ELSEVIER

Journal of Hazardous Materials 130 (2006) 33–41

Journal of Hazardous Materials

www.elsevier.com/locate/jhazmat

Harnessing data mining to explore incident databases

Sumit Anand [a], Nir Keren [a], Marietta J. Tretter [b], Yanjun Wang [a], T. Michael O'Connor [a], M. Sam Mannan [a,*]

[a] *Mary Kay O'Connor Process Safety Center, Chemical Engineering Department, Texas A&M University, College Station, TX 77843-3122, USA*
[b] *Department of Information & Operations Management, Mays Business School, Texas A&M University, College Station, TX 77843-4217, USA*

Available online 28 November 2005

Abstract

Large numbers of incident related databases have been established in the last three decades. The majority of attempts to explore these data marts were trials to identify patterns via first glance into the datasets. This study investigated a subset of incidents from fixed facilities in Harris County, TX, extracted from the National Response Center database. By classifying the information into groups and using data mining techniques, interesting patterns of incidents according to characteristics such as type of equipment involved, type of chemical released and causes involved were revealed and further these were used to modify the annual failure probabilities of equipments.

Keywords: Incident databases; Data mining; Failure rates

1. Introduction

A large number of incident related databases have been established in the last three decades. Federal agencies as well as private organizations collect data on incidents. These organizations differ from each other in their interests, data collection procedures, definitions and scope, and each of these organizations analyzes the data to achieve and accomplish both the goals and missions of the organization [1–3]. Thus, the information on the incidents is heterogeneous in both content and form.

The majority of incident databases consist of hundreds of thousands of records. But when there are so many trees, how is it possible to draw meaningful conclusions about the forest? Handling these data statistically, without robust analyzing platforms, becomes a complex and impractical mission. So what can be done?

Data mining is a field that was developed in the last two decades in order to handle and extract knowledge from databases that may contain terabytes of records. This field is a data-based process that uses a variety of tools to describe, present, build predictive models and validate them. Data mining is an extension of statistical methods. This means that while statistical methodologies rely on elegant theory and analytical methods, data mining uses new brute-force exploration techniques that are available due to recent developments in computational capabilities [4].

This paper presents several of the current and next generation data mining tools, and the knowledge discovered by mining the National Response Center (NRC) incident database with these tools.

2. Data mining and knowledge discovery

The major reason that data mining attracted attention in recent years is the need to analyze datasets with an enormous number of records and there is a lack of tools to extract knowledge from these sets. Data collection, storage, and management, progressed from primitive file processing systems in the 1960s to sophisticated database systems. Data can now be stored in many different types of structures. The most common architecture that has recently emerged is the

* Corresponding author. Tel.: +1 979 862 3985/845 3489; fax: +1 979 458 1493.
E-mail address: mannan@tamu.edu (M.S. Mannan).
URL: http://process-safety.tamu.edu.

0304-3894/$ – see front matter
doi:10.1016/j.jhazmat.2005.07.046

data warehouse, which is a repository of heterogeneous data sources, organized under a unified scheme at a single site in order to facilitate decision support systems.

The vast amount of data has far exceeded our ability for comprehension without powerful tools. As a result, important decisions are often made based on the decision maker's intuition and not based on information from rich-data, although the data are available. Data mining can be helpful in extracting knowledge by revealing important patterns that may not be revealed by other traditional tools.

Data mining consists of various techniques that, in a broad sense, can be categorized into two groups [5]:

- Classical techniques such as statistics and neighborhood.
- Next generation techniques such as decision trees and rules.

The following section briefly describes the techniques above. Further details are presented on the technique used to explore the National Response Center incident database.

3. Classical technique

3.1. Statistics

Current exploration of incident databases is mainly aimed at identification of patterns by grouping records with similar values (for example, counting the number of records in which the root cause of the incident is Human Error). Statistics may be helpful in the process of exploration of the database since it may reveal information on the following:

- identifying patterns in the database;
- predicting the probability that events will occur;
- identifying the significant patterns, such as linking causes and effects;
- developing prediction models.

Traditional techniques such as regression models aim at finding correlation between a target variable and other independent variables. Multiple non-linear regression models fit independent variables to the target variable using forms similar to the form given in Eq. (1):

$$Y = C_0 + C_1 \cdot X_1^{n_1} + C_2 \cdot X_2^{n_2} + \cdots + C_m \cdot X_m^{n_m} \qquad (1)$$

where Y is the target value, X_i the ith independent variable, C_i the coefficient of correlation of variable X_i and n_i is the power value of variable X_i.

The regression model calculates both, C_i and n_i. Indices such as least square and similar others are used to estimate the quality of the model. It is hard to find practical applications of regression models in the case of incident databases.

3.2. Nearest neighbor

Nearest neighbor is a prediction technique that predicts the value of a required variable according to the values of this variable in other cases that have similar characteristics. For example, the probability that the household income of a family that lives in a 3000 ft^2 house in a certain neighborhood may be higher than US$ 100,000 since the household incomes of neighbors who lives in 3000 ft^2 houses is higher than US$ 100,000. Generally, models will use a large number of parameters such as age and academic degree in these analyses.

