High Taxi Demanding Areas Prediction by Region and Time in NYC

Xuefeng Peng, Yiming Pan

Introduction

Heavy traffic, untampered energy waste, deteriorating public safety, illegal drug dealings, and etc., all of those issues are ineluctable for a growing city. With more and more concerns towards those issues, people are determined to seek solutions that are able to not only ease them but also eradicate them. This determination finally leads people to urban computing --- a solution empowered by massive computing capabilities. This new way is not limited to solely solve those common issues a developing city may encounter, it aims to plan the city to be more intelligent in various aspects. Many researches are conducted with the goal of making cities smarter, some of them endeavor to improve the city traffic by proposing a large-scale taxi/bike ridesharing service [1], some of them endeavor to monitor and predict the air/water quality of certain city regions [2], and some of them are interested in energy consumption control via supervising urbane refueling behaviors [3]. Those researches show numerous promising potentialities of urbane computing.

With the same goal of creating smarter cities, we are particularly interested in the traffic aspect. Many city dwellers would complain that sometimes in certain areas of the city, it is extremely hard to catch a taxi or it takes too long to wait a bus/subway to come. On the other hand, taxi drivers are complaining that sometimes they are not aware of which part of the city is demanding the rides so as to lose their potential revenues. Imaging what if we can predict which region of a city would encounter an increasing demand of taxi or transportation in advance, then we can solve the complains from both sides by allocating the taxi drivers or schedule more public rides to those areas beforehand. Here, we generalize our goal as predicting the high taxi demanding areas in NYC in accordance with time.

Method

Preliminarily, we think we can generalize the problem that we intend to solve as the followings:

Assumption I. Facilities that people decide to get to must serve some functionalities, in other words, people rarely get somewhere without any purpose.

We think the facilities that people usually go to can be classified into several groups, they are Health, Academic, Recreational, Residential, and Infrastructural. We further classify those groups into many subgroups.

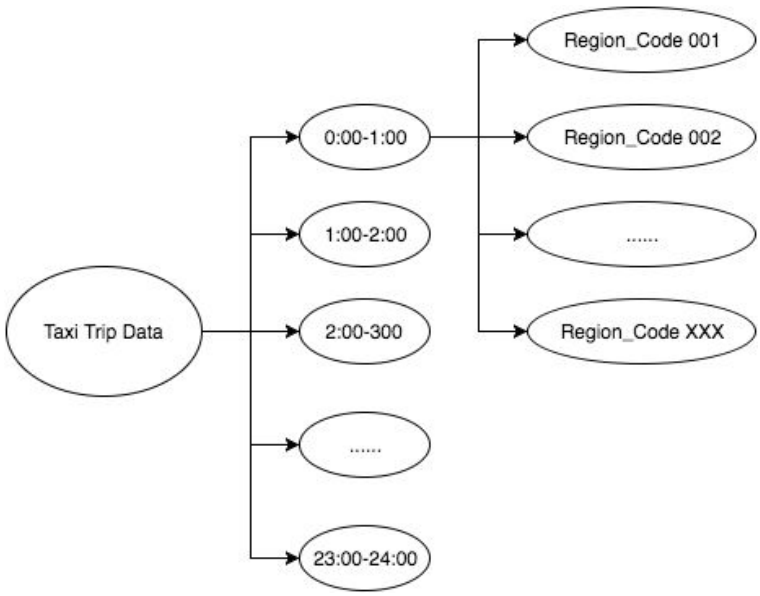
Health	Hospital, Nursing Center
Academic	Public school, Universities
Recreational	Theaters, Eateries, Travel spotlights
Residential	Apartment regions, dense living house region
Infrastructural	Airport, Train Station, Bus Station

The followings are the step abstract:

Step 1. Get the geolocation data of those above facilities in NYC

Step 2. Split the region of NYC and mark each with a region code

Step 3. Partition taxi trip data by date and further partition it in one-hour period, classify them by weekday



Step 4. Extract feature vectors in the form of *[region code, date, weekday, time, weather]*, relate them with ground truth demands

