Journal of The Korea Society of Computer and Information
Vol. 19, No. 12, December 2014
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.12.239

적합도 함수를 이용한 커뮤니티 통합에 필요한 추가에지수 결정 및 위치 선정 방법

전 병 현*, 이 상 훈**, 한 치 근**

A Method to Decide the Number of Additional Edges and Their Locations to Integrate the Communities by Using Fitness Function

Byung-Hyun Jun*, Sang-Hoon Lee**, Chi-Geun Han**

요 약

본 논문에서는 네트워크 내에 존재하는 두 개의 커뮤니티 $A, B(|A| \geq |B|$, $|\cdot|$는 커뮤니티의 노드 개수)를 통합하는데 필요한 에지 수 및 에지 위치를 결정하는 알고리즘을 제안한다. 제안된 알고리즘은 커뮤니티 내,외부로 향하는 에지들의 개수를 이용하여 커뮤니티의 성질을 나타내는 적합도 함수를 이용하고, 큰 값을 가질수록 커뮤니티로서의 성질이 크다는 것을 의미한다. 제안된 알고리즘은 그리디 방식으로, B의 하나의 노드에 대해 해당 노드를 A로 병합할 때 커뮤니티 A의 적합도 값이 증가할 수 있는 최소에지수를 결정한다. 최소에지수가 결정된 후, 새로 추가될 에지의 위치를 결정하기 위해 노드 중앙성을 이용한 커뮤니티 연결도 지표를 정의한다. 추가 에지의 위치는 통합된 커뮤니티 연결도 지표를 최대로 만들 수 있도록 결정한다. B의 모든 노드에 대해 이러한 과정을 적용하여 두 커뮤니티를 통합한다. Zachary의 가라데클럽 네트워크를 이용하여 제안된 알고리즘의 실효성을 검증하였다.

▸ Keywords : 커뮤니티 통합, 적합도 함수, 노드 중앙성, 커뮤니티 탐색

Abstract

In this paper, we propose a method to decide the additional edges in order to integrate two communitites $A, B(|A| \geq |B|$, $|\cdot|$ is the size of the set). The proposed algorithm uses a fitness function that shows the property of a community and the fitness function is defined by the number of edges which exist in the community and connect two nodes, one is in the community and the other is out of the community.

•제1저자 : 전병현 •교신저자 : 한치근
•투고일 : 2014. 10. 24. 심사일 : 2014. 10. 30. 게재확정일 : 2014. 11. 10.
* (주)아이컨택트(iContact Co., Ltd.)
** 경희대학교 컴퓨터공학과(Dept. of Computer Engineering, KyungHee University)

The community has a strong property when the function has a large value. The proposed algorithm is a kind of greedy method and when a node of B is merged to A, the minimum number of additional edges is decided to increase the fitness function value of A. After determining the number of additional edges, we define the community connectivity measures using the node centrality to determine the edges locations. The connections of the new edges are fixed to maximize the connectivity measure of the combined community. The procedure is applied for all nodes in B to integrate A and B. The effectiveness of the proposed algorithm is shown by solving the Zachary Karate Club network.

▸ Keywords : community integration, fitness function, node centrality, community detection

I. 서 론

사회 관계망 서비스, 공동 태깅 시스템, 또는 여론 조사 데이터 분석 시스템 등에서 시스템을 분석하고, 데이터의 특성, 흐름 등을 파악하는데 그래프를 이용한다. 이런 시스템에서 사용자 또는 데이터 사이에 많은 관계가 형성되고, 관계 설정의 과정이 누적되면서 형성된 그룹을 클러스터 또는 커뮤니티라고 부른다. 최근에는 커뮤니티 분석을 이용하여 인맥, 콘텐츠 관리, 유전자 분석, 질병 통제, 바이러스 확산, 또는 마케팅 등의 빅 데이터 분석에 이용하기도 한다.