4. Next generation techniques

The next generation techniques have been developed in the last two decades. Generally, when data mining is being mentioned in popular contexts, it is referring to these techniques.

The following sections will describe the decision trees (often referred to as classification) and association rules. Neural network (NNW) will not be addressed in this paper. Further study is required in order to find proper ways to apply NNW to incident databases.

4.1. Decision trees

A decision tree is an algorithm that classifies data according to a set of rules that is identified in the data. Decision trees are mainly used to examine the data and to induce a tree and its rules. Trees are developed through an iterative splitting (partition) of the data into discrete groups, with the goal to maximize the distance (i.e., the groups should contain as many items as possible with the same characteristics) [6].

Decision trees can be used for a prediction as well as for identifying significant variables. The prediction process is made based on historical data. Fig. 1 illustrates the process of a prediction on whether a candidate is considered a good or bad risk for a loan-granting process, based on historical data. The example in Fig. 1 uses a binary tree. However, decision nodes can be partitioned to more than two branches.

The distance will be maximized in the case above, if the data would be split in a way that the final two groups will consist of either bad or good risk cases. The following example can demonstrate how decision trees can be useful identifying significant variables in an incident database.

A 1000 records incident database consists of the following variables: cause of the incident, the type of equipment failure and the incidents, and a binary field (injury) that indicates

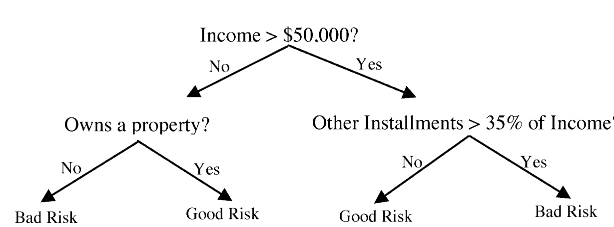

Fig. 1. Illustration of prediction process by decision tree.

whether the incident results in injuries (a value of 1) or not (a value of 0). Decision trees require that the user will determine a target value. Injuries are a major concern in incidents. Therefore, injury was assigned as the target value.

The tree in Fig. 2 is a typical presentation of the results of analyzing databases with decision trees. The database consists of 400 records on incidents that resulted in injuries. The analysis revealed that cause is the most significant variable in modeling the database. Records that consist of "Human Error" as a cause led to 250 incidents that resulted in injuries, and 100 incidents that did not result in injuries. The next contributing variable was type of equipment. The value "process vessel" was found in 270 of the incidents that resulted in injuries. In 150 of these incidents, Human Error was the cause. This example can be summarized as follows: "Human Error" is the most significant variable in contributing to injuries, and type of equipment is the second most significant variable.

The example above is a simple case and the results can be intuitively understood. However, databases can consist of several hundred variables. The National Fire and Incident Reporting System database consists of dozens of variables. Since the data should be cleaned and variables categorized, establishing models based on all these variables requires enormous efforts. A possible application of a decision tree would be identifying the few variables that are more significant in order to prepare the data for other modeling tools. A decision tree can be helpful in these types of assignments. The advantages of decision tree analyses can be summarized as follows:

- the analysis can identify significant relationships among variables;
- a decision tree can handle extremely large datasets easily;
- the results are easy to interpret.

This type of analysis has several insignificant disadvantages. However, those should be looked at prior to applying the analysis.

4.2. Association rules

Association rules (known also as rule induction) are mainly used in market basket analyses [7]. Market basket analyses aim at identifying associations among products purchased by a particular customer. For example, it is a common cliché in the data mining arena that customers that are purchasing diapers are likely to buy beer as well. If this rule would be pulled from a database it would have the following form:

"If diapers are purchased then beer is purchased 40% of the time (the Confidence), and this pattern occurs in 7% of all shopping transactions (the Support)".

That means, in addition to identifying the pattern, the analysis-extracted information on how strong the pattern is and how likely it is to occur again.

The following is a formal presentation of the rule. Below are explanations on the confidence and support. Moreover, the Lift value, which is a parameter that demonstrates the level of association, is presented as well.

