ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

실험계산을 통한 에지 한 개 추가에 따른 그래프의 중심성 및 순위 변화 분석

Effect Analysis of an Additional Edge on Centrality and Ranking of Graph Using Computational Experiments

한 치 근[1*] 이 상 훈[1]
Chi-Geun Han Sang-Hoon Lee

요 약

그래프에서 각 노드에 대해 그래프 내의 중요도를 나타내는 중심성(centrality)을 계산할 수 있고, 그 값에 따라 각 노드는 중요도 순위(ranking)를 갖는다. 중심성을 나타내는 방법으로는 여러 척도가 있는데, 본 연구에서는 연결도(degree) 중심성, 밀접도(closeness) 중심성, 특성벡터(eigenvector) 중심성, betweenness 중심성에 국한하여 연구를 수행하였다. 본 연구는 그래프에서 에지를 하나 추가할 경우, 그래프 내 노드 전체에 미치는 노드의 중심성 및 순위의 변화를 실험계산을 통해 확인한다. 그리고, 추가되는 에지가 노드 전체의 중심성 및 순위에 미치는 영향은 그래프의 형태에 따라 달라진다는 것을 PCA(Principal Component Analysis)를 통해 밝혔다. 이 사실은 그래프의 구조적 특성을 구분하는 방법으로도 사용될 수 있다.

☞ 주제어 : 그래프, 중심성, 순위, 커뮤니티

ABSTRACT

The centrality is calculated to describe the importance of a node in a graph and ranking is given according to the centrality for each node. There are many centrality measures and we use degree centrality, closeness centrality, eigenvector centrality, and betweenness centrality. In this paper, we analyze the effect of an additional edge of a graph on centrality and ranking through experimental computations. It is found that the effect of an additional edge on centrality and ranking of the nodes in the graph is different according to the graph structure using PCA. The results can be used for define the graph characteristics.

☞ keyword : centrality, ranking, graph, community

1. 서 론

세상에 존재하는 상호간의 관계를 나타내기 위해 그래프를 이용한다. 각 노드는 상호관계의 주체를 나타내고, 상호간에 관계가 있을 경우 에지로 나타낸다(방향성이 없다고 가정). 주체는 인간이 될 수도 있고, 사이버 공간에서는 홈페이지가 될 수도 있다. 또한 동일한 연구 분야의 참고문헌, 이메일 시스템의 계정이 될 수도 있다.

이러한 상호관계에서 어느 노드는 다른 노드에 비해 많은 관계를 갖게 되고, 상대적으로 중요 노드로 인식될 수 있다. 예를 들어, SNS 관계도에서 사회적으로 중요한 역할을 하는 사람에게는 많은 사람들이 연결되어 있고, 학문 분야의 주요 논문은 타 논문이 인용하는 횟수가 많다는 것을 확인할 수 있다. 따라서, 상호관계를 나타내고 있는 그래프에서 각 주체(노드)의 상대적인 중요성을 나타내는 방법에 대한 많은 관심이 있어 왔다. 노드의 상대적인 중요도는 노드의 순위로 나타내는데, 이 순위를 결정하는 중심성(centrality)을 측정하는 방법으로는 연결도(degree) 중심성, 밀접도(closeness) 중심성, 특성벡터(eigenvector) 중심성, betweenness 중심성, Lin's 중심성, Harmonic 중심성, Seeley's 중심성, Katz's 중심성, PageRank, HITS, SALSA 등 다양한 중심성 지표가 제안되었다[1]. 중심성 지표는 노드들의 순위를 결정하게 되는데, 중요 노드일수록 높은 순위를 갖게 된다.

상호관계를 나타내는 그래프는 정적이라기보다는 동적으로 변하는 것이 일반적이다. 사이버 공간에 있는 홈페이지들의 상호 연결 상태도 새로운 홈페이지가 만들어

[1] Dept. of Computer Engineering, Kyung Hee University, Gyeonggi-do, 446-701, Korea.
* Corresponding author (cghan@khu.ac.kr)
[Received 9 March 2015, Reviewed 30 March 2015(R2 23 June 2015), Accepted 18 August 2015]

지거나, 기존의 홈페이지의 구조가 바뀜에 따라 지속적으로 변경된다. 또한 SNS 상에서도 새로운 관계가 맺어질 때 마다 그래프의 구조가 바뀌게 된다. 그래프의 구조가 바뀌게 되면, 그래프 내의 노드의 순위(중요도)도 바뀌게 된다.

그래프의 특성을 나타내는 방법으로는 노드의 연결도 분포, 에지의 밀도, 두 노드간의 최장 거리 등이 사용될 수 있다. 그 값이 크고, 작은 것에 따라 그래프의 유사도를 설명할 수 있는데, 본 연구는 노드의 중심성, 순위의 값이 추가 에지에 따라 변경되는 양상이 그래프의 종류에 따라 다르다는 것을 확인한다. 그리고, 그 양상이 다르다는 사실이 그래프의 종류를 구분할 수 있는 척도로 활용될 수 있는지를 연구하는 목적을 갖고 있다.

연구 방법으로 그래프에 추가적인 에지가 한 개가 생성되었을 때 전체 노드의 중심성 및 순위에 어떠한 영향을 미치는지 실험계산을 통해 분석하고, 추가 에지 한 개가 미치는 영향은 그래프의 구조에 따라 확연히 달라진다는 것을 밝힌다. 주어진 그래프에서 추가적인 노드나 에지는 현재의 그래프의 구조, 성질을 변하게 한다. 본 연구는 추가 에지의 효과만을 도출해 내기 위해, 추가 노드는 고려하지 않고, 추가 에지에 따라 그 변하는 정도를 노드의 중심성, 순위 변화를 통해 관찰하고자 한다. 현실적인 의미로는 웹 또는 커뮤니티에서 순위 결정을 교란시킬 수 있는 링크 스패밍(link spamming)에 대해 안정되게 웹 시스템 또는 커뮤니티를 유지할 수 있는 기본 전략을 제시할 수 있다[2]. 또한 새로운 링크의 생성이 웹 시스템에 도움이 되는지를 미리 예상할 수 있다[3].

