2011

A Novel Data Mining Methodology for Narrative Text Mining and Its Application in MSHA Accident, Injury and Illness Database

Xiaoli Yang
Iowa State University

Recommended Citation

Yang, Xiaoli, "A Novel Data Mining Methodology for Narrative Text Mining and Its Application in MSHA Accident, Injury and Illness Database" (2011). *Graduate Theses and Dissertations*. 12074.
https://lib.dr.iastate.edu/etd/12074

A novel data mining methodology for narrative text mining and its application in MSHA accident, injury and illness database

by

Xiaoli Yang

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:

Sigurdur Olafsson, Major Professor

Gary Mirka

Peter Sherman

Iowa State University

Ames, Iowa

2011

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Mining is one of the most dangerous industries. Mine Safety and Health Administration (MSHA) maintains a database that records thousands of mining related accidents, injuries or illnesses every year with incident descriptions in narrative texts. How to uncover knowledge from these narrative texts is lacking. The goal of this study is to propose a new data mining methodology that incorporates or extends existing methods and is able to uncover useful information from massive amount of narrative texts in a streamline fashion. In our experimentation with data of 2008, we focus on 3 different types of common injuries and apply the new methodology to their narrative texts. Some interesting results are found that are worthy further investigations with the help of mining safety experts.

CHAPTER 1 INTRODUCTION

1.1 Background

Mining has been one of the most dangerous industries since the earliest days people started digging useful materials out from the earth. Although continuously emerging new technologies has allowed mining workers a much safer working environment than many years ago, mining accidents are still constantly happening all over the world. In the United States, mining industry has a fatality rate eight times that of the national average for all industries. The importance of effective safety enforcement in mining industry should never be overstated.

One way to improve safety in mining industry, as many studies suggested to other industries with high accident rates, is to make use of databases of related accidents to identify patterns and gain insights. As data collection and storage become increasingly easier and less expensive, such databases have been more prevalent and easier to access than ever before. Massive amount of data that contain valuable information of "what happened" are collected and stored in these databases only waiting to be explored and investigated. They can be of great help for people who want to better understand what have happened in the past and inspire future improvements. However it also becomes even more challenging for data analysis practitioners to properly handle those data in massive amount, various formats, being lack of concrete structure, and with no clear data analysis objective in place. Traditional hypothesis-driven data analysis methods falter in front of tasks such as discovering hidden and unexpected patterns in dataset consisted of various data types. They are also incapable of addressing data mining inquiries within datasets that are lack of concrete structure. Data mining technique emerged as a very prospective alternative solution to these challenges and has been shown both in academics and industries effective in transforming raw data into useful knowledge where traditional data analysis methods may be struggling with. In this study, data mining methods are applied to a subset of Mine Safety and Health Administration (MSHA)'s 2008 Accidents, Injury and Illness (AII) database. A new methodology is proposed under which all data mining methods used are integrated as one in a streamline fashion.

1.2 Motivation

MSHA AII database is a typical industrial incident database. It contains structural data with well defined contents and formats, and nonstructural data in the form of narrative texts to provide background information with regard to each incident recorded. Most existing data mining methods were initially devised to work with structural data, and a lot of efforts have been devoted to making the best use of them in various contexts. The opportunity of successfully applying these existing data mining methods to gain useful knowledge from MSHA AII database is least questionable. On the other hand, successful data mining practices with free-form narrative texts are relatively less common. Such practice in underground mining industry is even rarer. However, the value of narrative texts should not be underestimated. In the case of MSHA AII data, its narrative texts contain a good wealth of descriptions of all mining accidents occurred. These descriptions often contain very good details that are missing in the structural part of the database. We believe that a successful exploration of these narrative texts could yield interesting patterns that can be good supplements to what we can know about from the structural data, and therefore lead to better understanding of these accidents. How to extract useful information from these free-form narrative texts was then brought to our attention and became a field of study that we want to contribute to.

1.3 Objective

The objective of this study is to propose a new data mining methodology for uncovering useful knowledge from narrative texts in MSHA AII 2008 database. Through an implementation of this methodology we expect to improve our understanding of various types of mining accidents and injuries. More specifically our objective is to find natural grouping of specific types of mining accidents from free-form narrative texts of these accidents, and to characterize the natural grouping in a systematic manner that can provide useful information and serve as the baseline for further interpretation and investigation. We will structure the narrative texts in sparse binary data and define natural grouping in this format. Since existing methodology appears to be lacking, we will also need to incorporate and extend existing methods in order for them to be used for our purpose.

To achieve the above objective, the specific research tasks are thus two-fold:

First, we will extend existing methods in order to obtain and characterize meaningful grouping from the free-form accident descriptions in the MSHA AII database. These existing methods are integrated in the framework of our methodology in a streamline fashion. While not tested on other databases, the resulting new methodology is expected to be a general data mining methodology for narrative texts and can be easily extended to other similar usage scenarios.

Second, we will analyze 3 common types of mining accidents, namely “Burn or Scald”, “Crushing”, and “Cut, Laceration or Puncture”. For each type we will apply the above methodology, that is, finding and characterizing natural grouping within incidents of each kind to provide meaningful insights. The expectation is that these insights could then be interpreted by mining safety experts to implement safety improvements.

1.4 Thesis Structure

Chapter 2 Literature Review goes over some recent studies in incident data mining and text mining. Chapter 3 Data Mining is a general introduction of data mining. Chapter 4 Methodology discusses in details the methodology we propose. Chapter 5 Case Study presents some of the findings we found by applying the new methodology to the narrative texts in MSHA AII database with concentration on 3 different types of injuries. We conclude in Chapter 6 Conclusion through a summary our new methodology and some thoughts on future work.

Chapter 2 LITERATURE REVIEW

Data mining has been successfully applied to many areas in both industry and academia. However it’s still relatively new to many safety engineers. Since most databases in safety related fields are structurally similar with those from other areas of many industries, one can anticipate that data mining should also be able to work for safety engineers.

In recent years, data mining methods have gained increasing popularity in the field of safety, health and environment. Various data mining methods have been used to explore incidents in chemical engineering area. Interesting patterns were found that could help reduce safety issues in chemical processes (Anand et al., 2006). Similar data mining practices can also be found in drug safety area. It has been shown that with appropriate data mining methods it is possible to predict early signals of adverse drug reactions (Lindquist et al., 2000). In another study, a data mining methodology was proposed to integrate with active expert involvement to evaluate coronary heart disease risk (Gamberger et al., 2003). The study showed that data mining is able to detect risks and discover natural grouping of patients. The results can therefore provide useful information about the root cause of the disease and other insights.

Research in data mining application has notable contribution to the field of transportation safety in particular. Tree-based classification method has been shown effective in determining key factors for frequent accidents on Taiwan’s freeway (Chang et al., 2005). Similar methods have also been used to model the severity of injury resulted from traffic accidents (Chong et al., 2004). The results open new insights into these accidents and their causes. In another study of traffic accidents in Korea, various data mining methods were used to uncover hidden patterns behind those accidents and it was pointed out that all methods found protective driving devices play more important roles in traffic accidents in different ways and therefore should be brought up to even more in-depth investigations (Sohn et al., 2001).

Data mining was also introduced to introduced to the field of food engineering in a recent study. It has been shown capable of recognizing patterns in foodborne disease outbreaks (Maitri et al., 2009). Various data mining methods were discussed and put in practice to analyze a CDC database of foodborne diseases. This data-driven approach opens new opportunities for addressing issues that are not clearly known before in the area of food safety.

Since many incident databases in various industries incorporate large amounts of narrative text information, data mining began to approach them from a text mining perspective. A fuzzy Bayesian model was applied to categorize cause-of-injury codes in the narrative texts extracted from National Health Interview Survey (Wellman et al., 2004). This application was then further improved by using Singular Value Decomposition (Noorinaeini et al., 2007). One major challenge of applying data mining methods to narrative texts comes from the fact text information is far less structured than the more common table-like data.

The lack of structure also creates difficulties for linguistic research. The concept of “Bag of Words” was proposed in the linguistic context years ago to address this issue (Harris, 1954). More and more data mining practices adopt the idea of “Bag of Words” as a way to convert nonstructural narrative information to structural dataset so that they can be fed to common data mining methods. One particular area that has seen many successful applications of Bag of Words is web content search and mining (Kosala et al., 2000).

So far data mining methods have not yet been well made use of in mining industry. This is a surprising fact, given that databases such as MSHA database have long been well established and that mining industry has an accident rate much higher than the average across all industries. Having seen these many successful applications of data mining methods in various safety areas, we believe that such a successful application in mining industry is also very prospective. We can legitimately expect to see many interesting and useful results extracted from the MSHA database to help improve mining safety.

