Visualization and Pagerank Analysis

in Metropolitan Subway Network

|  |  |
| --- | --- |
| Mulkyeol Kim | Jihwan Oh |
| *Department of Industrial Engineering* | *Department of Industrial Engineering* |
| *UNIST* | *UNIST* |
| ccomo456@unist.ac.kr | dhwlghks00@unist.ac.kr |

***Abstract***

**Our goal is to analyze Seoul and metropolitan subway pagerank. First, pre-processing of subway data in Seoul and the metropolitan area and Exploration data analysis are performed. Calculate the page rank using only a graph with no direction. But the results were almost random.**

**To solve this problem, the second task is to calculate the weighted pagerank using a direct graph. To calculate the weighted pagerank, the number of people getting on and off is estimated, and the weighted pagerank is calculated through the maximum eigenvalue method. The results were quite consistent with our common sense.**

**Additionally, correlation and OLS analysis were conducted, and future research directions were considered.**

I. Introduction

*A. Problem Definition*

The metropolitan subway, which spreads from Seoul to Gangwon-do, Incheon, and Chungcheongbuk-do, etc., has a very large number of users so also very complicated. Due to the large number of passengers, they are very crowded, and safety and discomfort problems have been steadily raised. Considering these points, we thought that selecting highly important stations on the metropolitan subway would help us intensively take care of the station and solve those problems.

Therefore, our goal is to find stations of high importance. These stations play an important role between other stations and lines, and there should be a lot of passenger movements. We selected ‘pagerank’ as the standard. Thus, in summary, our project aims to analyze pagerank in metropolitan subways and visualize it for easy use(developing important subway stations become safer and more convenient) in the real world.

*B*. *Hypothesis*

Before starting the task, we hypothesized which inverse would be of high importance.

1. The pagerank value of the station in central Seoul, not the outskirts, will be higher.
2. The pagerank at the transit station will be generally high.

II. Literature Review

*A. Subway Stations Network Structure Analysis by Using Social Network Analysis*

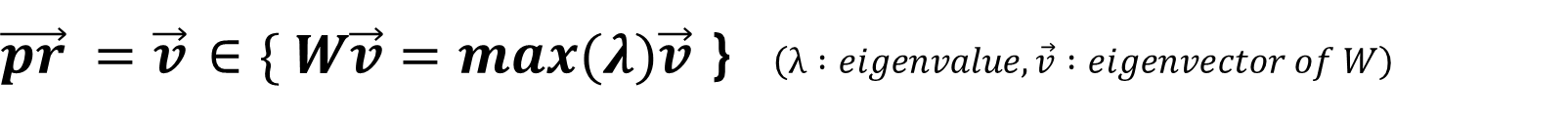
In this literature, degree centrality, closure centrality, and eignvector centrality were calculated using the data on the number of people getting on and off the metropolitan subway. As a result of Literature's analysis, ‘Gangnam’ Station was ranked high not only in the number of people getting on and off but also in the centrality analysis, which can be seen as a hub in the subway station network. In addition, ‘Jamsil’ Station showed a high ranking in the number of getting on and off rankings, but the ranking fell in the centrality analysis. This indicates that ‘Jamsil’ Station has a large number of people getting on and off, but is connected to stations that are relatively insignificant.

Literature mentioned that even if one centrality index is high or low, other centrality indices are low or high, so it is necessary to analyze the characteristics of each station in more detail.' Therefore, when analyzing the subway network, it seems necessary to calculate the centrality or importance by considering not only one indicator, but also the relationship with the neighbor stations and the information on the number of people getting on and off.

*B. Using Complex Numbers in Website Ranking Calculations: A Non-ad Hoc Alternative to Google’s PageRank - Keita Sugihara*

According to this paper, the power-iteration method is not available in complex networks. As an alternative to this, there is a method using weight matrix's eigenvector and eigenvalue. The page rank is also dependent on the maximum eigenvalue. Therefore, the page rank converges to the eigen vector when it has the maximum eigenvalue. Therefore, since our data is a complex network structure, we will use a method of finding the maximum value of the eigen value of the weight matrix instead of the power-iteration method that requires repetitive computation.

The PageRank formula as shown below, Equation 1.



(1)

III. Method

*A. Data*

We needed data with information about metropolitan subway lines (which should also include line order) and data with information on the number of people getting on and off. We were able to receive such data from ‘서울 열린 데이터 광장’, an open-source data site provided by the Seoul Metropolitan Government.



Fig. 1 Line data

**

Fig. 2 Data about number of people getting on and off

*B. Data Preprocessing*

(1) Selection of key lines

There are 24 subway lines in the metropolitan area, and some lines are centered in the outskirts (Gyeongchun Line), or have a special purpose(Eveline). So, we thought that using all 24 routes was inefficient and interfering with visualization and analysis, so we selected only lines 1-9, which people usually use for traffic purposes.

(2) Get information of pair between previous station and the next station is required

For the visualization of the graph of the metropolitan subway, we needed information on the pair of previous and next stations. The preprocessing was performed using the line data in Figure 1. Since column ‘외부코드’ contains information related to the line and its order, the station of the next row is designated in the ‘다음역’ column after sorting based on the ‘외부코드’. In addition, ‘다음역’ of the last station of each line was designated as the previous station so that they could be recognized as the end point of line.

