Dynamic Data Mining for Information Exploitation

Sherry Marcus
21st Century Technologies Inc., 8302 Lincoln Lane Suite 103, McLean VA 22102

Abstract

There is now an ever increasing threat to the nation's cyberspace infrastructure. Cyberterrorists can and have broken into power systems, banking systems, and defense systems, with relative ease. In this paper, we show how data mining technologies can be exploited in the identification of threats on the Internet. Currently, large data repositories are currently maintained by military organizations which contain Internet addresses of those who access (legitimately or illegitimately) military systems. This information together with additional information from other data sources can be mined so as to identify suspicious 'profiles'. A profile consists of sets of rules that define suspicious behavior. Knowledge bases consisting of these profiles can be developed in conjunction with data mining technologies such as case based reasoning, association, clustering, temporal, and similarity reasoning for the purpose of targeting -- in advance-- potential threats on the Internet.

In this paper, we report on a system called ProfileMiner(TM) that we have developed. Using ProfileMiner, users can describe profiles of interest to them. The ProfileMiner system automatically creates similar profiles and alerts the user when an individual or activity precisely matches the profile, or when he matches a "similar" profile. ProfileMiner may also identify other investigators looking at similar profiles, so that investigators are aware of other ongoing parallel investigations.

I. INTRODUCTION

Every day, law enforcement agencies conduct thousands of investigations pertaining to drug trafficking, weapons smuggling, NBC-related transactions, to illegal transportation of endangered species. In most cases, the perpetrators are individuals or criminal organizations, while in other cases, they involve governments or officials serving governments of rogue nations.

The Internet provides a medium for communication between the entities of such organizations – yet, unlike the case of phone systems where monitoring technology is well developed, not much is known about how to legally monitor suspicious Internet transactions.

For example, an investigative official monitoring the newsgroup *news.listserv.disarm* may wish to see who all are posting messages to this newsgroup. He may also wish to correlate this information with names of suspected terrorists, belonging to rogue nations, and/or to correlate this information with individuals who are also posting messages to the newsgroup *nuclear.news*. He may wish to find out all known e-mail addresses for a suspect who satisfies the above conditions. Expressing such correlation requests is potentially cumbersome and complex in classical languages like SQL. In time critical situations, the investigator should not have to waste his valuable time, fighting with cumbersome SQL query systems. Instead, he should be able to express conditions that he is interested in monitoring using a simple, graphical user interface.

The language in which these conditions are specified should be rich enough to access heterogeneous data sources [1] as well as rich enough to analyze temporal activity patterns. Furthermore, once he has specified one or more such conditions of interest to him, the system should automatically alert him when new individuals or organizations fit the profiles that he has specified. In addition, the system should be smart enough to infer new profiles that may be of interest to the user. We have developed a system called ProfileMiner that is capable of managing multiple profiles registered by authorized users, and providing the array of services described above.

The organization of this paper is as follows. Section 2 contains a birdseye view of the architecture of the ProfileMiner system. Section 3 provides a description of the data mining and link analysis components of the ProfileMiner system, while Section 4 describes a sample scenario of the working of the system.

0-7803-9914-5/98/$10.00 © 1998 IEEE

II. PROFILE MINER ARCHITECTURE

A. System Architecture

We describe in this section the current architecture of the ProfileMiner system. This architecture has been design to maximize the ability to "plug and play" with most Commercial Off The Shelf Products (COTS). Our architecture contains the following basic components:

Web-Compliant Interface: The user (such as an intelligence or law enforcement investigator) interacts with through any standard Web browser (such as Microsoft Internet Explorer or Netscape Navigator). This single interface may be used to specify the different profiles that the user is interested in monitoring as well as directly execute queries on data sources . It can also correlate the answers of different queries, as well as browse the results of automated monitoring and/or the results of specific queries.

Analytic Tools: The conditions that the user specifies may be executed by a variety of analytic tools. These tools include data mining, text retrieval and extraction, and relational database technologies.

Data Mining Tools: Data Mining technologies provide the user the ability to identify new links and relationships from data that were previously unknown. We rely heavily on case based reasoning tools, in the context of a relational database, that provide the means to find records similar to a specified record or records.

In addition, we exploit knowledge discovery tools that are specially designed to identify significant relationships that exist among variables. These tools are useful when there are many possible relationships. For example, if theanalyst has to track 200 variables about a particular individual, or 200 individuals about a particular variable, our knowledge discovery capability will provide the ability to point out what are the significant relationships.