CHAPTER 3 DATA MINING

3.1 Overview

Today's world is becoming increasingly data driven. Data is playing significant roles in almost every respect of our daily life. We use data to gain a better understanding of what was in the past, to know where we are currently at, and to predict what will be in the future. With the rapidly developing information technology, data collection and storage are becoming easier and cheaper than ever before. When the availability of data is in many cases is no longer a major concern, it is one that how we can utilize these data both efficiently and effectively.

Data mining, which is often inappropriately referred to as knowledge discovery, is a set of concepts, techniques and methods used for "mining" useful information, interesting patterns, or knowledge out of data. Data mining differs from traditional hypothesis-driven data analysis methods in that it is able to uncover often hidden, unexpected knowledge. As an inductive approach, it does not require an a priori hypothesis. Consider the case that patterns describing incidents of a certain type of accident need to be explored. Traditional hypothesis-driven methods require a criterion as the start of a query or search, looking for evident that are consistent with this criterion. One would naturally foresee that the results of the traditional hypothesis-driven methods are heavily dependent upon the pre-specified criterion. On the other hand, data mining does not require such criterion and is able to explore in a large problem space. This illustrates the fundamental difference between data mining techniques and traditional hypothesis-driven methods.

The most learned knowledge in data mining is classification, clustering and association rules mining. In this study, clustering and association rules mining lie within the core of the whole methodology. Attribute selection is also used to foster better results. All methods will be introduced in the next sections. One should be noted that even though the state-of-art techniques or algorithms play main roles in the field from a research standpoint, data mining itself is a process that involves many other practices, many of which are less technical but as critical as, if not more critical than, the technique or algorithm itself. Many unsuccessful data mining practices are actually caused by the people's misinterpretation of data mining as a fully automatic one stop solution. While an abundant of algorithms are available in the field of data mining and often novel algorithms draw a lot of attention, one should also be aware of that data mining is more a methodology than merely a particular algorithm. This methodology follows the same way as how people approach an inductive task. Figure 1 shows a complete data mining process.

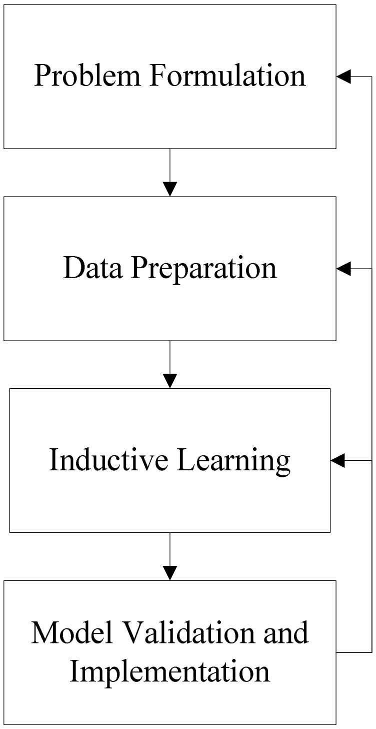

Figure 1 - Data Mining Process

- Problem Formulation: every data mining project needs a clearly defined objective. This requires a close cooperation between the data analyst and the customer. It is very critical that the data analyst fully understand what the customer wants and be able to convert the customer needs into a well defined data mining problem.

- Data Preparation: once the project objective is defined, the next step is to identify the appropriate data source. Data collection, integration and preprocessing constitute a complete data preparation, and could take up a large portion of time in the whole data mining project.

- Inductive Learning: inductive learning is the core of the whole data mining process. Classification, clustering and association rule are most learned knowledge in inductive learning. There are a lot of inductive learning algorithms available nowadays, and none of them outperform all others. Hence the how to choose most suitable algorithm also become an interesting topic in the field of research. However, most popular algorithms are able to yield good results if applied appropriately.

- Validation and Implementation: the knowledge we obtained from inductive learning needs to be validated so that it can be implemented for further usage. This usually calls for domain expertise. If it turns out to be invalid, the data mining project may be needed to be redone. A wrong data mining practice is no better than doing nothing.

In the follow sections, we will give a brief and basic introduction of the major methods in the field of data mining, namely Classification, Clustering, Association Rules Mining and Attribute Selection.

3.2 Classification

Classification is a supervised machine learning method that predicts the subgroup of the whole data space (class label) that an instance belongs to. Classification is supervised because it requires the availability of a set of data instances that are already labeled with classes called training set. Classification can be considered as a learning of mapping of function. Given an instance x, classification seeks to predicts its class label y based on a mapping or function f it constructed, so that $y = f(x)$, and function f is commonly called classifier.

Classification is a 2-step process. In the first step, classification constructs a classifier based on a training set. Assuming that all instances in the training set are labeled correctly, classification constructs a classifier to fit the associations between instances and classes. This step is mostly where different classification methods differ from each other. The classifier is also evaluated on different measures as a guidance of how reliable the classifier is.

In the second step, the classifier constructed and passed the evaluation in the first step is applied to new instances to predict their class labels.

Decision Tree is one of the most important classification methods. It constructs its classifier in a tree structure that is very easy to read and interpret and hence gains a lot of popularity. Other popular classification methods include Bayesian Classification and Support Vector Machine (SVM) etc.

3.3 Clustering

Clustering is an unsupervised learning method that split a whole dataset into groups (classes) based on their similarity or distance with each other. It is unsupervised because unlike Classification it does not need a training set for learning. Clustering helps us understand the intrinsic structure or hidden patterns within a dataset or between different instances. Many clustering algorithms have been devised to handle different types of data in different contexts. They can be divided into two major categories: hierarchical clustering and partitioning clustering.

A hierarchical clustering method creates a hierarchical structure that groups instances based on different similarity levels. It can work in a top-down fashion, which starts with each individual instance being a cluster and proceeds with merging them until certain criterion is met, or in a bottom-up fashion, which starts with treating all objects in one single cluster and proceeds with breaking it down into multiple groups. Popular hierarchical clustering algorithms include Single-Linkage, Complete-Linkage, BIRCH, etc.

A partitioning clustering method simply partitions the whole dataset based on the similarity between objects, and each partition represent a cluster. Popular partitioning clustering algorithms include k-Means and k-Medoids. The MinDisconnect algorithm we will introduce is also a partitioning clustering algorithm.

All clustering methods seek to group similar instances together, and assign dissimilar instances to different groups. Similarity is therefore an important concept in clustering. A clustering method itself does not guarantee a successful clustering practice. Given an appropriate clustering method, how to define the similarity between any two instances becomes critical to whether or not we can make sense out of the clusters to be found. Therefore for every clustering practice we must devise carefully the similarity measure based on the characteristics of the dataset and always keep in mind the objective. More than often similarity is also interpreted as distance. Two similar objects are considered to be close to each other, while on the other hand, two distant objects are unlikely to be similar. One of the most common of such measures is Euclidean distance measure, and in many situations when people refer to “distance” they mean Euclidean distance. The Euclidean distance d_{ij} between x_{ik} and x_{jk} is defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$$

where x has n coordinates.

3.4 Attribute Selection

Attribute Selection is a method that seeks to remove irrelevant or redundant attributes from the dataset. The resulted reduction of data set size can be beneficial to many computationally intensive data mining methods. The removal of noises being irrelevant or redundant information can also yield higher quality data mining results.

Most Attribute Selection methods determine good attributes using various statistical tests. Based on the statistical significance of how well an attribute interact with class labels, those don't yield high significance are then considered irrelevant or redundant. What statistical test is used differs one Attribute Selection method from another.

One other hurdle that any Attribute Selection method must overcome is how to efficiently navigate the space of different combinations of attributes and search the ones that yields significance in statistical tests. An exhaustive search can be prohibitively time consuming or even practically impossible. Heuristic search methods, for example Greedy search, are often utilized to improve the overall efficiency.

CHAPTER 4 METHODOLOGY

4.1 Overview

We start with data preprocessing to eliminate instances that provides limited or no useful information for us to understand the root causes of incidents. We proceed with the preprocessed dataset to build its bag-of-words matrix using all the words involved with Term-Matrix Generator. Our next step is to utilize a new clustering method called MinDisconnect to identify the natural groupings of instances with one particular injury type at a time. Among all the different groupings generated by MinDisconnect, we choose the one that has the most what we call "clusters of interest". This grouping is then fed to RELIEF attribute selection method to filter out less relevant words. With a group of "clusters of interest" and a set of words considered more relevant to the injury type being examined, we apply Apriori association rule mining method to systematically characterize each "cluster of interest" with words that have strong association with it. Finally, these characterizations are examined. Certain interesting ones will be reported to inspire corrective actions or further investigations.