1. Branches coming out from a single line

Some one station had two next stations at the same time, forming a branch. So these was handled.

1. Specify latitude and longitude using GoogleMap

To plot the station accurately, we used latitude and longitude as coordinates. For this, we used Googlemap. It finds latitude and longitude by name. By loading and utilizing Google Maps from Python, we could quickly and efficiently find latitude and longitude values for the names of numerous stations in data frame. We used it and created a data frame containing latitude and longitude.

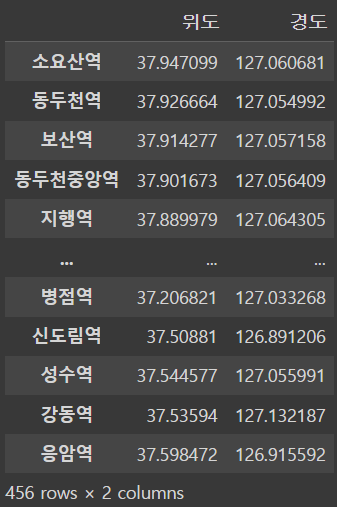
 ……..

Fig. 3 Dataframe of latitude and longitude

Afterwards, when the data frame of Figure 1 and the data frame of Figure 3 are combined, a data frame containing comprehensive information such as the next station, latitude, longitude, and arc can be generated.

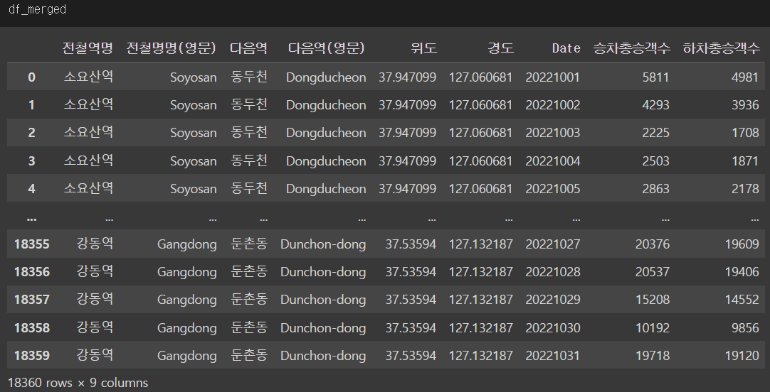
1. Adding getting in and out data



Fig. 4 Dataframe of number of people getting on and off

The data frame in Figure 4 is data that records the number of people getting on and off of each subway station by day. It is time series data recorded in a daily frequency.

In the data frame obtained in III.Methods-B-(4) above, value of ‘승차총승객수’ and ‘하차총승객수’ in dataframe of Figure2 is filled based on the name of the train station.

 Fig. 5 Final Dataframe

The data on the number of people getting on and off are completely combined.

There were also about five stations without getting on and off data, and according to our search, they were newly established in 2022 (within the last year). Since it was added to the ‘end’ of the subway line, they are in outskirts, and if the data on getting on and off is randomly filled or designated as a small number such as 1, they could affect the analysis of other stations, so these stations were removed.

*C. Global Clustering Coefficient*

Clustering coefficient refers to how connected some node’s neighbors are to each other (relative to how connected they could be). If the clustering coefficient is calculated for the entire network, the global clustering coefficient can be obtained.

(2)



Global cluster coefficient can be calculated by Equation (2).

*D. NetworkX*

NetworkX is a Python library for studying graphs and networks. NetworkX makes it easy to visualize graphs with large number of nodes and edges. In addition, the location (coordinate; ‘pos’) of nodes can be specified by the user, so it is possible to accurately plot the location of the stations using the latitude and longitude information obtained above. Also, it provides function which can calculate pagerank.

*E. Pagerank*

Pagerank, developed by Larry Page and Sergey Brin (founders of Google) is a method for rating the importance of web pages objectively and mechanically using the link structure of the web. Pagerank can be calculated by ‘Power iteration’.

1. Power iteration

Assume there is a graph with n nodes. First, an initial page rank is assigned each node. And it is repeated until convergence to equation (3).

(3)



Then when it converges, stop repeating and is final pagerank.



1. Weight estimation

Case1: Normal station

In normal station case, directed graph can be drawn as Fig6. Then, we focus on vector ba and bc that highlighted. We can estimate two values.

Ain and Aout means that number of passengers getting in and out at station A. We consider the weight using number of getting off passengers in both sides.

