Data Mining Customer-Related Subway Incidents

Hector Ramirez, Peter Cronin, Rujul Inamdar, Shawn Richard, Richard Washington, Layne Yeskey
Seidenberg School of CSIS, Pace University, White Plains, NY 10606, USA
{hr23715n, ly63414, pcronin,}@pace.edu, srichard@mcttelecom.com,
rujulinamdar@gmail.com, richard.washington@nyct.com

Abstract

This study analyzed customer-related subway incident data using data mining. The primary goal of the study is to find correlations between subway-related incidents and violent acts to MTA employees by reviewing all the incidents that occurred within a five hour window prior to the violent act. It allows for the possibility of drawing a correlation between, for example, a subway train delay and an act of violence. Initially, the subway data was analyzed for basic understanding of its structure and to form basic assumptions the data would provide. The assumptions were formulated into structured queries to produce initial results. Data mining tools like WEKA and SQL Server were used to process the data. WEKA was used to create prediction models based on selected attributes, and each of the models were trained on the 2006 data to determine how well they could predict the 2007 data.

1. Introduction

Raw data has the potential of becoming informative in relation to its stipulation of knowledge; however it requires data extraction specification and analysis to ensure the successful conversion of raw data into something valuable and functional. According to Kantardzic, "Data mining is the entire process of applying computer-based methodology including new techniques for knowledge discovery, from data" [1]. Raw data in its shear simplicity is exclusively data that has yet to have been processed, investigated, and learned from. "Data mining is about solving problems by analyzing data already present in databases" [2]. Data mining is not merely the effortless task of picking out observable relations in accordance with common and well known correlations. Instead, data mining involves a much more comprehensive examination of data sets aimed at identifying hidden or undiscovered interrelations. These connections are often found not by utilizing a single method of analysis, but with various methods of analysis. To illustrate, classification, attribute selection, and clustering are just a few of the techniques that can be used in data mining to support the revealing of new and useful information.

This project analyzes two databases provided by the NYC Metropolitan Transit Authority (MTA) that contain customer-related subway incidents for the years of 2006 and 2007. The data of each incident is broken down into specific categories including, incident ID, incident code, delay (measured in minutes), station code, and train line. For this project we take the raw data provided by MTA and pinpoint specific correlations between subway service-related delays and customer reactions which may have resulted in violent acts against a MTA employee.

2. Relevance in the context of the work

For the purpose of our project, the tools used to transform the raw data have yielded useful information. However, the results alone cannot indicate a single cause for the incidents recorded in the data sets, but the results can give some trends that lead us to ask more questions.

From the output, we have discovered clues that lead us to potential issues. Without using the data mining techniques, some basic information can be derived. The consequent information is used to further investigate possible indicators that might lead to causes of the acts recorded in the data. The other processes including SQL inquiries and excel spreadsheet manipulation, narrow the areas to be explored. This further provides a field of view that can be shown from the data set.

What cannot be pulled from the data are the alternative reasons for the violent acts that could not be figured into our conclusions. That being the case, the use of WEKA as a data-mining tool to dig into the data and show possible trends and patterns proved difficult at best. The data sets were not very conducive to the application because of the size requirements of the software.

Additionally, it was requested that a second tool be used in our data mining research. BayesianLab predictive modeling software was selected and utilized to determine if the results from WEKA could be replicated or provide other clues to further our investigations.

At this time, the results are inconclusive and need further analysis. The work at this point in the process consists of collectively reviewing the data produced for underlying patterns and trends not discovered in the first phase of the project.

3. SQL Database

Our customer, who works for the MTA has provided data about train incidences. This data is directly related to customer assaults on MTA employees, and includes all train incidents that happened within a five hour window prior to the specific assault.

