Journal of The Korea Society of Computer and Information
Vol. 18, No. 7, July 2013
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2013.18.7.101

모듈래러티를 이용한 소셜 네트워크의 커뮤니티 통합에 필요한 에지 수 결정 방법

전 병 현*, 한 치 근*

A Method to Decide the Number of Additional Edges to Integrate the Communities in Social Network by Using Modularity

Byung-Hyun Jun*, Chi-Geun Han*

요 약

본 연구는 소셜 네트워크 그래프에서 커뮤니티로 탐색된 2개의 커뮤니티를 하나의 커뮤니티로 통합하는 방법을 연구한다. 모듈래러티는 커뮤니티의 품질을 보여주는 측정치이다. 통합하여야 할 커뮤니티에 에지를 추가함에 따라, 커뮤니티의 품질은 증가하게 된다. 커뮤니티를 통합하기 위해서, 각 커뮤니티의 모듈래러티 값을 이용하여, 추가하여야 할 에지 수를 결정하는 방법들을 제안한다. 단순 그래프를 이용한 실험계산을 통해 통합된 커뮤니티의 모듈래러티 값이 통합하기 전의 각각의 커뮤니티의 모듈래러티 값보다 크게 만드는 방법이 유효한 커뮤니티 통합 방법임을 보이고, 그 방법이 적용될 수 있는 그래프의 조건을 확인한다. 이 결과를 이용하여 실제 소셜 네트워크 예에 대한 실험계산을 통해 본 방법의 유효성을 확인한다.

▸ Keywords : 커뮤니티 통합, 모듈래러티, 소셜네트워크

Abstract

In this paper, a method to decide the number of additional edges to integrate two communities in social network by using modularity is studied. The modularity is a measure to be used to describe the quality of the community. By adding additional edges to the communities, the quality of the communities is enhanced, To integrate two communities, we propose methods to decide the number of additional edges by calculating the modularity. Also, the conditions that the proposed method is valid is investigated in a simple test graph and the efficiency of the proposed method is

•제1저자 : 전병현 •교신저자 : 한치근
•투고일 : 2013. 4, 10. 심사일 : 2013. 5. 3, 게재확정일 : 2013. 6. 25.
* 경희대학교 컴퓨터공학과(Dept. of Computer Engineering, KyungHee University)

approved by integrating two communities in Zachary Karate Club network.

▸ Keywords : community integration, modularity, social network

I. 서 론

트위터, 페이스북, 미투데이 등의 소셜 네트워크 서비스(SNS)가 발전함에 따라 사용자 사이에는 많은 관계(relation)가 만들어지고 있다. 관계 설정의 과정이 누적되면서, 자연스럽게 커뮤니티(그룹)가 만들어 지게 된다. 많은 경우, 소셜 네트워크 서비스의 참여자들은 자신이 속한 커뮤니티의 실체를 파악하기가 어렵다. 그러나 소셜 네트워크 서비스의 관리자들은 관리적인 측면과 더 향상된 서비스를 위해 네트워크 내에 존재하는 커뮤니티의 구조를 그래프이론을 이용하여 파악을 하고자 한다. 사용자 a와 사용자 b는 그래프 상에 노드(node, vertex)로 표현되고, 사용자 a, b간에 관계(정보교환)가 있으면, 노드 a, b 사이에 에지(edge, link)를 연결하여 그래프로 표현할 수 있다.

이러한 연구는 소셜 네트워크 서비스가 발생하기 전에 이미 사회과학 분야, 생물학분야[1]에서 연구되어 왔었고, world wide web 환경에서 사이트들 간의 관계, 유사 정보 컨텐츠 검색 분야 등에서도 연구대상이었다[2]. 연구는 주로 주어진 네트워크에서 어떤 형태의 커뮤니티가 존재하는지를 알아내고자 하는 목표를 갖고 이루어졌다. 이러한 연구문제를 커뮤니티 탐색 문제(community detection)라 하고, 그 해결 방안으로 다양한 방법이 제안되었다. 특히 네트워크의 크기가 매우 큰 경우(수백만 개의 노드) 이 문제를 효과적으로 해결하는 방법에 대한 연구가 많이 진행되었다.

본 논문에서는 커뮤니티 발견과는 다른 방향의 문제, 즉, 커뮤니티 통합의 문제에 대한 해결방안을 연구한다. 커뮤니티 탐색 방법으로 탐색된 커뮤니티 정보를 이용하여 두 개의 커뮤니티를 통합하기위한 문제이다. 예를 들어, [그림 1]에는 3개의 커뮤니티가 있다는 것을 쉽게 확인할 수 있다.

그런데, 커뮤니티 A, B를 통합하기 위해서는 어떻게 해야 할까? 이러한 질문에 대한 단순하면서 명백한 답은 '커뮤니티 A, B 사이에 에지를 추가한다'라는 것이다. 물론, A, B 사이에서 생성할 수 있는 가능한 모든 에지를 추가하게 된다면, A, B는 하나의 커뮤니티로 될 것이 분명하다. 그러나 에지 추가에 대한 비용이 발생하는 환경이라면, 가능한 모든 에지를 추가하는 것은 효과적인 답변이 될 수 없다. 따라서 본 연구에서는 커뮤니티 A, B를 통합하기 위해 필요한 최소한의 에지 수를 결정하는 방법에 초점을 맞춘다.

