Data Mining of Historical Emergency Data to Improve Regional Response

by

Ruiji Qi

A Project Report Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

Supervised by Dr. Carol Romanowski

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences Rochester Institute of Technology Rochester, New York

December 2016
Abstract

Data Mining of Historical Emergency Data to Improve Regional Response

Ruiji Qi

Supervising Professor: Dr. Carol Romanowski

In this project, we attempt to study how historical emergency data plays an important role for region emergency management and how to improve regional response based on pre-identified incident footprints.

We would be exploring different data cleaning and preparation techniques, along with several association rules and data mining algorithms to determine the relationships between regional emergence events and their associate attributes.
## Contents

Abstract ........................................................................................................... ii

1 Introduction .................................................................................................... 1

2 Ethical Considerations .................................................................................. 4

3 Implementation .............................................................................................. 5
   3.1 Data Collection ......................................................................................... 5
   3.2 Data cleaning and Preparation ................................................................. 5
      3.2.1 Handling missing values ................................................................. 6
      3.2.2 Remove useless attributes .............................................................. 6
      3.2.3 Data preparation .............................................................................. 6
      3.2.4 Data Storage .................................................................................... 7
   3.3 Identification of fire events, jointly with domain experts ....................... 7
   3.4 Attribute selection, jointly with domain experts ..................................... 7
   3.5 Incident footprints identification ............................................................. 7
   3.6 Data mining algorithms implementation ............................................... 8

4 Analysis .......................................................................................................... 9
   4.1 Monthly Trend ......................................................................................... 9
   4.2 Hourly Distribution ................................................................................. 9
   4.3 Types of Fire Events ............................................................................... 10
   4.4 False Alarm Cases for Fire type AMTCA .............................................. 12
   4.5 Algorithms ............................................................................................. 14
      4.5.1 Zero Rule ....................................................................................... 14
      4.5.2 One-R Rule .................................................................................... 14
      4.5.3 Decision Tree Algorithm ............................................................... 14
      4.5.4 Nave Bayes ...................................................................................... 15
      4.5.5 Random Forest algorithm .............................................................. 15
4.6 Battalion workload distribution ........................................ 16
4.7 Association with the day of the week and weather .................. 17

5 Conclusions ................................................................. 20
5.1 Current Status .......................................................... 20
5.2 Future Work ............................................................ 20

Bibliography ................................................................. 21
List of Tables
List of Figures

1.1 The role of historical data in emergency management decision support [3] . 3
3.1 Sample data collection after cleaning . . . . . . . . . . . . . . . . . . . . . 8
4.1 Fire event monthly trend . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
4.2 Data visualization of fire event hourly distribution . . . . . . . . . . . . . 11
4.3 Fire event hourly distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 12
4.4 Types of fire events . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
4.5 Battalion distribution by using WEKA . . . . . . . . . . . . . . . . . . . . 15
4.6 Fire event jurisdiction prediction (jurisdiction vs. battalion) . . . . . . . 16
4.7 Battalion resources allocation distribution . . . . . . . . . . . . . . . . . . 17
4.8 Battalion fire event workload distribution . . . . . . . . . . . . . . . . . . 18
4.9 Sample cases in Year 2008 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
4.10 Sample cases in Year 2009 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
4.11 Sample cases in Year 2010 . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
Chapter 1

Introduction

The research conducted by Romanowski et al. (2014) discussed the importance of historical data and locally specific data, which have been used to build the foundation for an emergency management decision support system (DSS) [3]. The proposed DSS fuses real-time monitored data from the event with different types of scenarios based on historical data and predefined incident footprints. Then, the system provides recommendations to the central emergency operation center on how to respond to the event. To our knowledge, this work is the first to use locally specific data to build an emergency management decision support system (DSS).

The provided dataset has been used extensively as test case for research and testing purpose. Based on our in-depth study of the previous research, we can generally classify the previous work on this dataset into three major areas. The first area involves improving data cleaning and preparation techniques for the raw dataset. The second broad area involves determining the underlying relationships between major disaster events with time, season, local and more. The third and probably most worked on so far is how to provide better regional support and response by using the historical data.

The main goal of this project is to generate more thorough incident footprints and improve the existing emergency management decision support system (DSS), utilizing the historical data (2008-2010) along with their associated meta-data.