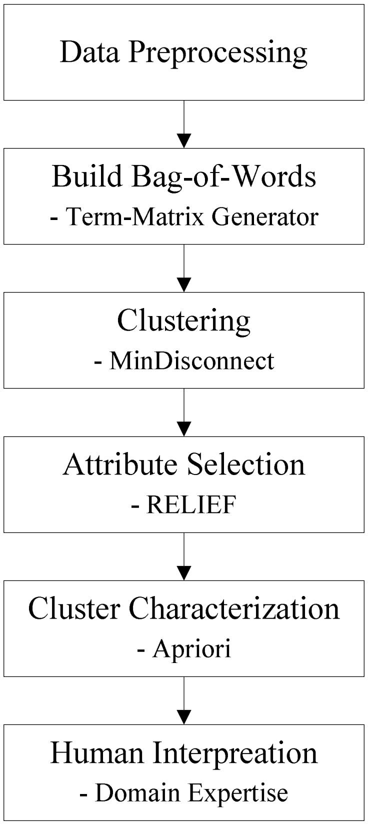

Figure 2 - Methodology Flowchart

4.2 Data Preprocessing

All data mining practices start with data preprocessing. Comes together with the availability of large amount of data is also a large amount of non-useful data or noises. These noises could be caused by undisciplined data collection practices, accidental data losses, data contaminations such as improper modifications, deletions or additions, or in many cases simply because the data contains instances that are irrelevant to the goal of the data mining task.

We can either remove these non-useful data, or convert them to useful one, depending on what kind of non-useful data we are dealing with or what goal we are trying to achieve. Note that the non-usefulness of data is scenario dependent. Non-useful data in one data mining task can be useful in another task, or sometimes in a different stage of the same data mining task.

In MSHA 2008 database, about 43.6% of all 13576 incidents resulted in injuries that have no work day losses (DAYSTOTL=0) which we consider as minor injuries. We think the causes of these minor injuries are too diversified or inconsistent to be able to yield any patterns that can be recognized systematically. In addition, given the still large amount of information available in the other 56.4% of data, we think they deserve most of our efforts to look for useful patterns to prevent more major injuries. So we decide to remove all minor injuries from the scope of this research and only focus on major injuries that yielded work day losses.

A bonus of removing 43.6% of 13576 instances from a dataset is that we end up with a much smaller dataset. Since many data mining algorithms are computational intensive, reducing the size of a dataset down to about its half could result in significant saving on algorithm run time and resources needed.

4.3 Building Bag-of-Words

A Bag-of-Words is a data table with words as its header and each data instance corresponds to one and only one row in the table where every word in the header is represented. For each data instance, a word can be represented by a binary (0-1) variable that tells whether or not it occurs, a numerical variable that says how many times the word occurs, a categorical variable that says how important or interesting the word is, or many other representations.

Building a Bag-of-Words is a process of identifying words that are relevant to what we are looking for in the Bag-of-Words. With no other particular reason to believe one word is more important than the other, in this study we treat all words equally, that is we attempt to extract as many words as possible as long as they exist in the narrative texts. One exception is that we eliminated common stop words because most of them in most cases do not contribute meaningful information to actual interpretations.

Term-to-Matrix Generator (TMG) is a powerful tool that can build Bag-of-Words quickly from a large amount of texts without much preprocessing or formatting required. As a MATLAB toolbox, TMG provides capabilities of handling text indexing, retrieval, dimensional reduction and other practices including simple text mining. In this study, we will utilize TMG solely as a text retrieval tool. TMG can extract words from texts based on different criteria and generate a Bag-of-Words in various forms, and eliminate stop words in a semi-automatic fashion.

Our Bag-of-Words will be built with words as the column headers and each instance corresponding to a row variable, where each of its member is a binary variable indicating whether or not a word in the column header occurred in the instance represented by the row variable. The table below shows the structure of our $m \times n$ Bag-of-Words.

Table 1 - An Example of $m \times n$ Binary Bag-of-Words

	word 1	word 2	word 3	…	word n
instance 1	1	0	0	…	0
instance 2	0	1	1	…	0
instance 3	1	0	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮
instance m	0	0	0	…	1

With 7660 instances (after removing those with no work day loss) of narrative texts, the Bag-of-Words returned by TMG turned out to be a huge and sparse matrix with over 8000 words/columns. In our implementation, this resulted in difficulty in handling data on a common personal computer, let alone when data of this size and sparseness is fed to many data mining algorithms that are computational resource sensitive. To overcome this difficulty in a context of academic study, we arbitrarily choose to include in our Bag-of-Words only words with a global frequency at least 10. This reduces the number of words/columns down to slightly more than 1300, which makes possible many data related operations in the next phases.

Now every instance is now represented solely by words from its original narrative texts, and is accordingly transformed into a structural format. This opens the door to many popular data mining methods most of which were devised for structural data only.

4.4 Clustering with MinDisconnect

With a Bag-of-Words on hand, we now seek to cluster all incidents based on the words they contain. We want to look for words that are representative of certain clusters. These words can therefore be helpful with uncovering and interpreting related patterns.

4.4.1 Similarity and Distance Measure

The very first question to be answered before initiating any data clustering practice is what kind of instances should be considered to be grouped together. Since in almost all cases we attempt to cluster similar instances together, this is actually a question of how to define "similarity" of any two instances. One can have multiple definitions of similarity for the same dataset. Which one of them is more preferable than the others is a decision that needs to be made carefully in a case by case fashion.

Let I_i and I_j are two instances of row variables in our Bag-of-Words. In this study, we propose the following similarity measure:

$$Sim_{ij} = \frac{\left\| I_i \cap I_j \right\|}{\left\| I_i \cup I_j \right\|}$$

Note that $Sim_{ij} \in [0,1]$, with Sim_{ij} being closer to 1 meaning more similar of the two instances I_i and I_j are to each other.

The above similarity measure is essentially a ratio of the number of the words in common to the total number of words present between I_i and I_j. The norm of the intersection or conjunction of I_i and I_j makes sure we focus only on the presences of words (1's) instead of the absences of words (0's) by ruling out all the 0's involved. The reason that only shared words count are two-folded: (a) the absence of a word in an instance does not necessarily indicate the absence of the meaning the absent word refers to, thus is not a pattern; (b) one should anticipate a lot more absent words than those in common between any two instances, and the former should not overwhelm the latter since the latter is where all the patterns could possibly exist in. In short, the similarity of any two instances should be measured by how many words they share in common, not by how many words they are different at.

Since the concept of similarity is often represented as distance in most clustering methods, we adopt the following distance measure:

$$d_{ij} = 1 - Sim_{ij}$$

Note that $d_{ij} \in [0,1]$, with d_{ij} being closer to 1 meaning the more distant the two instances I_i and I_j are to each other.

The clustering method, MinDisconnect, looks for the nearest neighbors of each instance based on the relative values of their distances. The absolute value of the distance between two instances is therefore less important. While we could have other options of defining distance measure, for example defining d_{ij} as the reciprocal of Sim_{ij}, aka. $d_{ij} = \frac{1}{Sim_{ij}}$, having a distance measure between 0 and 1 could have other potential benefits from a computational perspective.

4.4.2 Clustering Quality Measure

The next question to be answered is, given the similarity definition, what defines a good cluster. In this study, a cluster is a set of instances as row variables indicating the occurrences of certain words. Intuitively, a good cluster should be able to differentiate itself from others in terms of the words its members possess. For instance, if cluster X possess word A, and no other cluster possess word A, then we say X is a good cluster in regard with word A. However, this is a very strict requirement. A lot of times a word can be shared across instances in different clusters but has different degree of prevalence in different clusters. For instance, if more than half of the instances in cluster X possess word A, much more than in any other clusters, then we say cluster X is also a good cluster in regard with word A. When we are looking at a decently large number of words ("bag of words") that may distribute across the whole problem space sparsely, it may be difficulty to find such single word A. However, similar pattern could exist in terms of a word set consisted of more than one word. For instance, cluster X may possess more word set {A, B} than any other clusters do. We call cluster X a good cluster in regard with word set {A, B}.

Before we introduce the cluster quality measure, there are several important concepts we need to define beforehand.

Let W be a full set of words. $W = \{W_n\}$, $n = 1, 2, ..., N$ and N is the total number of words.

Let I be the set of all instances. $I = \{I_m\}$, $m = 1, 2, ..., M$ and M is the total number of instances. $I_i = \{w_{i1}, w_{i2}, ..., w_{iN}\}$.

We use a binary variable w_{ij} to denote the occurrence of word W_j in instance I_i, that is

$$w_{ij} = \begin{cases} 1, \text{ if } W_j \text{ is present in instance } I_i \\ 0, \text{ otherwise} \end{cases}$$

Now we are ready to define the cluster quality. Let C be a clustering result. $C = \{C_k\}$, $k = 1, 2, ..., K$ and K is the total number of clusters found. C_k is the set of indices of instances that are clustered to formed C_k. The quality of cluster C_k is defined as:

$$CQ_k = \left(\sum_j \left| \frac{\sum_{i \in C_k} w_{ij}}{|C_k|} - \theta \right|^+ \right) \times \big| |C_k| - \lambda \big|^+ \times |C_k|$$

Note that $|x|^+ = \begin{cases} 1, \text{ if } x \geq 0 \\ 0, \text{ otherwise} \end{cases}$. θ is the majority threshold value, and λ is the minimum number of instances that a cluster has to contain to be evaluated.