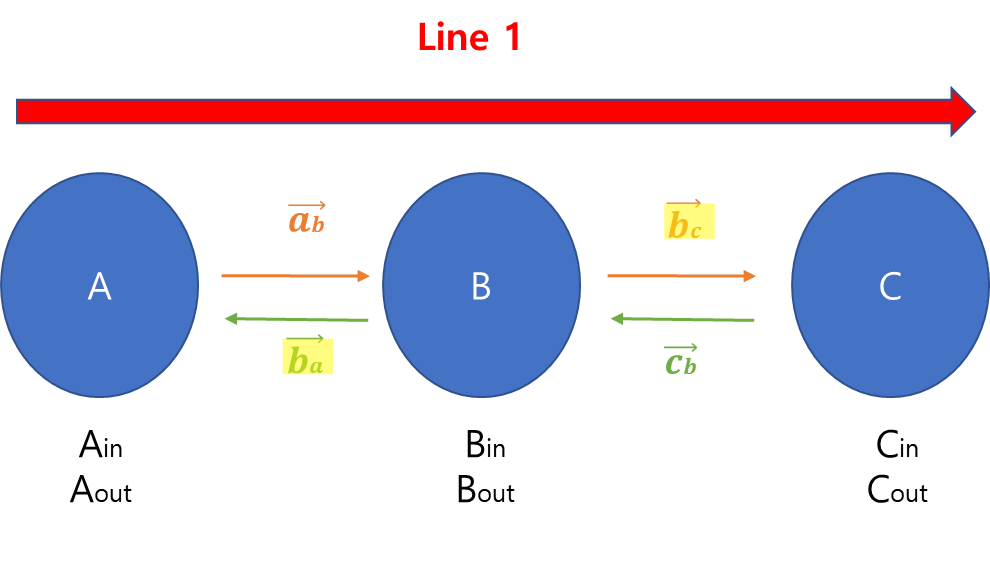
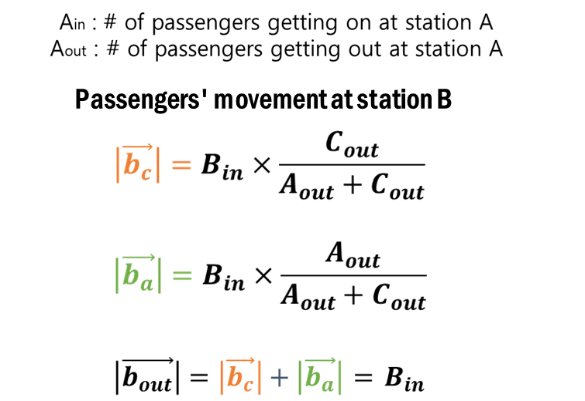


Fig6. The directed graph of normal station

We can estimate the passenger’s movement.

Ain and Aout means that number of passengers getting in and out at station A. The total amount of movement at station B is the same as Bin.

(4)

Equation 4 is the estimated weight equation in normal station.

Case2: Transfer station

In transfer station case, directed graph can be drawn as Fig8. Then, we focus on vector ba, bc, bd and be that highlighted. We can estimate four values.

We consider the weight using number of getting off passengers in both sides.

The total amount of movement at station B is the same as Bin.

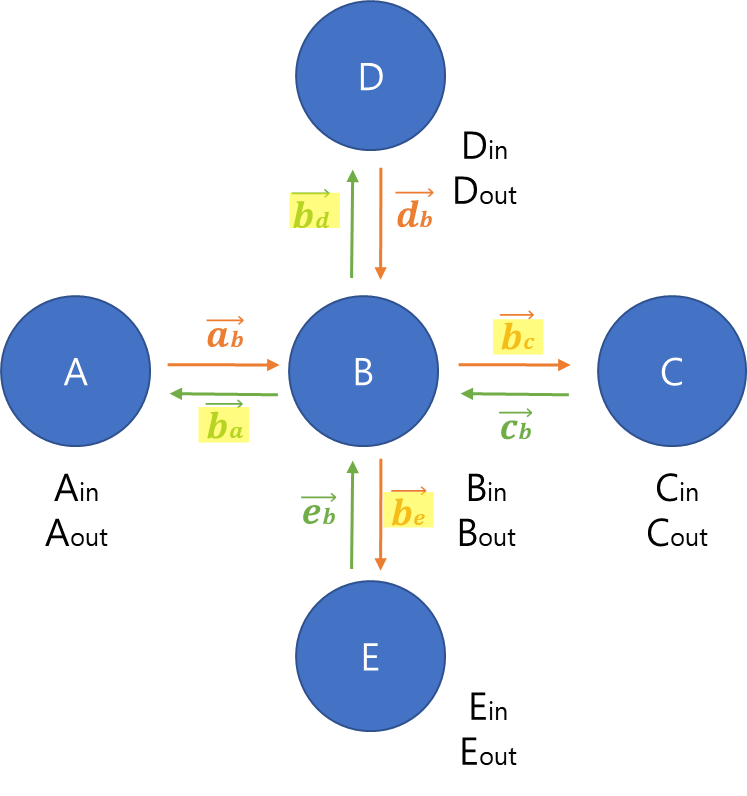
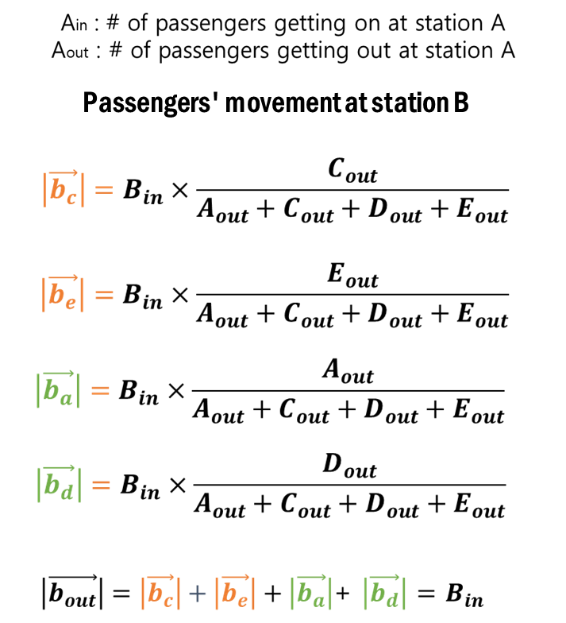