The data set used in this project was collected for Monroe County, NY 911 calls (2005-2012). After preparing and cleaning the data, the data now contains 8,404,363 instances and 100,025,206 pieces of continuation data for those instances. Each instance containing
This paper is organized in the following manner. In Chapter 2, we discuss the major ethical concerns that are related to this project. Chapter 3 discusses about the implementation details of the project, which includes the analysis of the dataset. In Chapter 4, we discuss the analysis and results of our project. In the last section, we present our conclusions followed by possible future work in this area.
Figure 1.1: The role of historical data in emergency management decision support [3]
Chapter 2

Ethical Considerations

There are several ethical considerations regard to this project. More specifically, the issues are mainly related to the data.

As mentioned before, this dataset was provided by Monroe county 911 call center. Hence, all records are one hundred percent confidential. The dataset contains numerous amount of personal information, such as history of diseases and locations. It is extremely to make sure all data are being protected all time.

Another concern is the accuracy of the data. We noticed there were many null values and miss entered data in the raw database. For those instances, we are not certain about the problem. Therefore, we could not trust the data.
Chapter 3

Implementation

For this project, we intend to study the provided dataset in depth and test multiple data mining algorithms for jurisdiction prediction. It is inevitable that there was no previous research that specifically focused on year phase between 2008 to 2010. There are many unknown challenges hidden. Upon having conducted intense research on the previous research related to this dataset, we are now aware of the problems we may run into.

This section entails our efforts in terms of the techniques and algorithms that we used for this project, how they performed under the given circumstance, and how they can possibly be improved.

3.1 Data Collection

The data set used in this project was collected for Monroe County, NY 911 calls (2005-2012). Because of the great work that has been done by professors and previous students, the data is cleaned, prepared and stored in Server 2008 SQL Management Studio (locates in CIMS room. The data now contains 8,404,363 instances and 100,025,206 pieces of continuation data for those instances. Each instance containing 52 fields and each continuation containing 14 fields [2] [1].

3.2 Data cleaning and Preparation

Data cleaning and preparation is one of the most important tasks in this project. The first step of our business was to extract our data from Server 2008 SQL Management Studio.
Since the limitation of memory space on the server, we must pre process the data and try to throw out any attributes that might not contribute to our research. Selecting only those features that will help in the following stages will help to ensure a high accuracy for our results.

Another important step that we had to carry out in this phase was to handle unclean data, which includes missing values, outliers and wrong data. Preparing the data for future use involve surveying, modeling, checking for sample bias and choosing the appropriate technique for cleaning the data.

3.2.1 Handling missing values

The missing values in the data were a result of data miss entries by the monitoring and reaction center. Due to the ethical issue and to ensure the accuracy of our research, we deleted features that had more 90 percent of its original data missing.

3.2.2 Remove useless attributes

Because of the "curse of dimensionality" and the limitation of hardware, we had to determine whether each attribute is useful for our research. This step is a very critical to our project since the provided dataset contained too many useless attributes and features. Our data mining results would be only as good as the features that are selected at this stage. Therefore, it is important to throw out useless features at early stage.

3.2.3 Data preparation

Data preparation helps to determine the underlying relationship in a broader view. Since one of our major interests was to see how the number of fire events distribute throughout the year. The attribute "JulianDay" was categorized into 12 months, which helped us tremendously in the later stage. Another major data preparation task was to reformat all time attribute and calculate the event duration by focusing on the attributes "CallRecipt" and "EventClosed". This step helped us to determine true positive cases versus false alarms.
3.2.4 Data Storage

After preparing and cleaning the data, the data was converted into CSV file. The data was read and parse using python. The newly preprocessed data was then stored in Server 2012 (which also locates in CIMS room).

3.3 Identification of fire events, jointly with domain experts

At the beginning, the assumption was that we could determine fire events by multiple ways. The first one is to look at the event id for each instance. However, we noticed that many events were mislabeled with wrong event id number. The second method is to focus on jurisdiction, in terms of only focus on the events that were handled by fire department. Interesting, we found that more 75 percent emergency events that were handled by fire departments are medical related. Only about 15 percent of the events are fire related. The last method is to look at final event type that were determined after the events closed. After intensive training and testing, we found the last method of looking at the final event type gives us the best result of all.

3.4 Attribute selection, jointly with domain experts

This step was critical to our research. It is impossible to select attributes without any domain knowledge. Therefore, for this stage, we carefully look at each attribute. Compared all attributes with existing research, and limited to several important ones as listed below.