The above measure requires that to contribute to CQ_k a word must satisfy (a) the majority criterion θ and (b) the minimum number of instances criterion λ. The majority criterion requires a word to appear in at least θ different instances in the same cluster or in other words be representative. The minimum number of instances criterion requires a cluster to contain at least λ instances to be considered as a valid cluster for both evaluation and practical purpose. Both measures are given subjectively based on expertise but can serve as a basis guideline to define the quality of a cluster in the context of this project.

In the end, we use Average Cluster Quality (ACQ) to measure how a clustering method performs over the dataset across all clusters it generates:

$$ACQ = \frac{CQ_k}{K}$$

4.4.3 MinDisconnect

Most existing clustering methods implicitly or explicitly assume compactness and convexity of clusters to be found, whereas this is not always true. The Bag-of-Words being examined this study is one good counter example. As was already mentioned in 4.3 our Bag-of-Words is a very sparse matrix. Also, since a Bag-of-Words is simply a representation of occurrences of words, if each word is a dimension, how instances distribute in this N dimensional space is random, and the relative position of each instance to others does not carry much practical meaning, so the convexity assumption can hardly hold.

Instead, MinDisconnect adopts the idea of connectivity and looks at how well objects in the same cluster are connected to each other. As is to be introduced, the idea of connectivity is how instances are “connected” to each other either directly or indirectly. When this is put in a linguistic context, two sentences (broken down into words) can be considered as close to each other because they share words in common, or there is one other sentence that shares words with each of them individually, and this is where the idea of connectivity and MinDisconnect fits well.

Before we get to the details of MinDisonnect, there are two important concepts need to be explained.

Cluster k-Nearest Neighbor Consistency (kNN) - for any data object in a cluster, its k-nearest neighbors should also be in the same cluster.

Cluster k-Mutual-Nearest Neighbors Consistency (kMN) – If object i is in the set of p nearest neighbors of object j, and object j is in the set of q nearest neighbors of object i, and $k = \max(p, q)$, then we say object i is a k-mutual-nearest neighbor of object j, and vice versa. If a cluster is kMN consistent, for any object in it, its k-mutual-nearest neighbors should also be in this cluster.

MinDisconnect seeks to maintain both kNN consistency and kMN consistency and therefore eliminate disconnectivity within every cluster to be found. The concept of disconnectivity is proposed to incorporate this idea.

Let σ_i denote a set of k-nearest neighbors ob object i, then

$$\sigma_i = \{ j \mid d(i, j) \leq d(i, k) \},$$

where $d(i, j)$ is the distance between object i and object j, and k is the kth nearest neighbor of object i. The disconnectivity $f^{(k)}(P)$ at a given partition $P = \{C_1, C_2, ..., C_g\}$ is defined as a penalty for the violation of both kNN consistency and kMN consistency:

$$f^{(k)}(P) = \sum_{r} \sum_{i \in C_r} \sum_{j \notin C_r} (b_{ij}^{(1)} + b_{ij}^{(2)}) \cdot \frac{1}{d(i, j)}$$

where

$$b_{ij}^{(1)} = \begin{cases} 1 & \text{if } j \in \sigma_i(k) \\ 0 & \text{otherwise} \end{cases}$$

and

$$b_{ij}^{(2)} = \begin{cases} 1 & \text{if } i \in \sigma_j(k) \\ 0 & \text{otherwise} \end{cases}.$$

MinDisconnect minimizes $f^{(k)}(P)$ in a heuristic fashion. For object i and object j, penalty will be yielded if either kNN and kMN consistency is not satisfied. In particular, penalty is doubled if kMN consistency is violated.

The value of enforcing kMN consistency in addition to kNN consistency can be shown in this example. Let there be three instances A, B and C. Assume B is one of the k nearest neighbors of A, and C is one of the k nearest neighbors of B. As is shown below, where arrow indicates "is one of the k nearest neighbors of". What can be said about the similarity between A and C?

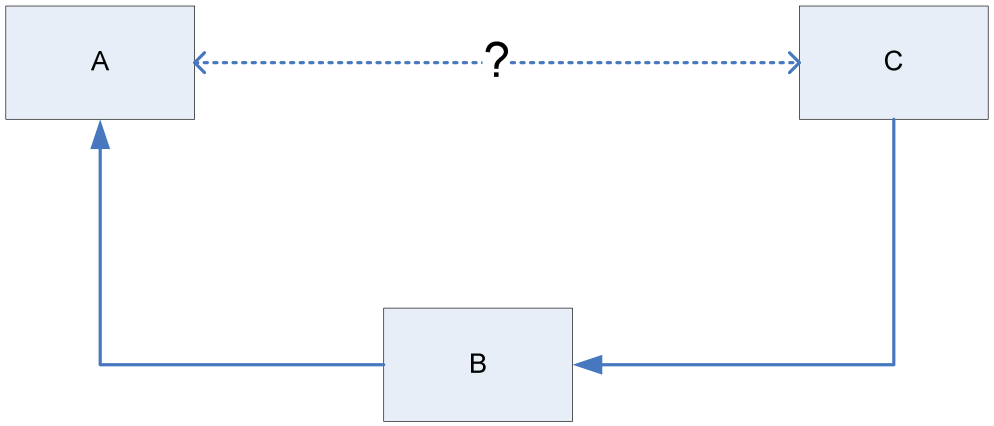

Figure 3 - Object Relations under kNN Consistency Only

The answer is it depends. They can be similar with each other, or they can completely dissimilar with each other. This example shows that kNN consistency does not guarantee we have a group of instances that are all similar with each other.

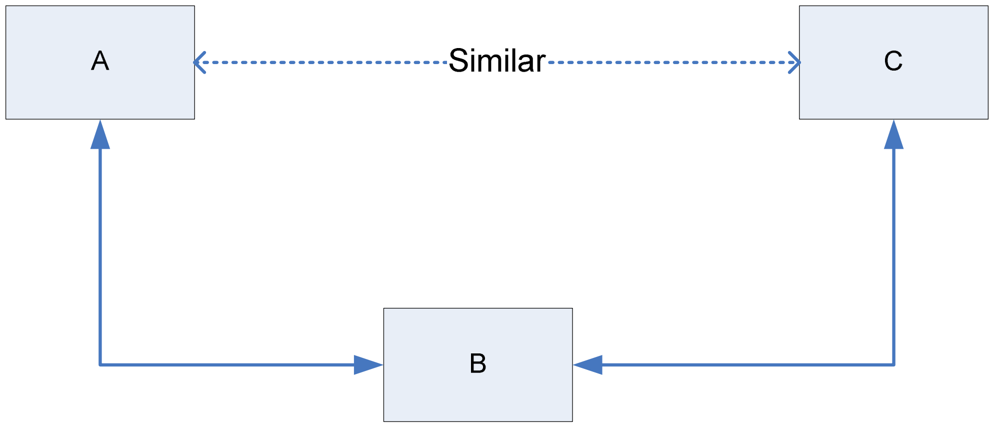

Figure 4 - Object Relations under kMN Consistency

Consider the case when kMN is enforced. Assume A and B are kMN consistent, B and C are kMN consistent, one can easily tells that both A and C are both among the k nearest neighbors of B, they are similar with each other in regard with B and hence “connected”. If all three instances are in one cluster, we know the concept in B will be enhanced.

One other advantage of enforcing kMN consistency in addition to only kNN consistency is the stability of clustering result. If only kNN consistency is required, and there is an object j that is one of the k most similar objects to object i, these two objects should be in the same cluster from object i’s perspective. However, object i may not be one of the k most similar objects to object j, so object i and object j do not have to be in the same cluster from object j’s perspective. Even from the algorithm’s standpoint, either case yields the same amount of penalty. Therefore, the clustering result can be unstable if only kNN consistency is required. Accordingly, patterns to be unveiled from the clustering result become inconsistent.

By enforcing kMN consistency, MinDisconnect is able to output clusters consisted of instances that are similar to each other and we can expect the clustering results to be more stable.

We can think of the above three instances A, B and C as three instances of row variables indicating occurrences of words. Suppose A and C may not “look” very similar to each other. They can be talking about the same thing hence carrying the same pattern, especially if they both “look” similar to one instance, say B, due to some wording style. One can expect the pattern in B can be enhanced if A and C are grouped together with B. With only kNN consistency, this is not guaranteed.

4.4.4 Comparison of MinDisconnect, Single-Linkage and k-Means

The following example shows that MinDisconnect outperforms k-Means and Single-Linkage. In this example we arbitrarily choose θ=0.5 and λ=1, that is, a word must appear in half of total instances in one cluster to be considered as representative, and each cluster must have at least 1 instance in it to be evaluated (due to the small amount of instances).