Fig7. The directed graph of normal station

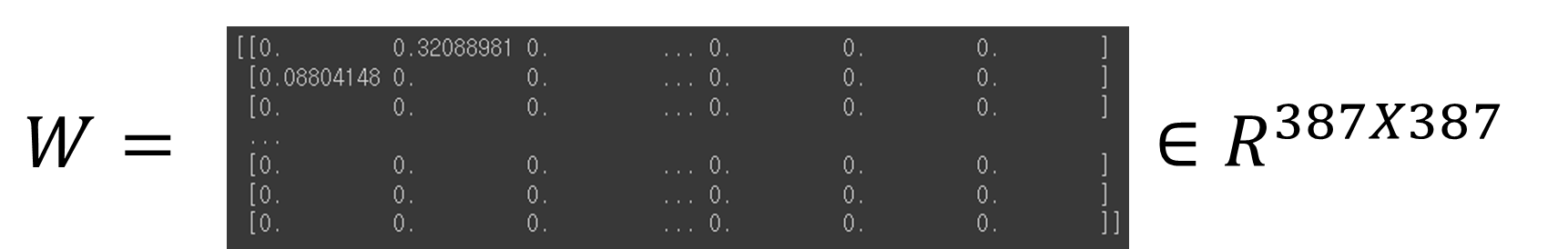
We can estimate the passenger’s movement as Equation 5.

 (5)

1. Weight matrix

We can calculate all weight through weight estimation. Now, we can derive weight matrix W. (Fig.8)

Dimension of W is (387,387) because the number of all station is 387.



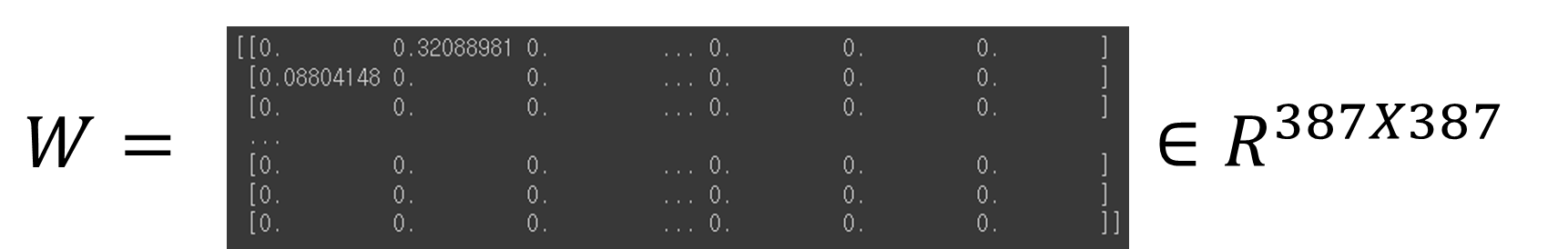
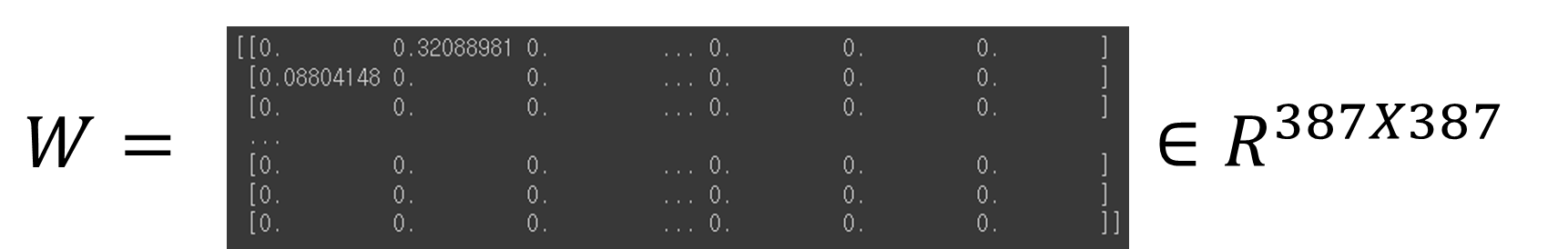
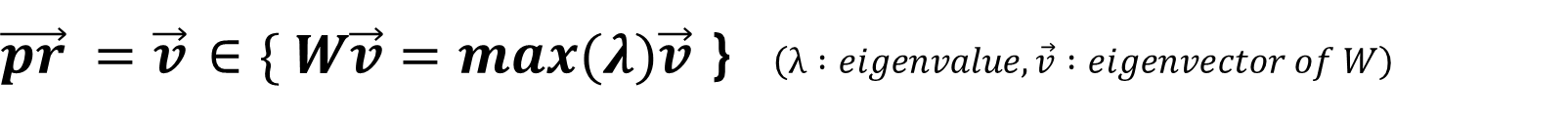


Fig8. Weight matrix for all station.

