RESEARCH ARTICLE OPEN ACCESS

Event Data Analysis Using Data Mining

[1]A.Arokia Marshal Roach ,[2]G.Raja Raja Cholan

[1] Research Scholar, Department of computer Science Prist University, Puducherry, India

[2]Asst. professor in Department of computer Science Prist University, Puducherry, India

--*************************------------------------------

Abstract:

There is a need to understand how industry as a whole is performing from a safety standpoint. To date, no one can really answer this question with certainty. People do a decent job collecting data on Events, but few take the analyses of the data past basic trending. Having the capability to collect enormous amounts of data is a feat in and of itself; however, it begs the question, "So what?" With the amount of resources spent to collect data, it seems logical to look at the data under extreme scrutiny to obtain as much knowledge about the data as possible. Data in a database is just that, data. By analyzing and understanding what is in the database yields knowledge. Passing this knowledge on to others can improve the understanding of what went wrong with Events from the past thereby greatly enabling the prevention of future Events.

Trending analyses do provide useful comparisons in the data, however, going beyond comparisons by using data mining techniques can enable one to build predictive models, unveil relationships within the data that are not necessarily intuitive, and perhaps answer the question, "How is industry's safety performance doing?" Marketers have successfully harnessed the power of data mining to build predictive models to increase profit by, for example, determining customer buying habits based on advertisement campaigns.

Keywords— **mining, market, data analysis**

--*************************------------------------------

Introduction:

The advantage of using data mining is its ability to analyze an enormous set of data . Using the data mining as an analysis tool applied to Event databases can make a huge, positive impact on industry and the public at large.

The benefits of performing a thorough analysis of Event databases include better understanding of safety performance, better understanding of how to focus efforts to reduce Events, and a better understanding of how people are affected by these Events.

This system collects data on Events where there was a release or threatened release of a hazardous substance that resulted in some public health action . It is unique in that it collects data for the purpose of analyzing the effects these Events have on the health of the public as opposed to other databases that focus on environmental impact.

The objective of this research is to use data mining and text mining to analyze the HSEES system data by identifying relationships among the variables, predicting variable(s) of interest, and assessing the value added by the text data. Furthermore, the results of this research will define what can be done with this type of data in terms of analyses and what types of questions more thorough analyses may answer.

Analyses of the information in databases help connect the dots between what went wrong and what people can do to prevent it—the relationship between the cause of an Event and its consequences.

Trevor Kletz time and time again reminds us that an essential part of not repeating mistakes from history is to make sure that lessons are learned, and to make sure these lessons are shared as new generations join in. Several people have put their mark on assessing what knowledge is available from Event databases and some of these are discussed in the following.

Eboni Trevette McCray compared several Event databases in an effort to formulate improvements for these databases as well as determine national safety goals to be implemented given these improvements. It was argued that comparing and trending data from the databases is impossible due to the overall discrepancy in data collection agendas, methods, and definitions from year to year.

The process of ERNS

As a result, it was proposed to create a single database with a thorough amount of information on Event details including the causes and effects of the Events. This proposed database will originate from the existing Emergency Response Notification System (ERNS) database and is expanded upon with questions from an Accidental Release Information Program (ARIP) survey . Although it is agreed that there are errors and discrepancies in data collection, it is disagreed that ascertaining any useful analysis from these databases is impossible. It is implicit through modeling these data that there is some level of variability and uncertainty, yet the overall trend and relationships will be foretelling enough to draw conclusions and make recommendations for safer practices regarding chemicals. In the case of my research, the data used from 2002 to 2004 have common definitions and the inherent nature of HSEES being an active system means the data are more reliable Others have made strides to analyze databases such as Fahad Al-Qurashi's work where the combined effects of accidental, failure rate, and reactive chemical databases were considered. Specifically, the Environmental Protection Agency's (EPA) Risk Management Program (RMP) database was used to decipher the most significant chemicals released and ultimately it was concluded that there is a need for more data with regard to failure rates and reactive chemicals. It was stressed that with the appropriate understanding of equipment reliability and the inherent hazards of chemicals used, the number of Events can be reduced . Although the focus of this research is to link different data sources together to find new learning's,

It still identified the most frequently occurring offenders and basic trends with the analysis, but did not incorporate predictive modeling. Looking at the benefits of using data mining, one can consider Sumit Anand's work where data mining techniques were applied to the National Response Center's database to uncover interesting patterns in data pertaining to fixed facilities in Harris County, Texas from 1990-2002. Example techniques applied to these data are decision trees where consequences of an Event are compared to the type of equipment failure and Event cause, and association analysis used to compare the type of equipment failure and the chemical involved. Using the data mining results, An and updated equipment failure probabilities and built a decision support system . Finding associative behaviors between variables, like type of injury and chemical released, might be a viable option for the HSEES dataset. It could show how likely the presence of some chemical X will result in some injury y. An alternative option is clustering events.

