Comp 4710 term paper

abstract

Public transportation is essential for the public's daily life. For passengers, an accurate transportation prediction, which means less bus waiting, and transit time. This leads to maximize their time utilization. Also, it is critical for bus companies, and accuracy prediction will help all them quick aware of the current states of the bus running. By devoting more bus, or short the bus departure gap, the delay of the bus could be minimized, which allows them provided more quality servers to the public. There are lots of researches that digs into this area, which focuses on the relationship between the delay time and time sections of a day (e.g., morning, afternoon, night), route and stations. However, in order to reach predict an accurate bus delay, this is not enough. In this paper, we would go beyond the conditions discussed above. and exam another important factor that causes delays - the weather condition. specifically, This approach uses open data from government open portal such as the historical bus delay datasheets, and historical weather datasheets. as the source. then generate a large datasheet by joining that two datasheets. then apply different data mining techniques on these data sheets to discover the relationships and frequent patterns. to achieve this, we used a Apriority algorithm as well as deep learning algorithm is designed to evaluate those frequent patterns and examine the relationship. base this. conclude the reason that lead to delay, also the data unbalancing issues that cause hard in form a correct module for prediction . The result is compared to the government open portal provided delay data as well as weather conditions and shows the effectiveness and correctness of our approach.

1. Instruction

It is a kind of fortunate that we are in the era of big data that provided us with an opportunity to examine the huge volume, a high-speed and abundant variety of data. In our paper, we will deal with certain precise data provided by the government of Canada and the streetcar delay information provided by the TTC.

Public transport is the most popular travel method in the world and billions of people rely on it. People choosing public transport because it is convenient, reliable, eco- friendly and most important it's feasible.

Bus delay common and almost unable to eliminate, but instead eliminate, we could try to find the relationship between delay and reasons, and guide people to adopt the delay and reduce waste time. Thanks to the big data, the bus delay data is collected with accurate time, together with a wide range of useful information. In Canada, lots of cities have an open portal and present the transport datasheets, in our research, we considered two sets of transport data:

the Toronto street bus delay dataset

and Winnipeg street bus delay dataset

We choose The Toronto street bus delay dataset as our research source.

The weather is probably one of the most important factors for the delay.

Unlike the time section, bus station location or route number, the weather is a sequential factor in the data analysis, that is, a time section may just effect on a certain time period, and each period is relatively independent. Weather as a sequential factor, such that previous weather conditions will heavily influence the later on traffics. For example, in the winter in Canada, yesterday's snowstorm may affect today's traffic, some roads may be closed. Or if we have snow and the temperature was greater than 0 degree. Then it is highly possible that there are lots of ice on the road. such things will influence the traffic. Weather condition not only influences the traffic but also affects people, in Canada's extremely cold winter, people are not suitable for a stay outside too long and may take a longer time to reach the bus station than expected.

So the key point in our approach is we analyzed the historical transport delay data set to join with the historical weather dataset. And find the relationship between current and previous weather conditions then combine with the traditional approaches and predicting whether the bus would delay or not to be specific. In our approach, two datasets are joined in a special method, each entry not only contains the delay information and weather conditions but also contains yesterday's weather condition. Base on this big dataset, a frequent pattern mining in conduct to find the potential relationship and also filter the dataset. to examine the accuracy of our data mining result, another copy of data set is sending to a deep learning network for further learning. which allow us to perform the cross matching with apriori approach. To evaluate the accuracy of the result of learning, we use a different dataset from another year provided by TTC. The evaluation shows the accuracy of the test on this year's (2019) delay data, on data mining on previous years (2014-2018) delay and weather condition.

The rest of the paper is organized as follows, the next section 3, we would like to introduce some related works. In section 4, explanation and discusses our data mining approach, and the deep learning method. section 5. is our experiment on the real data and the data analyzed.last in section 6. is our conclusion.

2.background and related work

In this section, we explore related works and and provide the background on the prediction model. Over the decades, many efforts have been done to build a reliable prediction model for bus arrival time, which can be divided into two types: analytical models and statistical models. The analytical models are complicated to describe the bus travel time and related factor as the data collected from the bus transit system is complex (Chen 2012). As for statistical models, there are mainly three types of models: 1) historical data-based model, 2) regression model, and 3) artificial neural networks models.

The first type of prediction model estimates the future bus travel time based on the past related data. Chien and Kuchipudi (2003) collected travel time data by AFC system to discuss the path-based and link-based travel-time models.

The second approach requires a mathematical models to represent the relationship between travel times and related factors, but there might be an issue that the model is inapplicable to or need to be reloaded for a new route (Tétreault and El-Geneidy 2010).