Our customer has specifically prepared this data from a database system the MTA is currently using. The first task entailed preparation of the prior mentioned data for manipulation. The data was then imported into a relational database called Microsoft SQL Server. Our customer gave us five text files, four of which are tab delimited files of MTA data. The files received are listed as follows:

- Data field descriptions – A list of 11 column headers for 2006 and 2007 data.
- 2006 incident data – 11 column list of incident data for year 2006.
- 2007 incident data – 11 column list of incident data for year 2007.
- Station list – two column list of station ID and station descriptions.
- Trouble list – two column list of trouble ID and trouble descriptions.

4. ARFF Files

In order to use WEKA, the raw data needed to be converted into ARFF (Attribute-Relation File Format) files which are ASCII text files, which describe a list of instances sharing a set of attributes.

ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.

The **Header** of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

An example header on our project dataset looks like this:

@relation 'tblSubwayData2006edit'
@attribute AssaultID Numeric

@attribute IncidentNo Numeric
@attribute IncidentDate {'2006-01-05 20:55:00'..}
@attribute Duration numeric
@attribute LateTrains numeric
@attribute TerminalCancel numeric
@attribute EnrouteCancel numeric
@attribute StationID numeric
@attribute TroubleCode numeric
@attribute TrainLine {1.0,2.0,3.0,7.0…}

The **Data** of the ARFF file looks like the following:

@data

104144,104148,'2006-01-05
19:53:00',91,1,0,3,295,7107,1.0
104144,104140,'2006-01-05
17:34:00',4,1,1,0,440,4018,3.0
104144,104149,'2006-01-05
17:32:00',8,14,2,1,353,504,2.0

Lines that begin with a % are comments. The **@RELATION**, **@ATTRIBUTE** and **@DATA** declarations are case insensitive

5. WEKA

Our task was to take a set of New York City subway incidents and manipulate the data with a software program called WEKA.

WEKA is a widely used toolkit for machine learning and data mining originally developed at the University of Waikato in New Zealand. It is a large collection of state-of-the-art machine learning algorithms written in Java. WEKA contains tools for classification, regression, clustering, association rules, visualization, and data pre-processing. WEKA is open source software under the GNU GPL. It is easily extensible, which allows researchers to contribute new learning algorithms to WEKA, keeping it up-to-date with the latest developments in the field. As a result, WEKA has become very popular with academic and industrial researchers, and is also widely used for teaching purposes.

The main focus of WEKA is on classifier algorithms. Simply put, a classifier maps a set of data instances onto a finite set of classes. Each data instance is described by its attribute values. Each data instance includes values of the observation attributes. The goal of a classifier learning (or training) process is to derive a classifier from a set of labeled data (i.e. a set of data instances together with their correct labels). The idea is that a classifier learned on a labeled data set can then be used to predict class labels for future (unlabelled) data instances.

The purpose of the task was to determine if there could be a way of looking at the data and deriving some causal reasoning as to the underlying issues behind the acts of violence relating to a subset of indicators.

First, we need to look at the software WEKA and the processes used to understand what the output represents. For our experiment, we used the latest version of the software obtained through a downloaded package that completes the installation process automatically.

5.1 Prediction Modeling

One of the objectives of the projects is to use classifiers available in data mining software and use them for prediction of future incidents.

The prediction modeling process used available datasets to train one another. The training data can be the complete data available or an attribute of the data. The application of algorithms on the data generates a model file. If the prediction modeling is to be obtained on the run then (instead of saving as a model file), the output can be applied directly to the test set, which is the database of dataset where some of the values or all the values of an attribute are missing. The process applied generates the predicted value.

5.2 Selecting an Algorithm

When it comes to choosing a specific algorithm, there are no obvious choices. Different algorithms perform differently depending on characteristics of the data. Some algorithms can be used for both regression and/or classification while others are only for a specific type. Some can handle only nominal attributes while others can handle both nominal and ordinal/continuous variables. Our choice of algorithm was dependent on the attributes. The dataset used in the project comprises of numeric, nominal, and string. Based on the attribute selected algorithms were selected in project. Preprocessing of attributes was also done during this process. Sometimes it is difficult to analyze a data set with non-standard data characteristics, such as combinations of numeric and alphanumeric sets to manipulate in modeling forms. One method to overcome this is additive regression applied to the multivariate data set. Additive regression utilizes mechanisms to smooth out non-standard variables by adding weight to approximate characteristics for certain variable attributes. This format has two distinct benefits, first each of the individual terms is estimated within predetermined parameters by using a smoothing mechanism that prevents the value from applying more weight to the data than in other methods. Second the

derived estimates provide a more refined product than if they were individually calculated and applied to the model. Using the additive regression method will provide outputs based on the mean of the dependant variables from information transformed through non-linear link functions.