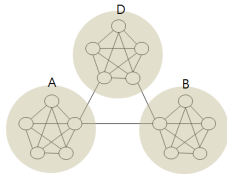

그림 1. 커뮤니티의 예
Fig. 1. A community example

본 연구에서 주어진 그래프 $G=(V,E)$는 연결된 무향그래프라 가정한다. V는 노드의 집합, E는 에지의 집합이다. 2장에서는 모듈래리티에 대한 기본적인 설명과 커뮤니티의 품질을 측정하는 측정치에 대해 알아본다. 3장에서는 커뮤니티를 통합하는 세 가지 방법에 대해 설명한다. 그리고 단순그래프를 대상으로 세 가지 방법을 비교하고, Zachary의 가라데 클럽 네트워크[3]에 대해 커뮤니티 통합문제를 제안한 방법을 이용하여 해결한 결과를 설명한다. 4장에서 결론 및 추후연구의 방향을 제시한다.

II. 관련 연구

1. 커뮤니티 탐색 및 통합

커뮤니티 탐색의 문제는 클러스터링 문제, 파티션닝 문제 등과 같이 사회과학, 자연과학의 분야에서 많이 다루어져 온 문제 유형이다. 커뮤니티를 정의할 수 있어야, 존재하는 커뮤니티를 찾을 수 있는 것이므로, 커뮤니티를 정의하는 것이 먼저 필요하다. 커뮤니티를 정의하는 다양한 방법이 제안되었고[4], 그 정의에 따라 커뮤니티를 탐색하는 방법도 달라질 수 있다.

커뮤니티 탐색 문제에 대해서는 많은 연구가 진행되어 왔고, 그 해결 방법으로 다양한 접근법이 제시되었다. 에지 근접성(betweenness)을 이용한 단순한 방법[5], spectral 알

고리즘, 다이내믹 방법, 모듈래러티를 이용한 방법 등 여러 방법이 존재한다. 그 중 Newman이 제시한 모듈래러티를 이용한 방법이 현실적인 방법[6]으로 사용되고 있다. 이 방법에서는 그래프에서 한 커뮤니티 내에 존재하는 에지들이 노드가 서로 무작위로 연결될 때에 비해 얼마나 많이 집중되어 연결되어 있는가를 나타내는 지표로 모듈래러티를 사용하고 있고, 이 값이 큰 경우에 집중도가 큰 커뮤니티로 간주한다. 그래프 G가 h개의 커뮤니티로 나누어진다면, 모듈래러티 Q는 다음과 같이 정의된다[6].

$Q=\sum_{i=1}^{h}(e_{ii}-a_i^2)$, e_{ii}는 그래프의 모든 에지 개수 대비 커뮤니티 i내부에 존재하는 에지의 비율, a_i는 그래프의 모든 에지 개수 대비 커뮤니티 i에서 내부에 존재하는 에지 + 외부로 나가는 에지의 비율을 나타낸다. 일반적으로 Q값이 0.3 이상인 경우 적절하게 탐색된 커뮤니티로 간주한다.

Clauset et al.은 모든 노드들을 독립적인 각 커뮤니티에 배정하고, Q값을 증가시킬 수 있을 경우, 두 개의 커뮤니티를 병합하는 알고리즘을 제시하였다[7]. 그리고 Blondel et al.은 모든 노드들을 독립적인 각 커뮤니티에 배정하는 방법은 [7]과 동일하지만, Q값을 증가시킬 수 있는 경우에는 노드가 속해있는 커뮤니티에서 빼서 다른 커뮤니티로 이동시키는 방법을 사용하고 있다[8]. 이 방법을 이용하여 수백만 개의 노드가 있는 문제를 해결하여, 이 방법의 우수성을 입증하였다. Chi-Geun Han et al.은 Blondel et al.의 방법을 향상시킬 수 있는 방안을 제시하였다[9].

반면, 많은 연구가 진행된 커뮤니티 탐색과는 다르게 커뮤니티 통합에 대한 연구는 많이 진행되지 않았다. 커뮤니티 통합 방법으로 노드 근접성(vertex betweenness)을 이용한 방법[11]이 있다. 이 방법은 통합하고자 하는 두 커뮤니티에서 노드 근접성이 큰 노드들을 차례로 연결하는 방법으로 커뮤니티를 통합하는 방법이다.

2. 커뮤니티의 품질을 측정하는 지표

두 개의 커뮤니티를 통합하여 하나의 커뮤니티를 구축하기 위해서는 어느 정도의 집중도를 갖는 커뮤니티를 구축하고자 하는 지에 대한 답변이 필요하다. 이를 위해 커뮤니티를 정의하는 방법을 설명한다[10].

무향그래프 $G=(V,E)$, $n=|V|$, $m=|E|$, C=커뮤니티 집합, c=하나의 커뮤니티, h=그래프 내의 커뮤니티의 개수, n_c=커뮤니티 c의 노드 개수, m_c=커뮤니티 c 내부 에지 개수, m_c^{out}=커뮤니티 c에서 외부로 나가는 에지 개수, 커뮤니티 간에는 공유되는 노드는 없음을 가정.