1. Weighted PageRank

We already know the power iteration method. But in our model, there are many edges and nodes. And the graph is very complicate. So, it does not converge. So, we can't derive the PageRank using power iteration.

So, I introduce alternative method for find PageRank using Eigenvector and eigenvalues of 𝑾. Another definition of PageRank is that eigenvector that maximum eigenvalue of 𝑾. The result maximizes eigenvalue method is almost same as iteration method.



(6)

Equation 6 is Weighted PageRank equation by maximizes eigenvalue.

Our PageRank result as Fig9.

The dimension of PageRank is (387,1).

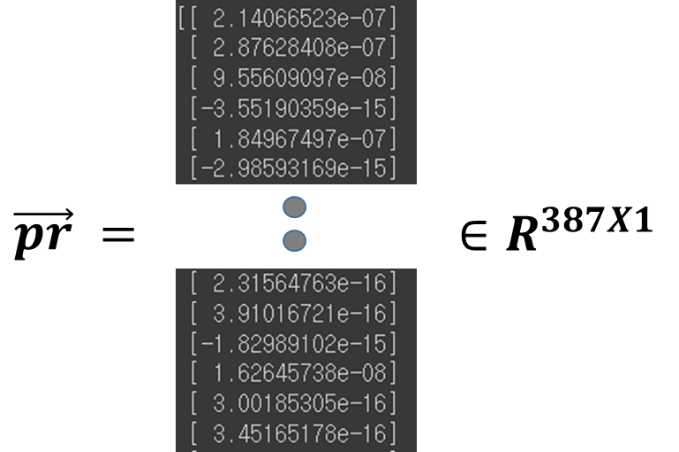


Fig9. Weighted PageRank results

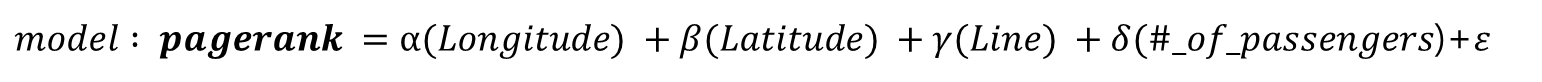
*F. Correlation Matrix*

Analyze the correlation between weighted PageRank and dependent variables. The dependent variables areLongitude,

Latitude, Line, Total number of passengers getting on, Total number of passengers getting off.

*G. Linear Regression (OLS)*

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable. Equation 7 shows that Weighted PageRank OLS equation.



(7)

IV. Application

First, we can find the importance of subway stations through pagerank and visualizing them to identify at a glance. After this, it is possible to focus on expanding or improving the congested transfer procedure in stations with high importance. In fact, stations such as ‘Sinchon’ and ‘Gangnam’, which are well known, have a wider and more interior than the outer stations(ex: in outskirts).

Second, we can find out lines that are packed with stations of high importance. If there are many stations with particularly high importance on a particular line, we can increase the number of subways running on that line. However, depending on the situation, only some stations on that line (ex: center of Seoul) may be concentrated, and in this case, additional subways that do not go to the end point and only run important stations can be operated for efficience. In fact, there are subways that run only to ‘Sadang’ Station, not to ‘Oido’ Station, which the last stop station of Line 4. This is because there are relatively more people getting on and off near ‘Sadang’ Station(in Seoul) than ‘Oido’ Station(the last stop of Line 4).

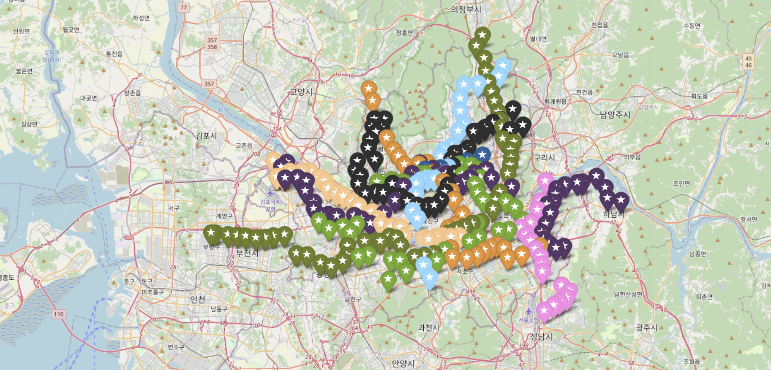
 Fig10. Example photo of a subway both operating to ‘Oido’ Station, the last stop of Line 4, and a train to ‘Sadang’ Station, the intermediate station

V. Results and Discussion

*A. Consider only transfer information (Neighbor nodes; pair of previous-next stations)*

** Fig11. Concatted dataframe of fig1 and fig3

This is data frame after completing (4) of the preprocessing mentioned in III. METHOD above.

 Fig12. Subway stations plotted on a real map

Using package 'Folium' provided by Python, user can plot the location he or she wants to find on Google Maps.We plotted the locations of subway stations on the actual map. We can find out that the latitude and longitude values we specified are correct.

