Traffic Relations

Extended Abstract*

David Nyberg
CU Boulder
Boulder, Colorado
dany3289@colorado.edu

Eric Ha
CU Boulder
Boulder, Colorado
erha5113@colorado.edu

Nicholas Sugarman
CU Boulder
Boulder, Colorado
shadowelecman@hotmail.com

ABSTRACT

This paper is in ACM SIG format and covers an overview of our data mining project that will be going on during Spring 2018 at University of Colorado, Boulder.

KEYWORDS

ACM, LaTeX, Traffic, CU Boulder

ACM Reference Format:
David Nyberg, Eric Ha, and Nicholas Sugarman. 2018. Traffic Relations: Extended Abstract. In *Proceedings of Traffic Relations Data Mining 2018 (BOULDER, 18).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Through the access of a large traffic data set, we seek to find knowledge involving a correlation between seasonal differences and its application towards traffic volume. An example of a finding would be that there is an increase in traffic during winter seasons. We hope to apply this knowledge towards building more efficient roadways that mitigate traffic that occurs during seasons of high traffic. Interesting knowledge that we also seek to find would be to find a correlation between traffic volume and the type of road in which it occurs. To extend our previous example, we may find that there is a larger traffic volume at rural roads during the winter compared to other types of roads. Another interesting attribute to look at along with the previous could be lanes of traffic, depending on how many lanes a road may have, does this impact how many people drive on this road, and when do they do it? All these are important questions to any commuter or city planner or even construction companies trying to repair and build roads.

2 PREVIOUS WORK

Traffic issues are an ongoing problem in the United States and lots of previous work has been done to research and try to fix traffic around the world.

https://www.zmescience.com/research/technology/google-maps-traffic-05443/

This article talks about Google maps and how it uses predictive technology in tandem with traffic data to predict when and where traffic jams will occur. It can predict future traffic jams by referencing past traffic data and comparing it with current conditions to see if there is a significant overlap. Similar work has been done to determine what days are more dangerous to drive on then others

https://www.bactrack.com/blogs/expert-center/35042821-the-most-dangerous-times-on-the-road.

This data is used to measure when drunk drivers are most likely to be on the road. https://www.nature.com/articles/srep37300

Research has been done which investigates what weather conditions are most likely to put people in danger and where. The study also tries to take factors such as socioeconomic status into account, which is a little beyond the scope of our project but still fascinating.

http://kalw.org/post/driving-apps-waze-are-creating-new-traffic-problems

This article writes about how traffic phone applications such as waze which is similar to Google maps showing users the fastest route to a location are actually creating more traffic issues. This is closely related to our project as we are going to look at when people use local roads depending on the time of year and how many lanes they have, this article mentions how apps like waze are sending people onto these small rural roads in order to find a 'faster' path but it is actually slowing down traffic according to their studies.

2.1 Proposed Work

Our data will require cleaning to ensure there are no null values or incomplete data which could affect the data results. We will also have to reduce the data to remove any variables that we will not want to study or that will not aid our investigation. This is important to do so we can speed up our analysis times as we have over 7 million data points. We will also have to transform the data to ensure that it is normalized and that the traffic volumes for different areas are on an even scale. We then need to create interesting models that show a correlation between some of our attributes that we can report as an interesting find. This will be the majority of the work. To achieve this, we will use association rule mining to find patterns in the data. This will hopefully prove, using frequent item-sets, that there is a correlation between the variables we wish to prove might be related. When we discover which variables are related, most likely using the apriori algorithm, we will be able to determine which variables will be worth investigating, and which do not appear to be related.

*The full version of the author's guide is available as `acmart.pdf` document

BOULDER, 18, March 2018, Boulder, Colorado USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

Once we have determined which individual variables are related, we will investigate further to see if a group of variables might be related to high or low traffic at a particular time. Finally, we would want to visualize our data in a meaningful way with Python or another open source tool. Once we can visualize the data, we will probably want to account for situations with abnormaly high or low traffic (sports events, concerts, city evacuations, etc.) and deal with them accordingly. Our study is similar to Google Maps in that we wish to create a predictive set of data that will help us determine when to drive on certain roads, but our data is more specific than GoogleâĂŹs broad strokes studies.

2.2 Data Set

https://www.kaggle.com/jboysen/us-traffic-2015/feed

The dataset is a comprehensive view of various factors which affect the traffic in a particular area at a particular time. Along with the date and times the data was taken at, the data assess factors such as the direction being traveled in, what kind of road is being driven on, the state code, and the traffic volume at various times. There are over thirty attributes and over seven million data points in this dataset that can be used for creating interesting models about our topic.