본 논문에서는 커뮤니티 탐색 방법을 이용하여 그래프를 커뮤니티로 구분하고, 필요에 의해 커뮤니티를 병합해야할 때 커뮤니티 사이에 에지를 추가하여 병합하는 커뮤니티 통합 문제에 대해 다룬다. 커뮤니티 통합 문제는 방향성이 없는 그래프에서 두 커뮤니티 사이에 새로운 최소 개수의 에지를 추가하여 하나의 커뮤니티로 통합하는 문제이다. 커뮤니티 통합 문제는 이미 분류된 빅 데이터의 커뮤니티를 통합하여 관리하거나, 유전자나 단백질 구조 분석에서 끊어진 연결(missing link)을 찾아 나누어진 구조를 병합하는 경우, 그리고 질병 통제나 바이러스 확산 등을 위해 여러 지역을 하나의 지역으로 통합하여 관리하는 경우 등 다양한 분야에 적용 가능하다.

커뮤니티 통합을 위한 단순한 방법은 통합하고자 하는 커뮤니티 사이에 추가할 수 있는 모든 에지를 추가하는 것이다. 하지만 추가 에지에 비용이 발생하는 경우라면 최소의 에지를 추가하여 커뮤니티를 통합하는 방법이 필요하다. 따라서 본 논문은 커뮤니티 통합에 필요한 최소의 에지 수를 결정하는 방법과 적은 수의 에지로 커뮤니티를 통합하기 위해 커뮤니티 내의 중요 위치에 있는 노드를 결정하는 방법에 초점을 맞춘다.

본 논문에서는 사용된 그래프 $G=(V,E)$는 연결된 무향(connected undirected)그래프라 가정한다. V는 노드의 집합, E는 에지의 집합이다. 2장에서는 적합도 함수를 이용하여 커뮤니티 탐색하는 방법과 기존 커뮤니티 통합 방법과 커뮤니티 내에서 노드의 중요도를 판단하는 노드 중앙성에 대해 알아본다. 3장에서는 적합도 함수를 이용하여 커뮤니티를 통합하기 위한 추가에지수를 결정하는 알고리즘과 에지의 위치 결정을 위한 커뮤니티 연결도 지표를 정의하고 이를 이용한 통합 알고리즘을 설명하고 Zachary의 가라데 클럽 그래프[1]에 대해 제안된 알고리즘을 수행한 결과를 제시한다. 그리고 4장에서 결론 및 추후 연구의 방향을 제시한다.

II. 관련 연구

1. 커뮤니티 탐색

커뮤니티 탐색(community detection) 문제는 탐색 방법에 따라 전역 탐색 방법과 지역 탐색 방법으로 나눌 수 있다. 전역 탐색 방법은 커뮤니티 연결 정보와 그래프의 전체 연결 구조를 파악하여 탐색하는 방법으로 학술논문지의 참고문헌 관계를 파악하여 연구자 집단을 구분하는 것과 같은 문제가 이에 속한다. 최근 커뮤니티 탐색 문제에서 주로 사용되는 에지 사이성(edge betweenness)을 이용한 방법, 모듈래러티를 이용한 방법, 고유벡터를 이용한 spectral 방법, 다이내믹 방법, 유사성을 이용한 방법 등 다양한 방법들이 있다[2]. 지역 탐색 방법은 페이스북과 같은 사회 관계망에서 특정 인사를 중심으로 구성된 하나의 커뮤니티를 찾는 것과 같은 문제를 해결하는 방법으로, 특정 노드를 포함한 지역 커뮤니티를

찾는 방법이다. 특정 커뮤니티가 갖고 있는 커뮤니티 내부 또는 외부로의 지역적인 정보만을 이용하여 적합도 함수(Fitness function) f를 만들고, 이 함수로 커뮤니티의 결합도를 나타낸다. 적합도 함수를 정의하는 방법에 따라 다양한 지역 탐색 방법이 존재한다[3, 4, 5, 6].