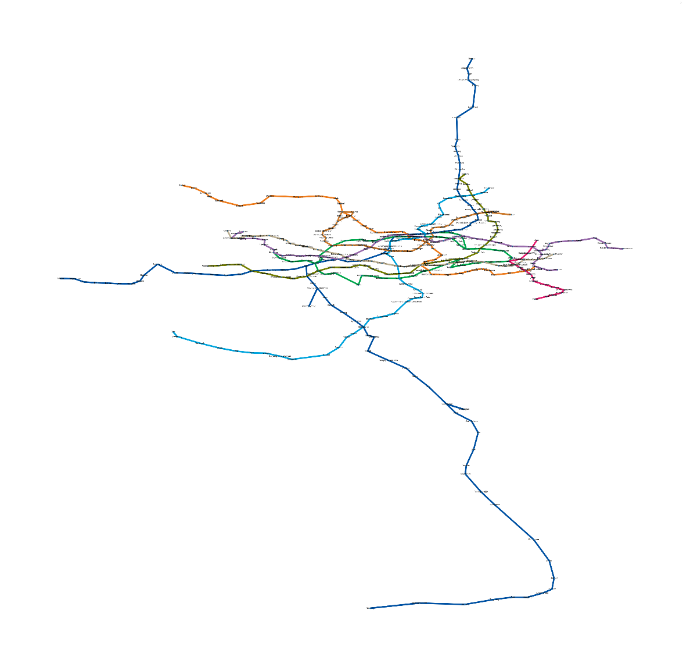
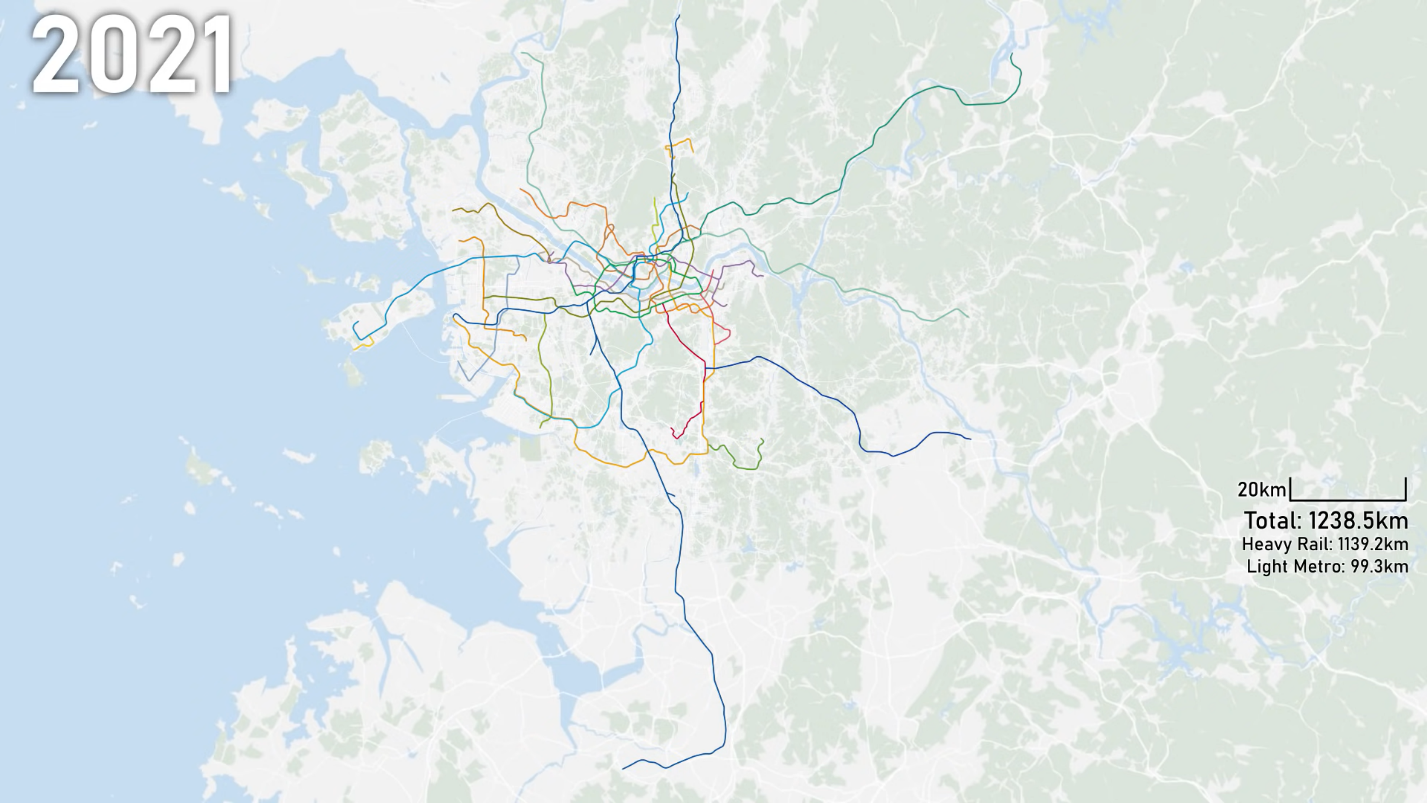


Fig13. Visualization of subway stations by differencing color by line

 Fig14. actual map containing subway lines

Comparing Figure13 and Figure14, they are similar. So our visualization and data are correct.

