Analysis of HSEES Chemical Incident Database Utilizing Data Mining

By Mahdiyati Syukri

Mary Kay O'Connor Process Safety Center - Texas A&M University

Chemical incident databases have been established since the occurrence of catastrophic incidents such as Flixborough, Seveso, and Bhopal (Meel, 2007). The process of reporting and collecting information on occupational incidents is a good practice in the industry because it provides as a source of valuable information that is beneficial to prevent the reoccurrence of incidents. Since 1970 there has been a continuous improvement in safety performance in the industry, and this is partly attributed to the sharing and learning lessons from incidents (Jones, 1999).

HSEES (Hazardous Substances Emergency Events Surveillance) is a chemical incident database administrated by ATSDR (Agency for Toxic Substance and Disease Registry) - CDC (Center for Disease Control) of USA. The objective of this chemical incident database is to capture the health effects of incident and use the findings to reduce subsequent morbidity and mortality. HSEES data consists of release type, chemicals released, and great details on victims' injury severity. Due to the broad coverage of incidents collected by HSEES as shown in Figure 1, this study focuses on a particular segment of the data, that is data covering chemical releases which occurred in manufacturing facility.

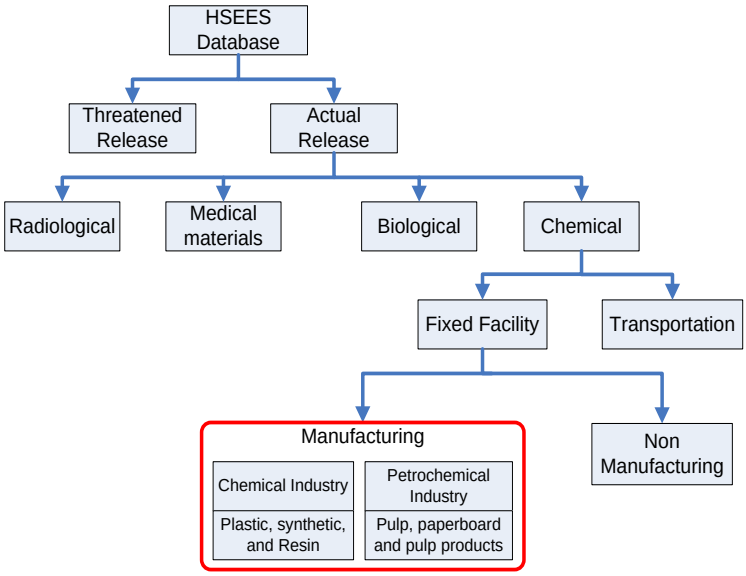

Figure 1 HSEES database

This research is aimed to analyze HSEES data using data mining methodology. Data mining is a process of multivariate analysis that is used to reveal pattern, association or relationship between variables in large datasets. Data mining comprises of different techniques, several which are used in this research are cluster analysis, decision tree and association rule.

As shown in Figure 2, the analysis starts by exploring HSEES data using statistical means to get the broad picture and the feel of the data. The statistic findings then will be used to direct the study on a particular segment of data or variables that are desired based on its significance or number of occurrences. Once the desired data has been selected, the next step is to prepare the data for analysis. This step involves cleaning data from duplicates entry and imputation to manage missing data. The step after is data transformation, where data is converted to a suitable form. Then data mining model is applied on to the transformed data. Data mining model application/building will identify pattern, trend or associations among the variables of the data. The results of data mining model will be evaluated accordingly before using it as knowledge for improving safety performance.

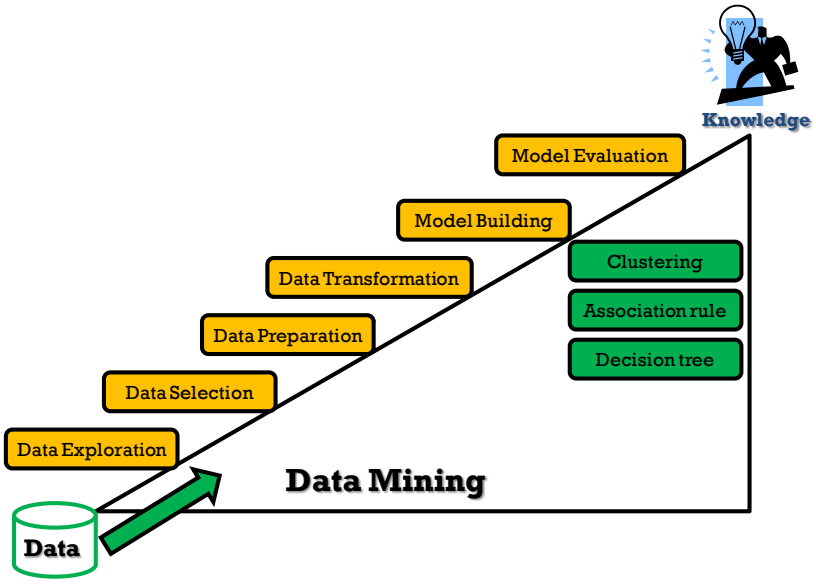

Figure 2 Data mining methodology

The statistical analysis resulted in several findings on the trend and number of chemical releases in manufacturing facility, industry where releases occurred, prevalent contributing cause to the incidents, states where the release located, and prevalent chemicals released as shown in Figure 3 and Table 2 through Table 5.