본 연구에서는 연결된 무방향 무가중치 단순(connected undirected unweighted simple) 그래프를 가정한다. 2장에서는 관련 연구를 설명하고, 3장에서는 에지 추가가 그래프 전체의 중심성 및 순위에 어떤 영향을 주는지 확인하고, 그 영향은 그래프형태에 따라 다르다는 것을 설명한다. 4장에서 결론을 맺는다.

2. 관련 연구

사회적인 관계 내에서 중요한 역할을 하는 사람은 누구인가, 한 학문분야에서 해당 분야에 가장 많은 영향을 주는 논문 또는 연구자는 누구인가? 등 사회학 분야에서 활동 주체의 중요도를 파악하고자 하는 노력은 이전부터 많이 진행되어 왔다. 중심성은 노드의 중요도를 나타내는 정량적인 지표이며, 중심성을 나타내는 지표로는 크게, 기하학적인(geometric) 지표, spectral 지표, 경로기반 지표로 나눌 수 있다[1]. 기하학적인 지표로는 연결도(degree) 중심성, 밀접도 (closeness) 중심성, Lin's 중심성, Harmonic 중심성 등이 있고, spectral 지표로는 left dominant 특성벡터를 이용한 방법, Seeley's 중심성, Katz's 중심성, PageRank, HITS, SALSA 등이 있다. 그래프의 경로를 기반으로 한 지표로는 betweenness 중심성 등이 있다[1][4].

Borgatti et al.은 그래프의 일부 데이타가 불확실할 경우, 연결도, 밀접도, 특성벡터, betweenness 중심성이 어떻게 변화하는지를 분석하였는데, 노드/에지의 추가/삭제를 다양한 비율로 변경하면서 네 가지의 중심성이 변화하는 형태를 연구하였다[5]. Segarra와 Ribeiro는 네 가지의 중심성을 이용하여, 각 중심성의 안정성(stability)을 정의하고, 연속성을 분석하였다. 그리고, 그래프의 일부 구성에서 변화가 있을 때 중심성이 어떻게 변화하는지를 분석하였다[6].

경기 승패의 결과에 따라 팀의 순위를 결정하는 알고리즘이 존재한다. Chartier et al.은 완전한(perfect) 시즌인 경우 선형대수 기반인 Colley, Massey, Markov 방법이 일부 게임의 승패가 달라졌을 때 전체 팀의 순위에 어떤 영향을 주는지, 즉, 이 순위결정 방법들의 민감도(sensitivity) 분석을 수행하였다[7].

Ng et al.은 HITS와 PageRank 방법에 대해 일부 연결 정보가 달라졌을 때 순위가 어떻게 달라지는가와 이 방법이 안정적이기 위한 조건을 제시하였다[8]. Lempel과 Moran은 일부의 연결 정보가 바뀌었을 때 동일한 순위를 제공하는 순위-안정성(rank-stability)과 순위를 결정하는 방법들이 같은 자료에 대해 유사한 순위를 제공하는 순위-유사도(rank-similarity) 측면을 HITS, PageRank, SALSA 방법에 대해 분석하였다[2]. 또한 PageRank 방법으로 순위를 정할 경우, 웹페이지간의 관계에서 링크가 추가되었을 때 전체 순위에 미치는 영향을 분석한 연구가 있다[3].

하나의 중심성 지표에서 중요 노드로 판정된 노드는 다른 지표에서도 중요 노드로 확인될 가능성이 매우 높으므로 다양한 중심성 사이에는 밀접한 연관관계가 있다. [4]에서는 여덟 개의 중심성 지표간의 상관계수를 계산하여 그래프 종류에 따라 상관계수가 서로 다르다는 것을 확인하였고, PCA를 이용하여 상관계수들의 관계가 그래프 구조의 특성을 대변하므로, PCA 결과가 중심성 상관계수 프로파일(centrality correlation profile)로 활용할 수 있다고 하였다.

[5]와 [6]은 각각 그래프 연결정보가 달라졌을 때 순위가 어떻게 달라지는 지를 연구한 반면, 본 연구에서는 특정 에지(연결되는 노드의 순위를 고려한)의 추가가 노드들의 순위에 어떠한 영향을 주는가에 주목한다. 즉, 추가되는 에지의 종류 별로 전체 노드의 중심성 및 순위에 미치는 영향을 분석하도록 한다. 또한 이러한 영향이 그래프의 구조에 따라 다르므로, 커뮤니티 그래프의 특성을 정의하는 profile로 사용할 수 있음을 보인다.

3. 본 론

3.1 그래프 중심성

본 절에서는 본 연구에서 사용한 네 가지의 중심성, 연결도, 밀접도, 특성벡터, betweenness 중심성을 설명한다. 본 연구는 추가 에지가 만들어 내는 중심성 및 순위의 변동에 초점을 맞추고 있으므로, 이 들 네 가지 중심성에 국한하여 연구를 진행한다. 그래프는 연결된 무방향 무가중치 단순 그래프 $G=(V,E)$ ($V=\{v_1,v_2,...,v_n\}$: 노드의 집합, E:에지의 집합)를 가정한다. 각 노드 i에 대해 $c_i \in R$, $r_i \in I$ (R은 실수 집합, I는 자연수 집합)는 중심성, 순위를 각각 나타낸다, 여기서 중심성은 큰 수가 높은 중심성을 나타내고, 순위는 작은 수가 높은 순위를 나타낸다.