본 논문에서는 Lancichinetti et al.이 제안한 적합도 함수를 이용한다[7].

$$f(c)=\frac{2m_c}{(2m_c+m_c^{out})^{\alpha}} \quad \text{[식 1]}$$

Lancichinetti et al.이 제안한 적합도 함수 [식 1]은 기존 Clauset 지역 탐색 방법의 커뮤니티 전체 에지 연결도에 대한 내부 에지 연결도 비율로 적합도를 구하고, α를 이용하여 커뮤니티의 크기를 결정한다. α값이 작으면 상대적으로 작은 크기의 커뮤니티가, α값이 크면 큰 크기의 커뮤니티가 생성된다. $\alpha=1$일 경우, Clauset 방법과 동일한 크기의 커뮤니티를 구성하게 된다. 그리고 주변 노드의 적합도 변화를 $f_c^v=f(c)_{+\{v\}}-f(c)_{-\{v\}}$로 계산하여 f_c^v 값이 가장 큰 주변 노드를 커뮤니티에 포함시켰다. 이때, $f(c)_{+\{v\}}$는 노드 v를 커뮤니티 c의 내부 노드로 포함할 경우의 계산된 적합도이고, $f(c)_{-\{v\}}$는 외부 노드로 지정할 경우 계산된 적합도이다.

2. 커뮤니티 통합

커뮤니티의 탐색은 주어진 그래프에서 커뮤니티를 파악하는 문제이고, 커뮤니티의 통합문제는 주어진 커뮤니티를 통합할 수 있도록 추가에지를 결정하는 문제이다. Tong et. al.은 고유값(eigenvalue)과 고유벡터(eigenvector)를 이용하여 정보의 확산 속도를 증가/감소시킬 수 있도록, 그래프 상에 에지를 추가/삭제하는 방법에 대한 연구를 수행하였다[8]. 이 방법은 그래프의 연결을 인접행렬로 표현하여, 고유값/고유벡터를 구한 후 알고리즘을 수행한다. 그런데, 그래프가 대규모인 경우는 인접행렬의 고유값/고유벡터를 구하는 것이 또 다른 문제가 될 수 있다. 또한 추가/삭제할 에지 수를 미리 알고 있는 상태에서 알고리즘을 수행해야 하는 단점이 있다.

Jun et. al.은 커뮤니티의 성질을 나타내는 모듈래러티를 이용하여 두 개의 커뮤니티를 통합하는데 필요한 에지 수를 결정하여 통합하는 방법을 제시하였다[9]. 이들은 모듈래러티를 이용하여 커뮤니티를 통합하는데 필요한 에지 수를 결정하는 방법을 제시하였다. 통합된 커뮤니티의 모듈래러티 값이 각각 존재할 경우의 모듈래러티 값의 합보다 커지는 최소 에지수를 찾는 방법이다. 하지만 모듈래러티를 이용하여 추가에지수를 결정하는 방법만을 제시하고, 추가에지의 위치를 결정하는 방법에 대해서는 설명하지 않고 있다. 또한 노드의 closeness를 이용한 커뮤니티의 통합 방법을 제안하였다[10].

3. 노드 중앙성(node centrality)

두 커뮤니티를 통합하기 위해서는 양측의 중요노드를 서로 연결해 주는 것이 가장 기본적인 방법이다. 따라서 효과적인 커뮤니티 통합을 위해 커뮤니티에서 중요 노드를 선택하여 서로 연결하여야 한다.

노드 중앙성은 커뮤니티에서 각 노드가 중앙에 위치한 정도를 판단할 수 있는 성질이다[11]. 노드 중앙성을 측정하기 위한 지표로 degree centrality, closeness centrality, betweenness centrality 등이 존재한다.

