), 256-276.

Kaplan, H., Strauss, M., & Szegedy, M. (1999). Just the fax—differentiating voice and fax phone lines using call billing data. *Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 935-936). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Klemettinen, M., Mannila, H., & Toivonen, H. (1999). Rule discovery in telecommunication alarm data. *Journal of Network and Systems Management*, 7(4), 395-423.

Krikke, J. (2006). Intelligent surveillance empowers security analysts. *IEEE Intelligent Systems*, 21(3), 102-104.

Liebowitz, J. (1988). *Expert System Applications to Telecommunications*. New York, NY: John Wiley & Sons.

Mani, D., Drew, J., Betz, A., & Datta, P (1999). Statistics and data mining techniques for lifetime value modeling. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 94-103). New York, NY: ACM Press.

Masand, B., Datta, P., Mani, D., & Li, B. (1999). CHAMP: A prototype for automated cellular churn prediction. *Data Mining and Knowledge Discovery*, 3(2), 219-225.

Mozer, M., Wolniewicz, R., Grimes, D., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunication industry. *IEEE Transactions on Neural Networks*, 11, 690-696.

Rosset, S., Murad, U., Neumann, E., Idan, Y., & Gadi, P. (1999). Discovery of fraud rules for telecommunications—challenges and solutions. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 409-413). New York: ACM Press.

Rosset, S., Neumann, E., Eick, U., & Vatnik (2003). Customer lifetime value models for decision support. *Data Mining and Knowledge Discovery*, 7(3), 321-339.

Sasisekharan, R., Seshadri, V., & Weiss, S (1996). Data mining and forecasting in large-scale telecommunication networks. *IEEE Expert*, 11(1), 37-43.

Sterritt, R., Adamson, K., Shapcott, C., & Curran, E. (2000). Parallel data mining of Bayesian networks from telecommunication network data. *Proceedings of the 14th International Parallel and Distributed Processing Symposium*, IEEE Computer Society.

Wei, C., & Chiu, I (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications*, 23(2), 103-112.

Weiss, G., & Hirsh, H (1998). Learning to predict rare events in event sequences. In R. Agrawal & P. Stolorz (Eds.), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 359-363). Menlo Park, CA: AAAI Press.

Weiss, G., Ros, J., & Singhal, A. (1998). ANSWER: Network monitoring using object-oriented rule. *Proceedings of the Tenth Conference on Innovative Applications of Artificial Intelligence* (pp. 1087-1093). Menlo Park: AAAI Press.

Weiss, G. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations*, 6(1), 7-19.

Winter Corporation (2003). *2003 Top 10 Award Winners*. Retrieved October 8, 2005, from http://www.wintercorp.com/VLDB/2003_TopTen_Survey/TopTenwinners.asp

KEY TERMS

Bayesian Belief Network: A model that predicts the probability of events or conditions based on causal relations.

Call Detail Record: Contains the descriptive information associated with a single phone call.

Churn: Customer attrition. Churn prediction refers to the ability to predict that a customer will leave a company before the change actually occurs.

Signature: A summary description of a subset of data from a data stream that can be updated incrementally and quickly.

Subscription Fraud: Occurs when a perpetrator opens an account with no intention of ever paying for the services incurred.

Superimposition Fraud: Occurs when a perpetrator gains illicit access to an account being used by a legitimate customer and where fraudulent use is “superimposed” on top of legitimate use.

Total Lifetime Value: The total net income a company can expect from a customer over time.