3.1.1 연결도 중심성

$c_v = degree(v)/(n-1)$, $degree(v)$는 노드 v에 연결된 에지의 개수이다. 정규화를 위해 (n-1)로 $degree(v)$를 나누어 중심성으로 사용한다. $degree(v)$가 높다는 의미는 다른 많은 노드(주체)들과 직접 연결되어 있다는 뜻이므로, v가 노드들의 중심이 될 가능성이 높다.

3.1.2 밀접도(closeness) 중심성

$c_v = \frac{1}{\sum_{t \in V} dist(v,t)}$, $dist(v,t)$는 노드 v로부터 노드 t까지의 최단거리를 나타낸다. 노드 간의 거리는 경로상의 에지수로 정의된다. 즉, 노드 v로부터 다른 노드들로 가는 최단거리의 합이 작다는 것은 v로부터 정보가 신속히 다른 모든 노드들로 전파가능하다는 것이고, 그래프의 노드들의 중심이 될 가능성이 크다는 것이다. 다른 노드들로 가는 최단거리의 합이 작을수록 밀접도 중심성은 큰 값을 갖게 된다.

3.1.3 특성벡터 중심성

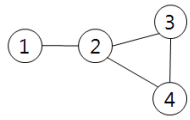
그래프를 인접행렬(adjacency matrix)로 표현하고, 그 행렬의 최대 특성값(eigenvalue)을 구한 후, 그 특성값에 해당하는 특성벡터(eigenvector)를 찾아 해당하는 노드에 대응시켜서 구할 수 있다. 연결도 중심성은 직접 연결된 노드의 수만을 고려하는데 반해, 특성벡터 중심성은 주변 노드들의 중심성을 고려하여 노드의 중심성을 계산하는 방법이다.

3.1.4 Betweenness 중심성

$c_v = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$, 여기서 σ_{st}는 s로부터 t로 가는 최단거리경로의 개수를 나타내고, $\sigma_{st}(v)$는 v를 경유하여 s로부터 t로 가는 최단거리경로의 개수를 나타낸다. 노드 v가 다른 노드 쌍의 최단거리경로 상에 많이 사용된다는 뜻은 그 만큼 노드 v가 그래프에서 중심에 위치하고 있다는 뜻이 되며, 이를 반영한 중심성이 betweenness 중심성이다.

3.1.5 중심성 계산 예

다음 단순 그래프 그림 1에 대한 전술한 네 개의 중심성 값은 표 1에 나타나 있다. 중심성마다 서로 다른 값을 나타내고 있지만, 노드 2가 그래프의 중심에 위치하고 있으므로, 노드 2에 대한 각 중심성 값이 가장 큰 것을 확인할 수 있다.



(그림 1) 단순 그래프 예
(Figure 1) a simple graph example

(표 1) 단순 그래프 (그림 1)의 중심성 계산
(Table 1) centrality of Figure 1

노드 번호	연결도 중심성	밀접도 중심성	특성벡터 중심성	betweenness 중심성
1	0.33	0.2	0.28	0
2	1	0.33	0.61	4
3	0.67	0.25	0.52	0
4	0.67	0.25	0.52	0

3.2 추가 에지에 따른 중심성 및 순위 변동

본 절에서는 그래프에 추가 에지가 발생하였을 때 전체 노드들의 중심성 및 순위에는 어떤 영향이 발생하는지를 확인하는 실험계산을 설명한다. 특정 중심성을 제외하고는 분석적 방법으로 추가 에지에 의한 중심성의 변화를 쉽게 계산하는 방법은 알려져 있지 않다. 따라서, 본 연구에서는 추가 에지를 기존 그래프에 추가하여 새로운 중심성과 순위를 계산한 후, 새로운 중심성, 순위가 기존의 중심성, 순위와 어느 정도 다른지를 확인하는 실험계산을 수행한다. $G=(V,E)$의 노드 i의 중심성, 순위를 각각 c_i, r_i라 하고, 에지 e 하나가 추가된 그래프 $G'=(V,E')$, $E'=E\cup\{e\}$, $e=(x,y)\notin E$의 노드 i의 중심성과 순위를 c'_i, r'_i로 표시한다.

먼저 추가 에지의 구분에 대해 설명한다. 추가 에지의 속성을 확인하기 위해 먼저 노드들을 현재의 순위 r에 따라 집합으로 나눈다. 순위 상위 α%에 속하는 노드 집합 S_1, 중간순위 α%에 속하는 노드 집합 S_2, 순위 하위 α%에 속하는 노드 집합 S_3를 찾는다. 그리고 다음과 같이 추가 에지 $e=(x,y)\notin E$들의 집합 $t_1,...,t_6$을 정의한다.

$e\in t_1$ if $x,y\in S_1$, $e\in t_2$ if $x,y\in S_2$, $e\in t_3$ if $x,y\in S_3$, $e\in t_4$ if $x\in S_1, y\in S_2$, $e\in t_5$ if $x\in S_2, y\in S_3$, $e\in t_6$ if $x\in S_1, y\in S_3$. 즉, 추가 에지가 t_1에 속한다는 것은, 상위 α%에 속하는 두 노드를 연결하는 에지라는 것이다. 여기서 세 개의 그룹으로 노드 순위그룹을 나눈 이유는 분석을 단순화하고, 각 그룹 내 또는 그룹 간에 걸치는 추가 에지의 영향을 추가 에지 종류별로 쉽게 도출하기 위해서이다. 만일 노드들을 g개의 순위 그룹으로 나눈다면, ${}_gC_2$ 개의 추가 에지 종류가 발생한다. 실험 계산 알고리즘에서는 N_q개의 추가 에지 종류를 가정하여 기술하였다.

추가 에지의 종류 $t_1,...,t_6$에 따라 전체 노드의 중심성 및 순위의 변화를 분석한다. 본 연구는 네 가지의 중심성에 대해 각각 수행된 것이므로, 중심성 종류에 따라 노드의 순위가 달라질 수 있으므로, 기술한 추가 에지의 집합은 각 중심성에 따라 각각 설정된다.

