Network Analysis of Seoul Subway Network

이예준 최수환

2021년 1학기 신입생세미나 중간에세이

Table of Contents

1. Introduction

수도권 지하철은 수도권 지역의 핵심 교통수단으로서 자리매김하고 있다. 본 에세이에서는 수도권 지하철 네트워크의 중심성(centrality)을 다양한 방법를 통해 분석한다. (예정: 나아가, 지하철 승하차객 수 데이터와 연관 지어 네트워크 분석을 진행한다.)

https://github.com/Wittgensteinian/Seoul_Metro에서 본 에세이에 관한 모든 코드를 찾아볼 수 있다.

2. Preprocessing Network Data

수도권 지하철 네트워크 데이터는 http://gangwon.github.io/subway-data/로부터 받아왔다. 이 데이터는 수도권 지하철의 역명과 역번호, 각 역 간의 이동 시간에 대한 정보를 담고 있다. 이 데이터를 바탕으로, 두 가지 다른 그래프를 만들어 분석에 활용했다. 하나는 역명을 노드로, 하나는 역번호를 노드로 삼는 그래프다.

역명을 노드로 삼는 그래프의 경우, 우리가 떠올리는 지하철 노선도와 가장 유사한 그래프 구조를 생성할 수 있다는 장점이 있다. 그러나 역과 역 사이의 최단거리(정확히는 최단시간)를 구하기가 복잡하다는 단점이 있다. 구체적으로 이야기하자면, 역명을 노드로 삼는 그래프에서 계산하는 최단거리는 노선 간 환승 시간을 전혀 고려하지 못한다는 뜻이다. 예를 들면, 이 그래프에서는 2호선 서울대입구역부터 2호선 방배역까지의 거리와 4호선 남태령역까지의 거리가 동일하다(모든 노드 사이의 거리를 1이라 가정). 그러나 이는 환승 시간을 고려하지 않은 결과로, 실제 이동 시간과는 괴리가 있다.

그림 1: 사당역과 그 주변 역들

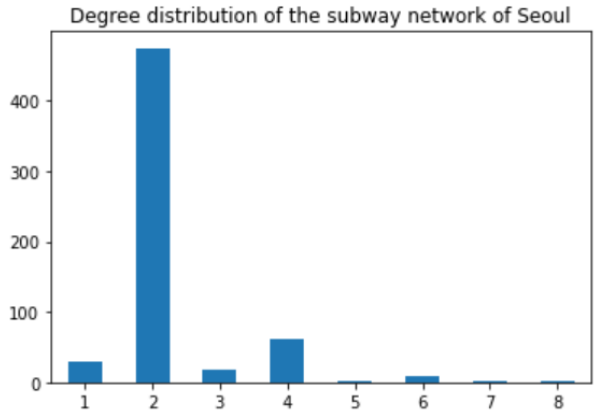
따라서 최단거리 계산을 위해 역번호를 노드로 삼는 그래프를 만들었다. 역번호를 노드로 삼는 그래프의 경우, 같은 역일지라도 노선이 다른 역(환승역)은 여러 개의 다른 노드로 취급한다. 같은 역이여도 노선이 다르다면 역번호 역시 다르기 때문이다. 예를 들면 2호선 사당역은 노드 '226'으로, 4호선 사당역은 노드 '433'으로 다르게 취급된다. 두 노드 사이의 거리에는 일괄적으로 '3'을 부여했다. 환승 시간을 3분으로 설정한 것과 같은 의미다.

본 에세이에서는 지하철 네트워크의 구조를 보존해야 하는 경우 역명을 기준으로, 최단거리를 계산해야 하는 경우 역번호를 기준으로 하는 그래프를 활용했다. 또한 거리를 계산해야 하는 경우 모든 엣지(edge)의 거리에 1을 부여하는 네트워크 거리(network distance)와 실제 시간(physical time)을 같이 활용했다.

3. Centrality Analysis

A. Degree Centrality

Degree란, 해당 노드와 인접한 노드의 개수를 의미한다. 역명을 기준으로 하는 그래프에서 총 노드의 개수는 598개이며, degree distribution은 다음과 같다. Degree centrality가 가장 높은 노드(degree = 8)는 공덕이다. 두번째로 높은 노드(degree = 7)는 왕십리다.



B. Closeness Centrality

$$c_{Cl}(v) = \frac{1}{\sum_{u \in V} \mathrm{dist}(v,u)}$$

Closeness centrality는 해당 노드에서 다른 노드까지의 거리의 합에 반비례하는 값으로 주어진다. 네트워크 거리를 따를 때는 이촌역(경의중앙선), 실제 시간을 따를 때는 서울역(4호선)이 제일 높은 centrality를 가진다.

Closeness centrality는 eccentricity를 활용하는 방법으로도 변형될 수 있다. Eccentricity란, 해당 노드에서 다른 노드까지의 최단거리의 최댓값이다. 따라서 closeness centrality를 eccentricity에 반비례하는 값으로 설정할 수 있다. 네트워크 거리를 따를 때는 노량진(1호선), 동작(4호선), 총신대입구(이수)(4호선), 남태령(4호선), 실제 시간을 따를 때는 총신대입구(이수)(4호선)가 제일 높은 centrality를 가진다.

Eccentricity의 최댓값을 그래프의 지름(diameter), 최솟값을 그래프의 반지름(radius)이라고 부른다. (Lee et al., 2008) 네트워크 거리를 따를 때는 지름이 75, 반지름이 39이며, 실제 시간을 따를 때는 지름이 242, 반지름이 121이다.

