그래프 이론과 중심성: 주요 지표와 응용에 대한 종합적 고찰

이태홍[1], 오형국[2], 노영태[3]

한국에너지공과대학교[1], 국방과학연구소[2], 한양대학교[3]

etehong@kentech.ac.kr[1], youngtaenoh@hanyang.ac.kr[3]

GraphTheoryandCentrality:AComprehensiveOverviewofKeyMeasuresand Appplications

TaeHong Lee, Hyungkook Oh, Youngtae Noh

요 약

본 논문은 그래프 상에서 각각의 노드의 중요도를 평가하는 기준인 Centrality에 대한 주요 개념과 지표들을 종합적으로 고찰하고, 이들의 비교 분석 및 다양한 응용 분야에서의 활용을 탐구한다. 먼저, Degree Centrality, Closeness Centrality, Betweenness Centrality, PageRank Centrality 등 대표적인 Centrality 지표들을 소개하고, 각 지표들의 이론적 배경과 계산 방법을 소개한다. 이후, 이러한 지표들 간의 상관관계 및 장단점을 비교 분석한다. 마지막으로, 기존의 계산 방식과는 달리 심층 강화학습을 통해 도출된 새로운 Centrality 지표를 제시하며, Centrality 측정의 잠재적 확장 가능성을 논의한다.

Ⅰ. 서 론

그래프 내의 노드들은 그래프의 구조에 따라 여러 가지 역할을 수행 할 수 있으며 이 역할들을 얼마나 잘 수행하는지 평가하는 대표적인 지표로써 여러가지 Centrality가 있다. 각 노드의 Centrality를 조사하는 것은 그래프 전체의 특성을 파악하는데 결정적인 역할을 하며 일반적으로 여러 가지 지표를 동시에 활용해 노드의 중요도를 평가하게 된다. 노드의 Centrality로는 Degree Centrality, Closeness Centrality, Betweenness Centrality, PageRank Centrality 등 다양한 Centrality가 존재하며, 이러한 지표들을 통해 그래프 상에서 각 노드들이 어떠한 역할을 하는지 판단할 수 있다. 그러나 이러한 Centrality를 구하기 위해서는 해당 노드와 다른 노드들과의 관계를 탐색한 후 구하는 것이 일반적이며 보통의 경우 NP-Hard 문제로 분류된다. 실제로 각각의 Centrality를 구하는데 상당한 연산시간이 필요하며 그래프 크기가 커질 수록 요구되는 메모리 자원은 비선형적으로 증가하게 된다. 최근 인공신경망연구과 심층 강화학습은 이러한 연산비용을 크게 줄여주는데 일조하고 있다. 신경망의 특성상 훈련을 시키는데에 드는 연산 비용은 크지만 훈련이 완료된 후에는 간단한 연산을 통해 결과를 바로 받아볼 수 있다는 장점이 있다. 본 연구에서는 기존의 Centrality와 더불어 심층강화학습으로 얻어낸 Centrality의 비교를 통해 각자의 장단점에 대해 조사한다.

Ⅱ. 본론

2-1DegreeCentrality

Degree Centrality는 수식으로 $C_D(v) = \deg(v)$로 표현하며 이는 노드v에 연결된 연결선 수로 정의된 중심도이다. Degree Centrality는 가장 단순한 Centrality임과 동시에 네트워크를 이해하는데 가장 직관적인 지표를 준다. 사회연결망에서 친구나 지인이 많은 인원은 그렇지 못한 인원들보다 확실히 더 영향력이 있고, 논문 인용 네트워크에서 높은 인용수를 기록한 논문은 상대적으로 적은 인용수를 가진 논문보다 더 큰 학술적 영향력을 행사하기 때문이다. 또한, Degree가 다른 노드에 비해 압도적으로 높은 Hub 노드는 그래프의 평균 최단경로를 줄이는 데 중요한 역할을 할 수 있다[1]. 이러한 특징은 Degree의 분포가 거듭제곱분포를 지니는 Scale-Free Graph에서 주로 일어난다. 이러한 그래프에서는 대부분의 노드가 낮은 Degree를 가지지만, 소수의 노드는 매우 높은 Degree를 가지게 되며, 이러한 Hub 노드는 그래프에서 다른 노드들 간의 경로를 단축시켜 준다. Degree가 낮은 노드라고 하더라도 Hub 노드에 연결되어 있을 확률은 높으며 Hub 노드를 통해 다른 노드까지 짧은 거리로 이동할 수 있다.

