ANALYSIS OF ROAD ACCIDENTS IN NYC AND THEIR CONTRIBUTING FACTORS

Rahul Ramesh Kumar

New York University
Tandon School of Engineering
Computer Science
New York
rrk310@nyu.edu

Girish Ganesh Prabhu

New York University
Tandon School of Engineering
Computer Science
New York
ggp234@nyu.edu

ABSTRACT

One of the key challenges in Accident Analysis is to correctly identify accident causing factors and various research has been done on finding these factors. However, the main challenge that an analytic faces is to overcome the heterogenous nature of data that yields accurate results without making any assumptions about the size of the data set.

In this paper we have developed an analytic that aims to extract valuable information about road accidents and how this could be related to external factors such as street conditions, vehicle types, season, time of day etc. We do this by applying a K-mode clustering algorithm followed by Association rule mining using Spark MLlib thereby achieving as much accuracy as possible from the data set.

The results provided here can be used by NYC Road safety and NYS Department of Transportation to open up new lanes or may be widen existing roads to reduce the number of accidents .

Index Terms— Big Data; Accident Analysis; Spark; Association Rule Mining

1. INTRODUCTION

There are a lot of accidents that occur in NYC leading to injuries and occasional fatalities. Each accident has two or more parties involved and either or all parties are punished based on the traffic disobeyment laws. However, most of officials reporting the accidents do not scrutinize the area or external factors such as road conditions, lighting, season, time of day etc. Since these factors are not considered, over a period of time the number of accidents keep increasing.

There is a motivation to understand and analyze these accidents that takes into account the various external factors that may play a role in these accidents. The paper describes an analytic that aims to extract valuable information about road accidents and how it could be related to factors such as street conditions, vehicle types, seasons etc.

We use 2 data sources provided by NYC OpenData - NYPD Motor Vehicle Collisions and NYC OpenData - Street Assesment Rating. The first data source contains the vehicle collisions in NYC and the other contains the assessment ratings of the various streets in NYC. Here, focus is on obtaining the right cause factors for accident using big data analysis tools so that the analytic code can scale well as time progresses.

Section 2 describes the motivation for the analytic. Section 3 describes the design which consists of Data profiling , ETL (Extract, Transform, Load), K-mode clustering and Association Rule Mining. Section 4 describes the results for clustering and association rule mining with the performance analysis. Section 5 describes literature that has worked on a similar problem. Section 6 describes future work that we wish to work on in our analytic.

2. MOTIVATION

There are a number of accidents in NYC whose reasons are not properly accounted for. In this analytic, we aim to provide an accurate description for the reason that attributed to the accidents so that it can be used further to improve the road safety conditions and bring safety awareness to people.

We are motivated by some of extensive research performed in [2],[3],[4] and [5]. We also have all accident reports from New York City OpenData. Hence we believe that we can apply the research findings on the accident reports from NYC to correctly identify relationship between different factors that attribute to an accident and come with a analysis using big data tools that are available in the open source world.

3. DESIGN

Figure 1 depicts the design achitecture of the analytic along with the Big data tools employed. Once the data profiling is done , both the tables accident and street are inserted as hive tables using beeline.

For our analytic we intend to use Apache Spark since the data needs to be iteratively processed to determine the clus-

ters. In order to start the clustering, the data is loaded from the accident and street table using Spark-sql and we iteratively compute the clusters. Once the clusters are identified , we run an Association rule mining algorithm to determine the rules that are highly prevelant in the data set.

3.1. Data Profiling

Before designing the analytic, we profile the data to find out the range of values, attribute type and the size of the data. We used Map Reduce Framework to profile each data source where each map task emits a key-value pair where the key corresponds to the attribute name and value corresponds to 1. The reduce task reduced each key and provided the count of the occurence of each unique attribute in the dataset.

We used the profiling information to get a better understanding of the data and prune data that may otherwise be of no help in the further stages of the analytic.

3.2. ETL (Extract, Transform, Load)

The data obtained after profiling and cleaning is loaded into Hive as two seperate Tables named accident and street. Once the data is loaded, we use pyspark to query each of the data as RDD objects from Hive. Here a map task is performed between the street table and accident table to get the rating of the street where the accident occurred. The map function call is a user defined function that does coordinate matching of the geo-coordinates within a threshold radius to determine if the accident happened in that street and extract its street rating. The mapped result created as a separte RDD will now contain the accident details and the street rating which is used in the further analysis. The ETL design is depicted in Figure 1. The attributes of the mapped result is depicted in Table 1.