Additive regression employs a theory that one segment of the data set is a non-standard variable that has dependence among other variables in the data set. Each component has some dependence on other values within the variable data set. To isolate the significance of the component variable without creating some artificially inflated product algorithm selection potentially decreases the probability of errors as the sample data set increases.

J48 is a decision tree classification algorithm used to create pruned or unpruned decision trees based on the data attributes presented. When the data set has been preprocessed the J48 algorithm creates a schematic of the attribute in a visual form that can be reviewed and analyzed for further processing based on highlighted variables derived from the original data set. This new derived information has essentially been created from the original information based on rules selected in the machine learning portion of the process. It is necessary that the data be in a nominal format for this algorithm to be effective.

IBK or k-nearest neighbor is a supervised learning algorithm where the result of initial manipulation is classified based on majority of k-nearest neighbor category. The purpose of this algorithm is to classify non-numerical data based on variable attributes and filtering selection. The classifiers do not use any specific model to fit variables and is only based on memory of the calculated base. When the data is analyzed a base point is selected and the subsequent data is derived from that base to calculate k-number of points closest to the base. The data set is then calculated based on a majority pattern among the classification of the k-base variables. Any commonalties can be relabeled at occurrence. The IBK algorithm uses neighborhood classification as the prediction value of the new query instance.

The M5 algorithm is a fast correlation based filter algorithm that can be used in discrete and continuous problems. In each manipulation the M5 builds a modeling tree that takes the leaf of the greatest weight from the derived attributes and uses that segment to create a rule that is applied to the data set. It generates a decision list for regression problems using separate-and-conquer methodology. This method produces a model that effectively selects the key features of the data set reducing the model dimensions without

negative effects on the outcome or accuracy of the predicted results. The size of the model trees produced in this manner is significantly smaller than in other processing formats. The result is that a large original data set is reduced under manipulation while maintaining the core information for the derived prediction model output.

5.3 New information: BayesiaLab

BayesiaLab is a software application used for data modeling and mining that has the potential to be more effective from a visual perspective than WEKA. BayesiaLab provides a complete set of learning methods based on Bayesian learning network algorithms and a unique adaptive learning style. The application rapidly assembles the data into associative models that can be changed at any time to reflect different conditions and yield result sets uncovering patterns that might otherwise go unnoticed. Robust tools adapt by applying unsupervised learning methods discovering potential new ideas hidden within the data. Operation can be structured in a way that complex patterns are revealed from otherwise common data sets linking them by associative schemes. Variables can be targeted in a linear fashion revealing the true identity of the underlying patterns then focused to provide a probable outcome matrix from many different angles.

BayesiaLab provides users with an analysis tool box to quickly discover the different connections and strengths of probabilistic relationships. Automatic update of learning mechanisms allows for quick manipulation of existing data or added components into the created network. The application offers a fully functional set of tools, both creative for experimental purposes and traditional for data mining evaluation that could potentially be used in this project by future groups. It can provide a more appealing visual representation of the attributes within the scope of discovery and how they interact. For the project, small amounts of the data were introduced with limited results, as WEKA was the main software utilized.

6. Customer Aggression Model

The underlying cause of the anger behind customer aggression was divided into two separate categories: service related customer aggression and non-service related customer aggression. A breakdown of both categories is shown in Figure 1. The initial hope was to have percentages verified by operating personnel (ex. conductors, train operators). Unfortunately we were not able to get the proper authorization to do so. So the numbers reflect the opinions of a few managers

from support areas (ex. capital construction, system safety area.) of the MTA.