Thus, when a user specifies a set of profile conditions that he wants to monitor, the system will intelligently expand this set of profile conditions to new, related conditions that the user may not have explicitly specified. For example, the system might notice that an NBC-suspect (say named *David Jones*) posts a message to the newsgroup *news.listserv.disarm* every Monday, and a response is posted within 24 hours to the same newsgroup by *Jonathan Smith*. The data mining system can discover such patterns on USENET postings or any type of structured textual data format.

Link Analysis Tools: Link analysis is the science of linking a suspect with other individuals or organizations. In the above example, the Link Analysis tool might conclude that the link between David Jones and Jonathan Smith is significant because Jonathan Smith is employed by an international shipping company – this may lead to a possibility of illegal shipment of NBC-related items.

Visualization Tools: Visualization Tools provide the user with a visual representation of the exploited data. When specifying his profile information, the user of ProfileMiner can also specify how he wants to visualize the results computed by ProfileMiner, e.g. through an Excel bar-chart, or just as a Web document containing hyperlinks to the postings, indexed by the name of the poster and date of the posting involved.

Information Exploitation Database (IED): The analytic tools described in the section above analyze views of raw data contained in an Oracle Information Exploitation Database created by ProfileMiner. For example, the IED may contain *pointers* to the various postings made by Jonathan Smith– the postings are raw data sources. In other words, the IED represents a view of the raw data sources, and thus only contains information that is timely and pertinent to the profiles being managed by ProfileMiner. Thus, even though postings to *news.listserv.disarm* may include postings by James Rubin, the State Department spokesman, these would not find their way into the IED because these are not deemed relevant to the profiles managed by ProfileMiner.

Information Extraction Tools: The IED is derived from a variety of "raw" data sources, such as USENET postings, GOTS databases, and video/audio sources (these are not currently integrated in the ProfileMiner system, but techniques for such integration tasks were developed by us in collaboration with the University of Maryland and the US Army [1,2]). We have developed algorithms to extract the key "content" of a text documents such as USENET postings, and then update/populate the IED with this content.

Raw Data Sources: All the tools described above access a variety of data sources, such as USENET data, and GOTS databases, which may be stored in a plethora of heterogeneous data formats, and also may be resident at distributed network sites. MAVIS has integrated most commercial relational and object oriented database systems. ProfileMiner can currently access a variety of such sources through our proprietary MAVIS system.

II. Data Mining and Link Analysis Modules

A. *Profile Descriptions*

Analysts will be able to determine user, newsgroup, date, subject, country of origin, country of destination, and content of USENET posting merely by typing in a specific field of interest into our graphical user interface. These fields (such as user, newsgroup, and date) were used because these are the fields found in USENET postings. Any analyst-defined fields can be implemented as required. For example questions of the sort:

Find all postings to the Internet made by anyone in Israel

Could easily be handled by typing into the appropriate newgroup and country of destination fields. We are able to identify any specific country through the e-mail address used to post to the newsgroup. This type of query used in conjunction with other profiling mechanisms can be used in the identification of focused targets. (See figure 1 for query description and figure 2 for an example output page.)

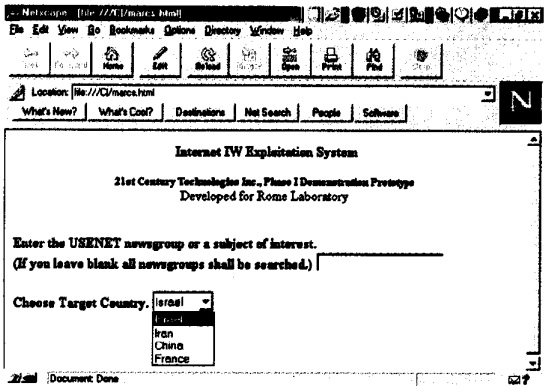

Figure One: Interface for IW Exploitation

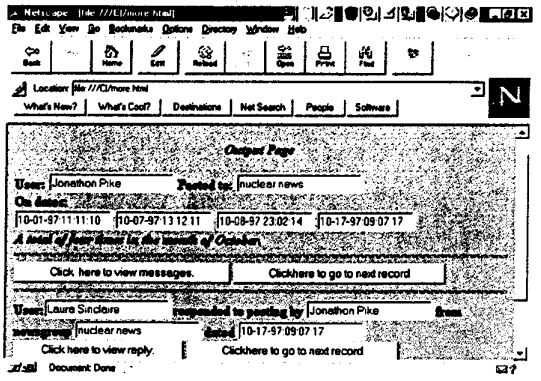

Figure 2: Sample IW Output

The query: "Find all e-mail accounts used by a specific person to all UseNet groups." could easily be handled by typing into the appropriate newsgroup posting and country of destination fields.