Table 1. Mapped Accident Attributes

Attribute Name	Values
Season	Spring, Summer, Fall, Winter
Time	Morning, Afternoon, Evening, Latenight
Borough	Manhattan, Brooklyn, Queens
Geopoint	Latitude, Longitude
Street Name	Streets in NYC
Accident Factor	(Driver inattention, Fatigued/drowsy, ...)
Vehicle Type	(Passenger Vehicle, SUV, ...)
Street Rating	(Good, Fair, Poor)

3.3. K-mode Clustering

For the analysis below we believe a clustering algorithm can provide better groups of clusters that have highly related data. The objects in a one cluster are more similar to each other and objects in different clusters are dissimilar to each other. Due to the heterogenous nature of data, we cannot use K-means algorithm since the entire dataset consists of categorical data. We use K-mode as the clustering algorithm as per [2] since K-mode works well with mixture of categorical and numerical data.

We found that K-mode algorithm implementation is unavailable in Spark MLlib and Apache Mahout. Hence we resorted to use the K-mode implementation in Python provided by [6]. We ran the K-mode clustering algorithm on the mapped RDD for six clusters. The choice of six as the number of clusters is due to the good spread of clustered data provided and as per [2]. The results of the clustering are discussed in the Results section.

3.4. Association Rule Mining

K-mode algorithm provides us with six clusters where each cluster contains similar type of accidents. In order to further fine tune the results, we would like to capture the frequent patterns that occur in each individual cluster. These patterns can then be used to generate association rules with very strong confidence thereby providing us crucial insights into the data that may not be visible by looking into individual accidents.

We used the implementation available in Spark MLlib to perform the frequent pattern growth and association rule mining. The results of applying the association rule mining algorithm show very distinct insights into the dataset and are summarized in the results section. Figure 2 depicts the K-mode clustering and Association rule Mining performed on the entire dataset.

4. RESULTS

Table 2 and Table 3 depicts the results after running adhoc queries on Hive-Spark. We run these queries to get an idea about the most common factors and vehicle types that are involved in an accident.

After running the cluster analysis we can use this data to obtain certain contradictions or challenge the factors mentioned for the accidents.

Table 2. Top 5 accident factors mentioned in the dataset

Accident Factor
Driver inattention/distraction
Other Vehicular
Failure to yield right-of-way
Fatigued/drowsy
Lost consciousness

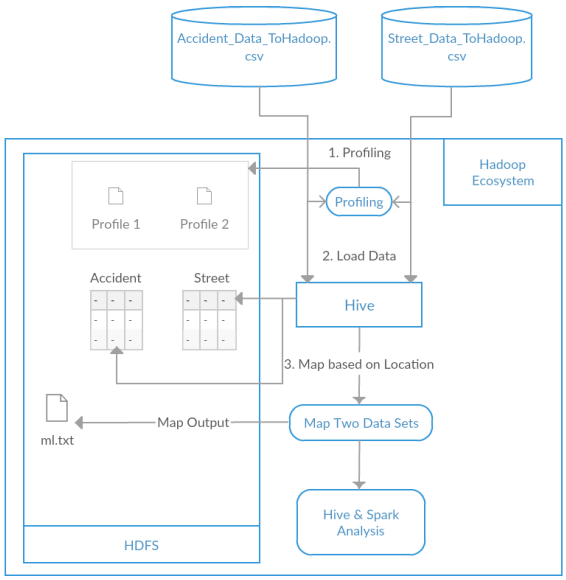

Figure 1. ETL Design Diagram

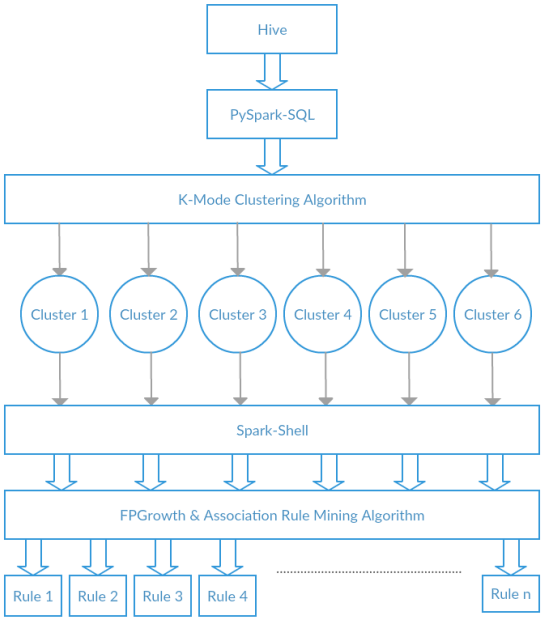

Figure 2. Analytic Design Diagram

Table 3. Top 5 vehicles involved in accidents

Vehicle Type
Passenger vehicles
Sport utility / station wagon
Taxi
Van
Other

4.1. Clustering Results

4.1.1. Cluster 1 Analysis

This cluster involves data containing high rate of accidents in Manhattan followed by Brooklyn. It also shows accidents that happened in streets with an overall good rating during winter season. As per the results, most of the accidents are caused by driver inattention/distraction especially during the evening and morning. The second factor leading to accidents is failure to yield right of way.