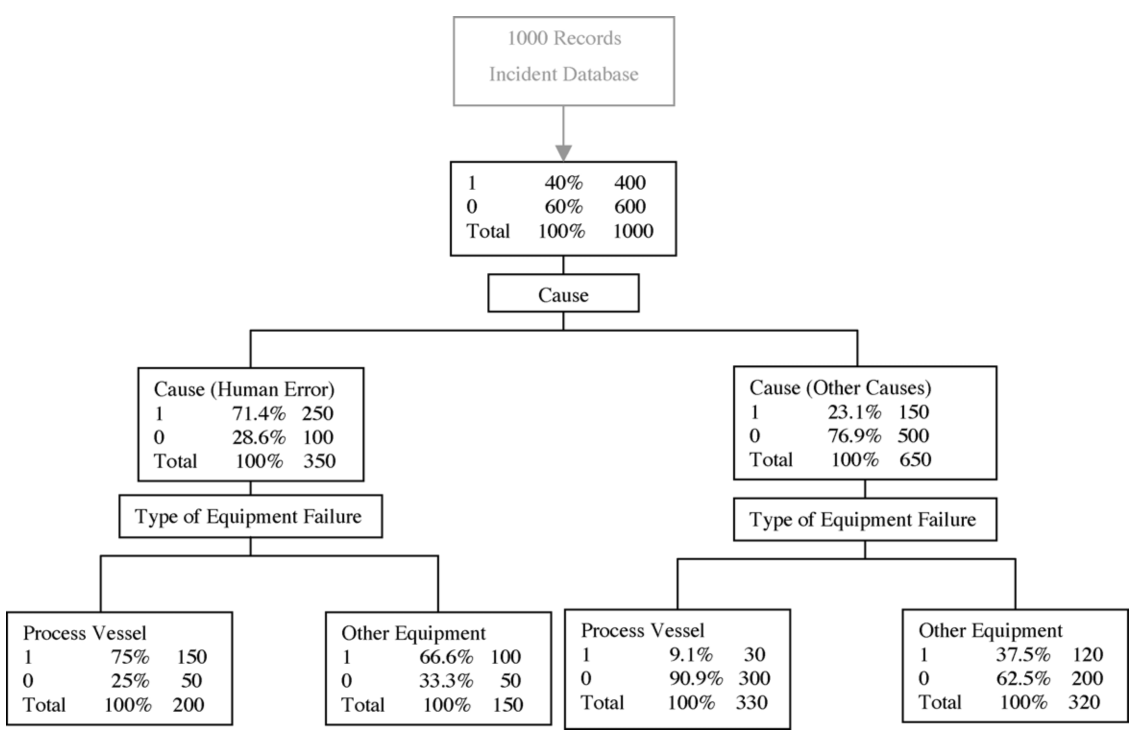

Fig. 2. Illustration of a decision tree analysis of an incident database.

The general form of the rule is as follows:

"IF event A occurs THEN event B occurs as well, in *X*% of the times, and this pattern occurs in *Y*% of all events in the dataset"

where *X* is the confidence and *Y* is the support.

Support represents the probability that both events A and B occurred simultaneously in the dataset. This value is calculated as presented in Eq. (2):

$$\text{Support} = \frac{\text{number of records in the dataset in which both events A and B occured}}{\text{total number of records in the dataset}} = P(\mathrm{A} \cap \mathrm{B}) \tag{2}$$

Confidence presents the probability that event B will occur given that event A occurred. This value is calculated as presented in Eq. (3):

$$\text{Confidence} = \frac{\text{support}}{\frac{\text{number of records in which event A occured}}{\text{total number of records in the dataset}}} = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{A})} \tag{3}$$

In the terminology of the set theory, confidence is the conditional probability of event B, given that event A has occurred [8].

The Lift value is the ratio between the probability that B will occur when A occurs to the general probability that B will occur. Lift is being calculated as follows:

$$\text{Lift} = \frac{\text{confidence}}{\frac{\text{number of records in which event B occured}}{\text{total number of records in the dataset}}} = \frac{P(\mathrm{A} \cap \mathrm{B})}{P(\mathrm{A}) \cdot P(\mathrm{B})} \tag{4}$$

The diaper and beer example can be used to emphasize the meaning of Lift. The value of Lift is the ratio of the probability that beer will be purchased when diapers are purchased, to the general probability that beer will be purchased.

A Lift value of 1 means that there is no difference between the probability that beer will be purchased when diapers are purchased, to the general probability that beer will be purchased (no association). A Lift value that is greater than 1 means that when beer is purchased it is more likely to be purchased with diapers (positive association). However, a Lift value of less than 1 means that if beer is purchased it is less likely that diapers will be purchased too.

The following is a formal presentation of the interpretation of values of Lift [6]:

Lift > 1: There exists a positive association between event A and event B of the rule.

Practically, if Lift = 2, it is twice as likely that event B will occur when event A occurs than the likelihood that event B will occur.

Lift = 1: There is no association between occurrence of events A and B.

Practically, if Lift = 1, it is neither more likely nor more unlikely that event B will occur when event A occurs, than the likelihood that event B will occur. In these cases, A and B are considered independent.

Lift < 1: There exists a negative association between event A and event B of the rule.

Practically, if Lift < 1 it is less unlikely that event B will occur upon occurrence of event A, than the likelihood that event B will occur. If Lift = 0, then event B will never occur simultaneously with event A (A and B are mutually exclusive [8]).

This paper presents the process and the results of analyzing the NRC databases with association rules.

An important point is that intensive efforts were invested in an attempt to extract knowledge from this database by simply applying a logistic regression analysis on the data. The results of this attempt were difficult to interpret and it was impossible to extract meaningful knowledge in this study.