C. Betweenness Centrality

$$c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Betweenness centrality는 두 노드를 잇는 최단 거리 중 해당 노드를 거치는 최단 거리의 수로, centrality를 구하는 대표적인 방법 중 하나다. 네트워크 거리를 따를 때는 왕십리(경의중앙선), 실제 시간을 따를 때는

구로(1호선)가 제일 높은 centrality를 가진다.

D. Eigenvector Centrality

$$c_{Ei}(v) = \alpha \sum_{\{u,v\} \in E} c_{Ei}(u)$$

Eigenvector centrality는 해당 노드와 인접한 노드의 중심성이 높을수록 높다. Eigenvector centrality의 경우, 각 edge의 weight를 실제 시간의 역수로 설정했다. Eigenvector centrality는 weight가 높은 edge의 연결성을 높이 평가하기 때문이다.

네트워크 거리를 따를 때는 공덕(공항철도), 실제 시간을 따를 때는 효자(의정부선)가 제일 높은 centrality를 가진다.

Eigenvector centrality를 변형한 centrality로는 Katz centrality와 PageRank가 있다. Katz centrality의 경우 네트워크 거리를 따를 때는 공덕(공항철도), 실제 시간을 따를 때는 왕십리(2호선)에서 제일 높다. PageRank의 경우 네트워크 거리를 따를 때는 왕십리(5호선), 실제 시간을 따를 때는 디지털미디어시티(6호선)에서 제일 높다.

E. Abnormal results of eigenvector centrality

Eigenvector centrality와 그 변형을 실제 거리에 따라 구할 경우 예상과 다른 결과가 관찰된다.

이는 eigenvector centrality에서 가장 두드러진다. Eigenvector centrality가 가장 높은 10개의 역은 **모두 의정부선**이다. 11번째가 되어서야 왕십리(5호선)가 등장한다. 의정부선이 서울 외곽에 위치해 있으며, 환승역도 단 하나라는 점을 고려하면 이상한 결과다.

그림 2: 의정부선 노선도

Katz centrality가 가장 높은 100개의 역 중 11개의 역이 인천2호선이다. 인천2호선이 서울 외곽에 위치한, 총 27개의 역과 2개의 환승역을 가진 노선임을 고려하면 이상한 결과다. 또한, centrality가 8번째로 높은 역(효자)과 9번째로 높은 역(경기도청북부청사)이 의정부선이다.

PageRank는 앞선 두 개의 centrality보다 더 예상 가능한 결과를 내놓는다. 그러나 centrality가 가장 높은 100개의 역 중 무려 21개의 역이 1호선이다.

각 centrality의 분포 양상을 살펴보자. Eigenvector centrality의 경우, 의정부선을 제외할 경우 centrality는 왕십리를 중심으로 점차 낮아지는 형태를 띤다.

Katz centrality와 PageRank는 비슷한 양상을 띤다. 어느 한 곳의 중심을 가지는 대신 centrality가 높은 역들이 이곳저곳에 분포한다. 이때 눈에 띄는 점은, 서울의 중심부에서 살짝 떨어진 역들 중 centrality가 높은 역이 꽤 있다는 점이다. 대표적인 역은 디지털미디어시티, 상봉, 도화다.

이와 같은 결과가 나오는 이유로 크게 두 가지를 생각해보았다. 첫째로는 지하철 네트워크 구조의 특성, 둘째로는 가중치다.

첫째로 지하철 네트워크 구조는 다른 네트워크 구조와 유의미하게 다르다. 노드가 직선적으로 연결되어 있는 부분이 많고, degree가 높은 노드들이 한 곳에 집중되어 있는 경향이 있다. 지하철 네트워크에서의 eigenvector centrality는 un-weighted edge로 계산된 적 있다 (Takadama et al., 2007). Weighted edge로도 계산된 적이 있으나, 그 결과가 일반적인 예상과 정확히 일치하진 않았다. (Majima et al., 2007)

둘째로 가중치(weight)가 부여된 네트워크에서 eigenvector centrality의 성능을 생각해 볼 필요가 있다. 실제로 본 연구에서는, eigenvector centrality와 그 변형들이 network distance를 기준으로 했을 때는 예상에 부합하는 결과를 내놓았다. 다만 선행 연구에서는, weight를 부여하더라도 여전히 eigenvector centrality가 잘 기능한 것으로 보인다. (Bihari & Pandia, 2015; Newman, 2004)

References

Bihari, A., & Pandia, M. K. (2015). Eigenvector centrality and its application in research professionals' relationship network. 2015 international conference on futuristic trends on computational analysis and knowledge management (ABLAZE),

Lee, K., Jung, W.-S., Park, J. S., & Choi, M. (2008). Statistical analysis of the Metropolitan Seoul Subway System: Network structure and passenger flows. *Physica A: Statistical Mechanics and its Applications*, *387*(24), 6231-6234.

Majima, T., Katuhara, M., & Takadama, K. (2007). Analysis on transport networks of railway, subway and waterbus in Japan. In *Emergent Intelligence of Networked Agents* (pp. 99-113). Springer.

Newman, M. E. (2004). Analysis of weighted networks. *Physical review E*, *70*(5), 056131.

Takadama, K., Majima, T., Watanabe, D., & Katsuhara, M. (2007). Exploring quantitative evaluation criteria for service and potentials of new service in transportation: Analyzing transport networks of railway, subway, and waterbus. International Conference on Intelligent Data Engineering and Automated Learning,