2-2 Closeness Centrality

특정 노드의 Closeness Centrality는 그 노드에서 다른 노드까지 평균 거리를 측정하는 지표이다. $d(v,w)$가 노드 v에서 w까지의 최단 거리일 때 총 노드 개수가 n개인 그래프에서 특정 노드 v에서 그래프상의 다른 모든 노드로 가는 평균 최단 거리는 $\frac{1}{n}\sum_{w} d(v,w), w \in V$ 이다. 다른 노드들과의 평균적으로 짧은 거리에 위치해 있을수록 이 값은 작아진다. 그렇기 때문에 이 값의 역수를 취해 근접 중심도를 구하며 정확한 수식은 아래와 같다.

$$C_C(v) = \frac{1}{\frac{1}{n}\sum_{w} d(v,w)}, w \in V.$$

정보의 확산 네트워크를 생각해 보았을 때 Closeness Centrality 값이 높다는 것은 해당 노드가 평균적으로 정보를 더 빨리 퍼트릴 수 있다는 것을 의미한다. Closeness Centrality를 계산할 때 문제점으로는 그래프가 여러개의 Cluster로 나누어져 있을 이 값을 구하기 힘들어 진다는 점이 있다. 특정 노드에서 다른 노드로의 경로가 없을 때 관례적으로 그값을 무한대로 취급하며 Cluster가 나누어져 있는 그래프의 경우 모든 노드의 Closeness Centrality의 분모항은 무한대로 취급받고 결국 그 역수는 0이 되어 버리기 때문에 Closeness Centrality의 의미가 없어진다. 이를 해결 하기 위한 방법으로는 각각의 Cluster에 속해있는 노드들과의 최단 거리들의 조화평균 거리로써 Closeness Centrality를 다시 정의하는 방법이 있다.

$$C_C = \frac{1}{n-1}\sum_{w(\neq v)} \frac{1}{d(v,w)}.$$

2-3 Betweenness Centrality

Betweenness Centrality의 경우 그 노드가 다른 노드가 얼마나 잘 연결되어 있는지를 측정하기 보다는 그 노드가 다른 노드들의 최단 경로에 얼마나 많이 속해있는지를 측정하는 지표이다. $\sigma_{ij}(v)$를 노드 i에서 노드 j로 가는 최단경로 중 노드 v를 거쳐 가는 경로의 수 라고 하고 σ_{ij}는 노드 i에서 노드 j로 가는 최단경로 전체 수라고 하면 노드 v의 Betweenness Centrality는 다음으로 정의 한다.

$$C_B(v) = \sum_{i \neq v \neq j} \frac{\sigma_{ij}(v)}{\sigma_{ij}} .$$

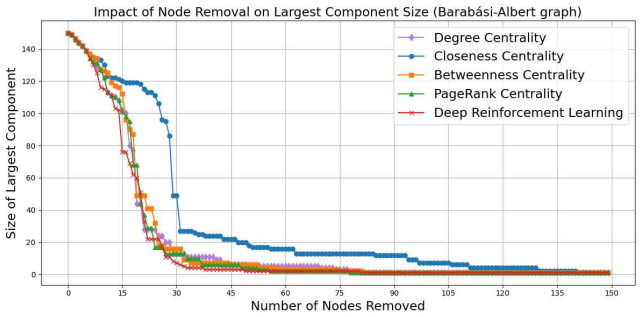
$\sigma_{ij}(v)$와 σ_{ij}가 모두 0인 경우 $\frac{\sigma_{ij}(v)}{\sigma_{ij}}$는 0으로 정의한다. Betweenness Centrality의 경우 그 노드를 지나는 최단경로의 수가 많으면 많을수록 그 값은 높아지고 이는 네트워크가 그룹으로 나눠져 있는 경우 두드러진다.