Within the service related grouping, violence triggers were further identified as being brought on by train delays (30%), overcrowding (20%), equipment malfunction (20%), non-courteous employees (10%), poor train conditions (10%), and high train fairs (10%). Within the non-service related grouping, violence triggers were identified as being motivated by homelessness (20%), group wildness (10%), youths (20%), mental illness (10%), intoxication (10%), and personal problems (30%).

Similarly, both classes, whether service or non-service related encouraged train customers to commit acts of violence. With manual and WEKA data mining, relationships correlating to acts of violence were isolated and used to promote aggression prevention strategies. Some deterrence tactics included increased police presence (30%), T.S.S (10%), station C/R (10%), cameras (20%), intercoms (10%), and public announcement systems (20%).

All prior violence prevention strategies would be instilled within the subway system with the purpose of influencing customer readiness to commit violence and in due course suppress overall incidents of violence.
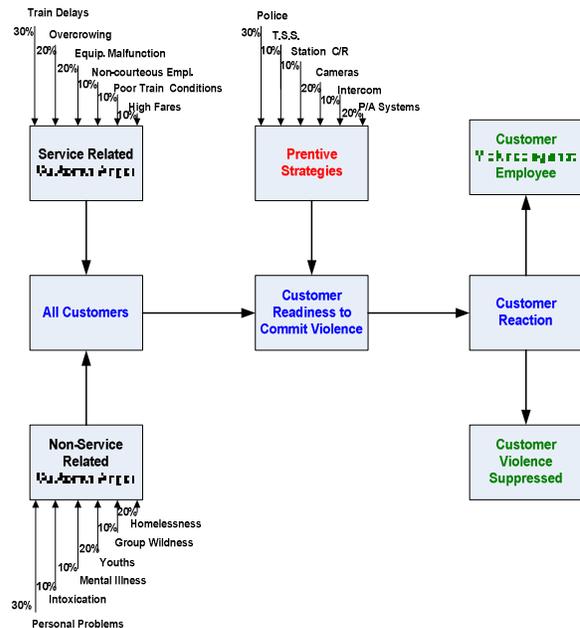

Figure 1. Customer Aggression Model

7. Results

Manual interpretation of data proved to be helpful because the output of information served as a

foundation for the WEKA and SQL techniques of the project.

The MTA is a 24 hour operation that is broken into three shifts. As a result, because of the non-stop operations there is no down-time for such jobs as track work that cause delays in service that sometimes result in an act of violence on an MTA employee. A breakdown of the three shifts was created to examine during what times the most incidents occur. The 2006 shift breakdown is shown in Figure 2 and the 2007 shift breakdown is shown in Figure 3.
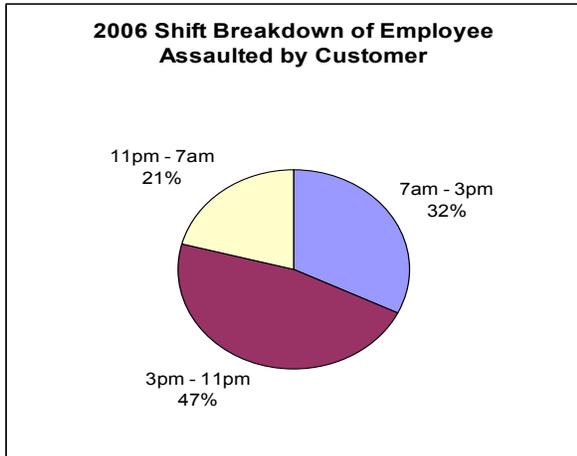