추가 에지가 노드 전체에 미치는 중심성 및 순위 변동을 측정하기 위해서는, 측정치(measure)의 정의가 필요하다. 다음은 전체 중심성 및 순위의 변동크기를 측정할 수 있는 측정치를 설명한다.

(m_1) 순위의 변동량

두 개의 순위 벡터 $r=(r_1,...,r_n)$, $r'=(r'_1,...,r'_n)$의 변동량을 다음과 같이 정의한다[2].

$$m_1=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}I_{r,r'}(i,j) \text{ where}$$

$$I_{r,r'}(i,j)=\begin{cases}1 & r_i<r_j \text{ and } r'_i>r'_j\\ 0 & otherwise\end{cases}$$

(m_2) 중심성의 변동량

두 개의 중심성 벡터 $c=(c_1,...,c_n)$, $c'=(c'_1,...,c'_n)$의 차는 두 벡터 차의 2-norm으로 정의한다.

$$m_2=\|c-c'\|_2=\sqrt{\sum_{i=1}^{n}(c_i-c'_i)^2}$$

(m_3) 1등 노드가 에지 추가 후 1등 유지 비율

(m_4) 1등 노드가 에지 추가 후 3등 이내 존재 비율

(m_5) 1등 노드가 에지 추가 후 10등 이내에 존재 비율

(m_6) 순위 10등까지의 유지 비율

$m_6=\frac{|A\cap A'|}{|A\cup A'|}$,그래프 G의 10등까지의 노드 집합 A, 그래프 G'의 10등까지의 노드 집합 A'. m_3~m_6는 [5]에서 사용한 순위의 변동성을 측정하기 위한 기준으로, 추가 에지에 따라 기존의 1등 순위 노드가 새로운 순위 내에서는 몇 등 이내에 있는지, 그리고 10등 이내의 노드들이 어느 정도 유지되는지를 나타내는 척도이다. 여기서, m_1,m_2는 값이 클수록 변동양이 많은 것을 의미하고, m_3~m_6는 값이 클수록 순위가 안정적이라는 것을 의미한다.

본 절의 실험계산은 추가 에지 한 개가 있을 경우, 그 유형 (t_1~t_6)에 따라 중심성 및 순위 변동량 측정치 (m_1~m_6)가 어떤 변화를 갖는지 확인하는 것이다. 실험계산을 위해 두 가지의 그래프 생성방법, Erdös & Rényl 방법[9]과 Barabási & Albert [10] 방법을 사용하여 그래프를 각각 20개 씩 생성하였다. Erdös & Rényl 방법은 생성 가능한 에지가 발생될 확률을 고정하여, random 그래프를 생성하는 방법이고, Barabási & Albert 방법은 scale-free 그래프를 생성하는 대표적인 방법이다. 생성된 그래프를 E그래프, B그래프로 칭한다. 이 외에도 Watts & Strogatz 방법[15] 등이 있으나, 본 연구에서는 타 연구와의 비교를 위해, 이 두 방법을 본 연구에서 활용하였다[4],[5]. 두 가지 서로 다른 방식으로 생성된 그래프를 이용하여, 추가 에지에 따라 측정치 들이 어떻게 달라지는지 확인한다. E그래프, B그래프는 동일한 노드 수($|V|=n$=40), 에지 개수($|E|$=115)가 되도록 하였다. 일반적으로, E그래프는 연결도가 노드 간에 큰 편차를 보이지 않고, 반면에 B그래프는 일부 노드에 연결이 집중되는 특성을 갖고 있다.

다음 ComputeVariation은 실험 계산의 의사코드이다. 여기서, 집합 $t_1,...,t_6$의 크기는 서로 다를 수 있다. 일반적으로 t_1은 작은 크기를 갖는데, 그 이유는 순위가 높은 노드에는 이미 많은 에지가 연결되어 있어서, 순위가 높은 노드 간에는 추가할 수 있는 에지 수가 작기 때문이다. 이와는 반대로, t_3의 크기는 상대적으로 큰 값을 갖는다. 여기서, N_p는 그래프의 개수(=20), N_q는 추가 에지 종류의 수(=6), N_s은 측정치 개수(=6), α=0.15, $I_k=\{1,...,k\}$ (k는 자연수)로 정의한다.

네 가지 중심성과 두 가지 생성 그래프 생성방법에 대해
procedure ComputeVariation {
그래프 $G_1, G_2, ..., G_{N_p}$에 대해 {
(1) 그래프 $G_i=(V_i,E_i)$, $|V_i|=n$를 생성한다.
(2) G_i의 중심성 $c\in R^n$및 순위 $r\in I^n$을 계산한다.
(3) $p_i^k(j)\leftarrow 0, n_k\leftarrow 0, k\in I_{N_q}, j\in I_{N_s}$
(4) while (아직 확인하지 않은 추가 에지 $e\notin E_i$ 존재){
(4.1) $G_i^{'}=(V_i,E_i^{'})$, $E_i^{'}=E_i\cup\{e\}$ 구축
(4.2) e가 속하는 집합($t_1\sim t_{N_q}$)을 확인한다. $e\in t_\tau$ 가정
(4.3) $n_\tau\leftarrow n_\tau+1$
(4.4) $G_i^{'}$의 중심성 c'및 순위 r'를 계산한다.
(4.5) c,c',r,r'를 이용하여 $m_1\sim m_{N_s}$를 계산.
(4.6) $p_i^\tau(j)\leftarrow p_i^\tau(j)+m_j, j\in I_{N_s}$ }
(5) $\overline{p_i^k}(j)=p_i^k(j)/n_k,\ k\in I_{N_q}, j\in I_{N_s}\}$
(6) $\overline{\overline{p^k}}(j)=\sum_{i=1}^{N_p}\overline{p_i^k}(j)/N_p, k\in I_{N_q}, j\in I_{N_s}\}$