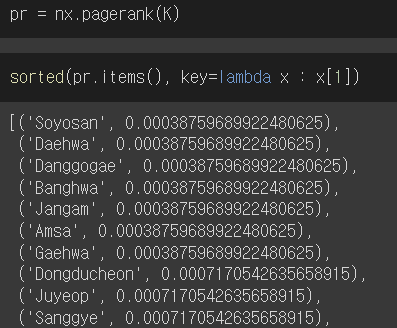
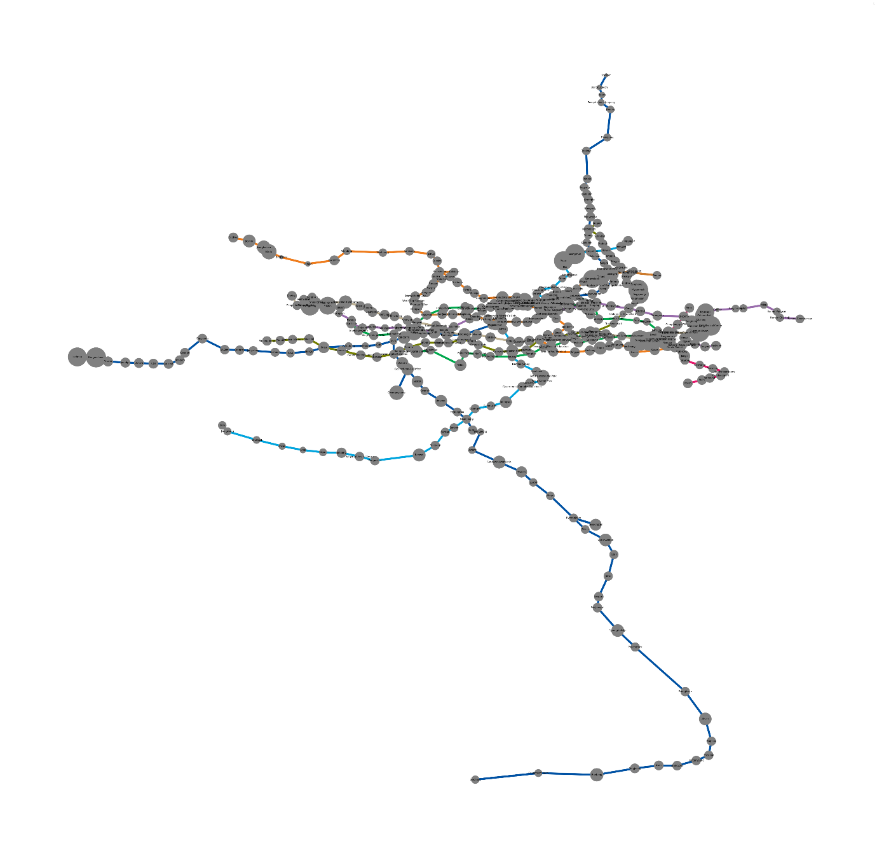


Fig15. Pagerank calculation

 Fig16. Pagerank visualization

Pagerank is calculated and visualized. But we can find somethings strange.

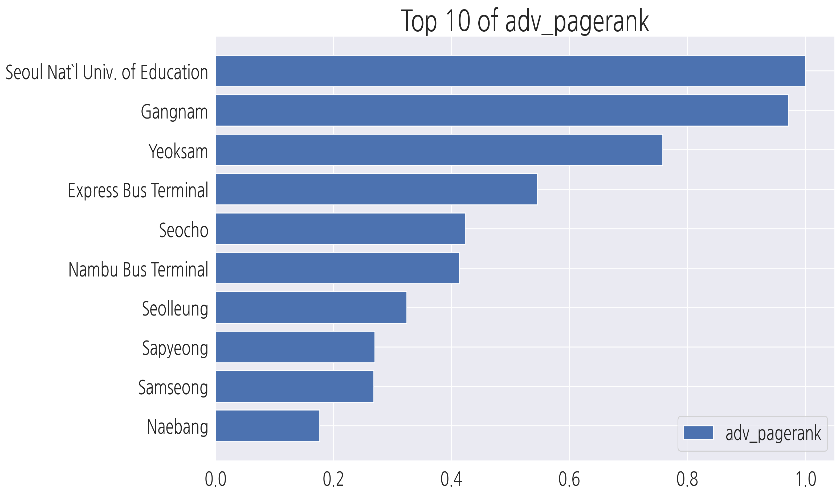
Figure15 shows in the order of stations where pagerank is high. ‘Soyosan’ and ‘Dangogae’ were ranked high even though they were not familiar to most people. In addition, ‘Sinchon’ and ‘Seoul station’, which are familiar stations to many people, which we often recognize to be of high importance, were even not included in the top 10. Plus, although ‘Gangnam’ was classified very important in the reviewed literature ‘Subway Stations Network Structure Analysis by Using Social Network Analysis’, this is not the case in our results.

