DATA MINING IN TELECOMMUNICATIONS

Gary M. Weiss
Department of Computer and Information Science
Fordham University

Abstract: Telecommunication companies generate a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers. This chapter describes how data mining can be used to uncover useful information buried within these data sets. Several data mining applications are described and together they demonstrate that data mining can be used to identify telecommunication fraud, improve marketing effectiveness, and identify network faults.

Key words: Telecommunications, fraud detection, marketing, network fault isolation.

1. INTRODUCTION

The telecommunications industry generates and stores a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network, and customer data, which describes the telecommunication customers. The amount of data is so great that manual analysis of the data is difficult, if not impossible. The need to handle such large volumes of data led to the development of knowledge-based expert systems. These automated systems performed important functions such as identifying fraudulent phone calls and identifying network faults. The problem with this approach is that it is time-consuming to obtain the knowledge from human experts (the "knowledge acquisition bottleneck") and, in many cases, the experts do not have the requisite knowledge. The advent of data mining technology promised solutions to these problems and for this reason the telecommunications industry was an early adopter of data mining technology.

Telecommunication data pose several interesting issues for data mining. The first concerns scale, since telecommunication databases may contain billions of records and are amongst the largest in the world. A second issue is that the raw data is often not suitable for data mining. For example, both call detail and network data are time-series data that represent individual events. Before this data can be effectively mined, useful “summary” features must be identified and then the data must be summarized using these features. Because many data mining applications in the telecommunications industry involve predicting very rare events, such as the failure of a network element or an instance of telephone fraud, rarity is another issue that must be dealt with. The fourth and final data mining issue concerns real-time performance: many data mining applications, such as fraud detection, require that any learned model/rules be applied in real-time. Each of these four issues are discussed throughout this chapter, within the context of real data mining applications.

2. TYPES OF TELECOMMUNICATION DATA

The first step in the data mining process is to understand the data. Without such an understanding, useful applications cannot be developed. In this section we describe the three main types of telecommunication data. If the raw data is not suitable for data mining, then the transformation steps necessary to generate data that can be mined are also described. Section 3 will show how data mining can be used to extract useful information from these data sets.

2.1 Call Detail Data

Every time a call is placed on a telecommunications network, descriptive information about the call is saved as a *call detail* record. The number of call detail records that are generated and stored is huge. For example, AT&T long distance customers alone generate over 300 million call detail records per day (Cortes & Pregibon, 2001). Given that several months of call detail data is typically kept online, this means that tens of billions of call detail records will need to be stored at any time.

Call detail records include sufficient information to describe the important characteristics of each call. At a minimum, each call detail record will include the originating and terminating phone numbers, the date and time of the call and the duration of the call. Call detail records are generated in real-time and therefore will be available almost immediately for data mining. This can be contrasted with billing data, which is typically made available only once per month.

Call detail records are not used directly for data mining, since the goal of data mining applications is to extract knowledge at the customer level, not at the level of individual phone calls. Thus, the call detail records associated with a customer must be summarized into a single record that describes the customer's calling behavior. The choice of summary variables (i.e., features) is critical in order to obtain a useful description of the customer. Below is a list of features that one might use when generating a summary description of a customer based on the calls they originate and receive over some time period P:

1. average call duration
2. % no-answer calls
3. % calls to/from a different area code
4. % of weekday calls (Monday – Friday)
5. % of daytime calls (9am – 5pm)
6. average # calls received per day
7. average # calls originated per day
8. # unique area codes called during P

These eight features can be used to build a customer profile. Such a profile has many potential applications. For example, it could be used to distinguish between business and residential customers based on the percentage of weekday and daytime calls. Most of the eight features listed above were generated in a straightforward manner from the underlying data, but some features, such as the eighth feature, required a little more thought and creativity. Because most people call only a few area codes over a reasonably short period of time (e.g., a month), this feature can help identify telemarketers, or telemarketing behavior, since telemarketers will call many different area codes.

