Larry Gates

CM1004

CSSE494-08

Senior Thesis Fall Summary

Introduction:

The purpose of the project is to find areas of congestion using traffic flow data and determine an appropriate advertisement to be placed at a given location, and if possible, at a given time of day. By using Hadoop Ecosystem, a common tool for big data analysis, the project should result in optimizing the use of advertisements for a given area based on traffic flow. The retrieved datasets contain traffic flow data, demographics for a given area, and advertisement data for a city.

Motivation:

The motivation for this is to find additional uses for big data. Many cities and states collect traffic statistics for roads, to show traffic flow on display boards or for statistical information. Utilizing public domain data and finding patterns from the datasets are essential for users for are interested in finding trends in patterns that are related to each other. By completing this successfully project, this will be an example for other cities to attempt to record traffic data to find patterns and optimizing selling locations.

For influencing the field of computer science, this would be another example of using Hadoop Ecosystem and corresponding services to filter, sort, clean and output data. Allowing for the data to be adding to model or be including in an aggregation function could be used in any city that collects this sort of data.

Literature Review:

Currently, there appears to be no academic research into using big data and coordinating traffic flow with advertisements. However, plenty of academic research and articles involving big data and traffic datasets, along with many application articles on using big data and advertisement.

Smith and Demetsky (1997) discuss the importance of "intelligent transportation systems," at a time for Google Maps and traffic flows on the internet. Their academic paper discusses the need to have forecasting models, using several time prediction techniques, such as machine learning and historical average, to determine traffic forecasting. With the datasets being collected, the project will use historical averaging with the available data. Daas, Puts, Buelens, and van den Hurk (2013) highlight that at the time, big data was heavily "IT-perspective" and "focus on soft- and hardware issues." In the Big Data case study of traffic loop detection data, a successful plot was shown of peak hours and vehicle flow. The case study shows the high potential of finding a trend in the data.

The big data and advertisement side of academic articles talk about the collection and use of user data. Couldry and Turow (2014) elaborate on personalized advertising constantly mining personalized data. The article also looks "more broadly at the consequences of embedding big data use in advertising," which is not concern for the project. Bughin, Chui, and Manyika (2010) discuss the opportunities companies take with using the data available for a web-based company. Allowing for better selective marketing for an area. The availability of data for marketing is not scarce due to the expanding amount of data, which will be useful in determining advertisements to display is a certain area.

Current Progress:

Using Dr. Mohan's lesson plans for *Introduction to Hadoop*, I have been teaching myself a condensed version of using Hadoop Ecosystem. Through the lesson plans, I am strongly familiar with the workings of Hadoop's internal systems, along with several of the services that work with Hadoop. I did get experience with various services, such as Pig and Hive, through some of the labs for the course.

Using the instruction manual from Apache, I set up a cluster that runs Ambari-Server, which is a Hadoop Ecosystem. The cluster runs with 5 nodes and can run various Hadoop labs. The server holds all available services that were initially provided by Apache and has been operational.

Since the beginning of the fall term, I have been collecting data from the Data Portal of Chicago. Using a bash script, every 5 minutes the script downloads 3 different layouts of traffic data for a given region of Chicago, along with traffic data for a given segment of Chicago. Since the website does not upload a new spreadsheet every 5 minutes, the script checks if the spreadsheet that was downloaded is not the same as the previous file that was downloaded. If the set of 3 files is the same, the new 3 files will be deleted. Currently, there are over 4,000 spreadsheets for both segment and region data from September 18th, 2016 to October 31st, 2016, with about 24 per day.

Data Description:

The two datasets currently collected are Segment Estimations and Regional Estimations.

Segment Estimations:

The segment estimations “contains the current estimated [vehicle] speed for about 1250 segments covering 200 miles of arterial roads” [S1] in Chicago. The spreadsheet apparently does not update every segment, which is due to an uploading problem on the data portal’s end. The segment estimations are supposedly collected every 20 minutes by the Data Portal. The important information from each row in the dataset are the location and area of the traffic flow, speed of traffic, and the timestamp the data was updated.

Region Estimations:

The region estimations “contains the current estimated congestion for the 29 traffic regions” [S3] in Chicago. The spreadsheet has updated all 29 regions. The region estimations are supposed to be collected every 10 minutes by the Data Portal. The important information from each row in the dataset are the boundary of a given region, speed of the traffic, and the timestamp the data was updated.

Future Progress:

The next step for this project begins with using the services of Hadoop and the actual data. I will continue to collect traffic data from the Data Portal of Chicago throughout the year, and I will look for additional sites that contain datasets pertaining to traffic data. Along with traffic, I will locate additional datasets to help determine locations of billboards in Chicago, along with the advertisements displayed on the billboard. The Data Portal of Chicago has many datasets that appear to have some relevant data about advertisements. Determining if the information is valid for the project along with the accuracy will be determine as the project progresses. Outside of the Data Portal of Chicago, I can use U.S. Census data to find demographics of a given area and overlay the information into segments and regions defined in the datasets I have acquired. These demographic data will help me determine some potential ideas of advertisements that would work.

The data will need to be sorted per segment and region, since there are data points for almost every hour of every day in a segment and region. Data will also have to be cleaned, as some rows have not been updated for years, i.e. one segment has the LAST_UPDATED for 2010-07-21. The data can be sorted by Pig, allowing the data to be aggregated and condensed. Once the data is sorted, I can find averages for a given period. I can use a service that will find the average for various stages along with finding common patterns in a segment or region among the data, or write a custom MapReduce application to do the work for me.

With advice from Dr. Wollowski for incorporating machine learning into my senior thesis, I could potentially use a simple form of machine learning to find typical traffic flow during certain times of day for a given direction. This can be done with finding the averages of speeds for times. If I am successful in finding data that contains location and type of advertisement, I can use the cluster to determine the type of advertisements to display for a segments or regions. The latter part of the suggestions for machine learning from Dr. Wollowski will be applicable if advertisement data is easily accessible.

The proposed architecture flow is displayed below, showing where I collect the data, tools used at what stages, along with stages and flow.

Proposed Architecture:

TODO

Sources: (What style formatting are the source?)

S1. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Se/n4j6-wkkf/about

S2: https://data.cityofchicago.org/api/assets/3F039704-BD76-4E6E-8E42-5F2BB01F0AF8?download=true

S3: https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Congestion-Estimates-by-Re/t2qc-9pjd/about

S4: https://data.cityofchicago.org/api/assets/88B2ABA5-BF4C-4A41-949C-2B11D725ADAB