① conductance $f(c)=\dfrac{m_c^{out}}{2m_c+m_c^{out}}$

② expansion $f(c)=\dfrac{m_c^{out}}{n_c}$

③ Internal Density $f(c)=1-\dfrac{m_c}{n_c(n_c-1)/2}$

④ Cut Ratio $f(c)=\dfrac{m_c^{out}}{n_c(n-n_c)}$

⑤ Normalized Cut

$$f(c)=\frac{m_c^{out}}{2m_c+m_c^{out}}+\frac{m_c^{out}}{2(m-m_c)+m_c^{out}}$$

⑥ Modularity $f(c)=\dfrac{2m_c}{2m}-\dfrac{(2m_c+m_c^{out})^2}{(2m)^2}$, $c\in C$

Expansion과 Cut Ratio 지표는 내부에 에지를 추가함에 따라 그 수치가 바뀌지가 않아 커뮤니티 통합 문제의 커뮤니티 지표로 사용하는데 적절하지 않다. 이 외에도, 커뮤니티 지표로 노드의 연결도(degree)를 이용하는 방법이 있다[10]. 본 연구에서는 이 중에서 커뮤니티 지표로 모듈래러티를 이용하는 방법을 사용한다.

III. 본 론

1. 모듈래러티를 이용한 커뮤니티 통합

노드 근접성을 이용한 통합 방법은 커뮤니티가 통합되기 전까지는 추가해야할 에지의 수를 알 수 없는 문제점이 있다. 따라서 본 연구에서는 두 커뮤니티를 통합하기 위해 필요한 에지 수를 결정하기 위해 2.2절에서 설명한 모듈레러티를 이용한다. 통합할 커뮤니티 $c_1,c_2\in C$에 대해 다음 식 (1)을 이용하여 각 커뮤니티의 모듈래러티 $Q(c_1)$, $Q(c_2)$를 구할 수 있다.

$$Q(c)=\frac{2m_c}{2m}-\frac{(2m_c+m_c^{out})^2}{(2m)^2},\ c\in C \qquad (1)$$

또한, c_1, c_2를 통합하였을 때 생성되는 커뮤니티를 $c_1 \oplus c_2$로 표시하고, 이것의 모듈래러티 $Q(c_1 \oplus c_2)$를 다음 식 (2)를 이용하여 구할 수 있다. $\oplus$ 연산자는 커뮤니티 두 개를 하나의 커뮤니티로 통합하였을 때의 커뮤니티를 나타낸다.

$$Q(c_1 \oplus c_2) = \frac{2m_{c_1} + m_{c_1c_2} + x}{2(m+x)} + \frac{2m_{c_2} + m_{c_2c_1} + x}{2(m+x)} - \left(\frac{2m_{c_1} + m_{c_1}^{out} + x}{2(m+x)} + \frac{2m_{c_2} + m_{c_2}^{out} + x}{2(m+x)}\right)^2 \quad (2)$$

x는 c_1과 c_2사이에 추가되는 에지 개수를 나타내고, m_{ij}는 커뮤니티 i와 j사이에 존재하는 초기 에지 개수를 의미한다. 그리고 $m_{c_1}^{out}, (m_{c_2}^{out})$,은 $c_1, c_2, (c_2, c_1)$,사이에 존재하는 에지 개수를 포함한다. 이 들 수식을 이용하여 다음과 같이 두 개의 커뮤니티를 통합하는 방법들을 생각할 수 있다.

〔방법 1〕 커뮤니티가 통합되기 위해서는 통합한 후의 통합 커뮤니티의 모듈래러티 값이 통합하기 전의 각 커뮤니티의 모듈래러티 값보다 커야 한다. 즉,

$$Q(c_1 \oplus c_2) > Q(c_1) + Q(c_2) \quad (3)$$

즉, 식 (3)을 만족하는 최소 x를 구한다.

〔방법 2〕 식 (2)의 값이 최대로 되는 x를 구하는 방법이다. 즉, 통합된 커뮤니티의 Q값이 최대로 되는 x를 찾는 방법이다.

$$Maximize \quad Q(c_1 \oplus c_2) \quad (4)$$

〔방법 3〕 두 개의 커뮤니티가 통합된 후의 전체 그래프 G의 Q값, 즉, $Q(G)$를 최대화 시킬 수 있는 x를 찾는 방법이다.

$$Maximize \quad Q(G) \quad (5)$$

이들 3가지 방법은 모두 x를 0부터 1씩 증가시키면서 식 (3), (4), (5)를 만족하는 x를 찾는 단순한 방법을 사용하여 문제를 해결할 수 있다. 추가되는 에지는 에지의 한 쪽은 c_1에 속한 노드에 연결되고, 또 다른 쪽은 c_2에 속한 노드에 연결된다. 본 연구는 그 에지의 개수를 결정하는 것만을 고려한다.

2. 실험결과 및 분석

2.1 단순그래프를 이용한 제안된 방법의 비교

단순한 그래프를 이용하여 〔방법 1〕, 〔방법 2〕, 〔방법 3〕의 특징을 살펴보기 위해, 〔그림 1〕과 유사한 형태면서 커뮤니티 A, B, D의 노드 수는 가변적이고, $m_A = m_B = 10$, m_D는 가변적으로 할 수 있는 〔그림 2〕와 같은 그래프를 가정한다. A, B, D간에는 하나의 에지만 존재한다.