○ Degree centrality : $C_D(v)=\frac{d_c(v)}{n_c}$

○ Closeness centrality : $C_C(v)=\frac{1}{\sum_{t\in V} dist(v,t)}$

○ Betweenness Centrality : $C_B(v)=\sum_{s\neq v\neq t\in V}\frac{\sigma_{st}(v)}{\sigma_{st}}$

여기에서 $d_c(v)$는 노드 v에 연결된 인접 노드의 수, n_c는 커뮤니티 내부의 노드 수를 나타낸다. 그리고 $dist(v,t)$는 v에서 t로 가는 최단거리를 나타내고, σ_{st}는 s로부터 t로 가는 최단거리경로의 개수를 나타내고, $\sigma_{st}(v)$는 v를 경유하여 s로부터 t로 가는 최단거리경로의 개수를 나타낸다. $C_D(v)$는 같은 커뮤니티에 많은 인접 노드를 가지고 있는 노드가 중요노드가 된다. $C_C(v)$는 특정 노드로부터 다른 노드들로 가는 최단거리의 합이 작을 때 커뮤니티의 중앙에 위치하게 되어 다른 노드에 비해 중요노드가 된다. 그리고 $C_B(v)$는 한 노드가 다른 노드들 사이의 최단 경로 위에 위치하면 할수록 그 노드의 중앙성은 높아진다.

추가적으로 커뮤니티 구조 분석에서 커뮤니티 내의 노드들과 연결의 유사성을 나타내는 coneighbors index, similarity index 등의 구조적 등위성(Structural Equivalence)으로 중요 노드 지표로 사용하기도 한다.

○ Coneighbors Index : $C_O(v)=\sum_{w\in V}|\Gamma(v)\cap\Gamma(w)|$

○ Similarity Index : $C_T(v)=\sum_{w\in V}\frac{|\Gamma(v)\cap\Gamma(w)|}{\sqrt{|\Gamma(v)|\times|\Gamma(w)|}}$

$\Gamma(v)=\{w\in V:(v,w)\in E\}\cup\{v\}$는 노드 v의 이웃노드집합을 의미한다. $C_O(v)$는 그래프에서 한 노드를 포함하는 다른 모든 노드와의 쌍에 대하여 인접 노드들의 교집합 크기를 더한 수로 두 노드 간에 상대적으로 공유하는 이웃 노드수가 클 때 중요노드가 된다는 개념이다. Coneighbors index는 Shared neighbors, common neighbors라고도 한다. 그리고 Similarity Index는 Jaccard 유사도 등 무수히 많은 지표가 존재하는데, 그 중 정보검색이나 텍스트 마이닝 분야 등에서 많이 사용되는 코사인 유사도가 대표적이다[12].

III. 본 론

1. 적합도 함수를 이용한 추가에지수 결정

[그림 1]과 같이 커뮤니티 c에 노드 w를 병합할 경우, 추가에지수를 t라고 가정하면, 2.1절에서 설명한 [식 1]의 적합도 함수 f를 이용한 새로운 적합도 함수 $f':(C,I)\rightarrow R$는 다음 [식 2]와 같이 정의할 수 있다. I는 음이 아닌 자연수 집합이다.

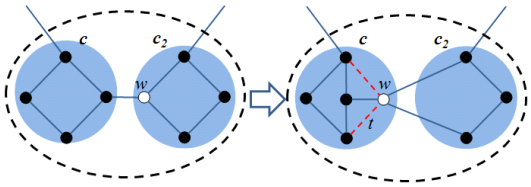

그림 1. 커뮤니티 c에 노드 w를 병합할 경우의 개념도
Fig. 1. Diagram for merging node w into cluster c

$$f'(c+w,t)=\frac{2m_c+2t+2d(w)_c^{in}}{(2m_c+2t+2d(w)_c^{in}+m_c^{out}-d(w)_c^{in}+d(w)_c^{out})^{\alpha}}=\frac{2m_c+2t+2d(w)_c^{in}}{(2m_c+2t+m_c^{out}+d(w))^{\alpha}} \quad [식\ 2]$$

이때, $d(v)_c^{in}$는 노드 v로부터 커뮤니티 c의 내부 노드로 향하는 에지 수, $d(v)_c^{out}$는 노드 v로부터 커뮤니티 c 외부의 다른 노드로 향하는 에지 수를 나타낸다.