4.1.2. Cluster 2 Analysis

This cluster covers majority of accidents that happen in Queens followed by Manhattan. The overall condition for streets are good. The accidents mostly involve Sport utility vehicles which occur in the morning time during the Fall season. It also reports many late night accidents that happen in Queens.

4.1.3. Cluster 3 Analysis

This cluster involves high rate of accidents in Brooklyn followed by Manhattan for fair quality roads during the fall season. Most of accidents here occur in the afternoon due to driver inattention/distraction involving passenger vehicles.

4.1.4. Cluster 4 Analysis

This cluster contains high rate of accidents in Brooklyn during the summer season. Most of the accidents occur in the morning times due to driver fatigue and loss of consciousness involving passenger vehicles and Sport utility vehicles.

4.1.5. Cluster 5 Analysis

This cluster contains high rate of accidents in Queens during the Spring season. The condition of the roads is generally good with most of the accidents occuring due to driver inattention/distraction followed by failure to yield right of way involving passenger vehicles during afternoon hours.

4.1.6. Cluster 6 Analysis

This cluster involves high rate of accidents in streets with fair condition roads in Queens borough followed by Manhattan during summer season. Most of the accidents occur during evening hours due to other vehicluar reasons followed by driver inattention/distraction. The vehicles types that are largely involved in accidents are Sport utility vehicles and passenger vechicles.

4.2. Association Rules

Each of the cluster output is passed to a FPGrowth algorithm in Spark MLlib that generates frequently occuring sets of attributes. These frequent subsets are then passed to the Association Rule Mining algorithm in Spark MLlib to generate association rules that exist in the data.

4.2.1. Association Rules for Cluster 1

The data belonging to cluster 1 contains high number of accidents belonging to Manhattan. Rules such as [evening] => [good] and [passenger vehicle,winter] => [good] were obtained. This indicates that during the evenings of the Winter season in Manhattan, more accidents involving passenger vehicles occured. The cluster contains high number of accidents with driver distraction as an accident factor. Also, association rules suggest that the quality of roads are usually very good in Manhattan which means that driver distraction was a more contributing reason to accidents rather than the quality of the roads.

4.2.2. Association Rules for Cluster 2

Rules obtained from cluster 2 analysis were very different to those obtained by cluster 1. Cluster 2 contained rules such as [fall] => [good],[morning] => [good], [sport utility / station wagon] => [good] and [queens] => [good]. As per the clustering results, we can see that cluster 2 contains accidents belonging mostly to Queens and Manhattan. These accidents have mostly occurred in the morning time during the Fall season. We can also infer from the rules that more accidents involving Sport utility vehicles or station wagons occurred. Majority of these accidents occured in the morning, despite the quality of roads being good.

4.2.3. Association Rules for Cluster 3

Applying Association Rule Mining to cluster 3, we have obtained results with confidence closer to 75% compared to cluster 1 and 2 which gave us results with confidence closer to 90%. From the clustering results, we can see that cluster 3 contains results involving roads which have fair conditioned roads. Rules such as [passenger vehicle] => [fair], [brooklyn] => [fair], [afternoon] => [fair], [driver inattention/distraction] => [fair] and [fall] => [fair] were obtained. We can infer from this that during the Fall season in Brooklyn, more passenger vehicles were involved in accident due to driver inattention and distraction. Seeing that most of the results occured in the afternoon, we can say that the average quality of road was definitely a factor that played a role in the accident. Due to fair quality of roads, many of the drivers may have lost their concentration therefore leading to accidents.

4.2.4. Association Rules for Cluster 4

Cluster 4 has provided many Association Rules. Rules such as [summer,brooklyn] => [good] and [morning,brooklyn] => [good] have confidence higher than 90%. Also, rules such as [summer,passenger vehicle] => [good] and [morning,passenger vehicle] => [good] were obtained with similar confidence. We can combine these rules to infer that accidents in Brooklyn occured in the summer mainly involved passenger vehicles. Most of these accidents occured in the morning. Thus, we can definitely say that Brooklyn is prone to accidents invlolving passenger vehicles in the summer season. Also, the quality of the roads did not help with the prevention of accidents as we had a high number of accidents in the streets with good quality roads.

4.2.5. Association Rules for Cluster 5

Cluster 5 provides us rules with confidence closer to 94%. The cluster mostly contains accidents in Queens borough. We obtain some rules such as [afternoon,queens] => [good], [driver inattention/distraction,passenger vehicle] => [good], [spring] => [good], [driver inattention/distraction,queens] => [good]. From this, we can infer that the spring season has had an impact on drivers leading to many accidents in Queens involving passenger vehicles. The quality of roads being good suggest that the Spring climate was very troubling to drivers which can be due to the rain and traffic conditions. Any slight amount of distraction coupled with the Spring climate has led to many accidents in Queens.