(3)의 $p_i^k(j)$는 그래프 G_i에서 추가 에지(에지 종류가 t_k) 한 개에 의한 측정치 m_j값을 누적하는 변수이고, n_k는 에지 종류가 t_k에 속하는 에지의 개수이다. 주어진 그래프 G_i에서 추가 생성 가능한 에지들 모두에 대해(4), 하나의 에지를 추가한 새로운 그래프 $G_i^{'}$를 생성한다(4.1). 추가 에지의 종류를 구분하고 ($e\in t_\tau$)(4.2), 그 그룹에 속한 에지의 개수를 하나 증가시킨다(4.3). 새로운 그래프에 대해 중심성 c'와 순위 r'를 계산하고(4.4), 이 값들을 이용하여 측정치 $m_1\sim m_{N_s}$를 계산한 후(4.5), 해당 누적 변수에 값을 누적한다(4.6). 추가 에지 종류별로 측정치의 평균치를 계산하고(5), 각 그래프별로 계산된 측정치의 평균을 구한다(6).

다음 표 2, 표 3은 그래프에서 에지 한 개 추가에 의한 중심성 및 순위 변동량 측정치 ($m_1\sim m_6$)의 평균 $\overline{\overline{P}}$를 나타내고 있다. 표에서 각 중심성, 측정치에 대해 측정치의 값을 최대로 하는 에지 그룹을 색 있는 셀로 표시하였다.

(표 2) 중심성 및 순위변화량, E그래프, n=40, $|E|$=115
(Table 2) centrality and ranking difference. E graph, n=40, $|E|$=115

		m_1	m_2	m_3	m_4	m_5	m_6
연결도 중심성	t_1	0.00137	0.036	0.79	0.97	1	0.76
	t_2	0.0073	0.036	1	1	1	1
	t_3	0.00455	0.036	1	1	1	1
	t_4	0.00478	0.036	0.84	0.97	1	0.86
	t_5	0.0064	0.036	1	1	1	1
	t_6	0.0032	0.036	0.84	0.97	1	0.86
밀접도 중심성	t_1	0.0044	0.0007	0.81	0.96	1	0.75
	t_2	0.0104	0.00076	0.98	0.99	1	0.95
	t_3	0.0079	0.00086	0.98	1	1	0.94
	t_4	0.0090	0.00082	0.85	0.96	1	0.85
	t_5	0.0106	0.001	0.977	0.99	1	0.93
	t_6	0.0108	0.00127	0.88	0.98	1	0.85
특성벡터 중심성	t_1	0.0116	0.06447	0.63	0.97	1	0.72
	t_2	0.0122	0.04206	0.99	1	1	0.96
	t_3	0.0031	0.01618	1	1	1	1
	t_4	0.0138	0.05666	0.88	0.99	1	0.85
	t_5	0.0072	0.03005	0.99	1	1	0.99
	t_6	0.0097	0.04774	0.95	1	1	0.92
bet ween ness 중심성	t_1	0.0076	34.63	0.74	0.97	1	0.78
	t_2	0.0110	25.95	0.93	1	1	0.92
	t_3	0.0139	28.13	0.89	1	1	0.89
	t_4	0.0099	30.74	0.75	0.94	1	0.85
	t_5	0.0142	32.37	0.89	1	1	0.89
	t_6	0.0127	39.62	0.69	0.92	1	0.79

(표 3) 중심성 및 순위변화량, B그래프, n=40, $|E|$=115
(Table 3) centrality and ranking difference, B graph, n=40, $|E|$=115

		m_1	m_2	m_3	m_4	m_5	m_6
연결도 중심성	t_1	0.0003	0.04	0.97	1	1	0.95
	t_2	0.0069	0.04	1	1	1	1
	t_3	0.0195	0.04	1	1	1	1
	t_4	0.0040	0.04	0.98	1	1	0.97
	t_5	0.0135	0.04	1	1	1	1
	t_6	0.0103	0.04	0.97	1	1	0.97
밀접도 중심성	t_1	0.0011	0.00056	0.97	1	1	0.94
	t_2	0.0042	0.00031	1	1	1	1
	t_3	0.0043	0.00054	1	1	1	0.99
	t_4	0.0068	0.00067	0.98	1	1	0.98
	t_5	0.0048	0.00044	1	1	1	1
	t_6	0.0119	0.00125	0.96	1	1	0.98
특성벡터	t_1	0.0087	0.04	0.93	1	1	0.9
	t_2	0.0078	0.02	1	1	1	0.99

중심성	t_3	0.0036	0.01	1	1	1	1
	t_4	0.0099	0.04	0.96	1	1	0.94
	t_5	0.0060	0.02	1	1	1	1.00
	t_6	0.0120	0.04	0.96	1	1	0.96
bet ween ness 중심성	t_1	0.0050	24.23	0.98	1	1	0.98
	t_2	0.0070	11.70	1	1	1	1
	t_3	0.0144	6.19	1	1	1	1
	t_4	0.0054	25.05	0.96	1	1	0.98
	t_5	0.0126	9.53	1	1	1	1
	t_6	0.0051	18.60	0.96	1	1	0.99

표 2, 표 3은 중심성의 종류와 측정치에 따라 다양한 결과를 보여주고 있다. 예를 들어, 표 2에서 m_1에 대해, 연결도 중심성인 경우 t_2에 속한 추가 에지가 가장 큰 변동성을 나타내는 반면, 밀접도 중심성에서는 t_6, 특성벡터 중심성에는 t_4, betweenness 중심성에서는 t_5에 속한 추가 에지가 가장 큰 변동성을 보여 주고 있다. 표 3의 m_1에 대해서는 표 2와 서로 다른 변동성을 보여 주고 있다. m_2에 대해서도 중심성에 따라 상이한 결과를 보여주고 있다. 연결도 중심성인 경우, m_2의 값은 정의에 의해 모두 같은 값을 보여 주고 있다. m_1,m_2의 결과로 부터, 변동성은 추가 에지가 S_1,S_2,S_3에 속할 때 보다는 S_1, S_2,S_3간에 걸쳐서 존재할 때 커지고, 특히 S_1과 S_3사이에 걸쳐서 존재할 때, 즉 상위와 하위 간에 걸쳐져 있을 때 극대화 된다는 것을 확인할 수 있다. 이것은 직관적인 기대와 부합되는 결과이다.