3.5 Incident footprints identification

The most important task for this project was to identify incident footprints for fire events that happened in Monroe county between year 2008 to 2010. In order to generate more thorough results, we looked at the data from different perspectives, which include monthly
Figure 3.1: Sample data collection after cleaning

<table>
<thead>
<tr>
<th>Battalion</th>
<th>Julian Day</th>
<th>Call Receipt</th>
<th>Call Dispatched</th>
<th>Event Closed</th>
<th>Dispatch OfLocation</th>
<th>Final Type</th>
<th>Incident Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0:10</td>
<td>0:10</td>
<td>0:42</td>
<td>X</td>
<td>WIREA</td>
<td>2255 E RIDGE RD IRO</td>
</tr>
<tr>
<td>B</td>
<td>15</td>
<td>1:00</td>
<td>1:00</td>
<td>NULL</td>
<td>Y</td>
<td>STRUCT</td>
<td>96/490</td>
</tr>
<tr>
<td>C</td>
<td>18</td>
<td>0:00</td>
<td>0:00</td>
<td>0:30</td>
<td>Z</td>
<td>VFIRE</td>
<td>ST/MAIN ST ERO</td>
</tr>
<tr>
<td>D</td>
<td>21</td>
<td>0:00</td>
<td>0:00</td>
<td>0:01</td>
<td>O</td>
<td>FIRE</td>
<td>3300 DEWEY AV GRE</td>
</tr>
<tr>
<td>E</td>
<td>22</td>
<td>0:00</td>
<td>0:00</td>
<td>0:01</td>
<td>P</td>
<td>AMTCA</td>
<td>RT 98 AT RT 490 PER</td>
</tr>
<tr>
<td>F</td>
<td>28</td>
<td>0:00</td>
<td>0:00</td>
<td>0:10</td>
<td>Q</td>
<td>BOXB</td>
<td>MAIN/PRINCE ST ROC</td>
</tr>
</tbody>
</table>

trend, hourly trends, fire event types distribute, jurisdiction distribution and more. The through incident footprints will be discussed in the next section.

### 3.6 Data mining algorithms implementation

To predict the jurisdiction and resource allocation of major fire events that are in Monroe county, several data mining algorithms were tested on our dataset. The details and results will be explained in depth in the following section.
Chapter 4

Analysis

4.1 Monthly Trend

At the beginning of our research, we were curious of the relationship between fire events and season. However, we found we were not certain about how to determine the season based on the month. Since we often see snow in March and April in Monroe county, and these two months are usually being considered as "Spring". Therefore, we decided to look at the monthly trend of all fire events that happened during year 2008 to 2010 in Monroe county [3].

Based on the result, we noticed that the trend reaches its peak at November and December. But surprisingly, the number of fire events of January is not as many as the number of events that happened in December. Even though, the temperature and weather are similar for these months.

Another interesting finding is the fact the February has the least amount of fire events happened in Monroe county. Although we considered the fact that it has a few less days compare to other months. The number of cases that happened in February is still significantly less that the other months.

4.2 Hourly Distribution

The next perspective for looking at our dataset is to see the distribution of fire events throughout a day. We categorized 24 hours into three major phases. The first one is day, which includes all events that happened during regular business hours (08:00AM to
04:00PM). The next time phase is evening, which contains all events that happened after people get home from work and before we all went to bed (04:00PM – 12:00 AM). The last one is at night, after most of us fall asleep (12:00AM – 08:00 AM). Based on the results, we found that more than 80 percent of fire events happened during Day and Evening, which indicates that most of fire events happen while we are awake and working around. Hence, we concluded that majority causes of fire events are highly associated with human mistakes.

### 4.3 Types of Fire Events

After determine the relationship between fire events with time and hours, we wanted to look at fire events itself in more depth. Based on the final type of fire event and type description that were provided in dataset, we categorized all fire events into seven major fire types.
Figure 4.2: Data visualization of fire event hourly distribution

- **FIRE.** This type contains all events that were labeled as FIRE, FIREA, FIREB. It refers to fire events that were being reported to 911 call center as emergency fire response. It is often related to house or apartment fire that were associated with human mistakes.

- **AMTCA.** This fire type refers to automatic fire alarm. As we noticed on the graph, more than 30 percent of fire events belongs to this type. However, there are many false alarm cases should be considered, which will be discussed in more details in the following section.

- **STRUCT.** This category represents the fire or smoke that are coming from structures.

- **BOXA.** This fire type is associated with fire box alarm. Like AMTCA, there is a possibility that this fire type also contains many false alarm cases.
• VFIRE. VFIRE often refers to vehicle fires that are caused by car accidents or overheating problems.

• WIREA. Many fire events that happened outside of household were being reported as WIRE fire type A.

• WIREB. Like WIREA, WIREB is also associated with event that happened outside, but more focus on burning trees, burning objects and transformers.