Consider 7 instances as follows:

$$\begin{aligned} I_1 &= \{1, 1, 0, 1, 1\} \\ I_2 &= \{1, 1, 0, 0, 0\} \\ I_3 &= \{0, 0, 0, 1, 1\} \\ I_4 &= \{0, 0, 1, 1, 1\} \\ I_5 &= \{0, 1, 1, 1, 0\} \\ I_6 &= \{1, 1, 1, 0, 0\} \\ I_7 &= \{0, 1, 1, 0, 0\} \end{aligned}$$

Based on our definition of similarity and our objective, we can tell the following is a good clustering result:

$$C_1 = \{ I_2, I_6 \},\ C_2 = \{ I_3, I_4 \},\ C_3 = \{ I_5, I_7 \},\ C_4 = \{ I_1 \}$$

The cluster quality of each cluster is shown below:

$$CQ_1 = 6,\ CQ_2 = 6,\ CQ_3 = 7,\ CQ_4 = 0$$

And ACQ = 4.75.

The clustering result yielded by MinDisconnect with k=1 is:

$$C_1 = \{ I_1, I_2, I_6 \},\ C_2 = \{ I_3, I_4 \},\ C_3 = \{ I_5, I_7 \}$$

The cluster quality of each cluster is shown below:

$$CQ_1 = 6,\ CQ_2 = 6,\ CQ_3 = 6$$

And ACQ = 6. Note that this is a higher ACQ than the previous one which is considered as a “good” clustering result. MinDisconnect successfully discovers C_2 and C_3, but did not discriminate I_1 from I_2 and I_6.

The clustering result yielded by Single-Linkage method at level 1 is:

$$C_1 = \{ I_2,\ I_6 \},\ C_2 = \{ I_3,\ I_4 \},\ C_3 = \{ I_5,\ I_7 \},\ C_4 = \{ I_1 \}$$

This is the same as our “good” clustering result. However, the result of Single-Linkage method becomes unstable when it enters level 2, because the distance from C_4 is the same for C_1, C_2 and C_3, so we could get different results that are of various quality. We think this is because Single-Linkage method creates clusters solely based on the distance between two instances, without much consideration of the shape or any other properties of the clusters it is looking at.

In the end, we applied k-Means to the above instances. One drawback of partitioning clustering methods like k-Means is the need to define a representative measure (such as the mean, as is in the case of k-Means) of clusters yielded in each iteration. In this example, we define the mean of a cluster as a new instance I', whose elements is consisted of the most frequent value on each position. When there are the same number of 1’s and 0’s on the same position, we choose to set it as 1 to preserve the concept represented by that word. For example, the mean of $C_2 = \{ I_3, I_4 \}$ is $I' = \{0, 0, 1, 1, 1\}$. With this definition and with arbitrarily set k=4 aiming at the good clustering mentioned above, kMeans yields the following clustering result:

$$C_1 = \{ I_1 \},\ C_2 = \{ I_2, I_6, I_7 \},\ C_3 = \{ I_3 \},\ C_4 = \{ I_4, I_5 \}$$

The cluster quality of each cluster is:

$$CQ_1 = 0,\ CQ_2 = 9,\ CQ_3 = 0,\ CQ_4 = 8$$

And ACQ=4.25. Note that this is the lowest ACQ we've got so far. The result yielded by k-means is not making much sense.

MinDisconnect outperformes Single-Linkage and k-Means in this example, though it yields a slightly different clustering result than the "good" one arbitrarily identified. It makes sense since a cluster with only one instance provides very limited information about itself, and should not be considered as a good one in this case.

4.4.5 Determining the Number of Clusters

MinDisconnect may yield singleton clusters – clusters that have only one member instance. These instances do not satisfy the kMN consistency with any other instances and therefore assigned to a separate cluster on its own. Since one instance by itself is not a strong enough evidence of certain pattern, we consider these instances as outliers as well as the singleton clusters they are in.

In our applications, it turns out that MinDisconnect yields many singleton clusters. We decide to group all singleton clusters together as one cluster (indicated as cluster #999 in our application). By doing this, we are actually characterizing one big cluster as “cluster of outliers”. Even though we do not expect strong patterns to be uncovered from the “cluster of outliers”, grouping all outliers together into one significantly reduces the number of class variables involved and is considered a computational benefit. In addition, this “cluster of outliers” can be investigated separately in a case by case fashion to better understand all varieties of exceptions.

After grouping singleton clusters, one may find a distribution very similar to the one shown below:

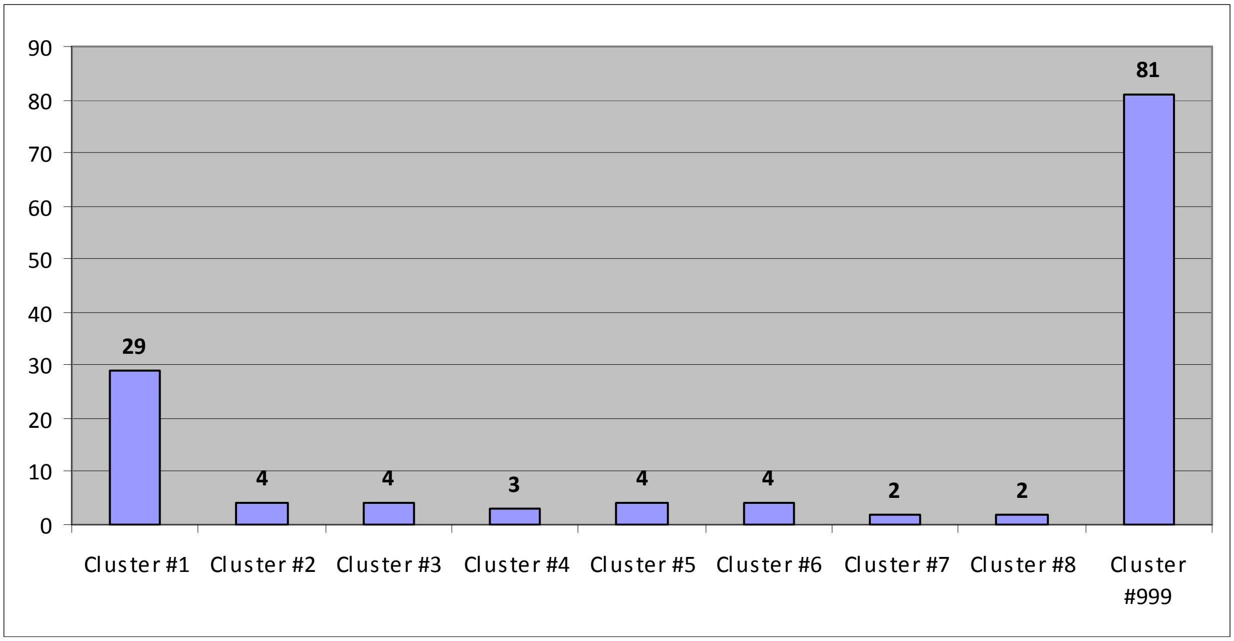

Figure 5 - An Example of Clusters Found by MinDisconnect after Merging Singleton Clusters as Cluster #999

MinDisconnect yields different clusterings with different k values. When k increases, the number of clusters increases at first, and decreases soon after. Among all the different clusterings, one may notice that there are always two clusters that have more instances than others. Most of the cases they are Cluster #1 and Cluster #999. When Cluster #999 is the “cluster of outliers” we created by grouping all singleton clusters together, Cluster #1 are the instances that well fits in the category of the injury type being investigated and is considered as “cluster of normality”. Clusters in between are the ones assigned to the same injury type as others but differentiate themselves from the “cluster of normality”. Since these clusters have multiple instances therefore be able to carry patterns that can be discovered systematically. These clusters are the ones we will focus on. We call them “clusters of interest”.

Since the requirement that instances in one cluster must satisfy kMN consistency remains the same regardless how many clusters are in the result, we choose the one that has the most “clusters of interest”. We expect discover interesting patterns in “clusters of interest” through characterization.

4.5 Attribute Selection with RELIEF

With over 1000 words in our portfolio, we believe not all of them are as useful for one type of injury as others. Filtering out those less relevant to the class variable reduces noises in the dataset and hence helps us focus on only the ones considered important. Furthermore, downsizing the dataset could also relieve the computational intensiveness in the next data mining phases.

We choose an attribute selection method called RELIEF to extract a subset of all the words. These attributes are then believed to be more informational than others when they are used to interpret the injury type being investigated.

RELIEF evaluates the estimate $W(A)$ of attribute A in a similar fashion as the following difference of probabilities:

$$\begin{aligned} W(A) = {} & P(\text{different value of A} \mid \text{nearest instance from different class}) \\ & - P(\text{different value of A} \mid \text{nearest instance from same class}) \end{aligned}$$

The rationale behind the estimate $W(A)$ is that good attribute should carry different values on instances from different classes, and same values on instances from the same class. This rationale is in consistent with our objective of clustering. We now utilize RELIEF to "polish" the results obtained from the clustering phase.

