**Prediction of the Daejeon Subway Congestion with Gradient Decent**

**JaeSeok Shin, YaeHyeon Kang**

**Dept. of AI Big Data, Woosong University**

**1. Abstract**

Daejeon Subway is efficient transportation that citizens usually use. And data on the subway line can be found in the Public Data Portal of the Korean government. However, unlike the Seoul subway line, Congestion of each subway is not offered in Daejeon. Already, through the Portal site, enough data to find out the congestion is not properly utilized. This paper tried to find the congestion at each subway station and find the best time to take the subway to reach the target station for every passenger. The solution is solved by gradient descent, one of the methods of linear regression in machine learning. The program is written in Python programming language and the data is origins from Daejeon Transportation Corporation. The result will show the time that has the least congestion from the departure station to the destination station.

Recommendation system in machine learning is an algorithm to suggest or recommend product or service to users or customers. This algorithm has been using in many areas like as E-commerce, Over-The-Top services, Social Network Services, and etc.

**2. Introduction**

**2-1. Background**

Digitalization affected many areas of our life. Transportation is one of the industries influenced by these changes. Now, we can use any transportation like bus, subway, taxi, and others with a single card. It not only gives convenience for human life but also converts into data. As the data is collected, we can use them to predict demand, recognize the danger, and prepare for future changes. Daejeon City, the location of Woosong University, provides these data through the Public Data Portal collected by the Daejeon Transportation Corporation.

The bus is also efficient and major transportation. However, it is markedly affected by external factors like traffic, bottleneck phenomenon, number of lanes, and rush hours. And they have many lines to ride to reach destinations, it is hard to quantify the congestion and times.

But, in the case of the Daejeon subway, it has one line. It means the analysis doesn’t have to consider the transfer to the other subway line. Also, there is no influence by other traffic. So, prediction can be more precise than the bus. And, the data provided by Transportation Cooperation is enough to process for the projects.

**2-2. Problem Statement**

Unlike Seoul subway lines, there are data of the most congestion but I didn’t utilize it within the actual operation. This may cause uncomfortable while using the transportation. Furthermore, people want not only the time that the subway comes but also to seek a seat or place less occupied by others. With mobile devices, we can easily access timetables of when the subway comes or leave. On the other hand, air pollution, climate change, and abnormally high or low temperature make us think about these condition. So, we think these service and data is not properly used.

**2-3. Contribution**

This project will serve to find out the approximate time considering the congestion of the subway. Subway is barely affected by external factors like weather, it will be more predictable than any other transportation. Also, calculating the congestion of the Daejeon Subway is only treated as data, not in a practical way. It will derive people and officials to think about congestion and pleasant transportation facilities.

**3. Proposed method**

We used Python programming language to progress the project. Python is one of the popular programming languages and it is specialized in machine learning and data processing. An additional advantage of Python is its diverse library. There are many mathematical calculations, data processing, visualization, etc. We used Pandas to call a Microsoft Excel file as the data frame format, NumPy as the array format, Matplotlib as the visualization, and OS inner module to call the directory of the source code file and Microsoft Excel file for a transplant.

**3-1. Congestion measurement**

Congestion can be the remaining number of passengers in the subway car unit out of capacity.[1]

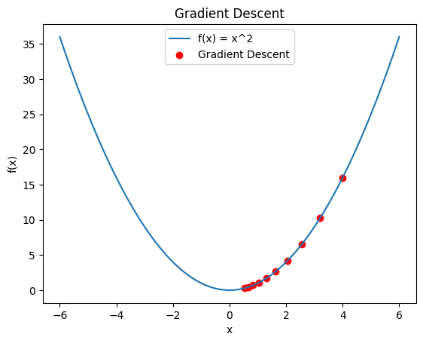
<Table 1> Eye Measurement of Congestion in Daejeon Transport Corporation [1]

|  |  |  |
| --- | --- | --- |
| Passenger  (Person) | Congestion  (%) | Passengers distribution in Subway |
| 90 | 32.8% | 50 % sit |
| 180 | 65.6% | 100 % sit |
| 270 | 98.5 | 100 % sit, 50 % handle |
| 274 | 100.0 % | 100% sit, 50 % handle, 2 people on two-door |
| 300 | 109.4 % | 100% sit, 100 % handle |
| 348 | 127.0% | 100% sit, 100 % handle, 2 people on each door |
| 444 | 162.0% | 100% sit, 100 % handle, 2 people on each door, 1 line in the middle |
| 516 | 188.3 % | 100% sit, 100 % handle, 2 people on each door, 1 line in the middle |
| 620 | 226.2 % | 100% sit, 100 % handle, 2 people on each door, 2 lines in the middle |
| 657 | 240.0 % | subway saturation. Failure of some passengers boarding |

Congestion is also influenced by the previous station passengers. Considering that, the congestion formula above be like this:[1]

d is Destination, d-1 is the previous station of Destination, N is capacity, is the number of passengers getting on a subway car, is the number of passengers who leave the subway car, and is Destination Congestion,

**3-3. Gradient Descent**

****

**[Fig. 1] Graph that presents the process of Gradient Descent**

Gradient Descent is an optimization algorithm that uses the first orders of approximation. By using the cost function, descending the value and finding out the least slope. This can minimize the error between the prediction value and the actual value. We use this method to find out the least congestion of certain times of the subway.