The above example demonstrates that generating useful features, including summary features, is a critical step within the data mining process. Should poor features be generated, data mining will not be successful. Although the construction of these features may be guided by common sense and expert knowledge, it should include exploratory data analysis. For example, the use of the time period 9am-5pm in the fifth feature is based on the commonsense knowledge that the typical workday is 9 to 5 (and hence this feature may be useful in distinguishing between business and residential calling patterns). However, more detailed exploratory data analysis, shown in Figure 1, indicates that the period from 9am to 4pm is actually more appropriate for this purpose. Figure 1 plots, for each weekday hour, *h*, the business to residential call ratio, *which* is computed as:

% weekday business calls during h / % weekday residential calls during h

Thus, this figure shows that during the period of 9am to 4pm, businesses place roughly 1.5 times as many of their total weekday calls as does a residence. Note that at 5pm the ratio is close to 1, indicating that the calls during this timeframe are not very useful for distinguishing between a business and a residence. However, calls in the evening timeframe (6pm – 1am) are also useful in distinguishing between the two types of customers.

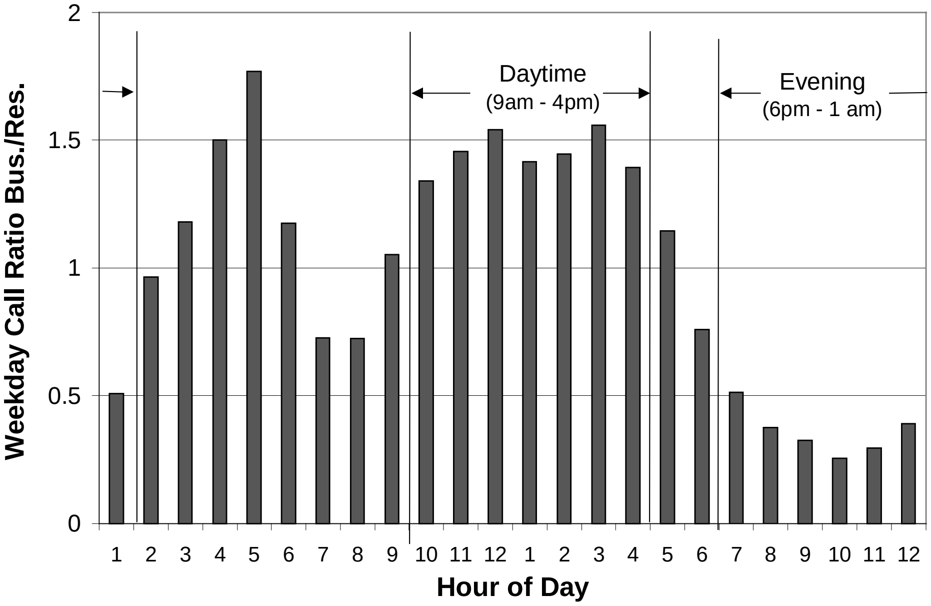

Figure -1Comparison of Business and Residential Hourly Calling Patterns

For some applications, such as fraud detection, the summary descriptions, sometimes called signatures (Cortes & Pregibon, 2001), must be updated in real-time for millions of phone lines. This requires the use of fairly short and simple summary features that can be updated quickly and efficiently.

2.2 Network Data

Telecommunication networks are extremely complex configurations of equipment, comprised of thousands of interconnected components. Each network element is capable of generating error and status messages, which leads to a tremendous amount of network data. This data must be stored and analyzed in order to support network management functions, such as fault isolation. This data will minimally include a timestamp, a string that uniquely identifies the hardware or software component generating the message and a code that explains why the message is being generated. For example, such a message might indicate that “controller 7 experienced a loss of power for 30 seconds starting at 10:03 pm on Monday, May 12.”

Due to the enormous number of network messages generated, technicians cannot possibly handle every message. For this reason expert systems have been developed to automatically analyze these messages and take appropriate action, only involving a technician when a problem cannot be automatically resolved (Weiss, Ros & Singhal, 1998). As described in Section 3, data mining technology is now helping identify network faults by automatically extracting knowledge from the network data.