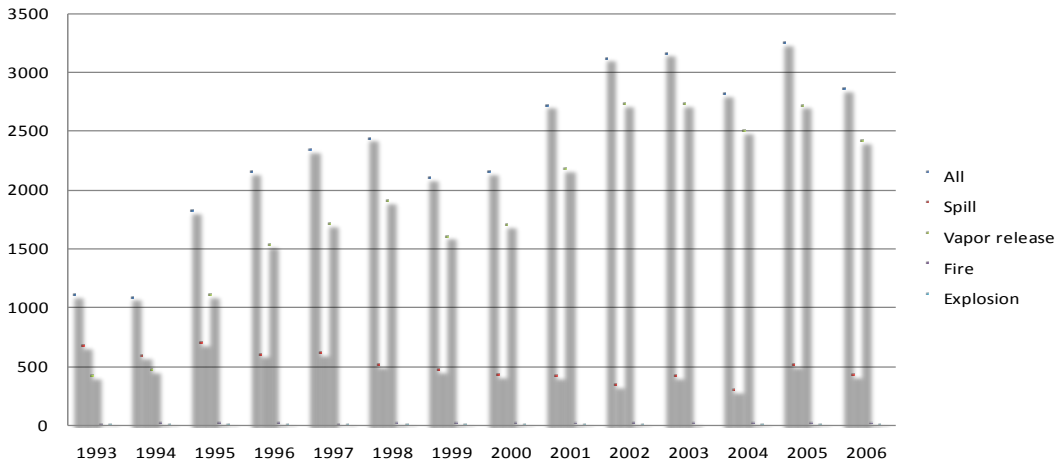

Figure 3 Chemical releases in manufacturing facility from 1993-2006

Table 2 Industry category of chemical releases in manufacturing facility

Industry	Count	Percentage
Chemicals	13404	39.1%
Petrochemicals	7602	22.2%
Plastics, synthetics, and resins	4321	12.6%
Pulp and paper product	1019	2.9%
Primary aluminum industries	295	0.8%
Others	428	1.2%
Missing	7224	21.1%

Table 3 Contributing category of chemical releases in manufacturing facility

Contributing factor	Count	Percentage
Equipment failure	20493	59.7%
Human error	2608	10.3%
Deliberate acts	2305	6.7%
System/process upset	901	2.6%
Maintenance	825	2.4%
Others	799	2.3%
Bad weather	732	2.1%
System startup/shutdown	571	1.6%
Power failure/electrical problems	358	1.0%
Improper procedure	239	0.7%
Factors beyond human control	143	0.4%
Missing	4319	12.5%

Table 4 States category of chemical releases in manufacturing facility

States	Count	Percentage
Texas	23603	68.8%
Louisiana	4157	12.1%
New York	1168	3.4%
New Jersey	966	2.8%
Minnesota	872	2.5%
Washington	758	2.2%
Alabama	516	1.5%
Utah	370	1.1%
Mississippi	323	0.9%
Michigan	269	0.8%
North Carolina	269	0.7%
Iowa	247	0.7%
Oregon	227	0.6%
Wisconsin	194	0.5%

Table 5 Chemical category of chemical releases in manufacturing facility

Chemical	Count	Percentage
Sulfur dioxide	3514	10.2%
Nitric oxide	2277	6.7%
Benzene	1253	3.6%
Butadiene	825	2.4%
Ammonia	801	2.3%
Hydrogen sulfide	725	2.1%
Ethylene	622	1.8%
Sulfuric Acid	595	1.7%
Carbon monoxide	540	1.6%
Vinyl chloride	512	1.5%
Chlorine	509	1.5%

Cluster analysis is used to group data into clusters that share distinct variables characteristic. Using Euclidian distance, the grouping is optimized iteratively with the objective to minimize the distance between members in each cluster and to maximize the distance between clusters. Data mining computation using STATISTICA™ resulted in two groups with characteristic shown in Table 6. Because of similarities among its member, patterns in these subgroups should be easier to be identified. Therefore, further analysis will use cases from these two groups.

Table 6 Cluster analysis result

Cluster	Industry	Equipment	Chemical	Release type	Count	%
1	Chemical Industry	Process vessel	Sulfur Dioxide	Vapor release	21105	78
2	Petrochemical Industry	Ancillary process equipment	Sulfur Dioxide	Vapor release	5932	22

Association rule is used to find association between categorical variables in large datasets. Specifically in this study, association rule is used to find co-occurrences of attributes of attributes that appear with the highest co-frequencies. Hence, it gives a quantification on how strong the association is and how likely it is to occur again. Table 7 shows the association rule results.

Table 7 Association rule result

No.	A	B	P(A∩B)	P(A|B)
1	Sulfur Dioxide	Process vessel	0.06	0.16
2	Spill	Process vessel	0.05	0.13
3	Vapor release	Process vessel	0.35	0.86
4	Spill	Piping	0.07	0.47
5	Vapor release	Piping	0.08	0.53
6	Vapor release	Ancillary process equipment	0.27	0.93
7	Process vessel	Sulfur Dioxide	0.06	0.59
8	Vapor release	Sulfur Dioxide	0.10	0.96
9	Process vessel	Spill	0.05	0.24
10	Piping	Spill	0.07	0.29
11	Process vessel	Vapor release	0.35	0.46
12	Piping	Vapor release	0.08	0.10
13	Ancillary process equipment	Vapor release	0.27	0.36
14	Sulfur Dioxide	Vapor release	0.10	0.14

The data mining resulted in several interesting pattern and association that still requires validation before being used in as an input improving safety performance. Further work in this study will includes more variables that describe the severity of consequences and the amount of chemical involved in the incidents.

Reference

Jones, S., C. Kirschsteiger, et al. (1999). "The importance of near miss reporting to further improve safety performance." Journal of Loss Prevention in the Process Industries **12**: 59-67.

Meel, A., L. M. O'Neill, et al. (2007). "Operational risk assessment of chemical industries by exploiting accident databases." Journal of Loss Prevention in the Process Industries **20**(2): 113-127.

Robert Nisbet, John Elder, Gary Miner. Handbook of Statistical Analysis and Data Mining Applications. Academic Press, 2009.