The last approach is to use artificial neural networks. Jeong and Rilett (2004) proposed an ANN model for predicting bus dwell time at bus stops, but in order to have better performance, ANN models require extra work to find the right network structure and determine the best parameter values.

3.Data minning

3.1 data cleaning

Two data sets were used in this experiment: bus delay data from 2014-2019 released by TTC and temperature monitoring data released by the government from 2014-2019.

The bus delay data is provided once a year and mainly contains the information about report date, route, time, day, location, incident, min delay, min gap, direction, vehicle. Among them, the min gap represents the interval between two vehicles running the same route, the vehicle represents the vehicle number, and they do not play a decisive role in analyzing the relationship between the bus delay and weather. Therefore, we eliminated these two columns in the subsequent analysis.

The weather data is recorded in a CSV file and published by the government once a month. It records weather data every hour, that is, 24 times a day, from 0 to 23 o'clock, and records information about longitude, latitude, station name, date, time, temperature, dew point temperature, relative humidity, wind direction, wind speed, visibility (km), humidity, wind chill, weather, etc. Because the weather factors that cause vehicle delays are mainly attributed to weather and temperature, we extract only the date, time, temperature, and weather columns for subsequent analysis.

1. Bus delay data cleaning

After reading the data of each sheet and storing it in a data frame, we processed the data as follows:

1) Clean up rows containing null values.
2) The min delay in the original data is accurate to minutes, so the distribution is extremely scattered, which will affect the mining results of a frequent itemset. Therefore, we divide it into 4 intervals based on the severity of the delay:
   - 0-5 minutes: consider as no delay. During long distances, delays caused by factors such as different stopping times at stations are inevitable.
   - 5-15 minutes: short delay, most delays are distributed in this interval.
   - 15-30 minutes: medium delay with delay time greater than 15 and less than 30.
   - Greater than 30 minutes: serious delay. Such delays are relatively rare, accounting for only 7% of the total 75,000 records.
3) Similarly, attribute "Time" discretize the data, so we divide it into 6 intervals:
   - Early morning, from 0:00 – 4:00
   - Morning, from 4:00-8:00
   - Noon, from 8:00-12:00
   - Afternoon, from 12:00-16:00
   - Evening, from 16:00-20:00
   - Midnight, from 20:00-24:00
4) To join with weather data form, we process the attribute "report time" based on the time interval in which the delay occurred. For example, if the delay occurs at 6:31 on January 2, 2014, the report time is 14-01-02 4:00.

5) Cause the data was entered manually, we found errors in the attribute "Direction". So, we convert it into a unified format.

2. Weather data cleaning

As mentioned above, weather data is updated every hour. In order to unify with the bus delay data, we calculate the average temperature within four hours as the temperature during this period. Also, we sort the severity of the weather and ranging it from sunny to "moderate ice particles" corresponding to 0-16. The larger the value, the worse the weather. After using this sorting method to weather data, the attribute "Weather "shows the worst weather in four hours, and yesterday's weather was the worst weather of the previous day. It should be noted that “Yesterday Weather” for the first day of each month is none.

3. Data merge

After getting the cleaned data, we use the attribute "Report Time" to merge the two forms and get bus delay data which includes weather information. An example view of the data provided in Table 1.

Table 1

Date/Time	Route	Time	Day	Location	Incident	Min Delay Ri	Min Delay	Direction	Temp (--=C)	Weather	YesterdayWeather
2017-01-02 16:00	504	16-20	Monday	Broadview S	Investigation	15~30	15	W/B	3.28	3	yesterday:4
2017-01-02 16:00	512	16-20	Monday	SCW Station	Investigation	5~15	5	W/B	3.28	3	yesterday:4
2017-01-02 20:00	504	20-24	Monday	Dundas Wes	Held By	5~15	12	E/B	3.47	4	yesterday:4
2017-01-02 20:00	504	20-24	Monday	Sumach / Kir	Mechanical	5~15	5	E/B	3.47	4	yesterday:4
2017-01-02 20:00	504	20-24	Monday	Neville Park	Mechanical	5~15	9	W/B	3.47	4	yesterday:4
2017-01-02 20:00	501	20-24	Monday	Long Branch	Held By	>30	68	E/B	3.47	4	yesterday:4
2017-01-02 20:00	504	20-24	Monday	dufferin	Mechanical	15~30	25	W/B	3.47	4	yesterday:4
2017-01-02 20:00	504	20-24	Monday	Dundas Wes	Mechanical	0~5	1	W/B	3.47	4	yesterday:4
2017-01-03 0:00	510	0-4	Tuesday	Spadina and	Mechanical	5~15	5	S/B	3.18	8	yesterday:8