Plus, according to our hypothesis, pagerank value of the station in central Seoul, not the outskirts, will be higher. However, in Figure16, there are many areas with high pagerank in stations located outskirts. This does not fit our hypothesis, and in reality, those stations are not recognized by people and are not importantly recognized.

We concluded that the above results were not properly calculated because only too little information was considered. Therefore, additional information that can be more accurate and directly related to importance should be further considered in calculating pagerank. Therefore, we searched more in "서울 열린데이터 광장" and were able to obtain data containing information on the number of people getting on and off by station(Figure 2). We thought that calculating the pagerank by also considering the number of people getting on and off would be more accurate because the actual traffic volume could be considered.

*B. Also consider information about people getting on and off*

Using weighted PageRank, we can sort the important stations. This is top 10 weighted PageRank results. The stations with high PageRank are same with those we commonly recognize as important, such as Express Bus Terminal and Gangnam. This result implies that our weight estimation is quite significant.

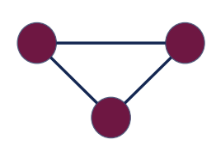
Fig17. Pagerank calculation

*C. Additional Analysis*

(1) Global Clustering Coefficient

(8)



 Fig18. Structure of closed triangle

We can derive Global Clustering Coefficient by equation 8. In subway network, there can’t be closed triangles because if there are some triangles, it circuits same stations. Since ‘number of triangles’=0, global clustering coefficients also becomes 0.

(2) Correlation Matrix

There is very strong correlation between Total # of passengers getting on and off.

There is slightly strong correlation as 0.35 between Total # of passengers getting on/off and weighted PageRank.

**

Fig19. Correlation Matrix

(3) Linear Regression (OLS)

Fig 14 shows that Weighted PageRank OLS results

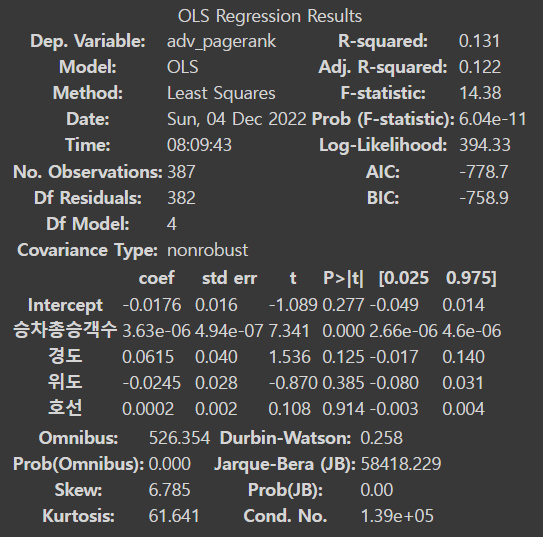


Fig20. PageRank OLS result

Only p-value of number of passengers are less than 0,05 It means that # of passengers is only significant.

VI. Conclusion

We visualized the metropolitan subway network using data on the relationship between subway stations in the metropolitan area and the data on getting on and off, and analysed it by calculating the PageRank. As a result of the Literature review, Gangnam Station and Jamsil Station showed high importance or centrality, and it is also highly recognized by most people. We conducted a task after establishing a hypothesis that the center of Seoul would be of high importance and that the transfer station would be of high importance. When PageRank was calculated using only pair information with surrounding stations, the results were different from both the Literature review and the hypothesis.

To solve this problem, we additionally estimated the weight using the data of the number of people getting on and off the subway. Using the estimated weight, the weighted PageRank was calculated using the maximum eigenvalue method. The top 10 PageRank included most of the stations frequented by people such as Express Bus Terminal and Gangnam. This means that the weight we estimated is appropriate and that the weight is well done. In addition, as a result of correlation analysis and OLS analysis, it was found that the weighted PageRank was somewhat proportional to the number of passenger's boarding and exiting.

References

1. Kyung Hoon Han, Su Seo, Geun Byung Park, “사회 네트워크 분석을 이용한 지하철역 네트워크 구조 분석”, 2015 년도 한국철도학회 춘계학술대회 논문집, no. 8, pp. 1-7, 2015
2. 서울 교통공사, “서울교통공사 노선별 지하철역 정보”, 2018.11.08., 서울 열린데이터 광장, <https://data.seoul.go.kr/dataList/OA-15442/S/1/datasetView.do>, 2022.11.28
3. 서울 교통공사, “서울교통공사 지하철 환승역 환승인원 정보”, 2015.02.27., 서울 열린데이터 광장, <http://115.84.165.39/dataList/OA-12033/S/1/datasetView.do>, 2022.12.01
4. Metro Liner, 2021. 6. 29., “지도로 보는 서울 수도권 전철 1974-2021”, <https://www.youtube.com/watch?v=eoyW-BMNnqg>, 2022.12.01
5. *Using Complex Numbers in Website Ranking Calculations: A Non-ad Hoc Alternative to Google’s PageRank - Keita Sugihara* <https://dl.acm.org/doi/abs/10.1145/775152.775190>