Terry L. Bunn et al analyzed tractor fatality data for the state of Kentucky focusing on the added benefits of analyzing the text given by way of Event investigation reports. They showed that analyzing text entries in addition to coded data provides far more information then looking at coded data alone. The advantage with the tractor fatality dataset is that the text entries are detailed Event investigation reports, not short comments on the nature of the event like what is contained in the HSEES dataset. Relationships in the tractor data were extracted about pre-event, event, and post-event conditions, namely the initiating event, the actual injury or outcome, and the response to the event.

HSEES DATA :

The Hazardous Substances Emergency Events Surveillance (HSEES) data includes information on events where:

- There was an uncontrolled/illegal release or threatened release of at least one hazardous substance NOT including petroleum (due to the Petroleum Exclusion clause of CERCLA) and the release of the hazardous substance(s) requires removal, clean up, or neutralization, or
- There was a threatened release of a hazardous substance that would have needed to be removed, cleaned up, or neutralized AND the threatened release resulted in a public health action. Although CERCLA has a Petroleum Exclusion, events where petroleum is released along with other hazardous substances are included in HSEES and petroleum is reported along with the other substances .

IN THIS PAPER SUPPORTING MODULES:

1. **Building Predictive Models**
2. **Decision Trees**
3. **Logistic Regression**
4. **Measuring Model Performance**
5. **Lift and Gain**

Action substance:

Hazardous Substance	Numbers
Subcategories	**# Events**
SC_OISC	5,119
SC_MIX	4,128
SC_VOC	3,743
SC_OXYORG	1,906
SC_OTHER	1,536
SC_ACID	1,512
SC_AMMONIA	1,404
SC_PESTAG	740
SC_BASES	627
SC_CHORLINE	576
SC_PANDD	429
SC_HYDROCARB	316

HSEES is unique compared to other databases since its focus is on public health whereas other databases focus on environmental impact. This focus on public health is aligned with ATSDR's mission "to serve the public by using the best science, taking responsive public health actions, and providing trusted health information to prevent harmful exposures and disease related to toxic substances" . The purpose for collecting the data is to assess the acute effects hazardous substance emergencies have on the morbidity and mortality of the first responders, general public, and employees, and thereby reduce these occurrences .

probabilities:

$$\text{Odd ratio} = \frac{\Pr\,(event\ occurs)}{\Pr\,(event\ does\ not\ occur)}$$

Conclusions

Data mining proved to be beneficial in both describing the HSEES events and building a fairly good model to predict the occurrence of injuries. The following are some conclusions drawn based on the analysis:

- Although HSEES data is collected to

describe the effects hazardous substance releases/threatened releases have on people, a fairly good predictive model was still obtained from the few variables identified as cause related.

- Visually exploring the data via bar graphs did not yield any noticeable patterns.
- Clustering the data identified groupings of categories across the variable inputs such as manufacturing events resulting from intentional acts such as system startup and shutdown, performing maintenance, and improper dumping.
- Text mining the data allowed for clustering the events and further description of the data, however, these events were not noticeably distinct and drawing conclusions based on these clusters was limited.
- Inclusion of the text comments to the overall analysis of HSEES data greatly improved the predictive power of the models. Interpretation of the textual data's contribution was limited, however, the qualitative conclusions drawn were similar to the model without textual data input.

References:

[1] S. Anand, N. Keren, M.J. Tretter, Y. Wang, T.M. O'Connor, M.S. Mannan, Harnessing data mining to explore Event databases, Journal of Hazardous Materials 130 (2006) 33-41.

[2] Agency for Toxic Substances & Disease Registry, Hazardous Substances Emergency Events Surveillance (HSEES) Protocol, http://www.atsdr.cdc.gov/HS/HSEES/protocol030804.html (last visited on 2008).

[3] E.T. McCray**,** Chemical accident databases: What they tell us and how they can be improved to establish national safety goals, MS Thesis, Chemical Engineering Department, Texas A&M University, College Station, TX, 2000.

[4] F. Al-Qurashi, Development of a relational chemical process safety database and applications to safety improvements, MS Thesis, Chemical Engineering Department, Texas A&M University, College Station, TX, 2000.

[5] S. Anand, Novel applications of data mining methodologies to Event databases, MS Thesis, Chemical Engineering Department, Texas A&M University, College Station, TX, 2005.