2.3 Evaluation Methods

After evaluating our dataset we hope to be able to have developed a few different models that will tell us important statistics about what we decide to model. We will try to find out which month that rural roads may be used the less, possibly dependent on the amount of lanes they have, what time of day, and how many cars are on the other roads. These specific results can be used by local governments to plan construction projects at the right time to impact less people. To evaluate this data, we will need to use pattern-recognition algorithms such as the Apriori algorithm and partitioning. We will also most likely use FP-trees to visualize our data. Once we have applied pattern-recognition algorithms, we will need to create contingency tables and assess the support, confidence, association, and lift using pre-determined support and confidence threshold of potentially related variables. Once we have determined whether the variables of interest have solid support for potentially being correlated, we can break away from single-dimensional correlations and try to determine if any multi-dimensional correlations exist. These multi-dimensional correlations could prove just as, if not more, useful than only determining single-dimensional correlations, as they would help paint a clearer picture of the factors which might be affecting the traffic on any given road.

2.4 Tools

Our project will utilize Python as its primary programming language to perform the data mining processes. We will use Anaconda and ipython notebooks as an easy way to create scripts and manage them. Python is extremely useful in its collection of open source libraries for data mining. This includes pandas, matplotlib, and numpy as they provide programming to manipulate the data and display it visually. Python is also excellent at handling mathematical equations, especially when open source libraries are used. Since a lot of our data analysis will require mathematical formulas, Python will be a good choice for handling this data. Python is also flexible enough to store data which has previously been acquired, which will be useful for calculating the multi-dimensional correlations. Github will be an important asset towards achieving our goal because it serves as our version control as well as our software for project tracking. It will be a place to know what portions of the project needs to be done and what has been done. Other tools we intend to employ would be any other external source we use to help guide and tutor us in the data mining process.

3 MILESTONES

We hope to have the data cleaned and pruned of unnecessary attributes and ready for our analysis to begin by the end of spring break. From there we plan on sorting the data in time for the progress report. Normalization should be done the week after that, along with the creation of our initial models with some preliminary results. From there, we can attempt to find any single-dimensional correlations between our chosen interesting variables and test the support and confidence of these correlations. Multi-dimensional correlation testing will follow soon after to see if there are any links worth noticing. In the final weeks we will perfect our models and find our interesting results, as well as finding a way to visualize our results in a meaningful way.

3.1 Completed Work

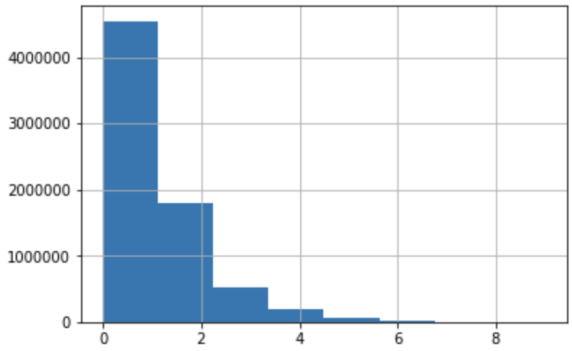
So far we have managed to start cleaning our data and remove a majority of rows with too many null values. We have also removed columns of extraneous data in our program. Such as columns that show the year of the data, as this is redundent information. This allows us to analyze the data faster because the program is not focusing attention on values which will not prove helpful to our inquiry. Some small graphs and looking at basic statistics about the data has been done as well.

3.2 To-Do

While we have a good start on getting our data prepped, we still need to normalize the data to give us a reference point on what counts as a high-traffic scenario and what counts as a low-traffic scenario. We also wish to create graphs of our normalized data, if only for the sake of having it as a reference. We also need to apply pattern-mining algorithms like Apriori in order to discover single-dimensional correlations between our variables. Once we have our single-dimensional references, we need to confirm they fit our support and confidence thresholds. Once we have single-dimensional relationships which pass those thresholds, we will need to test for multi-dimensional correlations and their support thresholds as well. Once we've perfected our analysis of these correlations, we will need to create graphs and visualizations of the data which will convey our findings in a manner which will be easy to read and visualize. Once we are ready to start this analysis we may choose to use sampling to run tests on as we may still have many millions of data points which would be useful to use random sampling to speed things up and get a faster analysis completed.

4 RESULTS

It is too early for us to see any interesting results in our data. Here is a small graphic displaying raw data about how much traffic is recorded divided into bins of how many lanes the roads have available. Kind of interesting to see most of the traffic numbers are actually in the bins of 0-3 lanes of travel.



1	2849316
2	1799970
0	1695513
3	527434
4	185915
5	53827
6	21265
7	6176
8	905
9	70

5 REFERENCES

https://www.kaggle.com/jboysen/us-traffic-2015/feed
https://www.nature.com/articles/srep37300
https://www.bactrack.com/blogs/expert-center/35042821-the-most-dangerous-times-on-the-road
https://www.zmescience.com/research/technology/google-maps-traffic-05443/
http://kalw.org/post/driving-apps-waze-are-creating-new-traffic-problems

6 CONCLUSIONS

At this time our group has a good layout of what we want to complete this semester and how we will do it. We plan to find interesting results from our data mining project and be able to create a real report that a city could potentially use to aid in construction projects or city planning around traffic networks. For example, if we find that the season of the year is somehow related to traffic on a particular road, we can advise the city to do the bulk of construction work during the season where the road is not used as much. Depending what we find in our analysis we might be able to extend our future plans with our results.