A System for Association Rule Discovery in Emergency Response Data

George Dimitoglou, Shmuel Rotenstreich

Abstract— **Association rule mining is used to find association relationships in data. Our work describes the use of association rule discovery as a basis for creating an early warning bio-terror attack system. The system establishes a baseline of "normal" behavior by mining historical emergency response (911) data. Using probabilistic models, we generate spatial and temporal statistics to correlate incident frequency and location in order to identify if a variation in future incidents carries an outbreak signature consistent with the effects of a biological warfare attack. Using three years of real emergency response data for experimentation, this work is focused on the activities relating to the processing and generation of detection rules. Preliminary results indicate that the system can provide reasonable detection rules but there is also more work to address inherent issues of both emergency response and biological warfare such as data quality during incident reporting and population mobility as it relates to outbreaks.**

Index Terms— Data mining, knowledge discovery, association rules

I. Introduction

EMERGENCY response data is widely used for analysis in various disciplines, ranging from criminal justice and sociology to political science and public health administration. The most common studies in these disciplines evaluate incident frequency and geographical location in search of patterns to reveal the formulation or evolution of trends that could then be addressed with changes in emergency response resources or public policy. This type of "post-mortem analysis" is very beneficial in understanding and profiling typical incident activity such as increased traffic accidents around national holidays or assaults, robberies and murders in crime-ridden certain neighborhoods. An interesting alternative is the use of emergency response data as a tool for incident identification for improved preparedness and prevention. Obviously, this type of use is limited to only a certain type of incidents, mainly those with a fairly prolonged and staged manifestation. Such incidents would be identified within certain symptoms in the population and could be the result of the intentional release of a biological agent or, an industrial accident that caused a toxic spill.

G. Dimitoglou is at the Department of Computer Science, Hood College, Frederick, MD 21701-8575 USA (phone: 301-696-3980; e-mail: dimitoglou@hood.edu).

S. Rotenstreich is at the School of Engineering and Applied Science of the George Washington University, Washington, DC 20052.

This work is focused on using association rule mining to create a bio-terror attack diagnosis system that establishes a baseline of "normal" behavior by mining historical emergency response (911) data. This baseline serves as the threshold of normalcy and any significant deviation from the baseline during "real-time" operations becomes a possible indicator of an attack.

II. Background and System Architecture

Data mining is a widely used analysis tool in many scientific and industrial applications. Association rule mining [1], [6] tries to find association relationships in data. The association relationships are described by rules. Each rule has two measurements: support and confidence. Confidence is a measure of the rule's strength, while support corresponds to statistical significance. Classification is a data mining operation that has been studied extensively in the fields of statistics, pattern recognition, decision theory, machine learning literature, neural network, etc. Clustering [4] is often an important initial step of several in the data mining process. Some of the data mining approaches which use clustering are *database segmentatio*n, *predictive modelin*g, and *visualization* of large data-bases [7]. Typical pattern clustering activity involves the following [7]: (a) Pattern representation (optionally including feature extraction and/or selection), (b) Definition of a pattern proximity measure appropriate to the data domain, (c) Clustering or grouping, (d) Data abstraction (as needed), and (e) Assessment of output (as needed). Sequential pattern mining is the process to analyze a collection of data over a period of time to identify trends [1]. Sequential pattern mining is closely related to association rule mining, except that the events are linked by time.

To mine, process and use 911 response data, we developed a software bio-terror attack diagnosis system. The main function for this system is to generate association rules from an incident response database and second, to attempt to identify a geographic pattern in the instances. The rationale behind the system is by identifying the occurrence of certain symptoms during a certain time period in a geographic area may be an indication of a health anomaly pattern within the population (i.e. a virus spreading).

The system uses 911 call data as input and generates (a) a set of baseline rules that reflect normal conditions and (b) alerts

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 193–199.
© 2007 *Springer*.

that are triggered when new, incoming calls deviate from the existing baseline rules. The intention is for these rules to be unambiguously formulated and easily readable so it is easy to be either parsed electronically or read by humans. For example, the following, although simplistic rule:

IF respiratory incidents > 20 per month in area *G* THEN alert;

accommodates both requirements first, by being unambiguous as to differentiating the condition-action parts of the rule and second, by being easily readable so anyone can understand there is action to be taken.