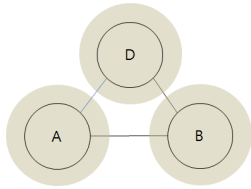

그림 2. 실험 계산을 위한 단순그래프(A, B, D:Community).
Fig. 2. Simple test graph

(식 2)에서 노드 수는 모듈래러티에 영향을 주지 않는 것을 알 수 있다. 그리고 커뮤니티 A와 커뮤니티 B를 통합한 커뮤니티 $A \oplus B$와 커뮤니티 D의 모듈래러티는 전체 그래프가 두 커뮤니티로 나누어진 형태이므로, 그 값이 동일하다는 것을 쉽게 확인할 수 있다.

다음 〔그림 3〕은 총 에지의 개수가 25인 그래프가 주어진 경우, 커뮤니티 A, B간에 에지를 추가하였을 때 변화하는 $Q(A)$, $Q(D)$, $Q(A)+Q(B)+Q(D)$, $Q(A \oplus B)+Q(D)$ $(=2 \times Q(D))$값을 보여 주고 있다.

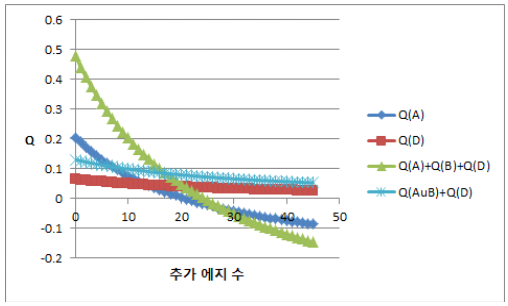

그림 3. 총 에지 수=25 인 경우 다양한 Q값
Fig 3. Q values for the graph(m=25)

〔그림 3〕은 그래프의 총 에지 수가 25개인 경우로, 커뮤니티 D가 2개의 에지를 갖는 경우이다. 통합되는 두 개의 커뮤니티의 초기 에지 수가 21개로, 그래프의 대부분 에지들을 갖고 있어, 모든 Q값이 모두 감소하는 것을 관찰할 수 있다.

이 때, 〔방법 2〕, 〔방법 3〕은 $Q(A \oplus B) = Q(D)$와 $Q(G) = Q(A \oplus B) + Q(D)$가 x에 대해 감소함수가 되어 식 (3)과 식 (4)를 만족하는, 즉 식이 최대로 되는 해를 갖지 못한다.

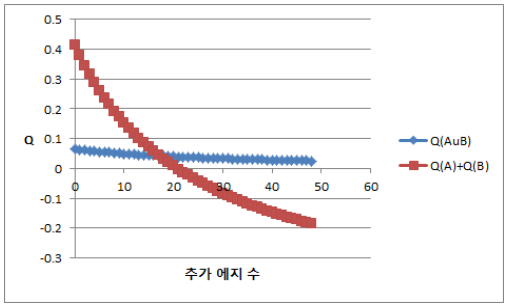

그림 4. 총 에지 수=25일 때 [방법 1]의 해
Fig. 4. Solution of [Method 1](m=25)

[방법 1]을 이용할 경우 [그림 4]를 보면 , 추가 에지 수 $x=18$ 임을 알 수 있다. 그러나 [그림 3]에서 추가 에지 수가 18인 경우 커뮤니티가 나누어진 그래프의 모듈래러티 $Q(G)=Q(A\oplus B)+Q(D)$가 0.083 정도가 되어 커뮤니티의 구조가 취약하다는 것을 알 수 있다. 이는 커뮤니티 A, B의 에지 개수가 전체 그래프에서 상대적으로 많은 비중을 차지하고 있기 때문이다.

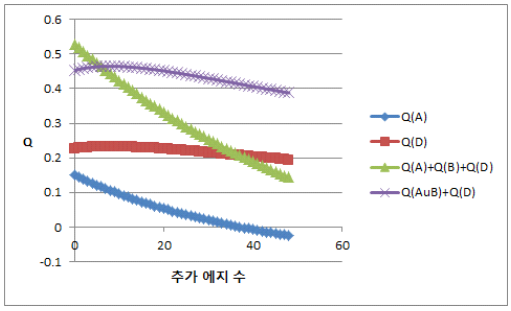

그림 5. 총 에지 수=50 인 경우 다양한 Q값
Fig. 5. Q values for the graph(m=50)

[그림 5]는 그래프의 총 에지 수가 50개이고, 커뮤니티 D가 27개의 에지를 갖는 경우이다. 즉, 통합되는 두 개의 커뮤니티가 갖고 있는 초기 에지의 수가 전체 그래프에서 42%를 차지하고 있다.

총 에지 수가 25일 때와 달리, [그림 5]와 [그림 6]에서 $Q(A\oplus B)+Q(D)$와 $Q(A\oplus B)$가 추가 에지 수의 증가에 따라 그 값이 초기에 증가하다가 감소하는 것을 관찰할 수 있다. 이는 통합할 두 커뮤니티에 에지를 추가하면 전체 그래프에서 잘 구성된 커뮤니티 구조로 변화하다 통합되는 커뮤니티에 필요 이상의 에지가 추가될 경우 커뮤니티 구조가 취약해짐을 나타낸다.