5. Incident databases—an overview

There is an increased interest in using data on incidents to improve safety in the last 20 years. During the late 1980s, Marshal consolidated incident data from 60 or so years and harnessed it toward loss reduction, and loss prevention, in his book "*Major Chemical Hazards*" [9]. Today, the interest is bigger than ever, because of the development of information technologies that look promising in their abilities to see what "the unarmed human eye" cannot see. Major efforts are being invested toward collection of incident related data. The US Department of Health and Human Services, The Agency for Toxic Substances and Disease Registry (ATSDR), maintain the Hazardous Substances Emergency Events Surveillance system (HSEES) and publishes annual and cumulative reports [10]. This project is only one among many other types of data collection projects that are maintained by the Center for Disease Control (CDC). The Department of Transportation repository consists of a large number of transportation safety-related databases. A large number of reports are available on their website [11]. The last are only two from dozens of sources of information of incident-related data that are available. However, the main challenge in using incident-related data only begins when the data are available.

Table 1 consists of a list of more than a dozen databases from 10 different sources. The form of the data in each of the databases reflects the interest, purpose and scope of the organization collecting the data. Therefore, integrating these sources to a single source of information is a mega task as described by Keren et al. [3], and much still has to be done to complete this task.

The National Response Center IRIS database was selected as the source to apply data mining analysis. The following sections will add details on the type of data that is available in this source.

Table 1
Sources of information and databases

Source	Database
Federal Emergency Management Agency (FEMA)	National Fire Information Reporting System (NFIRS)
US Consumer Product Safety Commission (CPSC)	National Electronic Injury Surveillance System (NEISS) Death certificates Investigation summary Incident summary
Mary Kay O'Connor Process Safety Center (MKOPSC)	News clipping database
States Associations	State of Iowa State of Florida
State Agencies	State of Texas
National Response Center (NRC)	Incident Reporting Information System (IRIS)
US Department of Health and Human Services, Agency for Toxic Substances and Disease Registry	Hazardous Substances Emergency Events Surveillance (HSEES)
US Department of Transportation (DOT)	Hazardous Material Incident Reporting System (HMIRS) Integrated Pipeline Information System (IPIS) also known as Hazardous Liquid Accident Data (HLAD)
US Environmental Protection Agency (EPA)	Risk Management Program (RMP), 5-year accident history Accidental Release Information Program (ARIP)
US Department of Labor, Occupational Safety and Health Administration (OSHA)	Accident Investigation System and several other databases

5.1. National Response Center Incident Reporting Information System (IRIS)

The National Response Center is the Federal point of contact to report all oil, chemical, radiological, biological and etiological discharges into the environment, anywhere in the US and its territories [12]. The NRC operates 24 h a day, 7 days a week and 365 days a year. The NRC database consists of a large number of fields and a large number of records (about 33,000 records in 2003 only).

The law mandates that any case of a chemical release (above a predetermined reportable threshold) to the environment should be reported to the National Response Center [13]. However, a major disadvantage of the NRC IRIS system is that it collects reports very close to the time of the incidents. Therefore, records quite often do not consist of good documentation. Moreover, the data consist of a large number of repetitions, because updates of earlier reported incidents are submitted as new records. Also data from drills is documented as real incidents, and other incidents that are not fixed facilities-based events are reported, as if they are fixed facilities incidents.

A large concentration of chemical, petrochemical, pharmaceutical facilities are located in Houston, TX, and around the city. This concentration of facilities became a major concern in recent years, especially in light of the September 11, 2001 events. Therefore, a subset of NRC IRIS that consists of incidents from fixed facilities in Harris County, TX, during the years 1990–2002, was selected as a case study. This subset consists of a total of 7265 incidents.

Modification and preparation of the dataset to enable mining with data mining tools required several activities as specified below:

- Integrating a variety of tables in the source into a flat file.
- Removing duplications, records on drills and records of incidents that did not occur in fixed facilities.
- Developing an equipment classification taxonomy (similar to the taxonomy proposed by Chung and co-workers [14]) and applying this taxonomy to the dataset.
- Establishing auxiliary fields such as a field consisting of information on the amount released in the incident, in uniform units.
- Grouping types of chemicals according to the following criteria:
  - releases of gasoline, diesel, transformer oil, fuel oil and similar cases are categorized as an oil release;
  - releases of all types of acids are presented as an acid release;
  - releases of water contaminated with chemicals above the reportable thresholds are presented as contaminated process water.

Upon completion of the tasks listed above, the dataset was in an acceptable form for analysis purposes.

6. Pattern identification in first glance

The most common form of pattern identification is identifying distributions of a number of incidents by the variable that is under investigation. Fig. 3 presents the distribution of number of incidents in the dataset as a function of the cause of the incident. As Fig. 3 revealed, equipment failure is a leading cause for incidents.[1]

[1] One of the limitations of the NRC data is that it does not include a detailed analysis of the incident and the remedial actions taken. The root causes involved in the incident cannot be concluded from this data.

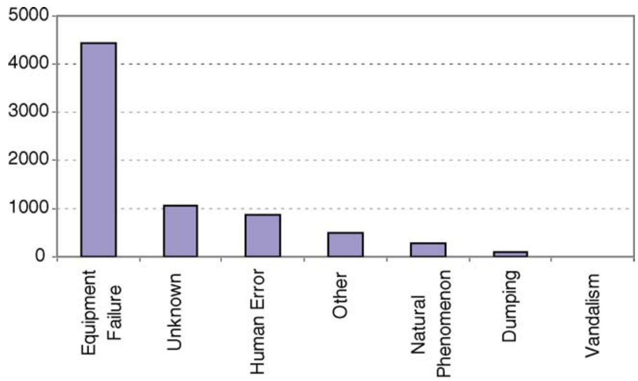

Fig. 3. Distribution of number of incident by cause.