m_3인 경우, 추가 에지가 S_2,S_3 내에 있을 때 1등의 노드가 바뀌지 않을 가능성이 크다는 것을 보여 주고 있고, m_4인 경우, B그래프에 대해서는 모든 경우, 에지 추가 하나가 추가 전의 1등 순위 노드를 에지 추가 후에 3등 바깥순위로 만들 수 없는 반면, E그래프에서는 m_3에서 관찰한 바와 같이, 추가 에지가 S_2,S_3그룹 내에 있을 때 1등의 순위 노드가 추가 후 3등 이내에 있을 가능성이 크다는 것을 관찰할 수 있다. m_5인 경우는 모든 중심성에서 에지 추가 이 후 추가 전의 1등 순위 노드를 에지 추가 후에 10등 바깥 순위로는 만들 수 없다는 것을 알 수 있다. m_6인 경우에도 B그래프가 더 10등 이내의 노드들을 유지한다는 것을 알 수 있다.

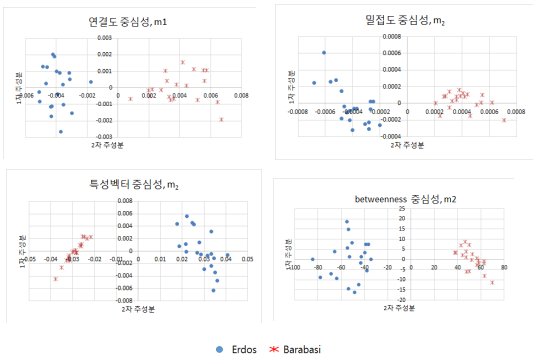
결과적으로, 그래프 종류와 중심성에 따라, 추가 에지에 대한 중심성 및 순위 변화량 평균값이 특이한 패턴을 나타낸다는 것을 알 수 있다. 그리고, 측정치의 최대, 최소값이 어느 에지 종류에서 발생하는 지의 단순 분석보다는, 각 측정치의 N_q차원 벡터(t_1~t_{N_q}에 대한 측정치)에 내재되어 있는 패턴을 분석하는 것이 그래프의 종류에 따라 중심성 및 순위의 변화 특성을 확인하는 방법이 될 수 있다. 따라서, 모든 그래프에 대한 평균값이 아닌, 그래프i에 대해 에지를 한 개 추가할 때 얻어진 $\overline{p_i^k}(j)$, $k \in I_{N_q}$, $j \in I_{N_p}$ 값들을 이용하여, 그래프의 종류에 따라 에지를 추가하였을 때 중심성 및 순위의 변화 특성이 존재하는지 확인하도록 한다.

3.3 그래프 별 추가 에지에 따른 중심성 및 순위 변동 분석

본 절에서는 주성분분석 (PCA: Principal Component Analysis)방법을 이용하여 그래프 별로 에지를 추가하였을 때 달라지는 중심성 및 순위에 특징이 있는지를 확인하는 방법을 설명한다. 결론적으로 그래프 별로 특성은 존재하고, 추가 에지를 이용한 분석을 그래프의 구조를 구분하는 profile로 사용할 수 있는 근거를 제시한다.

주성분분석 방법은 다 차원의 상호 연관되어 있는 벡터를 같거나 작은 차원의 선형적으로 상관관계가 없는 벡터로 수직 변환(orthogonal transformation)하는 방법이다 [11]. 본 연구에서는 그래프 G_i, 측정치 m_h에 대해 $\overline{p_i^j}(h)$, $j \in I_{N_q}$를 하나의 벡터($\in R^{N_q}$)로 만들어 변형하기 전의 벡터로 한다.

$X = \begin{pmatrix} U \\ W \end{pmatrix}$, $U, W \in R^{N_p \times N_q}$, 여기서 U와 W는 그래프의 문제 (E그래프, B그래프)를 이용하여 얻어진 $\overline{p_i^j}(h)$ 값들로 구성된다. 즉, $U(i,j)$, $W(i,j)$는 각 그래프의 계산 결과 $\overline{p_i^j}(h)$, $i \in I_{N_p}$, $j \in I_{N_q}$를 각각 저장한다. 여기서, N_p=20, N_q=6이다. 각 중심성 별로 측정치 m_1~m_6에 대해 PCA를 수행하였고, 그림 2는 이 중 그래프 종류에 따라 주성분이 확연히 구분되는 조합의 결과를 보여 주고 있다. 연결도 중심성인 경우 m_1이 m_2에 비해 분명한 구분을 만들었다. 또, 밀접도, 특성벡터, betweenness 중심성인 경우는 m_1이 두 그래프 종류에 따라 1차, 2차 주성분이 어느 정도의 구분을 만들었으나, m_2보다는 미미하였다. 모든 중심성에 대해 m_3~m_6의 경우 두 그래프 간에 분명한 구분을 보여 주지 못했다.