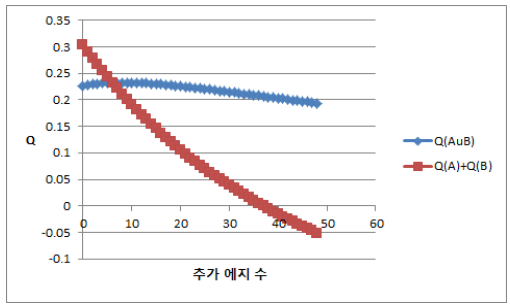

그림 6. 총 에지 수=50일 때 [방법 1]의 해
Fig. 6. Solution of [Method 1](m=50)

그리고 [방법 1]은 추가 에지 수 7개, [방법 2]와 [방법 3]은 추가 에지 수가 8임을 알 수 있다. 두 개의 커뮤니티를 통합한 후의 $Q(G)$의 값이 0.38~0.46의 값을 갖고 있으므로, 커뮤니티의 구성은 적절하다고 판단할 수 있다.

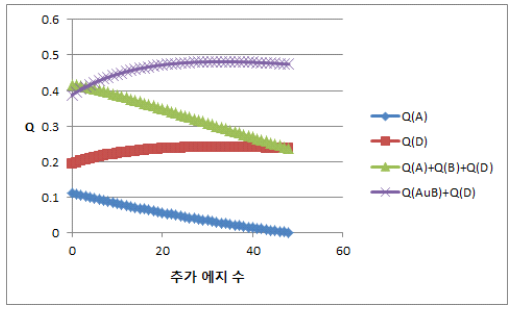

그림 7. 총 에지 수=75 인 경우 다양한 Q값
Fig. 7. Q values for the graph(m=75)

[그림 7]은 그래프의 총 에지 수가 75개이고, 커뮤니티 D가 52개의 에지를 갖는 경우이다. 즉, 통합되는 두 개의 커뮤니티가 갖고 있는 초기 에지의 수가 전체 그래프에서 28%를 차지하고 있다.

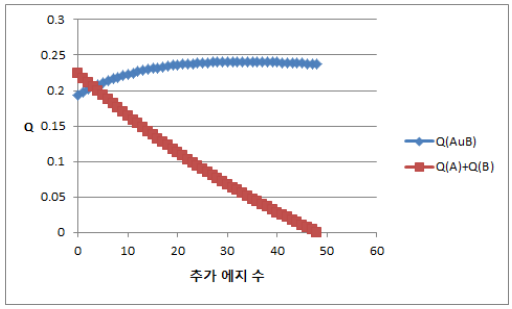

그림 8. 총 에지 수=75일 때 [방법 1]의 해
Fig. 8. Solution of [Method 1](m=75)

〔그림 7〕과 〔그림 8〕에서도 $Q(A \oplus B)$가 총 에지 수가 50일 때와 마찬가지로, 추가 에지 수의 증가에 따라 그 값이 초기에 증가하다가 감소하는 것을 관찰할 수 있다. 그러나 〔방법 1〕의 해는 4, 〔방법 2〕와 〔방법 3〕의 해는 33으로, 총 에지 수가 50일 때와는 다른 현상을 보이고 있다. 즉, 커뮤니티 $A \oplus B$가 전체 그래프에서 차지하는 에지 수의 비중이 감소하면서, $Q(A \oplus B)$의 값이 완만하게 증가하고 있는 것을 알 수 있다. 이는 커뮤니티 D의 에지 수가 증가하여 통합하는 커뮤니티에 이전 예보다 더 많은 수의 에지가 추가되어도 커뮤니티 구조가 취약해지지 않음을 나타낸다. 두 개의 커뮤니티를 통합한 후의 $Q(G)$의 값이 0.38~0.48의 값을 갖고 있으므로, 커뮤니티의 구성은 적절하다고 판단할 수 있다.

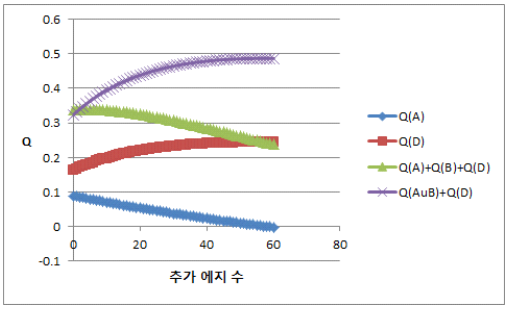

그림 9. 총 에지 수=100 인 경우 다양한 Q값
Fig. 9. Q values for the graph(m=100)

〔그림 9〕는 커뮤니티 D가 77개의 에지를 갖는 경우이다. 즉, 통합되는 두 개의 커뮤니티가 갖고 있는 초기 에지의 수가 전체 그래프에서 21%를 차지하고 있다.

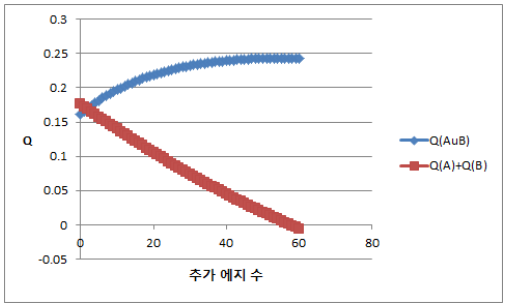

그림 10. 총 에지 수=100일 때 〔방법 1〕의 해
Fig. 10. Solution of 〔Method 1〕(m=100)

〔그림 9〕와 〔그림 10〕에서는 $Q(A \oplus B)$가 추가 에지 수의 증가에 따라 그 값이 증가하다가 감소하는데, 그 증가 속도가 매우 느린 것을 관찰할 수 있다. 이는 커뮤니티 D가 통합하려는 커뮤니티에 비해 상대적으로 내부 에지 수가 많아, 많은 수의 에지가 추가될 때까지 커뮤니티 구성이 좋아짐을 의미한다.