Figure 1 provides an overview of the system architecture. The system contains three processing engines and four repositories.

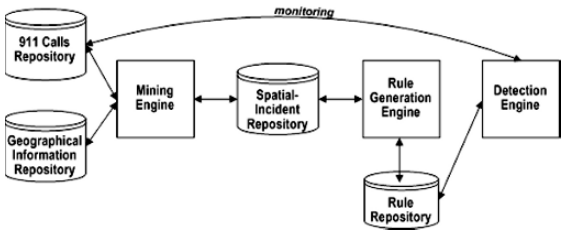

Figure 1: Bio-terror detection system architecture.

The system is comprised of:

- A Mining Engine processes the 911 calls stored in the Calls Repository and generates association rules that are in turn correlated with the contents from a Geographical Information Repository [7]. The Mining Engine generates data and populates a Spatial-Incident Repository by statistically, methodically and mathematically trying to identify unifying patterns and conditions
- A Rule Generation Engine processes the generated association rules and generates baseline rules that are stored in a Rule Repository. Baseline rules are the rules that have been identified by the data mining computations to reflect as the potential occurrence of bioterrorism events.
- The Detection Engine works as a monitor that constantly examines in real-time incoming 911 calls and applies the relevant rules from the Rule Repository.

III. METHODOLOGY AND EXPERIMENTATION

Using data from Regional Emergency Medical Communication Systems (REMCS), the 911 Calls Repository is populated with entries reflecting dispatch requests. The 911 REMCS Call Center receives an average of approximately 15,000 calls per year. The dataset used in our study contained 30,664 calls for a two-year period between January 1997 to December 1998. A sample segment of the data set is displayed in Table 1.

The columns in Table 1 contain information such as an internal identification number (**CALL.ID**), the date (**DATE.REC**), address (**LOCATION**), grid location information (**GRID**) and the patient's condition description (**PAT.COND.DESCR**).

CALL.ID	DATE.REC	LOCATION	GRID	PAT.COND.DESCR
57644	01/01/02	129 WRIGHT ST-NK.3 FL	516	BREATHING PROBS/ABNORM BREATHI
57645	01/01/02	748 S 10 ST-NK	512	TRAUMA/POS DANGEROUS BODY AREA
57646	01/01/02	354 PARK AV-NK.305	212B	PREGNANCY/2nd TRIMESTER BLEED
57647	1/1/2002	100 RT 1&9-NK SOUTH	189	TRAFFIC ACCIDENT W/ INJURIES
57648	01/01/02	MERCHANT ST-	316	HEMORRHAGE/POS DANGEROUS BLEED
:	:	:	:	:
:	:	:	:	:
86578	01/01/02	182 COURT ST-NK	411	SICK PERSON/NO PRIORITY SYMPTO

Table 1:Sample Data

The REMCS system has a grid designation that provides information about the general area of the incident location. Analysis allows us to view incidents grouped within the confines of a larger geographical area and patterns of incidents may also be diagnosed in relation to incidents in adjacent grids. Figure 2 illustrates how mock grid information is overlaid on top of a Newark area map. Patient address information is matched with the corresponding grid number during the response call in order to provide information for ambulance or emergency medical staff deployment.

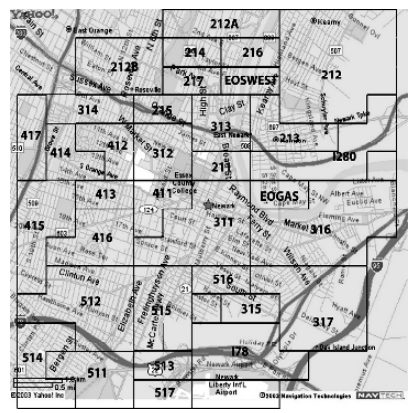

Figure 2: Conceptual Map Overlay. The grid overlay for a metropolitan area served by the REMCS.