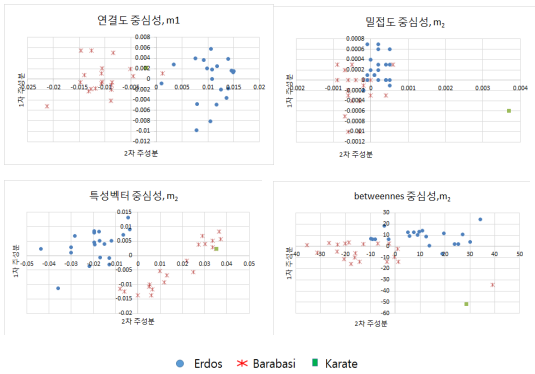
(그림 2) E그래프와 B그래프 별 PCA 결과
(Figure 2) PCA results of E graph and B graph

결론적으로, 두 그래프 종류에 따라 에지를 하나 추가하였을 때 변화하는 순위 및 중심성 변화량 m_1,m_2에는 분명한 차이가 있다는 것을 확인할 수 있었다. 즉, m_1, m_2에 대한 N_q차원 벡터(t_1~t_{N_q}에 대한 측정치)는 두 그래프 종류에 따라 내재적인 차이가 있다는 것을 보여 주고 있다. 따라서, 이러한 차이는 그래프의 특성을 설명할 수 있는 profile로 사용될 수 있다는 것을 의미한다.

3.4 실제 그래프에서 추가 에지에 따른 중심성 및 순위 변동 분석

본 절에서는 Zachary 가라데 클럽 회원의 관계를 나타낸 그래프[12] (n=34, $|E|$=79)와 돌고래 커뮤니티의 그래프[13] (n=62, $|E|$=160)가, 3절에서 사용한 두 종류의 그래프와 그래프 구조가 유사한지를, 에지를 하나 추가하였을 때 변화하는 중심성 및 순위 변화량의 PCA를 통해 확인한다. 그래프는 생성방법 별로 20개씩 생성하였고 α =0.15이다.

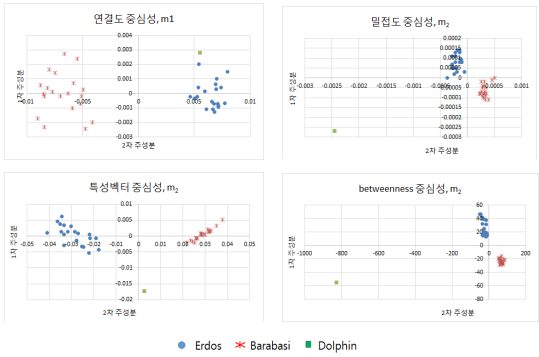
E그래프, B그래프와 가라데 클럽 그래프의 노드 수는 동일하다. 그런데, E그래프 생성 시에는 에지의 생성 비율을 조정하여 가라데 클럽 그래프의 에지 개수를 갖도록 하는 것이 가능하지만, B그래프 생성 방법에서는, 추가 생성되는 노드가 초기에 갖는 연결도가 자연수로 고정된다. 따라서, 가라데 클럽과 동일한 에지 수를 설정할 수 없으므로, 가장 근접한 66개의 에지를 갖도록 하였다. 그림 3은 PCA 결과인데, 3절에서 분석한 바와 같이 분명한 주성분 구분을 보여 주는 조합에 대해서만 결과를 제시하였다.



(그림 3) 가라데 클럽 그래프와 생성 그래프 PCA 결과
(Figure 3) PCA results of Karate club, E graph and B graph

연결도 중심성에 대해서는 두 그래프의 중간 정도에 위치하고 있고, 밀접도 중심성에서는 두 생성 그래프와 분명하게 구분되었으며, 특성벡터 중심성에서는 B그래프와 유사한 패턴을 보였다. Betweenness 중심성에서는 두 그래프와는 구분이 되었으나, B그래프의 하나의 이탈점과 비슷한 패턴을 보였다.

그림 4에서는 돌고래 커뮤니티 그래프를 이용하여 같은 분석을 수행하였다. 이미 기술한 바와 같은 이유로, E그래프는 돌고래 그래프와 동일한 노드 수, 에지 수를 갖고, B그래프는 돌고래 그래프와 동일한 노드 수를 갖지만, 에지 개수에서는 돌고래 그래프와 다른 181개의 에지를 갖는다.



(그림 4) 돌고래 커뮤니티 그래프와 생성 그래프 PCA 결과
(Figure 4) PCA results of dolphin community graph, E graph and B graph

연결도 중심성인 경우는 E그래프와 유사한 패턴을 보였지만, 다른 세 가지 중심성에 대해서는 분명히 다른 패턴을 보여 주고 있는 것을 관찰할 수 있다. 그림 3, 그림 4를 통해 가라데 클럽 그래프와 돌고래 커뮤니티 그래프는 본 연구에서 사용한 두 종류의 생성 그래프와 적어도 추가 에지에 대해 중심성, 순위가 달라지는 패턴 측면에서 서로 다른 구조를 갖는다는 것을 알 수 있다.

4. 결 론

실제의 커뮤니티를 나타내는 그래프의 구조는 어떤 특성이 있는지, 그러한 그래프를 생성하는 방법은 무엇인지 등은 흥미로운 연구가 될 것이다. 본 연구에서는 그래프에서 에지 한 개를 추가하였을 때 전체 노드의 중심성 및 순위의 변동성을 실험 계산을 통해 분석하였다. 에지 추가에 의한 변동성은 그래프의 종류, 중심성의 종류, 측정치에 따라 달라지는 것을 생성 그래프 두 종류를 통해 확인할 수 있었다. 주성분 분석을 통해 추가 에지에 의한 중심성 및 순위 변동에는 E그래프, B그래프, 가라데 클럽 그래프, 돌고래 커뮤니티 그래프 간에는 차이가 있음을 확인할 수 있었다. 따라서, 가라데 클럽 그래프나 돌고래 커뮤니티 그래프의 특성을 설명하기 위해서는 E그래프나 B그래프 이외의 방법이 필요한 것을 알 수 있다.