Step 5. Train regression model to predict

Step 6. Find the high taxi demand regions by time patterns in NYC

Datasets

Foursqaure check in data in NYC

2014 Yellow Taxi Trip Data: https://data.cityofnewyork.us/view/gn7m-em8n

2015 Yellow Taxi Trip Data: https://data.cityofnewyork.us/view/ba8s-jw6u

Facilities Geolocation Data

2014 and 2015 Daily Weather Data in NYC

Data Preprocessing

For the taxi trip data, there are several attributes that are not related to our target like *payment_type*, *total_amounts* and etc. Those could be eliminated and would not affect any further prediction. Then we integrate the dataset based on the partition rules above, we would integrate all the transactions with the same region code and time period but add a new attribute *demand* which is the sum of all the *passenger_count* in those transactions. However, during this process we would eliminate transactions that have the same pick_up region and

drop_off region since those trips don't give any information. Therefore, the partition rule of how to divide NYC into regions would matter a lot on the results, and this would be eventually determined by the performance of models that are built by different partition rules.

How we train the model and how we test the model?

The feature vector format for training the model is presented above, we prepare to use 2014 taxi trip data and ¾ of the 2015 taxi trip data as our training set, whereas the rest of the 2015 taxi trip data is used as our testing set. At this point, we haven't decide what specific model we will use, but the most basic model we would employ should be a regression model.

Goals

Basically, what we want to accomplish first is constructing a model that could predict the demands on taxi by given an one-hour period in a particular day, a certain region and the weather. Then we could tell which area in which time would be very popular in NYC or which area is the hottest in NYC. According to the facility function and location data, we could tell what kind of facilities make a region popular in a specific period and weather. However, there are lots of other factors which could affect the demands like weekends or holidays, we could integrate these attributes into our database for exploration. Those factors would determine the difference between dates and somehow improve the accuracy of the predictor. Further, we could figure out the correlation between the demands of taxi and any popular event like concert or NBA match. By taking the events into consideration, we could optimize the prediction even more.

Conclusion

We will first use over 200,000 checkins in NYC data to cluster the popular regions in NYC, and in order to identify them, each region will be marked with a unique region code. Then, we assign each facility into a region where it belongs to according to its geolocation. Secondly, we need to preprocess the 2014/2015 yellow-cab trip data. For each trip, what we need are geolocations of the pick-up and drop-off points, the start and end time, and the passenger number. We gauge the demand mainly by the number of passengers. The pick-up locations and time help us to identify when and where in the NYC there will have a high demand on taxi. The information mentioned above along with the facility locations will help us to explain why people will concentrate in some certain areas in some specific time. In addition, we also consider the weather condition since it could impact people's commuting choices a lot. For instance, in hot/cold day or windy/snowy day, people may more likely to take taxi instead of walking or waiting for public transportation.

In a nutshell, we preliminarily think facilities within the regions, the time slots, the weekday, the weather conditions will affect people's behaviors on choosing taxi. Thus, we train our predictor with those factors along with the ground truth demands; and eventually, by given certain area at certain time under certain weather condition, the demands of taxi in this region then could be estimated. To evaluate our predictor, we prepare to use 2014 taxi trip data and ¾ of the 2015 taxi trip data as our training set, whereas the rest of the 2015 taxi trip data is used as testing set. Besides, we also could learn interesting patterns about people's behaviors in NYC. For instance, one possible finding could be in Friday night many people will gather into a region full of bars. Another fun fact we can explore is querying Uber API to detect active cabs in a region at a specific time point where our predictor estimates as having a high taxi demand. In addition, it is also possible that the events like NBA event in Madison

Square, Concert at Lincoln Center hold in NYC might influence the traffic, so we will consider the events as an additional factor into our predictor in our future work.

Reference

[1]Shuo Ma, Yu Zheng, and O. Wolfson. 2013. T-share: A large-scale dynamic taxi ridesharing service. *2013 IEEE 29th International Conference on Data Engineering (ICDE)* (2013).

[2]Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. 2013. U-Air. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13* (2013).

[3]Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, and Yong Yu. 2014. Inferring gas consumption and pollution emission of vehicles throughout a city. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14* (2014).