〔방법 1〕의 해는 2, 〔방법 2〕와 〔방법 3〕의 해는 58로, 총 에지 수가 75일 때 보다 〔방법 1〕의 해는 작은 수를 갖고, $Q(A \oplus B)$의 최대값을 갖는 추가 에지 수가 많이 증가되었다. 〔방법 2〕, 〔방법 3〕의 경우, 커뮤니티 D의 많은 내부 에지로 인해 해당 방법의 조건식이 최대가 되는 값이 커지고 있음을 알 수 있다. 두 개의 커뮤니티를 통합한 후의 $Q(G)$의 값이 0.32~0.48의 값을 갖고 있으므로, 커뮤니티의 구성은 적절하다고 판단할 수 있다.

〔표 1〕은 단순 그래프에서의 커뮤니티 통합을 위해 제안된 방법들의 결과를 요약한 표이다.

표 1. 결과치 요약
Table 1. Test results summary

총에지개수	총에지개수 대비 $m_A + m_B$의 비율	〔방법 1〕의 해	〔방법 2, 3〕의 해	$Q(G)$
25	84%	18	-	0.05~0.13
50	42%	7	8	0.38~0.46
75	28%	4	33	0.38~0.48
100	21%	2	58	0.32~0.48

〔방법 2〕와 〔방법 3〕의 해는 본 연구에서 제안한 단순 그래프에서는 동일하게 나온 것을 알 수 있다. 그 이유는 그래프의 단순한 형태 때문에 커뮤니티의 통합 후 전체 그래프가 2개의 커뮤니티로 구성되어, $Q(A \oplus B)$와 $Q(D)$의 값이 동일하게 되었기 때문이다.

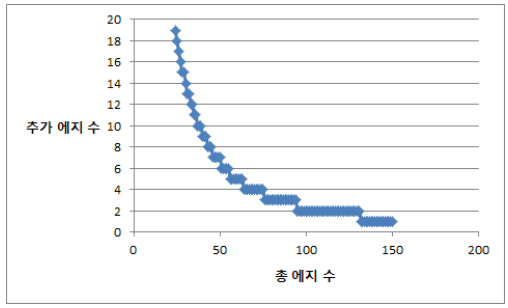

그림 11. 그래프 크기에 따른 〔방법 1〕의 해
Fig. 11. Results of 〔Method 1〕 for total # of edges

〔그림 11〕은 그래프의 크기에 따른 〔방법 1〕이 제시하는 추가 에지 수가 어떻게 변하는지를 보여 주고 있다. 총 에지 수가 증가하면, 전체 그래프에서 커뮤니티 A와 B가 차지하는 에지 수의 상대적인 비율이 줄어들면서, 작은 수의 에지가 추가됨에 따라 $A \oplus B$가 하나의 커뮤니티로 통합될 수 있다는 것을 알 수 있다.

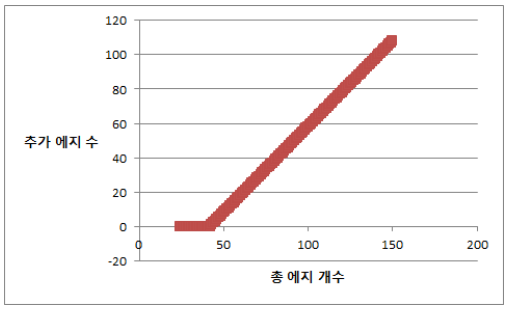

그림 12. 그래프 크기에 따른 〔방법 2,3〕의 해
Fig. 12. Results of 〔Method 2,3〕 for total # of edges

〔그림 12〕는 그래프 크기에 따른 〔방법 2, 3〕의 해를 보여 주고 있다. 전체 그래프의 총 에지 수가 42개 까지, $Q(A \oplus B)$와 $Q(A \oplus B) + Q(D)$는 추가 에지의 수가 0일 때 그 값이 최대이다. 또한 전체 그래프가 42보다 더 많은 에지를 갖게 되면, 그래프의 총 에지 수에 따라 더 많은 에지가 추가되어야 최대가 됨을 알 수 있다.

단순 그래프 문제에서와 같이 전체 그래프의 에지 수에 비해 통합하려고하는 두 커뮤니티의 에지 수 비율이 낮은 네트워크에 대해서는 〔방법 2〕와 〔방법 3〕이 적절한 방법이 아님을 알 수 있다.

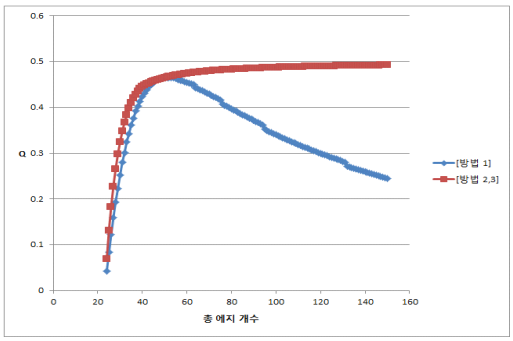

그림 13. 총 에지 개수에 따른 해에서 Q의 변화
Fig. 13. Q values at solutions of 〔Method 1,2,3〕

〔그림 13〕은 그래프 크기에 따른 〔방법 1〕과 〔방법 2,3〕의 해가 얻어질 때의 그래프의 모듈래러티값($Q(A \oplus B) + Q(D)$)을 나타내었다.