$$f'(c,t)=\frac{2m_c}{(2m_c+m_c^{out}+t)^{\alpha}} \quad [식\ 3]$$

[식 2]는 노드 w에 t개의 에지를 커뮤니티 내부에 추가로 연결하여 커뮤니티 c에 w를 병합했을 때의 커뮤니티 c의 적합도 값이고, [식 3]은 t개의 추가에지가 외부로 연결된 상태의 c의 적합도 값이다. 적합도를 이용한 지역적 커뮤니티 탐색 방법에서 w가 c에 병합되려면 w의 병합이 커뮤니티 c의 적합도 값을 증가시켜야 한다. 즉, [식 2]에서 [식 3]을 뺀 [식 4]의 값이 양수 값을 가져야 한다.

$$\Delta f'=f'(c+w,t)-f'(c,t) \quad [식\ 4]$$

따라서 t의 값을 1부터 증가시키면서 $\Delta f'$의 값을 양으로 만드는 최소값을 찾으면, 그 값이 w가 c에 병합되기 위한 최소 추가에지수가 된다.

그리고 커뮤니티 c_j 내부의 모든 노드를 다른 커뮤니티 c_i에 병합할 때, 병합하는 노드의 순서도 중요하다. 2.1절의 커뮤니티 탐색 방법에서도 적합도가 가장 큰 노드를 먼저 선택하여 탐색하게 된다. 따라서 [식 2]의 값을 크게 만들기 위해서는 분모에 사용되는 노드 w와 관계된 식 $d(w)_c^{in}-d(w)_c^{out}$ 값이 커야한다. 따라서 다음 [식 5], [식 6]에 따라 노드 w를 선택하여 추가한다.

$$\Delta d(v)=d(v)_{c_i}^{in}-d(v)_{c_i}^{out} \quad [식\ 5]$$

$$w=\arg\max_{v\in c_j}\Delta d(v) \quad [식\ 6]$$

즉, 병합하려는 커뮤니티로 향하는 에지 수가 커뮤니티 외부로 향하는 에지 수보다 상대적으로 많은 노드들을 먼저 선택하는 규칙을 사용한다.

이렇게 노드 선정 과정을 거치고, 그 노드 w를 병합하기 위한 최소 추가에지수를 결정하여 노드 w를 포함하도록 커뮤니티 c_i를 갱신한다. 그리고 다시 커뮤니티 c_j에 남은 노드에 대해 위 과정을 반복하여 커뮤니티 c_i에 모든 노드가 병합될 때까지의 추가에지수 합을 구하면 두 커뮤니티 c_i, c_j를 통합하기 위해 필요한 최소 추가에지수가 된다.

[그림 2]에 적합도를 이용한 추가에지수 결정 알고리즘의 의사코드가 정리되어 있다.

For given two communities c_i, c_j to integrate
$k=0$
repeat
  for each $v \in c_j$ do
    calculate $\Delta d(v) = d(v)_{c_i}^{in} - d(v)_{c_i}^{out}$
    set $w = \arg\max_{v \in c_j} \Delta d(v)$
  end for
  find $\Delta f' = \min\{t \geq 0 | f'(c_i + w, t) - f'(c_i, t) > 0\}$
  if $\Delta g > 0$ then
    $k = k + \Delta f'$
  end if
  $c_i \leftarrow c_i + \{w\}$, $c_j \leftarrow c_j - \{w\}$
  if $c_j = \phi$ then
    Terminate
  end if
until

그림 2. 적합도를 이용한 추가에지수 결정 방법
Fig. 2. Method to determine the number of additional edges using fitness

예를 들어 [그림 2]의 추가에지수 결정 알고리즘을 이용하여 [그림 3]의 그래프 예에서 추가에지수를 결정하는 과정을 살펴보자. 현재 $A=\{1,2,3,4,5\}$, $B=\{6,7,8,9,10\}$로 커뮤니티가 나누어 있는 초기 상태에서 커뮤니티 A, B를 통합하고자 할 때, 커뮤니티 A에 병합할 노드를 커뮤니티 B에서 선택하여 추가에지수를 결정하는 과정이다.