As was the case with the call detail data, network data is also generated in real-time as a data stream and must often be summarized in order to be useful for data mining. This is sometimes accomplished by applying a time window to the data. For example, such a summary might indicate that a hardware component experienced twelve instances of a power fluctuation in a 10-minute period.

2.3 Customer Data

Telecommunication companies, like other large businesses, may have millions of customers. By necessity this means maintaining a database of information on these customers. This information will include name and address information and may include other information such as service plan and contract information, credit score, family income and payment history. This information may be supplemented with data from external sources, such as from credit reporting agencies. Because the customer data maintained by telecommunication companies does not substantially differ from that maintained in most other industries, the applications described in Section 3 do not focus on this source of data. However, customer data is often used in conjunction with other data in order to improve results. For example, customer data is typically used to supplement call detail data when trying to identify phone fraud.

3. DATA MINING APPLICATIONS

The telecommunications industry was an early adopter of data mining technology and therefore many data mining applications exist. Several typical applications are described in this section. These applications are divided into three application areas: fraud detection, marketing/customer profiling and network fault isolation.

3.1 Fraud Detection

Fraud is a serious problem for telecommunication companies, leading to billions of dollars in lost revenue each year. Fraud can be divided into two categories: subscription fraud and superimposition fraud. Subscription fraud occurs when a customer opens an account with the intention of never paying for the account charges. Superimposition fraud involves a legitimate account with some legitimate activity, but also includes some "superimposed" illegitimate activity by a person other than the account holder. Superimposition fraud poses a bigger problem for the telecommunications industry and for this reason we focus on applications for identifying this type of fraud. These applications should ideally operate in real-time using the call detail records and, once fraud is detected or suspected, should trigger some action. This action may be to immediately block the call and/or deactivate the account, or may involve opening an investigation, which will result in a call to the customer to verify the legitimacy of the account activity.

The most common method for identifying fraud is to build a profile of customer's calling behavior and compare recent activity against this behavior. Thus, this data mining application relies on deviation detection. The calling behavior is captured by summarizing the call detail records for a customer, as described earlier in this chapter. If the call detail summaries are updated in real-time, fraud can be identified soon after it occurs. Because new behavior does not necessarily imply fraud, one fraud-detection system augments this basic approach by comparing the new calling behavior to profiles of generic fraud—and only signals fraud if the behavior matches one of these profiles (Cortes & Pregibon, 2001). Customer level data can also aid in identifying fraud. For example, one sample rule that combines call detail and customer level data for detecting cellular fraud is: "People who have a price plan that makes international calls expensive and who display a sharp rise in international calls are likely the victim of cloning fraud."

This same basic approach has been used to identify cellular cloning fraud, which occurs when the identification information associated with one cell phone is monitored and then programmed into a second phone (cloning fraud was a very serious problem in the 1990's, until authentication methods were developed to eliminate this type of fraud). This data mining application analyzed large amounts of cellular call data in order to identify patterns of fraud (Fawcett & Provost, 1997). These patterns were then used to generate monitors, each of which watches a customer's behavior with respect to one pattern of fraud. These monitors were then fed into a neural network, which determined when there is sufficiently evidence of fraud to raise an alert. Data mining can also help detect fraud by identifying and storing those phone numbers called when a phone is known to be used fraudulently. If many calls originate from another phone to numbers on this list of “suspect” phone numbers, one may infer that the account is being use fraudulently (Cortes & Pregibon, 2001).

Fraud applications have some characteristics that require modifications to standard data mining techniques. For example, the performance of a fraud detection system should be computed at the customer level, not at the individual call level. So, if a customer account generates 20 fraud alerts, this should count, when computing the accuracy of this system, as only one alert; otherwise the system may appear to perform better than it actually does (Rosset, Murad, Neumann, Idan & Pinkas, 1999). More sophisticated cost-based metrics can also be used to evaluate the system. This is important because misclassification costs for fraud are generally unequal and often highly skewed (Ezawa & Norton, 1995). For this reason, when building a classifier to identify fraud, one should ideally know the relative cost of letting a fraudulent call go through versus the cost of blocking a call from a legitimate customer.