Figure 2. 2006 Shift Breakdown of Employee Assaulted by Customer

In 2006, the highest percentage of incidences occurred during the 3:00pm - 11:00pm (47%) with the least occurring during the 11:00pm – 7:00am (21%).
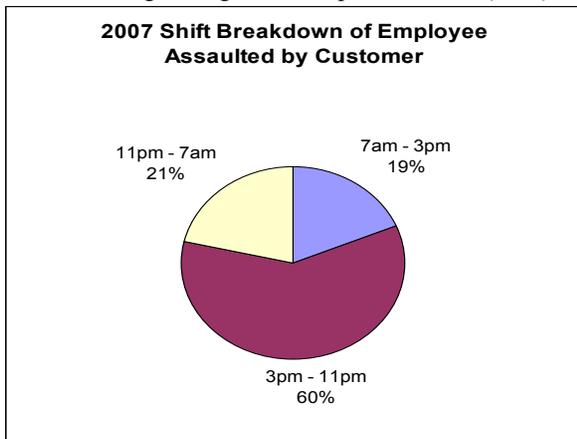

Figure 3. 2007 Shift Breakdown of Employee Assaulted by Customer

In 2007, the highest percentage of incidences occurred during the 3:00pm - 11:00pm (60%) with the least occurring during the 11:00pm – 7:00am (21%).

Based on the statistics, a correlation can be made that during the 3:00pm – 11:00pm the highest volume of passengers are riding the subways and there are more subways running to accommodate the volume. As a result, if they are too many trains running on the same track delays will occur because of the time between trains which can result in customers becoming frustrated and taking it out on an MTA employee

Figure 4 and 5 will show the correlation or relevance of the predicted attributes by WEKA to the attributes available through the customer.

Figure 4. The graph on the left is predicted by WEKA processing and the graph on the right is the actual 2007 data.

The above result in Figure 4 was obtained using additive regression prediction modeling. The training set used was 2006 data. The test set used was 2007 data. The above graph which predicted the EnrouteCancel attribute signifies that during year 2007 208 incidents took place where one train was cancelled and 48 incidents led to two train cancellations

The graph with EnrouteCancel as the attribute is the actual data available and it shows that during the year 2007, 197 incidents took place which led to one train getting cancelled and 42 incidents took place which led to two train cancellations
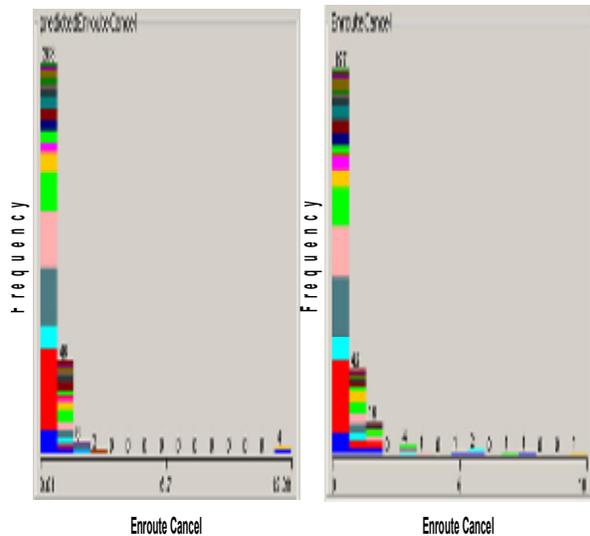
The correlation achieved using additive regression was 0.81 or 81%

Figure 5. The graph on the left is predicted by WEKA processing and the graph on the right is the actual January 2007 data.

Figure 5 demonstrates the number of incidents predicted by WEKA during January 2007 and the actual number of incidents that took place during the month of January 2007

As illustrated in Figure 5, 6 incidents were predicted and reported on the A Line and 3 incidents were reported on 6 line accurately.

Other results that were retrieved as a result of WEKA processing (Figure 5) are:
- 16 incidents took place during the month of January.
- Out of which, 6 incidents took place on line A.
- Out of which, most occurred as a result of employee assaulted by customers.

7.1 Basic Assumptions of Data

The main files provided to us included ten fields of data. Each one of these fields could be extracted to make basic assumptions. The two main pieces of information which most everything could be correlated to are the assault Id and incident number. This information could be considered a composite key in our relational database and would be included in every assumption we make.