4.6 Cluster Characterization with Apriori

Association Rules Mining is a type of unsupervised learning method that aims at discovering interesting correlation between any attributes in unlabelled data. Such correlation is usually represented as an association rule in the form of $X \Rightarrow Y$, where both antecedent X and consequent Y are sets of attribute-value pairs. This suits our needs to characterize clusters with words represented by binary variables. One can anticipate rules in the form of $X \Rightarrow Y$ where X is a set of words that appear in a major number of instances in cluster Y. An interpretation of the set of words X within the context of the original texts will lead to characterizations of cluster Y.

In addition, association rules are measured in terms of the quality of correlations they uncovered. These measures can be easily applied within the context of this research to evaluate how we the cluster is characterized. Among many such measures, Support, Confidence and Lift are the most widely used ones.

Support, is the probability that an instance contains all items in both the antecedent and consequent parts of the rule, that is, satisfies the rule.

$$Support(X \Rightarrow Y) = P(X \cap Y)$$

Confidence is the probability that an instance satisfies the rule, given that it contains all items in the antecedent part of the rule.

$$Confidence(X \Rightarrow Y) = P(Y \mid X) = \frac{P(X \cap Y)}{P(X)}$$

Lift is the ratio of the confidence to the expected confidence, where the expected confidence is the expected probability of an instance having all the items in the consequent part of the rule.

$$Lift(X \Rightarrow Y) = \frac{Confidence(X \Rightarrow Y)}{P(Y)} = \frac{P(X \cap Y)}{P(X) \cdot P(Y)}$$

Note that Lift evaluates how likely the antecedent occurs given that the consequent occurs. A Lift value being greater than one means it’s more likely, and less likely when its less than one. This is the measure we will be use to evaluate the quality of rules found in this study.

Apriori is used in this study to characterize clusters. Apriori is an important association rules mining method. It uses the prior knowledge of frequent itemsets and continuously searches frequent itemset in a level-wise fashion, where k-itemsets are used to explore (k+1)-itemsets. We expect Apriori to generate rules $X \Rightarrow Y$ where Y is the cluster indicator as the class variable, and X is a set of words at present in the cluster. This way the association rules can be interpreted as if the concept represented by words in X occurs it's likely that cluster Y also occurs. This correlation may reveals some interesting causes of certain injuries within a particular injury type. The likelihood of the occurrence can be specified by the measurements (Support, Confidence and Lift) mentioned above so we only focus on the ones that meet our requirements.

Given that our "clusters of interest" are mostly small in size containing no more than 10 instances, we will only look at the ones that yield very high Lift values. While how to define "interestingness" can be a whole another research topic, without further explanations, we arbitrarily define "interesting" as Lift being 25 or more.

CHAPTER 5 CASE STUDY

In this chapter, we will present some interesting findings from the application of our methodology to narrative texts in MSHA AII 2008 database. We will focus on three different types of injuries, namely Bur or Scald (NATINJ=120), Laceration or Puncture (NATINJ=180), and Crushing (NATINJ=170). Characterizations of clusters are first presented to create some hints for further thoughts, and the original narrative texts are quoted to provide more details. We hope by presenting such results to people with sufficient domain knowledge will inspire more concrete understanding of root causes of different types of injuries, and eventually lead to safety improvement measures.

5.1 Case Study of Burn or Scald Injuries (NATINJ=120)

There are a total of 107 incidents of burn or scald injuries after data preprocessing. When k=3, MinDisconnect discovers 7 “clusters-of-interest”, more than any other k values. RELIEF returns 509 attributes.

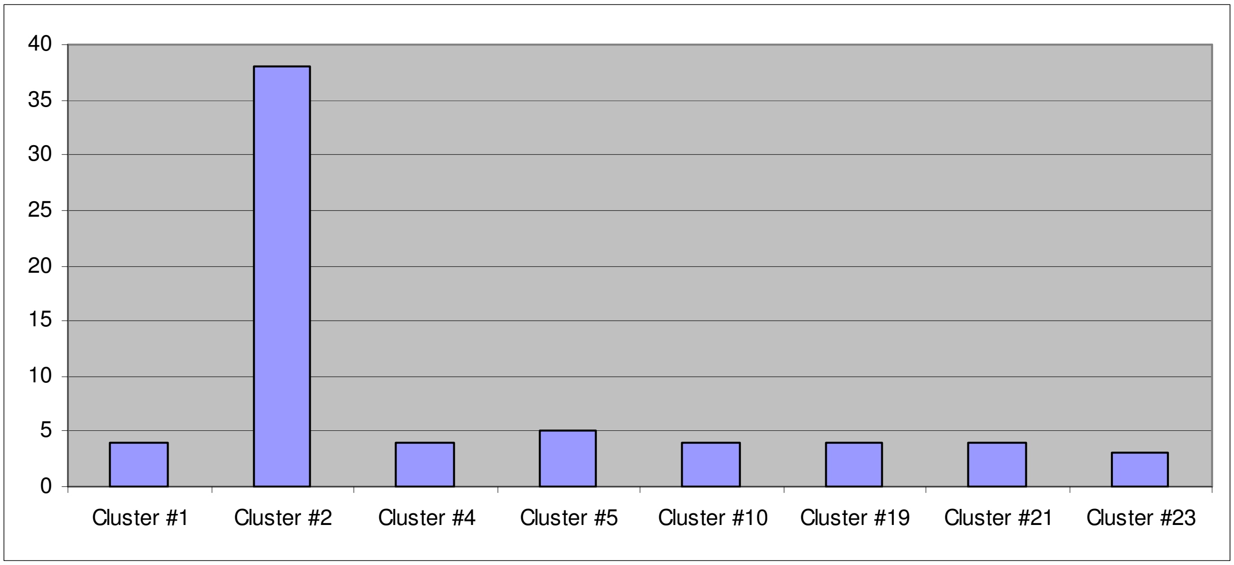

Figure 6 - Histogram of Clusters Found by MinDisconnect for Instances of Burn or Scald Injuries, k=3

The following pattern of the 4 instances in cluster #19 has a relatively high lift value (26.75), and was therefore brought to our attention.

$$\text{“pin”} \Rightarrow \text{Cluster \#19}$$

As a general pattern, “torch” appears in many instances of burn or scald injuries, which is not surprising. The following two examples show how pins are involved in these injuries. Here pins helped trigger the injuries by releasing heats in unexpected ways. In the first incident, pins and bores constitutes a structure that transmits heats in a less straightforward way. In the second incident, a pin hole in the oxygen hose became the direct cause of the injury by releasing heats to bare hands.

“In process of repairing upper hoist cylinder pin & bore, EE was preheating bore with a propane torch. Prior to final bushing placement, grease surrounding the bore sleeve super

heated & blew out of the bore onto the employees hands. EE had just removed his welding gloves to measure the bore temperature with a hand held heat gun.”

“Employee was cutting out a steel donut to weld onto the dump ropes that attach to the dragline bucket. While cutting he noticed the oxygen hose on the torch had a pin hole. He attempted to t aped the hose instead of replacing or fixing the hose. Employee continued to use cutting torch and his gloves became saturated with oxygen gas and a spark set his gloves on fire.”

It seems that there is a lack of instructions on preventive actions in torch jobs when pins and holes are at presence. Inspections of should always include finding pins and holes that are not easily noticeable. Also clear instructions on remedy actions when pins and holes are visually captured should be put in place to avoid risky attempts to proceed with potentially damaged torch equipments.

Another interesting finding looks into the issues from another perspective. This comes from the 4 instances in Cluster #21, which also yields a relatively high lift value (26.75).

“did”, “report” ⇒ Cluster 21

The two words appear in the antecedent do not reveal much about the incidents at the first lance. After reviewing the instances they are involved with, the story behind emerges. All 4 instances in Cluster #21 are incidents that were not reported immediate after the incidents occur. While why they were not reported timely remains unknown to us, we did notice that in the second instance the incident was only reported when the injury on the employee worsens and becomes intolerable.

"Employee did not report incident until 10/15/2008. Employee was working in the field all day and got grease on his boots. Employee used a citrus degreaser to clean his boots. The degreaser also soaked the employee's socks. Employee attempted to dry his boots with a lighter. Employee's boots caught fire and burned his leg."

"Employee alleges that he was welding inside a chute and received a couple small welding burns on his legs. He did not report it. One week later he stood up from a desk and felt a sharp pain in his left knee. The next day he was admitted to the hospital with an infection in his knee. He alleges the infection came from the small welding burns."