〔방법 1〕의 해는 초기 에지 수가 32~119일 때 Q값이 0.3 이상인 것을 알 수 있다. 즉, 이 정도의 문제 크기에서 〔방법 1〕이 유효하다고 할 수 있다. 〔방법 2〕와 〔방법 3〕인 경우는 문제의 크기가 증가함에 따라 최대값을 만드는 추가 에지 수가 증가하고, Q의 최대값이 0.5에 근접하는 것을 알 수 있다. 〔방법 2〕와 〔방법 3〕이 Q가 최대가 되는 추가 에지 수의 최대값(그 이상을 추가하면 $Q(G)$가 감소)을 제시하고는 있지만, 주어진 그래프의 에지 수와 통합할 커뮤니티의 에지 수에 따라 〔그림 12〕와 같이 다양한 답을 제시한다.

따라서 위의 단순 그래프를 이용한 실험계산을 통해, 〔방법 1〕이 커뮤니티 통합을 위해 적은 수의 추가 에지를 제시하는 것으로 결론지을 수 있다.

2.2 Zachary 가라데 클럽 네트워크 실험 계산

Zachary〔3〕는 1970년대에 가라데 클럽에 있는 34명의 회원 간의 관계를 이용하여 네트워크를 구성하였다. 그 관계의 네트워크는 〔그림 14〕와 같다〔4〕.

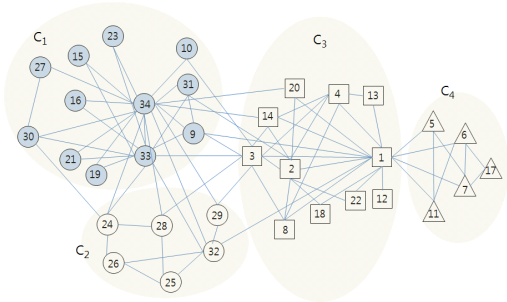

그림 14. Zachary의 가라데 클럽 회원 네트워크
Fig. 14. Zachary's karate club community network

가라데 클럽 네트워크는 커뮤니티 탐색 문제에서 사용되는 대표적인 네트워크의 예로, 다양한 탐색 방법을 통한 커뮤니티 탐색 해를 찾을 수 있다. 따라서 본 연구의 모듈래러티를 이용한 커뮤니티 통합 방법을 위해 이미 알려진 커뮤니티 탐색 해를 이용하는데 적절하다.

가라데 클럽 네트워크는 총 4개의 커뮤니티로 구성되어 있고, 총 78개의 에지가 존재한다. c_1커뮤니티는 내부에 20개, c_2, c_3와는 각 7개의 에지로 연결되어있다. c_2커뮤니티는 내부 7개, c_1과 7개, C3와 3개의 에지로 연결되어있다. c_3커뮤니티는 내부 24개, c_1과 7개, c_2와 3개, c_4와 4개의 에지로 연결되었고, c_4의 경우 내부 6개, c_3와 4개의 에지로 연결되

어있다.

〔표 2〕는 이 네트워크에 대해 두 개의 커뮤니티를 통합하는데 필요한 추가 에지 수를 〔방법 1〕, 〔방법 2〕, 〔방법 3〕에 따라 구한 것이다. 예를 들어, c_1과 c_2를 통합할 때 〔방법 1〕의 결과로 구한 필요한 추가 에지 수가 4, 〔방법 2〕, 〔방법 3〕의 결과는 각각 10, 0임을 나타낸다.

표 2. 〔방법 1,2,3〕을 이용한 Zachary의 가라데 클럽 회원 네트워크의 해
Table 2. The Results of Zachary's karate club community network using 〔Method 1,2,3〕

$\oplus$	c_2	c_3	c_4
c_1	4,10,0	28,0,0(a)	10,29,11
c_2		7,15,1	4,54,37(b)
c_3			4,13,2

(a) $c_1 \oplus c_3$인 경우 $c_1 \oplus c_3$의 내부 에지 수는 61개로 그래프 전체 에지 수의 78%를 차지한다. 〔그림 4〕와 유사한 형태가 되면서, 〔방법 2〕와 〔방법 3〕의 해는 0이 되었다.

(b) $c_2 \oplus c_4$인 경우 $c_2 \oplus c_4$의 내부 에지 수는 13개로 그래프 전체 에지 수의 17%를 차지한다. 〔그림 10〕과 같은 형태를 보이면서, 〔방법 2〕와 〔방법 3〕의 해가 큰 값을 갖는다.

〔표 3〕은 〔방법 1〕에서 구한 해의 수만큼 에지를 추가하였을 때의 그래프 모듈래러티 값, 즉 $Q(G)$를 나타낸 표이다.