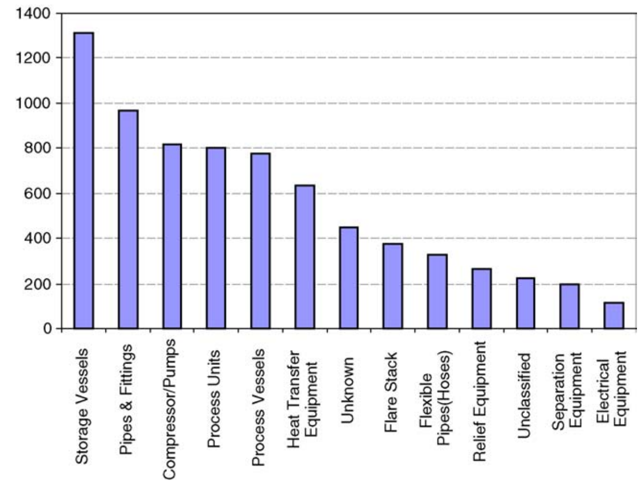

Fig. 4. Distribution of number of incident by type of equipment.

The classification of types of equipment was established early in the study. This classification system was used to identify patterns within the equipment failure category. Fig. 4 presents this distribution.

The decision, with regard to which of the variables should be incorporated in the development of association rules, is based on the distribution presented in Fig. 3. Since Fig. 3 emphasizes that equipment failure is a leading cause, the type of equipment and the type of chemical involved in an incident were selected as the variables that will participate in the association rules analysis. However, the NRC database consists of a large number of chemicals.

7. Association rules

The Lift value was calculated (using expressions 2–4) for all combinations of types of equipment and types of chemical involved in the incidents. Figs. 5–7 present the Lift values of these combinations (12 of the chemicals that were involved in the majority of the incidents were considered).

7.1. Interpretation of the results

A review of Fig. 5 reveals that the Lift value of electrical equipment incidents in which NOx was involved is

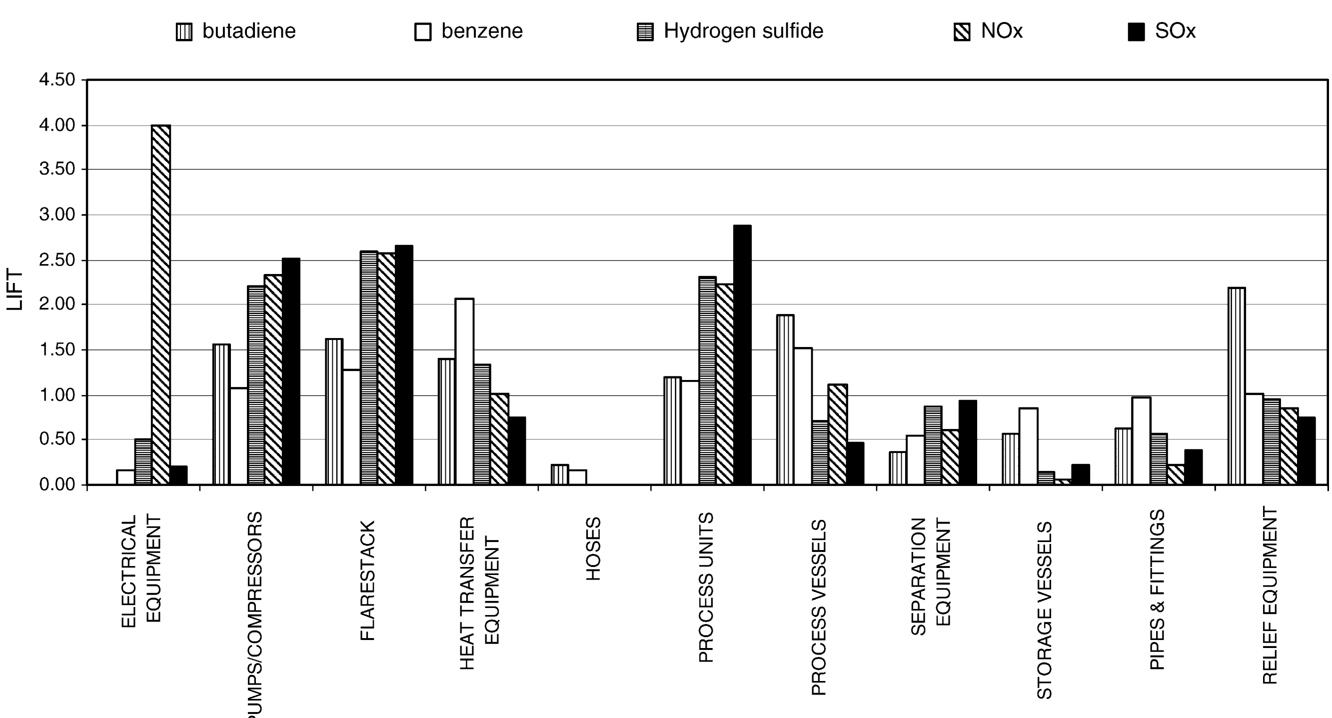

Fig. 5. Lift values—part 1.