Given those instances, we are tempted to believe that there are more incidents still outside of our scope. Disciplinary actions should be taken to prevent delayed report or no report of accidents. This is to help investigating root causes and improve working conditions, also make sure necessary medical treatments are applied on time to prevent more serious health risks.

5.2 Case Study of Cut, Laceration or Puncture Injuries (NATINJ=180)

There are a total of 733 incidents of cut, laceration or puncture injuries after data preprocessing. When k=2, MinDisconnect discovers 87 "clusters-of-interst", more than any other k values. RELIEF returns 1136 attributes.

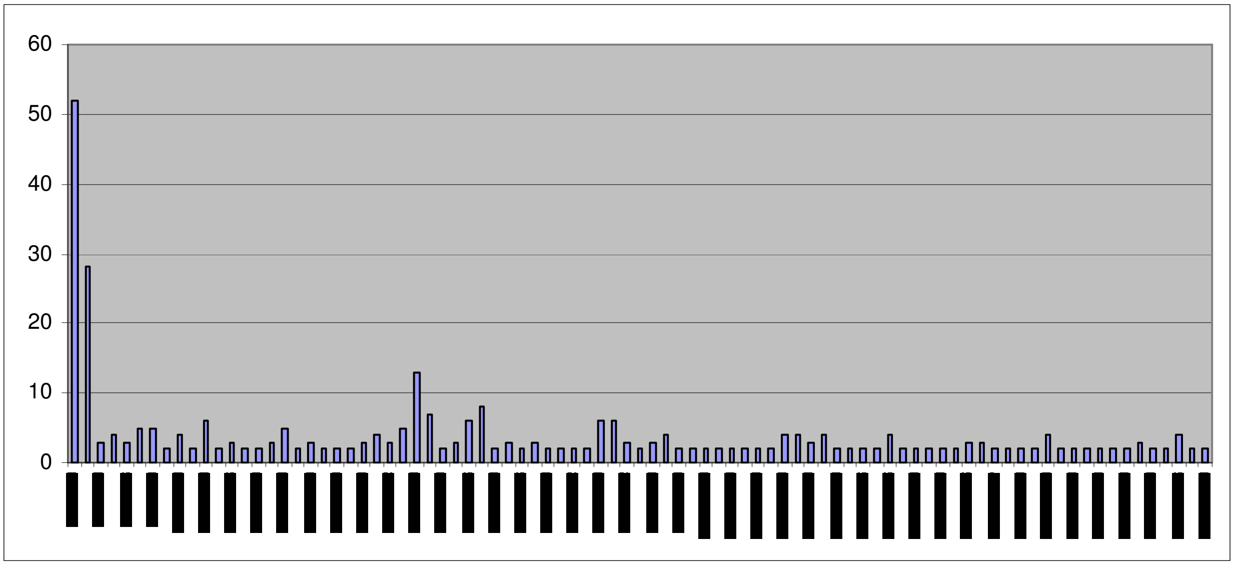

Figure 7 - Histogram of Clusters Found by MinDisconnect for Instances of Cut, Laceration or Puncture Injuries, k=2

A pattern about windshield was discovered by cluster #237. This cluster has 4 instances and yields a very high lift value (183.25).

“windshield” ⇒ Cluster #237

Another pattern pointing to cluster #237 is also found to yield high lift value (162.89). In this case there are 3 instances involved.

“highwall” ⇒ Cluster #237

A third pattern with regard to cluster #237 also yields high lift value (244.33), though covers only 2 instances.

“collapse” ⇒ Cluster #237

All the above 3 patterns seem to be correlated to each other. Looking back to the original text confirmed our guess. They are all about accidents where collapsed walls hit windshields in mining vehicles, broken glasses from windshields then resulted in cut injuries on employees in the vehicle. Two such examples from cluster #237 are shown below.

"Operator was mining wall throughout the day, When about 4:20 pm operator was backing out from wall when wall collapsed. Material fell on front end of loader and hit windshield, causing windshield to break and laceration to operator left and right arm. High wall is about 25 to 30 feet high."

"At 8 pm on 4/10/08 a section of high wall collapsed. The materials covered half of the 992 c loader a rock broke thru the windshield causing a laceration of the scalp and right hand."

As always, safety awareness when driving all kinds of vehicles should never be understated. Necessary protection of windshields could be another opportunity of improvement, especially if the vehicle works in an environment that may have higher chance of hitting unexpected objects. In addition, it should also be investigated that if sufficient support has been given to the high walls being mined. In case of presence of walls, no work activity should be allowed before a thorough wall inspection to ensure they are safe to be around. Also, broken windshields tend to hurt driver's hands, arms and other part of upper body. It may be a good idea to make suggestion to mining vehicle drivers to wear gloves or sleeves.

A pattern about cluster #10 was also brought to our attention. With a high lift value (244.33), all 2 instances in cluster #10 have to do with electricians.

"electrician" ⇒ Cluster #10

Looking into the original descriptions, interesting finding emerged.

"The employee was assisting the electrician in making a SLC cable splice. He was using a utility knife to remove the outer insulation from a phase when he cut the top of his left index finger that required sutures."

"Employee stated he was helping the electrician make a splice in the shuttle car cable and his knife slipped off of the cable cutting his little finger on his left hand. Had to have stitches put in finger."

In both instances, the injured employees were all involved in some electrical work which is outside of their functions. Getting involved with work that one did not get trained properly beforehand is more likely to cause errors. In this case, this cross job function assistance results in injuries. Disciplines should be put in place to disallow people working in functions they are not familiar with or have not been properly trained on, even if it is only an easy favor they can do to perhaps their buddies in other functions. Also both cases occurred when the injured employees were trying to make cable splices. This job procedure or instructions may also need to be reviewed.

5.3 Case Study of Crushing Injuries (NATINJ=170)

There are a total of 133 incidents of crushing injuries after data preprocessing. When k=6, MinDisconnect discovers 7 "clusters-of-interest", more than any other k values. RELIEF returns 626 attributes.

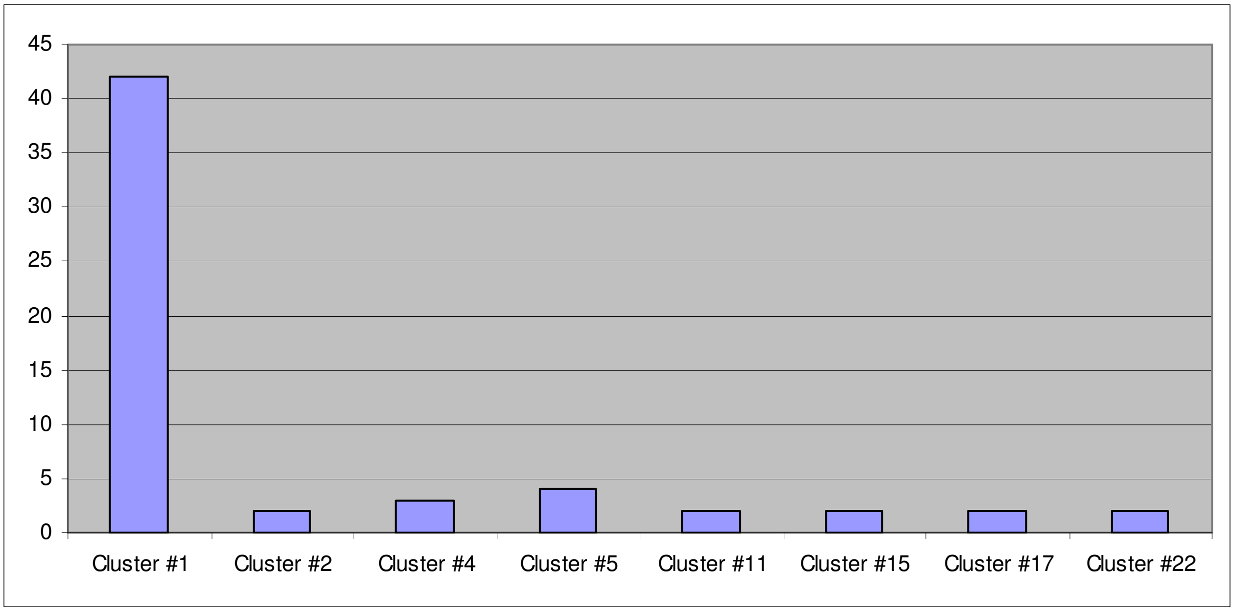

Figure 8 - Histogram of Clusters Found by MinDisconnect for Instances of Crushing Injuries, k=6

Cluster #2 has a relatively high lift value (66.5). Its two instances all have to do with fired engines.

“reverse”, “lunge”, “fire”, “engine” ⇒ Cluster #2

In both cases, fired engines caused vehicles moving in the wrong direction unexpectedly, resulting crushing injuries.

One such example is shown below.

The victims (previous owners) were on my property attempting to start a retired loader. The engine fired causing the loader to lunge in reverse running over the two victims.