The GUI that is used to express profiles generates PL/SQL code that may be executed on the IED Oracle database, yet, the user does not even have to be aware that SQL is being used by ProfileMiner.

B. *Data Mining and Link Analysis Techniques*

ProfileMiner's mining and link analysis tools are, in reality, a suite of data mining tools and an information architecture acting in concert to generate and identify unknown electronic relationships and activities of a target or set of targets.

We refer to ProfileMiner suite of data mining tools as "Digital Fingerprint" because we wish to draw an analogy between a "classic fingerprint" and an "Digital Fingerprint". A digital fingerprint is broadly defined as those sets of actions that can identify a target based on electronic information generated by the target and may be detectable only through data mining tools. Analogously, a "classic fingerprint" can identify a target based on physical information (fingerprints) generated by the target and is detectable only by the appropriate law enforcement methods.

Using the current version of ProfileMiner one receives a concise report and visualization of computer FTP/Telnet logs, e-mail, Web Server access and publicly available Usenet postings generated from the suite of tools briefly described in the next paragraphs.

Digital Temporal Patterns is a data mining tool contained within the current version of ProfileMiner designed to track and reference electronic temporal

activities. The result of such a temporal data mining activity identifies and graphically depicts when specified targets perform certain operations on the Internet, and correlates these activities with other potentially illegal or damaging events. For instance, every time a particular target accesses the site www.abcde.com, it may turn out that within one day, the target receives an electronic funds transfer from the Cayman Islands.

If such a temporal pattern is detected over a period of time, then one may be tempted to infer that when the target logs into www.abcde.com, then the site www.abcde.com recognizes this as a transaction requesting an electronics funds transfer. There are a variety of operational scenarios where such temporal information is crucial. Recognizing such patterns is a challenging task.

Digital Miner is a tool within ProfileMiner that uses a variety of database and data mining technologies such as case based reasoning, knowledge discovery, rule bases, and heterogeneous data access to exploit the structure of a relational database in order to determine new and unknown relationships. Digital Miner can identify and link seemingly unrelated facts. For example, the electronic record recording the fact that John Smith posts to newsgroups about computer security is innocuous. Another electronic record describing Jane Johnson's attempts to illegally penetrate a government computer system is apparently unrelated. However, the link that Jane Johnson used techniques found in John Smiths' ties these two records together. This is Digital Miners function. In this case, Digital Miner would have assisted law enforcement authorities in locating an additional suspect (in this case John Smith) or a network of suspects.

Virtual Network is another tool within the current version of ProfileMiner which identifies and graphically depicts indirect digital communications between two targets. For example, Targets A and B may communicate via e-mail to two distinct cutouts C and D. Cutouts C and D in turn, communicate to cutouts E and F who communicate with each other on Internet Chat. Virtual Network can identify varieties of Internet communications (such as Internet postings, FTP/Telnet logs, etc) and use such information to generate networks.

One fundamental aspect of communications chain of command in intelligence, paramilitary, narcotics, and terrorist networks, is that the principals usually maintain a considerable distance from the "low-level" operatives, so that arrest of the latter does not compromise the former. It is therefore very important to analyze not only *direct* communication links between networked individuals, but also the *indirect* links that arise when target A communicates with target B who then communicates with target C and so on. Thus, one is able to identify a "virtual network" of operatives each performing unique operations over the Internet.

IV. CONCLUSIONS

While many law enforcement agencies have extensively studied such link analysis in the context of phone systems, no comparable link analysis methods exist on the market today for Internet link analysis. Internet link analysis is considerably different from phone link analysis because of the infinite opportunities to spoof, assume false identities, and break through firewalls. The need for such tools in investigative capacities will increase as commercial use of the Internet increases in the consumer and corporate arenas. The *Virtual Network* component of ProfileMiner provides a first advance in this emerging area.

IV. REFERENCES

[1] A. Brink, S. Marcus and V.S. Subrahmanian. *Heterogeneous Multimedia Reasoning*, **IEEE Computer,** September 1995.

[2] S. Marcus and V.S. Subrahmanian. *Foundations of Multimedia Information Systems,* Journal of the ACM, November 1996.

[3] Towards a Theory of Multimedia Database Systems Multimedia Database Systems: Research Issues and Directions, Edited by V.S. Subrahmanian and S. Jajodia, Springer-Verlag 1995, with V.S. Subrahmanian