표 3. 〔방법 1〕 해의 $Q(G)$
Table 3. Q(G) at solutions of 〔Method 1〕

$\oplus$	c_2	c_3	c_4
c_1	0.40	0.19	0.35
c_2		0.37	0.46
c_3			0.39

$c_1 \oplus c_3$인 경우는 그래프(총 에지 수 78개)에서 두 커뮤니티 에지 합(에지 수 44개)의 비율이 56%로 〔방법 1〕을 적용할 수 없는 경우이다. 이를 제외하고, 모두 0.3 이상의 값을 나타내고 있다. 이는 〔방법 1〕의 결과로 통합된 커뮤니티 그룹이 유효한 결과를 제시한다고 할 수 있다.

IV. 결 론

본 연구에서는 모듈래러티를 이용하여 두 개의 커뮤니티를 하나로 통합하기 위한 추가 에지 수를 결정하는 방법으로, 통합된 커뮤니티의 모듈래러티 값이 통합하기 전의 각각의 커뮤니티의 모듈래러티 값보다 크게 만드는 방법, 통합된 커뮤니티의 모듈래러티 값을 최대로 하는 방법, 전체 그래프의 모듈래러티 값을 최대화 시키는 방법을 제안하였다.

단순 그래프를 이용한 실험계산에서 두 커뮤니티의 에지 수 합이 전체 에지의 50%를 넘을 경우, 통합된 커뮤니티의 모듈래러티 값을 최대로 하는 방법과 전체 그래프의 모듈레러티 값을 최대화시키는 방법은 추가해야할 에지의 수가 증가하는 현상을 보였다. 그리고 통합된 커뮤니티의 모듈래러티 값이 통합하기 전의 각 커뮤니티의 모듈래러티 값보다 크게 만드는 방법은 다른 두 방법보다 적은 수의 에지를 추가하여 두 커뮤니티를 통합할 수 있었다.

또한 제안된 방법을 실제 소셜 네트워크의 예인 가라데 클럽 네트워크에 적용하여 두 커뮤니티 통합에 필요한 에지 수를 결정하였고, 결정된 결과에 따라 커뮤니티를 통합하였을 경우 계산된 그래프의 모듈래러티를 통해 유효하게 통합되었음을 보였다.

제안된 방법은 해를 구하는데 계산량이 매우 작으므로, 수백만 개의 노드를 갖고 있는 대규모 소셜 네크워크를 대상으로 활용하는데 효과적으로 사용될 수 있다. 2.2절에서 설명한 단순한 커뮤니티 지표들과 모듈래러티를 결합한 수치를 향후 개발하여 더 고도화된 통합방법을 연구하고자 한다. 또한 정해진 추가 에지 수를 어느 노드 쌍에 연결해야만 커뮤니티의 품질이 좋아질 지의 연구도 수행될 예정이다.

참고문헌

〔1〕 Girvan, M. and Newman, M.E.J., "Community Structure in Social and Biological Networks", Proceedings of the National Academy of Sciences, Vol. 99, No. 12, pp. 7821-7826, 2002.

〔2〕 Papadopoulos, S., Zigkolis, C., Kompatsiaris, Y., and Vakali, A., "Cluster-Based Landmark and Event Detection for Tagged Photo Collections", IEEE Multimedia, Vol. 18, No. 1, pp. 52 - 63, 2011.

〔3〕 Zachary, W.W., "An Information Flow Model for Conflict and Fission in Small Groups", J. of Anthropological Research 33, 452-473, 1977.

〔4〕 Fortunato, S., "Community Detection in Graphs", Physics Reports 486, pp. 75-174, Jan. 2010.

〔5〕 Freeman, L. C., "A Set of Measures of Centrality Based on Betweenness", Sociometry 40,

pp.35-41, 1977

〔6〕 Newman, M.E.J. and Girvan, M., "Finding and Evaluating Community Structure in Networks", Phys. Rev. E 69, 026113, 2004.

〔7〕 Clauset, A, Newman, M.E.J., and Moore, C., "Finding Community Structure in Very Large Networks", Phys. Rev. E 70, 066111, 2004.

〔8〕 Blondel, V. D., Guillaume, J-L., Lambiotte, R., and Lefebvre1, E., "Fast Unfolding of Communities in Large Networks", J. of Statistical Mechanics: Theory and Experiment, P10008, 2008.

〔9〕 Chi-Geun Han, Moo-Hyoung Jo, "An Enhanced Community Detection Algorithm Using Modularity in Large Networks", J. of Korean Society Internet Information, Vol. 13, No. 3, pp.75-82, 2012.

〔10〕 Leskovec, J., Lang, K.J., and Mahoney, M., "Empirical Comparison of Algorithms for Network Community Detection", Proceeding of WWW '10 Proceedings of the 19th International Conference on World Wide Web, pp. 631-640, 2010.

〔11〕 Byung-Hyun Jun, Chi-Geun Han, "A Study on a Community Integration Algorithm Using Vertex Betweenness", Proceedings of the Korea Information Processing Society Conference, Vol. 19, No. 2, pp. 323 - 325, 2012.

저 자 소 개

전 병 현

1996: 경희대학교 전자계산공학과 공학사.
1998: 경희대학교 컴퓨터공학과 공학석사.
현 재: 경희대학교 컴퓨터공학과 박사과정.
(주)아이컨택트 책임연구원.
관심분야: 알고리즘, 유전자알고리즘, 빅데이터, 커뮤니티통합
Email : bhjun@khu.ac.kr

한 치 근

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