When looking at the main data files (tblSubwayData2006 and tblSubwayData2007) there are ten separate pieces of data fields to use. After reviewing the data that we should make basic assumptions on, the first two were glaringly obvious. The two validation tables (tblStationList and tblTroubleList) seemed like a good place to start.

Possible assumptions would include:
1. Number of assaults and incidents per subway station.
2. Most frequent trouble codes.
3. Length of delays per station and trouble code.
4. Average length of delay per incident per trouble code.

Some other assumptions would include:
1. Number of assaults and incidents per train line.
2. Number of incidents that lead to an assault by train line and station.
3. Total and average number of assaults and non-assault incidents per year.
4. Total number of incidents by time frames (Each hour of day).

7.2 Results based on SQL Server data

Here are descriptions of queries that produced results on the subway data:

Trouble Code – List all trouble codes with occurrence count and average duration of train delays for year 2007 and having more than one occurrence.

Train Line – List all train lines with occurrence count and average duration of train delays for the trouble code of "Person Holding Doors" and year 2007.

Stations – List all stations with incident occurrence count for trouble code of "Employee Assaulted by Customer." and year 2007 and having more than one occurrence.

Here are more results based on similar queries from SQL Server:
- 343 incidents lead to 96 assaults in 2006
- 266 incidents lead to 70 assaults in 2007

Figure 6. Most Common Incidents other then Assaults

- Track work delays (19%) and customers holding train doors (10%) are the most common incident other than actual assaults. (Figure 6)

As shown in Figure 7:
- Train line #2 had the most incidents (45) in 2007.
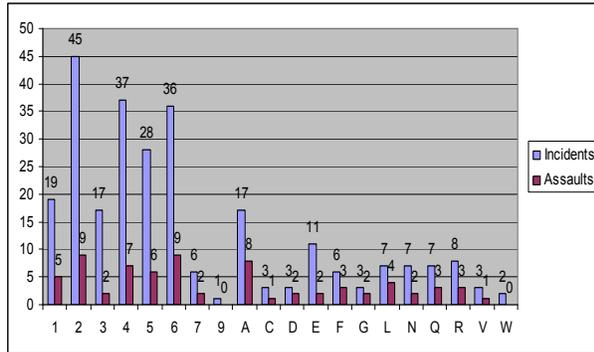- Train line #2 and #6 had the most assaults (9) in 2007.

Figure 7. 2007 Incidents and Assaults by Train Line

- Grand Central Station had the most incidents (8) happen in 2007.
- 125th Street at Lexington had the most assaults (4) happen in 2007.
- 125th Street had the highest ratio of incidents (6) to assaults (4) in 2007.
- Train line #2 had the most track work delays (15) in 2007 which resulted in an average of 201 minutes of delays.
- Train line #5 had the most incidents (8) of customers holding the train doors which lead to an average of 10 minutes in delays.
- Jackson Ave Station had the most incidents (4) of customers holding the train doors which lead to an average of 16 minutes in delays.
- In 2006, 61% of all assaults happen in the PM.
- In 2007, 71% of all assaults happen in the PM.

After looking at the number of incidents of violence per month in the year 2007, we noticed that the month of October had a higher occurrence rate than any other month. To investigate this finding we looked at three specific data categories that we feel could cause a high rate of assaults on employees. We first looked at the volume of riders by month and found that on average October had the highest volume. Next, we checked the total number of non violent incidents by month and found that October had almost double the amount of incidents than the second most months. During our research, we found that the particular incident of track work gangs was the most common incident amongst the data. So, we looked at this incident by month and