It should be investigated whether mechanic examinations have been done regularly to ensure all underground mining machines can function properly and safely. It may also worth investigating whether a specific type or brand of machines failed in both incidents so that more details regarding root causes of similar incidents can be discovered.

Another finding is more of a common sense. Two incidents of crushing injury in Cluster #11 all seem to be caused by slipped sledge hammer. Life value is also relatively high (66.5).

“sledge” ⇒ Cluster #11

One example is shown below:

“Employee was holding a loader tooth while another employee was using a sledge hammer to pound in the pin. The hammer glanced off the pin, striking the employee's left hand.”

In the above example, it’s the problematic method that employees used to pound in the pin caused the injury. Some assisting equipment may be introduced to avoid the involvement of a second employee in pounding jobs. Also the glancing off of the hammer may also be caused by slippery hammer handle. Requiring only hammers with anti-slippery handles may be a good idea as well.

The last but not the least, cluster #22 yielded a relatively high lift value (66.5) and revealed the only two crushing incidents that resulted in restricted duty assignments were all caused by stones.

“stone”, “restrict”, “duty”, “assign” ⇒ Cluster #22

One of the two cases is shown below.

"Employee was adjusting stone on conveyor belt; he pulled the stone back with his right hand and smashed his thumb between the stone he was adjusting and another stone. Restricted duty assigned."

Though it may not be surprising to see crushing injuries in a work environment that has a lot of stones at presence, and restricted duty assignments may indicate the injury's low severity, is it really OK to assign restricted duty without looking into the root causes? What type of restricted duty was assigned? A lot of questions may worth asking before we get to the conclusion that restricted duty assignment is the best way to handle these accidents

CHAPTER 6 CONCLUSIONS

In this study, we proposed a new methodology of applying a set of data mining methods, including clustering, attribute selection and association rules mining, to look for interesting patterns in the narrative texts in MSHA accident/injury database. We investigated 3 types of injuries and uncovered some interesting findings. We believe that the methodology introduced here can be extended to other injury types, other MSHA accident/injury/illness databases of other years, or other similar industrial databases having similar data structure. The main contribution of this study is that we incorporated clustering, attribute selection and association rules mining methods in one general methodology to uncover patterns in narrative texts in a streamline fashion. This generality ensures that our methodology can also be used in many other similar usage scenarios. In particular, we suggested using MinDisconnect, RELIEF and Apriori as the components of our methodology.

However it also opens more opportunities that are yet to be realized.

(1) Most of other attributes in the original MSHA 2008 database have been shown in early preliminary study that they are able to provide meaningful findings when fed to various data mining methods with appropriate data preprocessing. We are also interested to see how to integrate those results with the ones uncovered in this research in the same framework, to yield more in-depth understanding of those incidents.

(2) The evaluation of MinDisonnect can be extended to other scenarios with other data types and can also be placed in comparison with more advanced clustering methods in terms of effectiveness and efficiency.

(3) A more advanced MinDisconnect may be possible if our cluster quality measure can be incorporated in the algorithm systematically to yield results that are already internally optimized towards the cluster quality measure.

Finally, one important limitation we have not overcome in this study is that we may have eliminated many infrequent but possibly useful words. Our main purpose of doing so is to make it possible to feed the large but sparse dataset to various computational resource intensive data mining methods. However this may have reduced useful patterns that can be found. Future work should also try to extend this study to a more powerful and resource affluent computing platform, for example one with 64-bit computing capability and large memory size. We think there will be even more interesting patterns to be found if handling the full dataset is possible.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere thanks to those who helped me with conducting this study, writing this thesis and many others. I thank Dr. Sigurdur Olafsson for his patience, understanding, guidance and support all the way along. I couldn’t have accomplished this without his kindness and considerateness. I would like to thank Dr. Gary Mirka and Dr. Peter Sherman for their time and efforts involved in this thesis and the classes I enjoyed from them. I would also like to thank the faculty and staff of Department of Industrial and Manufacturing Systems of Iowa State University for their instructions, inspirations and all the favors I owe much.

Finally, I would like to thank my fiancé Shan Jin for her encouragement and support which have become a part of my life and will always be.

REFERENCES

Anand, S., Keren, N., Tretter, M. J., Wang, Y., O'Connor, T. M. and Mannan, M. S., 2006. Harnessing Data Mining to Explore Incident Databases. Journal of Hazardous Materials, Vol. 130, pp. 33-41.

Chang, L. and Chen, W., 2005. Data Mining of Tree-based Models to Analyze Freeway Accident. Journal of Safety Research, Vol. 36, pp. 365-375.

Chang, L. and Wang, H., 2006. Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Technique. Accident Analysis and Prevention, Vol. 38, pp. 1019-1027.

Chiu, T., Fang, D., Chen, J., Wang, Y. and Jeris, C., 2001. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 263.

Chong, M. M., Abraham, A. and Paprzycki, M., 2004. Traffic Accident Analysis using Decision Trees and Neural Networks. IADIS International Conference on Applied Computing, Portugal: IADIS Press, Vol. 2, pp. 39-42.

Chong, M. M., Abraham, A. and Paprzycki, M., 2005. Traffic Accident Analysis using Maching Learning Paradigms. Informatica, Vol. 29, pp. 89-98.

Gamberger, D., Lavrac, N. and Krstacic, G., 2003. Active Subgroup Mining: A Case Study in Coronary Heart Disease Risk Group Detection. Artificial Intelligence in Medicine, Vol. 28, pp. 27-57.

Grabmeier, J. and Rudolph, A., 2002. Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery, Vol. 6, pp. 303-360.

Han, J. and Kamber, M., 2006. Data Mining: Concepts and Techniques. Morgan Kaufmann, ed. 2.

Karypis, G., Han, E. and Kumar, V., 1999. Chameleon: Hierarchical Clustering using Dynamic Modeling. Computer, Vol. 32, pp. 68-75.

Kecojevic, V., Koomljenovic, D., Groves, W. and Radomsky, M., 2007. An Analysis of Equipment-related Fatal Accidents in U.S. Mining Operations: 1995-2005. Safety Science, Vol. 45, pp. 864-874.

Lee, J., Olafsson, S., 2011. Data Clustering by Minimizing Disconnectivity. Information Sciences, Vol 181, pp. 732-746.

Li, X. and Olafsson, S., 2005. Discovering Dispatching Rules using Data Mining. Journal of Scheduling, Vol. 8(6), pp. 515-527.

Lindquist, M, Stahl, M., Bate, A., Edwards, I. R. and Meyboom, R. H., 2000. A Retrospective Evaluation of A Data Mining Approach to Aid Finding New Adverse Drug Reaction Signals in the WHO International Database. Drug Safety, Vol. 23(6), pp. 533-542.

MSHA Fact Sheet 95-8 – Historical Data on Mine Disasters in the United States. http://www.msha.gov/MSHAINFO/FactSheets/MSHAFCT8.HTM (last visited Nov 08, 2009)

Nazeri, Z., Lehto, M. and Wu, S., 2007. Hybrid Singular Value Decomposition: A Model of Human Text Classification. In Smith, M. and Salvendy, G. (eds.), Human Interface, Part I, HCII 2007, LNCS 4557, pp. 517-525.

Noorinaerini, A., Lehto, M. R. and Wu, S., 2007. Hybrid Singular Value Decomposition: A Model of Human Text Classification. Lecture Notes in Computer Science, Vol. 4557, pp. 517-525.

Olafsson, S., Li, X. and Wu, S., 2008. Operations Research and Data Mining. European Journal of Operational Research, Vol. 187, pp. 1429-1448.

Parhizi, Sh., Shahrabi, J. and Pariazar, M., 2009. A New Accident Investigation Approach Based on Data Mining Techniques. Journal of Applied Sciences, Vol. 9(4), pp. 731-737.

Sohn, S. Y. and Shin, H., 2001. Pattern Recognition for Road Traffic Accident Severity in Korea. Ergonomics, Vol. 44(1), pp. 107-117.

Thakur, M., Olafsson, S., Lee, J. and Hurburgh, C., 2009. Data Mining for Recognizing Patterns in Foodborne Disease Outbreaks. Journal of Food Engineering, Vol. 94(2), pp. 213-227.

Wang, H., Parrish, A., Smith, R. and Vrbsky, S., 2005. Variable Selection and Ranking for Analyzing Automobile Traffic Accident Data. ACM South-East Regional Conference, pp. 268-273.

Wang, H., Parrish, A., Smith, R. and Vrbsky, S., 2005. Improved Variable and Value Ranking Techniques for Mining Categorical Traffic Accident Data. Expert Systems with Applications, Vol. 29, pp. 795-806.

Wellman, H., Lehto, M., Sorock, G. and Smith G., 2004. Computerized Coding of Injury Narrative Data from the National Health Interview Survey. Accident Analysis and Prevention, Vol. 36, pp. 165-171.