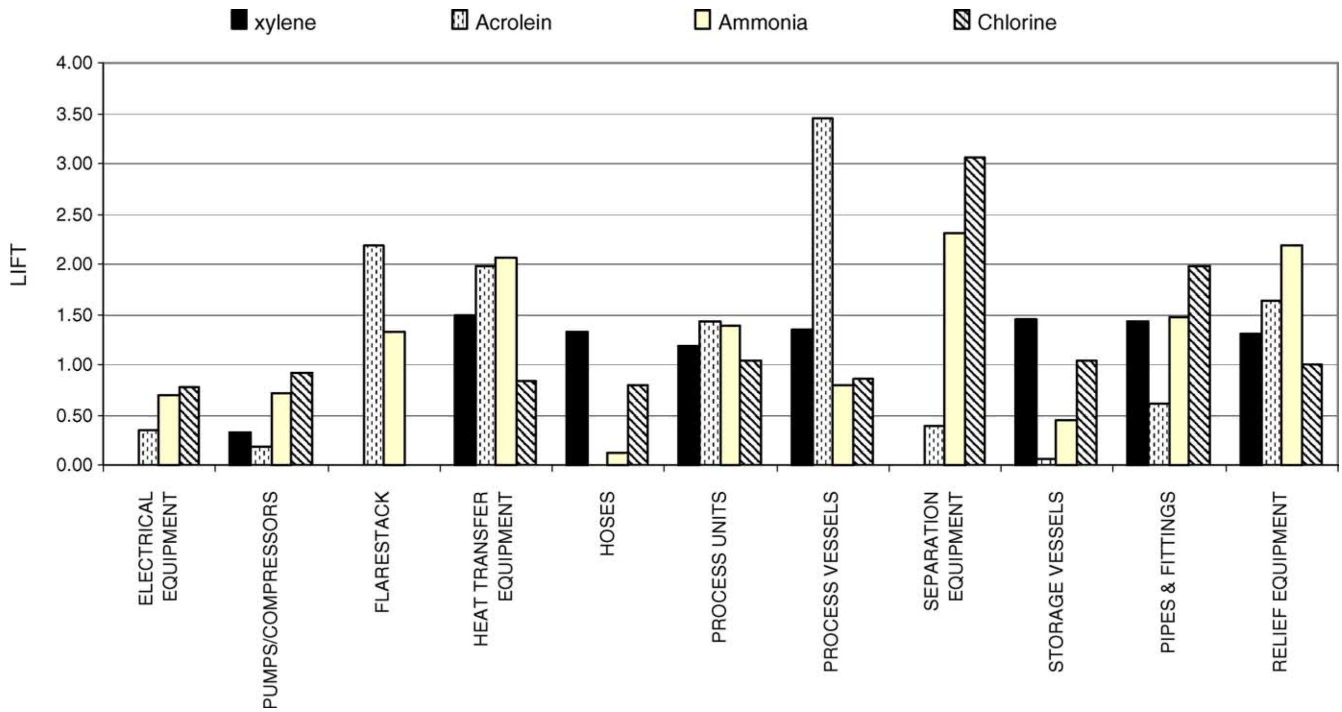

Fig. 6. Lift values—part 2.

approximately 4.0. That means that the probability that electrical equipment will be involved in NOx incident is 4.0 times higher than the general probability of electrical equipment incidents in the database.

Similarly, the Lift value of hose incidents in Fig. 7 emphasizes that the probability of hose incidents in which oil is released is 4.84 higher than the general probability of hose incidents in the database. Review of Figs. 5–7 will give indications on the susceptibility of types of equipment to the chemical involved in the process.

The following section presents the assumptions that have been employed in the examples below.