Another issue is that since fraud is relatively rare—and the number of verified fraudulent calls is relatively low—the fraud application involves predicting a relatively rare event where the underlying class distribution is highly skewed. Data mining algorithms often have great difficulty dealing with highly skewed class distributions and predicting rare events. For example, if fraud makes up only .2% of all calls, many data mining systems will not generate any rules for finding fraud, since a default rule, which never predicts fraud, would be 98.8% accurate. To deal with this issue, the training data is often selected to increase the proportion of fraudulent cases. For example, Ezawa & Norton (1995) increase the percentage of fraudulent calls from 1-2% to 9-12%. However, the use of a non-representative training set can be problematic because it does not provide the data mining method with accurate information about the true class distribution (Weiss and Provost, 2003).

3.2 Marketing/Customer Profiling

Telecommunication companies maintain a great deal of data about their customers. In addition to the general customer data that most businesses collect, telecommunication companies also store call detail records, which precisely describe the calling behavior of each customer. This information can be used to profile the customers and these profiles can then be used for marketing and/or forecasting purposes.

We begin with one of the most well-known and successful marketing campaigns in the telecommunications industry: MCI's Friends and Family promotion. This promotion was initially launched in the United States in 1991 and, although now retired, was responsible for significant growth in MCI's customer base. The promotion offered reduced calling fees when calls are placed to others in one's calling circle. This promotion purportedly originated when market researchers noticed small subgraphs in the call-graph of network activity—which suggested the possibility of adding entire calling circles rather than the costly approach of adding individual subscribers (Han, Altman, Kumar, Mannila & Pregibon, 2002). It is worth noting that MCI relied primarily on its customers to bring in members of their calling circle, even though MCI could have utilized its call detail data to generate a list of the people in each calling circle. The most likely reason for this is that MCI did not want to anger its customers by using highly personal information (calling history). This demonstrates that privacy concerns are an issue for data mining in the telecommunications industry, especially when call detail data is involved.

The MCI Friends and Family promotion relied on data mining to identify associations within data. Another marketing application that relies on this technique is a data mining application for finding the set of non-U.S. countries most often called together by U.S. telecommunication customers (Cortes & Pregibon, 2001). One set of countries identified by this data mining application is: {Jamaica, Antigua, Grenada, Dominica}. This information is useful for establishing and marketing international calling plans.

A serious issue with telecommunication companies is *customer churn*. Customer churn involves a customer leaving one telecommunication company for another. Customer churn is a significant problem because of the associated loss of revenue and the high cost of attracting new customers. Some of the worst cases of customer churn occurred several years ago when competing long distance companies offered special incentives, typically $50 or $100, for signing up with their company—a practice which led to customers repeatedly switching carriers in order to earn the incentives. Data mining techniques now permit companies the ability to mine historical data in order to predict when a customer is likely to leave. These techniques typically utilize billing data, call detail data, subscription information (calling plan, features, contract expiration data) and customer information (e.g., age). Based on the induced model, the company can then take action, if desired. For example, a wireless company might offer a customer a free phone for extending their contract. One such effort utilized a neural network to estimate the probability h(t) of cancellation at a given time t in the future (Mani, Drew, Betz & Datta, 1999).

In the telecommunications industry, it is often useful to profile customers based on their patterns of phone usage, which can be extracted from the call detail data. These customer profiles can then be used for marketing purposes, or to better understand the customer, which in turn may lead to better forecasting models. In order to effectively mine the call detail data, it must be summarized to the customer level as described earlier in this chapter. Then, a classifier induction program can be applied to a set of labeled training examples in order to build a classifier. This approach has been used to identify fax lines (Kaplan, Strauss & Szegedy, 1999) and to classify a phone line as belonging to a business or residence (Cortes & Pregibon, 1998). Other applications have used this approach to identify phone lines belonging to telemarketers and to classify a phone line as being used for voice, data, or fax.