**4. Experiment and result analysis**

**4-1. Data collection and preprocessing**

**4-1-1. Data collection**

**텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명**

**[Fig. 2] Data.go.kr webpage of datasets**

The data is collected from the Public Data portal administrated by the Korean government. We used “Daejeon Transportation Public Cooperation\_hourly number of get on/off”, “Daejeon City Subway Timetable”, and “Daejeon Transportation Public Cooperation\_Daejeon City Rail One\_Lead time , distance and fee”. All of them was CSV(Comma Separated Value) format.

**4-1-2. Data Preprocessing**

**텍스트, 스크린샷, 번호, 소프트웨어이(가) 표시된 사진

자동 생성된 설명**

**[Fig. 3] Original subway passengers get-on and off data set**

**텍스트, 스크린샷, 번호, 메뉴이(가) 표시된 사진

자동 생성된 설명**

**[Fig. 4] Pre-processed subway passengers get-on and off data set**

Compare to [Fig. 3] and [Fig. 4], After pre-processing implemented sorting with certain column and separated by the sheet of each stations.

**텍스트, 스크린샷, 번호, 평행이(가) 표시된 사진

자동 생성된 설명**

**[Fig. 5] Original timetable data set**

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

**[Fig. 6] Pre-processed time table data-set**

Compared to [Fig .5] with [Fig. 6], The column was reversed and the unnecessary index is deleted. So, the data can be seen as what is needed.

**4-2. Data Analysis**

The big flow of data analysis is as followed:

First, Get input data from the user about weekdays, Target Time, Departure station, and Destination Station. To outcome the exact time and congestion, we have to define a certain value for the test. We set input data as Thursday, 9 a.m., Jung-Ang-Ro (two stations to Dae-Dong(Woosong Uni)) and destination as Dae-Dong(Woosong Uni).

From the preprocessed Excel file, load to memory the whole data of the input weekdays. we have data, for 2023, Jan, 1st to 2023, March, 31st. As we set the test for Thursday, we load from Jan, 5th to March, 23rd. During this period, we also filter with the target time to save time and memories.

Read the timetable of the Departure station, and separate them by the minute. Each station has a timetable as most of the subways follow the time.

With the get-on and get-off data, Calculate the congestion starting from the departure station to the destination station. Using the formula introduced in advance, find the final congestion from the departure station to the destination station. Then collect the all congestion at every time on the timetable up to the target time.

Separate time in minutes as the x value and congestion as the y value, put a virtual equation, and process the gradient descent to find the least congestion among the x values. With several processes to find the best epochs and learning rates. Finally, we found out that 800 epochs and 0.025 learning rate represent the best consequences.

**스크린샷, 텍스트, 그래프, 라인이(가) 표시된 사진

자동 생성된 설명스크린샷, 라인, 그래프, 텍스트이(가) 표시된 사진

자동 생성된 설명스크린샷, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명스크린샷, 라인, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명라인, 스크린샷, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명라인, 스크린샷, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명**

**[Fig. 7 - 12] Gradient Decent Progress every 200 epochs (Total 800 epochs)**

From [Fig.7 -12], the linear shape of the graph is operating to find the minimum value. As a result, you can recognize that riding the early subway has the least congestion on Thursday at 9 a.m. from Jung-Ang Ro to Daedong(Woosong Uni)

**5. Conclusion & Limitations**

This project tried to find out the exact time with the least congestion with the machine learning method using the Daejeon subway data. We could learn from this project First. we tried to find out the congestion not provided by the current Daejeon subway system. Second, we tried to calculate each number of people who ride or leave the subway car unit with the trend of passengers between the front and rear time intervals. Lastly,

Last but not least, still, there are limitations to this project. In the subway, there are two last stops and the subway departs from each end. It has directions in the opposite way. With this direction, the number of people who ride or leave can be different. Also, ironically, the influence of other transportation is not considered. The subway can cover an area of transport rather than the bus. People use the bus to their destinations to reach closer. The other factor that we didn’t consider can be weather, holiday, events like city festivals, and governmental policy that control people’s migration like lockdowns or social distancing during the coronavirus pandemic.

**6. References**

[1] Jun-su Kim, Subway Congestion Prediction and Recommendation System using Big Data Analysis, Journal of Digital Convergence, Page No. 289 – 295, 2016

[2] KeunWon Kim et al., An Exploratory Study on improvement Method of the Subway Congestion Based Big Data Convergence, Journal of Digital Convergence, Page No. 35 – 42, 2015