3.2 Data analysis

3.2.1 raw data analysis

3.2.2 Data Analytics via Frequent Pattern Mining

The objective of this paper is to develop a methodology to predict bus delay time under various weathers by analysing the relationships between bus delays and weather. After data cleaning is done, the Apriority Algorithm is applied to find frequent datasets which includes the delay time and weather since the selected dataset should be strictly related to our goal. In particular, Apriority algorithm generates the dataset and then count for each dataset. If the count of the dataset is less than minimum support, then this dataset is not considered as a frequent dataset. This algorithm requires multiple database scans and each scan generates n+1 datasets. However, it is inefficient since the database contains mass of data. To have a better performance, parallelism is applied. The idea is to apply task parallelism alone with master slave approach in multicore architecture. The main thread spawns slave threads and each slave thread is assigned a different task in a different core (Rustogi and Swati 2017).

Thus, ideally, the frequent datasets should contain all delay buses due to different weather. In order to have more precise prediction, we also take into account yesterday’s weather as delays may depend on weather information from the past

days. For example, if yesterday's weather was Heavy Snow, then it is highly possible that the bus would delay due to terrible road conditions even though the current weather is Clear. It would be categorized as a bus delay in the Clear day which is not the case.

Once these frequent patterns were computed, we obtain the table as follows:
Table 2

Weather	Yesterday W	Delay	Support	Frequency
1	yesterday:2	0~5	0.020021425	1535
1	yesterday:2	5~15	0.045409931	3471
2	yesterday:2	5~15	0.073032202	5547
2	yesterday:3	0~5	0.022402229	1703
2	yesterday:3	5~15	0.048994726	3718
2	yesterday:2	0~5	0.0345309	2623
3	yesterday:3	5~15	0.03292117	2500
8	yesterday:8	5~15	0.038320018	2911

Observed from Table 2, we can conclude that the bus is mainly delayed in the good weather (Clear, Mainly Clear, Mostly Cloudy) and the only exception is Freezing Rain. The delay range is around 0~15 as expected since 80% of the total delayed buses were late for 0~15 minutes. Yesterday's weather has no much difference with the current weather.

The bus is likely delayed in the good weather is not a reasonable conclusion. As mentioned in section 3.1, the database is unbalanced because weatheres are not equally represented. For example, 75% of the total bus delays are in the good weather (Clear, Mainly Clear, Mostly Cloudy, Cloudy) and the highest total bus delay rate in the rest of weathers is 8.5%. Obviously, the Apriority Algorithm cannot extract the relationships between bus delays and terrible weather because the total amount of bad weather days is really small.

a. relationship learning
   An artificial neural network approach will be applied to this part. We built a supervised learning model. The model takes a date, time, weather, temperature, bus route, bus stop as input, and tries to predict delay time in minutes or degree of delay (how bad is this delay). After finishing training, this model's output will be used to compare with the results from other parts of this project.
b.

experiment

c. process
   i. data mining
   ii. learning
d. result
   i. data mining
   ii. learning

conclusion

Reference:

Jeong, R, and Rilett, R. “Bus Arrival Time Prediction Using Artificial Neural Network Model.” *Proceedings. The 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No.04TH8749)*. IEEE, 2004. 988–993. Web.

Tétreault, Paul R, and El-Geneidy, Ahmed M. “Estimating Bus Run Times for New Limited-Stop Service Using Archived AVL and APC Data.” *Transportation research.* 44.6 390–402. Web.

Chien, Steven I-Jy, and Kuchipudi, Chandra Mouly. “Dynamic Travel Time Prediction with Real-Time and Historic Data.” *Journal of transportation engineering /* 129.6 (2003): 608–616. Web.

Chen, Guojun, et al. “Bus-Arrival-Time Prediction Models: Link-Based and Section-Based.” *Journal of Transportation Engineering /*, vol. 138, no. 1, American Society of Civil Engineers, Jan. 2012, pp. 60–66, doi:10.1061/(ASCE)TE.1943-5436.0000312.

Sun, Dh et al. “Predicting Bus Arrival Time an the Basis of Global Positioning System Data.” *Transportation Research Record* 2034.2034 (2007): 62–72. Web.

Rustogi, Swati, Sharma, Manisha, and Morwal, Sudha. “TID Based Data and Task Parallelism for Frequent Data Mining.” *Journal of Information and Optimization Sciences*38.6 (2017): 961–970. Web.