Two sample rules for classifying a customer as being a business or residential customer are shown below (using pseudo-code). These rules were generated using SAS Enterprise Miner, a sophisticated data mining package that supports multiple data mining techniques. The rules shown below were generated using a decision tree learner. However, a neural network was also used to predict the probability of a customer being a business or residential customer, based solely on the distribution of calls by time of day (i.e., the neural network had 24 inputs, one per hour of the day). The probability estimate generated by the neural network was then used as an input (i.e., feature) to the decision tree learner. Evaluation on a separate test set indicates that rule 1 is 88% accurate and rule 2 is 70% accurate.

Rule 1: *if* < 43% of calls last 0-10 seconds *and* < 13.5% of calls occur during the weekend *and* neural network says that P(business) > 0.58 based on time of day call distribution *then* business customer

Rule 2: *if* calls received over two-month period from at most 3 unique area codes *and* <56.6% of calls last 0-10 seconds *then* residential customer

It is worth noting that because a telecommunications company generates a call detail record if the calling (paying) party is its customer, the company will also have a sample of (received) calls for non-customers. If a company has high overall market penetration, this sample may be large enough for data mining. Thus, telecommunication companies have the technical ability to profile non-customers as well as customers. However, there are legal restrictions on the use of this data, some of which are described in Section 4.

3.3 Network Fault Isolation

Telecommunication networks are extremely complex configurations of hardware and software. Most of the network elements are capable of at least limited self-diagnosis, and these elements may collectively generate millions of status and alarm messages each month. In order to effectively manage the network, alarms must be analyzed automatically in order to identify network faults in a timely manner—or before they occur and degrade network performance. A proactive response is essential to maintaining the reliability of the network. Because of the volume of the data, and because a single fault may cause many different, seemingly unrelated, alarms to be generated, the task of network fault isolation is quite difficult. Data mining has a role to play in generating rules for identifying faults.

The Telecommunication Alarm Sequence Analyzer (TASA) is one tool that helps with the knowledge acquisition task for alarm correlation (Klemettinen, Mannila & Toivonen, 1999). This tool automatically discovers recurrent patterns of alarms within the network data along with their statistical properties, using a specialized data mining algorithm. Network specialists then use this information to construct a rule-based alarm correlation system, which can then be used in real-time to identify faults. TASA is capable of finding episodic rules that depend on temporal relationships between the alarms. For example, it may discover the following rule: "if alarms of type *link alarm* and *link failure* occur within 5 seconds, then an alarm of type *high fault rate* occurs within 60 seconds with probability 0.7."

Before standard classification tasks can be applied to the problem of network fault isolation, the underlying time-series data must be re-represented as a set of classified examples. This summarization, or aggregation, process typically involves using a fixed time window and characterizing the behavior over this window. For example, if *n* unique alarms are possible, one could describe the behavior of a device over this time window using a scalar of length *n*. In this case each field in the scalar would contain a count of the number of times a specific alarm occurs. One may then label the constructed example based on whether a fault occurs within some other time frame, for example, within the following 5 minutes. Thus, two time windows are required. Once this encoding is complete, standard classification tools can be used to generate "rules" to predict future failures. Such an encoding scheme was used to identify chronic circuit problems (Sasisekharan, Seshadri & Weiss, 1996). The problem of reformulating time-series network events so that conventional classification-based data mining tools can be used to identify network faults has been studied. Weiss & Hirsh (1998) view this task as an event prediction problem while Fawcett & Provost (1999) view it as an activity monitoring problem.

Transforming the time-series data so that standard classification tools can be used has several drawbacks. The most significant one is that some information will be lost in the reformulation process. For example, using the scalar-based representation just mentioned, all sequence information is lost. Timeweaver (Weiss & Hirsh, 1998) is a genetic-algorithm based data mining system that is capable of operating directly on the raw network-level time-series data (as well as other time-series data), thereby making it unnecessary to re-represent the network level data. Given a sequence of timestamped events and a target event T, Timeweaver will identify patterns that successfully predict T. Timeweaver essentially searches through the space of possible patterns, which includes sequence and temporal relationships, to find predictive patterns. The system is especially designed to perform well when the target event is rare, which is critical since most network failures are rare. In the case studied, the target event is the failure of components in the 4ESS switching system.