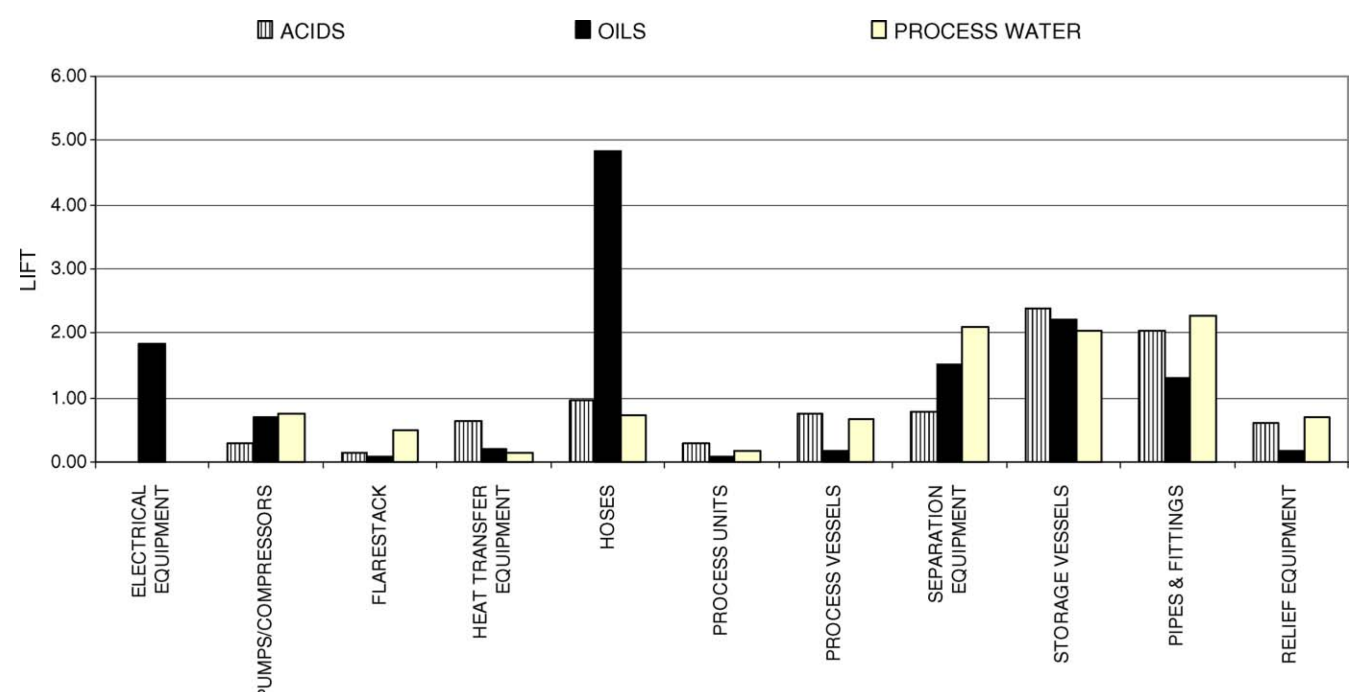

Fig. 7. Lift values—part 3.

Table 2
Modified annual hose failure probabilities

Chemical	Lift value	Recommended annual failure probability [failures/year]
Butadiene	0.23	0.008
Xylene	1.32	0.046
Acids	0.97	0.034
Oils	4.84	0.169
Process water	0.71	0.025

The general term of probability of failure of equipment is given in Eq. (5):

$$P(f) = 1 - e^{-\lambda t} = 1 - \left[1 - \lambda t + \sum_{n=2}^{\infty} \frac{(-1)^n (\lambda t)^n}{n!}\right] \quad (5)$$

where $P(f)$ is the failure probability, λ the failure rate [failure/year], t the time of exposure and n is an auxiliary variable.

However, when $\lambda \ll 1$, then the approximation in Eq. (6) is valid:

$$1 - \left[1 - \lambda t + \sum_{n=2}^{\infty} \frac{(-1)^n (\lambda t)^n}{n!}\right] = \lambda t \quad (6)$$

Therefore, failure probability can be estimated as given in Eq. (7), and then λ represents the annual probability of failure.

$$P(f) = \lambda t \quad (7)$$

Under the limitation given in Eq. (6), Lift values can be used to modify values of annual failure probabilities. Thus, λ (failure rate) is representing the annual probability in these cases. The majority of failure rates in the literature consist of averaged data. Since the Lift value[2] represents the elevation of probability of an incident of equipment with respect to a specific chemical, multiplying the averaged failure rates by the Lift value will produce an annual failure probability that represents the chemical in the process as well.

The following sections will emphasize the usefulness of this aspect of Lift values.

7.2. Example 1

Green and Bourne [15] recommended a hose failure rate of 4 [failures/10^6 h], which is equivalent to 0.035 [failure/year]. Table 2 consists of the annual hose failure probabilities considering a set of chemicals. As mentioned earlier, the modified value was gained by multiplying the value in the literature by the Lift value for the appropriate chemical.

7.3. Example 2

The figures and tables above presented classification of equipment and chemicals to rough categories. In large datasets, it is possible to establish sub-classifications. The following example presents recommended failure rates for gaskets in our dataset for the different types of equipment. Fig. 8 presents the distribution of these Lift values for butadiene, benzene and oils.

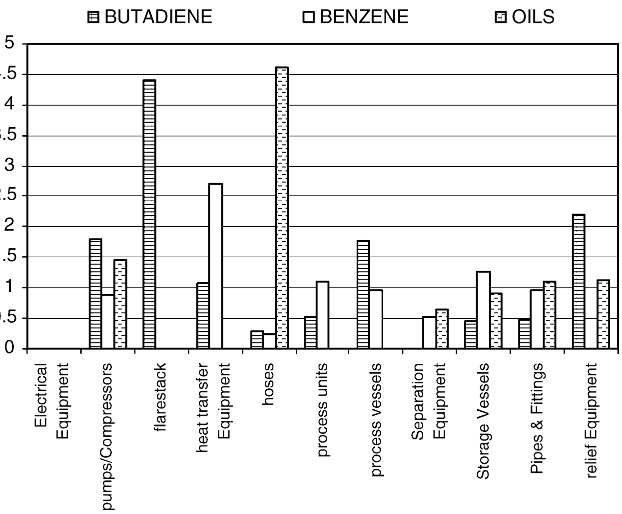

Fig. 8. Gasket Lift values for butadiene, benzene and oils.