found that October was second among all months. Our conclusion is that there is a lot of traffic in the month of October which has lead to a high rate of assaults on the subway system by both high volume of customers and high levels of track work delays. A breakdown of 2007 monthly volume is shown in Figure 8 and a breakdown of 2007 assaults, incidents and track work is shown in Figure 9
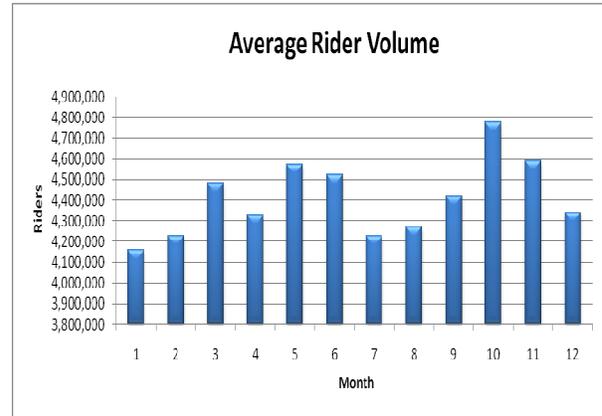

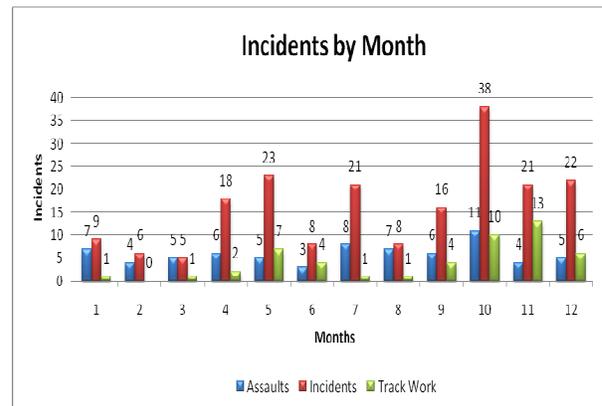Figure 8. 2007 Monthly Subway Rider Volume by Train Line

Figure 9. 2007 Monthly Assaults, Incidents and Track Work by Train Line

8. Summary and Conclusions

Throughout the process of this research project, we conducted several data mining experiments on a SQL database with the purpose of pinpointing any violence related problems involving train customers and employees. We were specifically looking for service linked train tribulations which may have resulted in customer violence.

Data from the 2006 and 2007 time period was examined manually and with WEKA software. Data extraction criteria were restricted to possible correlations between the number and types of incidents

over a time (periods of five hours), locations, and trains.

Our intentions for this project were to establish possible cause and effect relationships between service related incidents and customer related violence to allow for preventative measures to be taken. Once these violence relationships are identified, not only can improvements be made to help ensure employee and customer safety on trains, but to further decrease employee downtime in transportation service and decreased legal litigation by customers and/or employees.

To help create a possible violence prediction chart, the results captured manually and produce by the WEKA software were used to identify the potential pathways of customer aggression. The customer aggression model helped to identify the initial incident, feasible explanations causing the incident (service or non-service related), and preventative strategies to deter said incidents, and whether the incident led to a violent act or if the incident was ultimately suppressed.

9. Implications and Recommendations

Since this project is the first of its kind, numerous algorithms and methodologies were utilized throughout the project progression to gain further incite on the what's and how's of violence indicators. The data examined through mining techniques spanned the years of 2006 and 2007 subway incidents of violence. However, in the end, data sets were examined mainly by manual sifting and the data mining WEKA software in relation to a team created SQL database. This user friendly software produced various types of models and charts to display algorithm specific data which will supply the foundation of further study into customer and service related violence.

We suggest that future research is geared towards applying other diverse types of data mining software similar to WEKA. It would be of great interest to validate current WEKA results with other programs producing similar results or more in-depth results. Since limitations were specifically set for this project, specific incidents related to violence against subway employees were studied. However, further investigations can be made to make predictions of other factors that result in violence producing incidents in general. Generally, information provided by mining software will then be used to help reduce violent acts on trains and to help improve violence prevention and handling procedures.

10. References

[1] Kantardzic, Mehmed, Data Mining: Concepts, Models, Methods & Algorithms, 1nd Edition, Wiley-IEEE Press, New York, 2003, p.24.

[2] Witten, I. H. and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufman Publishers, San Francisco, 2005, p.5.