4.4 False Alarm Cases for Fire type AMTCA

By looking at the above graph, we noticed that among of fire events, the type AMTCA (automatic fire alarm) was rated the second highest on the graph. There are 30 percent of
fire events belong to this category. Due to the fact, this type of fire events generate notification by automatic fire alarm. We were curious about the fact how many cases were false alarm cases. Therefore, we focused on the data of AMTCA fire events, and calculated the duration of each event by looking at the time of ”callReceipt” and ”eventClosed”. We decided to consider the cases that need more than 30 minutes of help to be true positive cases. Surprising, we found that among all AMTCA fire events that happened during year 2008 to 2010, only 20.62 percent are true positive cases, which means that almost 80 percent of this type of fire events are false alarm. In fact, for all these fire events, the 911 monitoring and reaction center send out multiple resources and helps. Therefore, in one way we could consider this to be a waste of resource. In the future study, we should focus on how to
differentiate the true positive cases with false alarm cases, and allocate our resources more logically and accurate.

4.5 Algorithms

4.5.1 Zero Rule

Zero Rule is a relatively simple association rule that does not look at any attribute while classification the target variable. During the process of classification, zero rule would only focus on the class with highest counts.

In our project, we find the 37.87 percent was classified under CTYF. If we apply this rule to our dataset, it would always assign Monroe county fire events to CTYF jurisdiction and is a bad classifier.

4.5.2 One-R Rule

The next method we used is the One-R Rule. Based on the results shown in the figure below, the jurisdiction could be assigned based on the location of dispatch. This classifier method produced a high percentage of correctly classified instances (99.01 percent). However, the location of dispatch is one attribute that are highly associate with jurisdiction and the local of fire departments. Hence this classifier could not be used for the prediction model.

4.5.3 Decision Tree Algorithm

Based on the results, we noticed the classification accuracy of this model is also 99.01 percent. Like the previous method One-R rule, the tree determined the location of the dispatch to be the most important node. The tree chooses this attribute to split on first. We also noticed that both pre and post pruning have been used on the tree hence the classification accuracy is so high. Overall, we conclude that under this circumstance, the decision tree algorithm generated the same results as One-R rule.
4.5.4 Nave Bayes

The advantages of Naive Bayes algorithm is easy to implement. The runtime of this algorithm is relatively short. However, we noticed that the correctly classified instances accuracy is 96.97 percent, which is relatively too high to believe. One possible explanation is that the Nave Bayes algorithm performs does not perform very well with categorical data.

4.5.5 Random Forest algorithm

The final algorithm is the Random Forest algorithm. This algorithm construct decision trees based on random attributes. We noticed that this algorithm also gives us a high percentage (98.67 percent) classification accuracy.

Another suggestion proposed by Dr. Romanowski is to categorize all jurisdiction into major battalions.

Based on the fire districts and fire station locations map that was provided by Monroe County official website, we carefully categorize all jurisdictions, we created 8 major battalions as follow:

![Figure 4.5: Battalion distribution by using WEKA](image)

Other than CTYF (city fire department) and AIRF (airport fire department), other fire departments were categorized into 5 major battalions per their locations. We also found that there are 124 cases of fire events that were being handled by other fire departments.
that are not belong to Monroe county. One assumption was that fire departments that are in adjacent county had helped us out many times. After categorical our entire dataset, the same data mining algorithms were applied to the data. We noticed that the classification accuracy had significantly changed. Although, the accuracy percentage is not high as it was before. However, we found these results to be more believable.

![Bar chart showing fire event jurisdiction prediction](image)

Figure 4.6: Fire event jurisdiction prediction (jurisdiction vs. battalion)

### 4.6 Battalion workload distribution

As mentioned in previous subsection, we divided all fire departments that are in Monroe county into 8 major battalions. Based on the statistics, CTYF (Rochester city fire department) has the most amount of stations and resources among all the other battalions. Hence, they are responsible for majority of fire events that happen in Monroe county.

Like previous graph, the distribution of fire events is similar with the distribution of resources. Other than CTYF (Rochester city fire department), Battalion 2 and Battalion 3
also have lots of resources. Hence, they oversee more fire events compare to Battalion 1, Battalion 4 and Battalion 5.