4. CONCLUSION

This chapter described how data mining is used in the telecommunications industry. Three main sources of telecommunication data (call detail, network and customer data) were described, as were common data mining applications (fraud, marketing and network fault isolation). This chapter also highlighted several key issues that affect the ability to mine data, and commented on how they impact the data mining process. One central issue is that telecommunication data is often not in a form—or at a level—suitable for data mining. Other data mining issues that were discussed include the large *scale* of telecommunication data sets, the need to identify very *rare* events (e.g., fraud and equipment failures) and the need to operate in real-time (e.g., fraud detection).

Data mining applications must always consider privacy issues. This is especially true in the telecommunications industry, since telecommunication companies maintain highly private information, such as whom each customer calls. Most telecommunication companies utilize this information conscientiously and consequently privacy concerns have thus far been minimized. A more significant issue in the telecommunications industry relates to specific legal restrictions on how data may be used. In the United States, the information that a telecommunications company acquires about their subscribers is referred to as Customer Proprietary Network Information (CPNI) and there are specific restrictions on how this data may be used. The

Telecommunications Act of 1996, along with more recent clarifications from the Federal Communications Commission, generally prohibits the use of that information without customer permission, even for the purpose of marketing the customers other services. In the case of customers who switch to other service providers, the original service provider is prohibited from using the information to try to get the customer back (e.g., by only targeting profitable customers). Furthermore, companies are prohibited from using data from one type of service (e.g., wireless) in order to sell another service (e.g.. landline services). Thus, the use of data mining is restricted in that there are many instances in which useful knowledge extracted by the data mining process cannot be legally exploited. Much of the rationale for these prohibitions relates to competition. For example, if a large company can leverage the data associated with one service to sell another service, then companies that provide fewer services would be at a competitive disadvantage.

The telecommunications industry has been one of the earliest adopters of data mining technology, largely because of the amount and quality of the data that it collects. This has resulted in many successful data mining applications. Given the fierce competition in the telecommunications industry, one can only expect the use of data mining to accelerate, as companies strive to operate more efficiently and gain a competitive advantage.

5. REFERENCES

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

Cortes, C., Pregibon, D. Giga-mining. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining; 174-178, 1998 August 27-31; New York, NY: AAAI Press, 1998.

Ezawa, K., Norton, S. Knowledge discovery in telecommunication services data using Bayesian network models. Proceedings of the First International Conference on Knowledge Discovery and Data Mining; 1995 August 20-21. Montreal Canada. AAAI Press: Menlo Park, CA, 1995.

Fawcett, T., Provost, F. Adaptive fraud detection. Data Mining and Knowledge Discovery 1997; 1(3):291-316.

Fawcett, T, Provost, F. Activity monitoring: Noticing interesting changes in behavior. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 53-62. San Diego. ACM Press: New York, NY, 1999.

Han, J., Altman, R. B., Kumar, V., Mannila, H., Pregibon, D. Emerging scientific applications in data mining. Communications of the ACM 2002; 45(8): 54-58.

Kaplan, H., Strauss, M., Szegedy, M. Just the fax—differentiating voice and fax phone lines using call billing data. Proceedings of the Tenth Annual ACM-SIAM Symposium on

Discrete Algorithms. 935-936. Baltimore, Maryland. Society for Industrial and Applied Mathematics: Philadelphia, PA, 1999.

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

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

Roset, S., Murad, U., Neumann, E., Idan, Y., Pinkas, G. Discovery of fraud rules for telecommunications—challenges and solutions. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 409-413, San Diego CA. New York: ACM Press, 1999.

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

Weiss, G. M., Hirsh, H. Learning to predict rare events in event sequences. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. 359-363. AAAI Press, 1998.

Weiss, G. M., Provost, F. Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research 2003; 19:315-354.

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