The Rasmussen report [16] uses 2.3×10^{-4} [failures/year] as a value for gasket failure rate. Table 3 consists of the modified annual failure probabilities for the variety of combinations of type of equipment. The values are presented for butadiene, benzene and oils. The Lift values for gaskets in the cases of these three chemicals are presented in Fig. 8.

7.4. A word of caution

A necessary step, in order to validate the results, is verifying that the values are based on a sufficient number of incidents. For example, the dataset in this study consists of only three hydrogen-fluoride incidents. Therefore, information extracted on hydrogen-fluoride is misleading. However, any commercial data mining software can present the number of records for a given value in the dataset by a “click of

Table 3
Modified annual gaskets failure probabilities

Equipment type	Butadiene [$\times 10^{-4}$]	Benzene [$\times 10^{-4}$]	Oils [$\times 10^{-4}$]
Electrical equipment	0	0.04	4.3
Pumps and compressors	3.6	2.5	1.6
Flare stack	3.7	2.9	0.02
Heat transfer equipment	3.2	4.7	0.05
Hoses	0.5	0.04	11.1
Process units	2.7	2.6	0.02
Process vessels	4.3	3.5	0.04
Separation equipment	0.9	1.3	3.5
Storage vessels	1.3	1.9	5.1
Pipes and fittings	1.5	2.2	3.0
Relief equipment	5.1	2.3	0.04

[2] It is important to emphasize that this dataset consists information on fixed facility, from 1990 to 2002, in Harris County, TX, only. More accurate values can be gained by incorporating data on all fixed facilities in the US.

a button”. Therefore, verifying sufficient data to support the values is an accomplishable task.

8. Summary

Data mining is a powerful set of tools that definitely can be harnessed to explore incident databases. The work herein is demonstrating the usability of a single data mining methodology. However, data mining can be useful in other areas of process safety. For example, development of data-driven monitoring process based on neural networks and classification trees [17].

A significant advantage of data mining is its ability to analyze enormous data sets. Integration of several data sources into a warehouse will establish a dataset that will enable producing highly validated results, and will create opportunities to use data mining to support other safety arenas.

The efforts following this study will be devoted to establishing a decision-supporting platform that uses results from data mining of incident databases.

References

[1] F. Al-Qurashi, G. Sharma, W.J. Rogers, M.S. Mannan, Applications of relational chemical process safety databases for lowering mean failure rate, Process Saf. Prog. 20 (4) (2001) 280–285 (New York).
[2] M.S. Mannan, T.M. O’Connor, H.H. West, Accident history database: an opportunity, Environ. Prog. 18 (1) (1999) 1–6.
[3] N. Keren, T.M. O’Connor, M.S. Mannan, Use of incident data collection from various sources for industrial safety performance assessments, in: Proceedings of Workshop on Investigation and Reporting of Incidents and Accidents, VA, USA, September 16–17, 2003, pp. 51–67.
[4] Introduction to Data Mining and Knowledge Discovery, Two Crows Corporation, third ed., 1999, http://www.twocrows.com/intro-dm.pdf (last visited on 2004).
[5] A. Berson, S. Smith, K. Thearling, Building Data Mining Application for CRM, McGraw Hill, 1999, pp. 2–13.
[6] P. Giudici, Applied Data Mining, John Wiley & Sons, West Sussex, England, 2003.
[7] M.J.A. Berry, G. Linoff, Data Mining Techniques, John Wiley & Sons, NY, USA, 1997, pp. 141–147.
[8] R.E. Walpole, R.H. Myers, Probability and Statistics for Engineers and Scientists, 1989.
[9] V.C. Marshall, Major Chemical Hazard, Halsted Press: A Division of John Wiley & Sons, NY, USA, 1987.
[10] Division of Health Studies, Epidemiology and Surveillance Branch, Hazardous Substances Emergency Events Surveillance (HSEES): Annual Reports, http://www.atsdr.cdc.gov/HS/HSEES/ (last visited on 2004).
[11] Department of Transportation—The Bureau of Transportation Statistics, http://www.transtats.bts.gov/ (last visited on 2004).
[12] National Response Center, http://www.nrc.uscg.mil/ (last visited on 2004).
[13] Code of Federal Regulations, Title 40, vol. 25, 2003, http://www.access.gpo.gov/nara/cfr/waisidx_03/40cfr302_03.html (last visited on 2004).
[14] R.E. Iliffe, P.W.H. Chung, T.A. Kletz, M. Preston, The application of active databases to the problems of human error in industry, J. Loss Prev. Process Ind. 13 (2000) 19–26.
[15] A.E. Green, A.J. Bourne, Reliability Technology, Wiley/Interscience, New York, 1972.
[16] WASH-1400 (NUREG-75/014), Reactor Safety Study, An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants, U.S. Nuclear Regulatory Commission, October 1975.
[17] Y. Zhou, Data driven process monitoring based on neural network and classification trees, Dissertation, Chemical Engineering Department, Texas A&M University, College Station, TX, 2004.