4.7 Association with the day of the week and weather

One of our major task was to see the relationship between fire event and its meta-data. We often consider meta-data as the data of data. In our case, many attributes were already provided to us, such as data, time, and location. Other than the results we found so far, we were extremely curious about the relationships between fire events with the weather and the day. Unfortunately, we were not able to obtain a detailed historical weather database that could be merged with our current dataset. Therefore, we looked at each individually and selected 7 significant cases for each year. All these dates happened to have significantly large amounts of fire events recorded in our data.
Figure 4.8: Battalion fire event workload distribution

![Graph showing battalion fire event workload distribution for 2008-2010 in Monroe County, NY. The graph displays the number of fire events across different battalions, with CTYF having the highest at 37,661, followed by B1, B2, B3, B4, B5, AIRF, and OTHF with 124, 602, and 6256, respectively.]

<table>
<thead>
<tr>
<th>Julian Day</th>
<th>Actual Dates</th>
<th>Day</th>
<th>Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>2008 - 01 - 09</td>
<td>Wednesday</td>
<td>Rain, Snow</td>
</tr>
<tr>
<td>363</td>
<td>2008 - 12 - 29</td>
<td>Monday</td>
<td>Rain, Snow</td>
</tr>
<tr>
<td>259</td>
<td>2008 - 09 - 15</td>
<td>Monday</td>
<td>Mostly Cloudy</td>
</tr>
<tr>
<td>205</td>
<td>2008 - 07 - 23</td>
<td>Wednesday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>30</td>
<td>2008 - 01 - 30</td>
<td>Wednesday</td>
<td>Rain, Snow</td>
</tr>
<tr>
<td>117</td>
<td>2008 - 04 - 26</td>
<td>Saturday</td>
<td>Rain, Fog</td>
</tr>
<tr>
<td>158</td>
<td>2008 - 06 - 06</td>
<td>Friday</td>
<td>Rain, Thunderstorm</td>
</tr>
</tbody>
</table>

Figure 4.9: Sample cases in Year 2008
<table>
<thead>
<tr>
<th>Julian Day</th>
<th>Actual Dates</th>
<th>Day</th>
<th>Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>2009 – 03 - 11</td>
<td>Wednesday</td>
<td>Rain, Snow</td>
</tr>
<tr>
<td>114</td>
<td>2009 - 04 - 24</td>
<td>Friday</td>
<td>Thunderstorm</td>
</tr>
<tr>
<td>129</td>
<td>2009 - 05 - 9</td>
<td>Saturday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>148</td>
<td>2009 - 05 - 28</td>
<td>Thursday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>202</td>
<td>2009 - 07 - 21</td>
<td>Tuesday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>221</td>
<td>2009 - 08 - 09</td>
<td>Sunday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>343</td>
<td>2009 – 12 - 09</td>
<td>Wednesday</td>
<td>Rain, Snow</td>
</tr>
</tbody>
</table>

Figure 4.10: Sample cases in Year 2009

<table>
<thead>
<tr>
<th>Julian Day</th>
<th>Actual Dates</th>
<th>Day</th>
<th>Weather</th>
</tr>
</thead>
<tbody>
<tr>
<td>93</td>
<td>2010 – 04 - 04</td>
<td>Sunday</td>
<td>Cloudy</td>
</tr>
<tr>
<td>128</td>
<td>2010 - 05 - 08</td>
<td>Saturday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>146</td>
<td>2010 - 05 - 26</td>
<td>Wednesday</td>
<td>Cloudy</td>
</tr>
<tr>
<td>188</td>
<td>2010 - 07 - 07</td>
<td>Wednesday</td>
<td>Cloudy, Haze</td>
</tr>
<tr>
<td>189</td>
<td>2010 - 07 - 08</td>
<td>Thursday</td>
<td>Cloudy</td>
</tr>
<tr>
<td>202</td>
<td>2010 - 07 - 21</td>
<td>Wednesday</td>
<td>Rain, Thunderstorm</td>
</tr>
<tr>
<td>204</td>
<td>2010 – 07 - 23</td>
<td>Friday</td>
<td>Rain, Thunderstorm</td>
</tr>
</tbody>
</table>

Figure 4.11: Sample cases in Year 2010
Chapter 5

Conclusions

In conclusion, our results indicate that historical data and through incident footprints can provide a general framework for a regional emergency decision support system (DSS).

5.1 Current Status

Not only did we obtain several interesting observations of fire events that happened during year 2008 to 2010 at Monroe County, NY, but we also developed a deep understanding of how historical data plays an important role while building regional emergency management decision support system. This project allowed us to know about the resources we have and methods needed to achieve the goal made us more organized in our work-flow. Finally, a key takeaway is learning about the failure modes of certain techniques and to figure out a work around for such cases.

5.2 Future Work

Future work should consider adding meta-data of regional weather system and locally-specific disaster events. This will provide us an opportunity to generate a more complete picture of the damage caused by major fire events.
Bibliography