결론적으로 추가 에지에 의한 중심성 및 순위 변동이 그래프의 특성을 설명할 수 있는 요소라는 것을 밝혔다. 그래프의 구조 및 특성을 설명할 수 있는 요소는 여러 가지가 될 수 있는데, 추가 에지에 의한 중심성 및 순위 변동성도 그래프의 profile로 사용할 수 있음을 알았다. 즉, 주어진 그래프 G와 다양한 그래프 생성 방법 A_1~A_i에 의해 생성된 그래프 군 G_1~G_i으로부터 얻은 측정치 m_1, m_2의 t_1~t_6 별 측정치 벡터에 대해 PCA를 수행하고, 만일 G의 구조적 특징이 G_1~G_i 중 어느 하나의 그래프 군 G_k와 유사하다면, PCA 결과는 G와 G_k는 하나의 군집을 이룬 집단을 보여 줄 것이다. 또는 G_1~G_i에 대한 PCA 결과가, 예를 들어, B_1={G_{p_1}~G_{p_j}}와 B_2={$G_{p_{j+1}}$~G_{p_i}}의 두 군집으로 나뉜다면, 적어도 추가 에지에 의한 노드의 중심성, 순위의 변화량 측면에서 그래프들을 B_1, B_2 그룹으로 나눌 수 있다. 이 방법을 통해 주어진 그래프를 그룹핑할 수 있는 profile 수단으로 사용할 수 있다.

추후 연구로 실제 존재하는 많은 그래프에 대해 본 연구에서 제시한 방법으로 그래프를 종류별로 구분할 수 있는지를 확인하는 것도 흥미로운 연구가 될 것이다.

참 고 문 헌 (Reference)

[1] P. Boldi and S. Vigna, “Axioms for Centrality”, Social and Information Networks, Nov. 2013. http://dx.doi.org/10.1080/15427951.2013.865686

[2] R. Lempel and S. Moran, “Rank-Stability and Rank-Similarity of Link-Based Web Ranking Algorithms in Authority-Connected Graphs”, Information Retrieval, Vol. 8, No. 2, pp. 245-264, 2005. http://dx.doi.org/10.1007/s10791-005-5661-0

[3] K. Avrachenkov and N. Litvak, “The Effect of New Links on Google Pagerank”, Stochastic Models, 22 (2), pp. 319-331, 2006. http://dx.doi.org/10.1080/15326340600649052

[4] J. Ronqui and T. Gonzalo, “Analyzing Complex Networks Through Correlations in Centrality Measurements”, Social and Information Networks, June 2014. http://dx.doi.org/10.1088/1742-5468/2015/05/P05030

[5] S. Borgatti, K. Carley and D. Krackhardt, “On the Robustness of Centrality Measures under Conditions of Imperfect Data”, Social Networks 28.2, pp. 124-136, 2006. http://dx.doi.org/10.1016/j.socnet.2005.05.001

[6] S. Segarra and A. Ribeiro, “Stability and Continuity of Centrality Measures in Weighted Graphs”, Social and Information Networks, Oct. 2014. http://arxiv.org/abs/1410.5119

[7] T. Chartier, E. Kreutzer, A. Langville, and K. Pedings, “Sensitivity and Stability of Ranking Vectors”, SIAM J. Sci. Comput., 33(3), pp. 1077–1102, 2011. http://dx.doi.org/10.1137/090772745

[8] A. Ng, A. Zheng and M. Jordan, “Stable Algorithms for Link Analysis”, SIGIR ’01 Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 258-266, 2001. http://dl.acm.org/citation.cfm?id=384003

[9] P. Erdös and A. Rényl, “On Random Graphs, I”, Publicationes Mathematicae 6, pp. 290–297, 1959. http://ftp.math-inst.hu/~p_erdos/1959-11.pdf

[10] A.-L. Barabási and R. Albert, “Emergence of Scaling in Random Networks”, Science, 286, pp. 509-512,

1997. ISBN: 978-140084135-6;0691113572;978-069 111357-9

[11] I. Jolliffe, “Principal Component Analysis”, Springer Series in Statistics, 2002. http://dx.doi.org/

[12] W. Zachary, “An Information Flow Model for Conflict and Fission in Small Groups”, J. of Anthropological Research 33, pp. 452-473, 1977. http://www.jstor.org/stable/3629752

[13] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson, “The Bottlenose Dolphin Community of Doubtful Sound Features a Large Proportion of Long-Lasting Associations”, Behavioral Ecology and Sociobiology, Vol. 54, No. 4, pp 396-405, Sept. 2003. http://dx.doi.org/10.1007/s00265-003-0651-y

[14] C. Correa, T. Crnovrsanin, and K. Ma, "Visual Reasoning about social networks using centrality sensitivity", IEEE Trans. on Visualization and Computer Graphics, Vol. 18, No. 1, pp. 106-120, 2012. http://dx.doi.org/10.1109/TVCG.2010.260

[15] D. Watts and S. Strogatz, "Collective Dynamics of 'Small-World' Networks", Nature, Vol. 393, No. 6684, pp. 440 - 442, 1998. http://dx.doi.org/10.1038/30918

◐ 저 자 소 개 ◑

한 치 근 (Chi-Geun Han)
1983: 서울대학교 산업공학과 공학사.
1988: 펜실베니아주립대학교 Computer science. 공학석사.
1991: 펜실베니아주립대학교 Computer science. 이학박사.
현 재: 경희대학교 컴퓨터공학과 교수
관심분야: 알고리즘, 계산이론, 유전자알고리즘, 커뮤니티통합
Email : cghan@khu.ac.kr

이 상 훈 (Sang-Hoon Lee)
2010: 경희대학교 컴퓨터공학과 공학사.
2012: 경희대학교 컴퓨터공학과 석사.
현 재: 경희대학교컴퓨터공학과 박사과정.
관심분야: 알고리즘, 유전자알고리즘, 커뮤니티통합
E-mail : a01b01c01@khu.ac.kr