2-4 PageRank Centrality

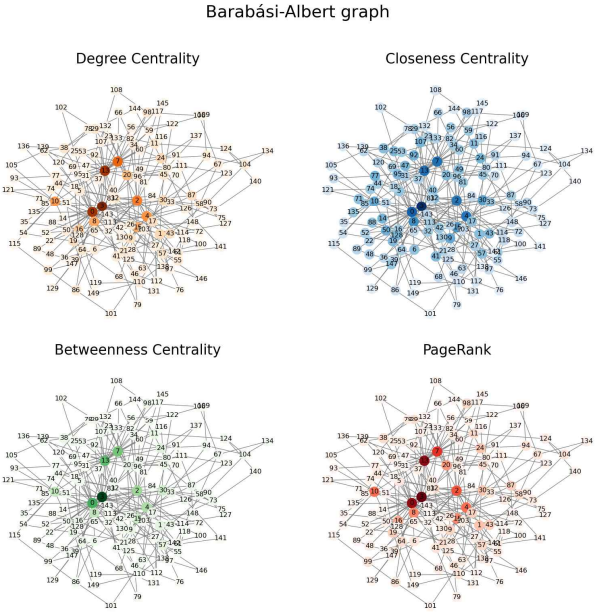
PageRank Centrality는 구글의 공동 창립자인 래리 페이지(Larry Page)와 세르게이 브린(Sergey Brin)에 의해 1996년에 제안된 알고리즘이다[2]. 이 알고리즘은 구글 검색 엔진의 핵심이 되었으며, 웹 페이지의 중요성을 평가하고 순위를 메기는 데 사용된다. PageRank의 기본 아이디어는 권위가 높은 노드가 가르키는 노드는 권위가 낮은 노드가 가르키는 노드보다 더 높은 권위를 갖는다는 점이 있다. 즉 어떠한 노드가 가지고 있는 PageRank 점수를 그 노드가 가르키는 노드의 개수만큼으로 나눈 값을 그래프 상의 다른 노드들로 전파하는 과정을 거쳐 수렴하는 값을 PageRank값으로 정하게 된다. PageRank를 구하는 대표적인 두가지 방법으로는 Random Surfer 알고리즘과 행렬의 고윳값을 통해 구하는 두 가지 방법이 있다. 여기서는 행렬의 고윳값을 통해 PageRank를 구하는 방법을 소개한다. 일단 그래프의 인접행렬 A로부터 열정규화된 인접행렬 H를 구한다. H의 j번째 열을 H_j라고 했을 때 $H_j = A_j / \sum_{k=1}^{n} A_{kj}$로 정의된다. 이제 H로부터 다음과 같이 정의된 stochastic행렬 $S = H + \frac{e \cdot a^T}{n}$을 구한다. 여기서 벡터$e$는 모든 성분이 1인 열벡터이고, a는 $\sum_{i=1}^{n} H_{ij} = 0$이면 $a_j = 1$이고 아니면 $a_j = 0$인 열벡터이다. S로부터 구글행렬 $G = mS + (1-m)E$을 유도한다. 여기서 m은 $0 \leq m \leq 1$ 이고 $E = \frac{e \cdot e^T}{n}$이며 보통 $m = 0.85$을 사용한다. 거듭제곱법을 사용하여 G의 가장 큰 고윳값을 구하게 되고 그 값이 바로 각 노드의 PageRank가 된다. [그림 1]은 총 노드 개수가 150개이며 새로 추가되는 연결선수 $(m = 2)$인 Barabási-Albert 그래프에서 각각의 Centrality들을 나타낸다.

2-3 심층강화학습

위에서 설명한 방법들은 모두 컴퓨터를 통해 계산하게 되며, 상당한 컴퓨터 자원을 요구하게 된다. 특히 대규모 그래프에서 Centrality를 계산하는 연산의 경우 현실적으로 불가능한 경우도 존재한다. 인공신경망과 강화학습의 발전으로 그래프 자체의 특성을 심층강화학습을 통하여 학습하는 방식[3]이 연구되고 있다. 심층강화학습의 장점이라고 한다면 한번 학습을 마친 이후 Centrality를 구하는데 요구되는 연산량이 현저하게 적다는데에 있다. [그림 2]는 총 노드 개수가 150개이며 새로 추가되는 연결선수$(m = 2)$인 Barabási-Albert 그래프에서 각각의 Centrality 순서대로 노드를 제거하여 Giant Component Size의 변화를 조사한 그림이다.



[그림 1] Centrality별 노드 제거를 통한 Giant Component Size의 변화



[그림 2] Barabási-Albert 그래프상에서 각각의 Centrality

Ⅲ. 결론

본 논문에서는 각 Centrality를 구하는 방법과 장단점에 대해 알아보았다. 특히 주목할 점은 심층강화학습을 통해 얻어진 Centrality가 기존의 Centrality 지표들을 충분히 대체할 수 있을 정도로 뛰어난 성능을 보여주었다는 점이다. 심층강화학습으로 얻은 Centrality는 그 활용도와 잠재력이 충분하다고 보여지며 향후 그래프 분석 분야에서 심층강화학습 기반 방법론이 중요한 역할을 할 수 있음을 시사한다.

ACKNOWLEDGMENT

이 논문은 방위사업청의 재원으로 국방과학연구소의 지원을 받아 수행된 연구임 (411JJ5-912967201)

참 고 문 헌

[1] Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world' networks." Nature 393.6684 (1998): 440-442.

[2] Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine." Computer networks and ISDN systems 30.1-7 (1998): 107-117.

[3] Fan, Changjun, et al. "Finding key players in complex networks through deep reinforcement learning." Nature machine intelligence 2.6 (2020): 317-324.