4.2.6. Association Rules for Cluster 6

In cluster 6, we obtain rules with confidence closer to 70%. Cluster 6 provides us with rules such as [passenger vehicle] => [fair], [evening] => [fair] and [queens] => [fair]. We can infer most of the accidents involved Passenger vehicles in Queens during the evening hours. Also, the condition of the street being fair indicates that quality of roads in Queens does play a vital role in the cause of Accidents impacting Passenger Vehicles.

4.3. Performance

Here we analyze the performance of our analytic and also mention the version of big data tools used to perform the execution of our analytic.

We mapped *60527* Accident records with *70104* Street assesment records to perform the geolocation mapping. We limited the output to *38343* mapped Accident records due to huge computational time taken in NYU HPC. This task of mapping was executed on PySpark using Hive(engine=MR) and took around *5* hours.

The K-mode clustering algorithm ran on these *38342* mapped Accident records with a total run time of *11* minutes on Cloudera Quickstart VM. The FPGrowth and Association Mining Spark MLlib job for all clusters took *42* seconds to execute on Cloudera Quickstart VM.

Table 4. Big Data Tools Version Information

Name	Version
Spark	1.6.0
Scala	2.10.5
Java	1.7.0
Python	2.7
Hive	1.1.0-cdh5.8.0
Hadoop	2.6.0-cdh5.8.0

5. RELATED WORK

Lots of research has been done in identifying the cause of accidents from a large dataset. We present some findings from literature that has worked on similar problems and their approach.

[2] aims to identify the main factors associated with a road and traffic accident. Since the data is heterogeneous in nature, meaning it is not normalized in a way an analytic would require for processing, a segmentation is performed on the data to overcome the heterogeneity of the data.

A K-mode clustering algorithm is used to group the data into different homogenous segments. Once the clusters are ready, an associative rule mining is performed on the individual clusters using Apriori algorithm.[3] similarly aims to identify factors of traffic injury severity by using Classification and regression tree method and the output here is called Variable Importance Measurement that predicts what was the major cause of accident and the confidence score associated to it .

[4] illustrates spatio-temporal visualization results through two case studies in multiple road segments, and the impact of weather on crash types. Using 2 case studies, a table was created summarizing the results based on commuting hours, tourist attractions, schools and shopping malls, accident-prone segments and expressways. Through spatio-temporal visualizations, they found that weather is likely to have more impact on accidents for bridges, next for branchroads, but doesnt have obvious influence on arterial roads.

[5] involves usage of logistic regression to estimate the influence of accident factors on accident severity. Through logistic regression, they were able to determine the coefficients that make the observed outcome i.e. a fatal or non-fatal accident using the maximum-likelihood technique.The backward selection process identified location (LOC) and accident cause (CAUS) as being significantly related to accident severity. Odds were provided through the statistic in a matrix format for better interpretation, to establish priorities for programs to reduce serious accidents.

6. FUTURE WORK

The street assesment data also contains the width information of each street and we would like to use that as an external factor to determine if it played a role in cause of accidents. The width can be useful to determine whether narrower roads were a contributing factor to an accident as compared to wider roads.

Currently our implementation of location mapping in Spark uses a heuristic approach to map any accident with in a threshold radius which can generate certain false positives. Hence we wish to improvise the mapping by using geo-apis that incorporate filtering using Manhattan distance or Poly Line distance.

7. CONCLUSION

We have demonstrated how an analytic can be implemented using heterogeneous data. The results have been carefully assessed and we have performed various iterations to provide the final results. The tools used to develop the analytic are Hadoop, Hive, Spark, MLlib which are highly recognized in the Big data open source world.

8. ACKNOWLEDGEMENTS

We thank Professor McIntosh for introducing us to the various big data tools and also providing us the various usecases and feedback for our analytic during the implementation.

We would also like to thank NYU HPC for providing the Dumbo cluster for performing computation of our Analytic.

9. REFERENCES

1. https://en.wikipedia.org/wiki/MapReduce
2. A data mining framework to analyze road accident data, 2015

3. A data mining approach to identify key factors of traffic injury severity, 2010
4. Big Data Analytics and Visualization with Spatio-Temporal Correlations for Traffic Accidents
5. Using logistic regression to estimate the influence of accident factors on accident severity
6. Extensions to the k-modes algorithm for clustering large data sets with categorical values, Data Mining and Knowledge Discovery
7. Spark: Cluster Computing with Working Sets
8. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing
9. Apache Hive. http://hadoop.apache.org/hive.