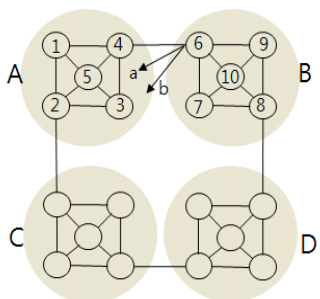

그림 3. [그림 2] 알고리즘을 위한 테스트 그래프
Fig. 3. Test graph for the algorithm [Fig. 2]

먼저 [식 6]을 이용해 커뮤니티 A에 병합할 노드를 선정하면, 노드 6은 [식 5]의 Δd의 값이 1-3=-2이고, 노드 7,9는 -3, 노드 8, 10은 -4의 값을 갖는다. 따라서 노드 6이 처음으로 병합할 노드로 선택된다.

노드 선정 후, 이 노드에 대해 최소 추가에지수를 계산한다. [그림 3]에서 노드 6의 $f'(A+\{6\},t) = \frac{(16+2\times t+2)}{(16+2\times t+2+4)}$이고, $f'(A,t) = \frac{16}{(16+2+t)}$가 된다. $t=0$인 경우 [식 4] $\Delta f'$는 음수 값을 갖게 되어, 2.1절의 지역 탐색 방법으로는 노드 6이 A에 병합되지 않는다. 따라서 추가에지가 필요하고, [식 4]의 $\Delta f'$ 값이 $t=1$일 때 -0.0087, $t=2$일 때 0.046이 된다. 따라서 그림 3처럼 노드 6이 커뮤니티 A에 병합되기 위해서 커뮤니티 A 내부로 에지 두 개를 추가로 연결해야한다.

이 과정을 반복하여 커뮤니티 B의 노드가 병합되는 순서를 정하고, 추가에지수를 결정하여 정리하면 표 1과 같다. 그림 3의 예에서 커뮤니티 A, B를 병합할 경우, 노드 6, 7, 10, 9, 8의 순서로 병합되고 최종적으로 추가해야할 에지의 총 수는 3개가 된다.

표 1. 그림 3의 병합 순서와 추가에지수
Table 1. Merging Sequence and the Number of Additional Edges in [Fig. 3]

추가순서	노드 #	Δd	$\Delta f'$	t
1	6	-2	0.046	2
2	7	-1	0.024	1
3	10	0	0.018	0
4	9	1	0.038	0
5	8	2	0.058	0

2. 커뮤니티 연결성 지표

커뮤니티 내에서의 노드 중요도 지표는 그래프 G나 커뮤니티 c_i 하나에 대해서 특정 노드와 다른 노드, 특정 노드와 이웃 노드들, 또는 특정 노드의 최단 경로 등을 이용하여 계산된 값이다. 하지만 커뮤니티 통합 문제의 경우, 새로운 에지가 추가될 경우 통합되는 커뮤니티 전체 노드들의 노드 중요도 지표에 영향을 준다. 그러므로 단순히 특정 노드의 중요도 지표 값만을 비교해서는 안 되고, 통합할 두 커뮤니티 c_i, c_j의 전체 노드에 대한 노드 중요도 지표 값을 향상시키는 두 노드를 선택해야한다. 따라서 중요 노드 선택을 위해 노드 중앙성 지표의 합을 이용하여 다음 4가지의 측정 방법을 정의한다.

○ Closeness 지표 : $D_C(c_i, c_j, v) = \sum_{v \in c_i \cup c_j} C_C(v)$

○ Vertex betweenness 지표 :

$$D_B(c_i, c_j, v) = \frac{1}{\sum_{v \in c_i \cup c_j} C_B(v)}$$

○ Coneighbors 지표 : $D_O(c_i, c_j, v) = \sum_{v \in c_i \cup c_j} C_O(v)$

○ Similarity 지표 : $D_T(c_i, c_j, v) = \sum_{v \in c_i \cup c_j} C_T(v)$

Vertex betweenness 지표 D_B는 Vertex betweenness centrality가 커지면 노드의 중요도는 커지는 반면, 최단 경로의 합도 증가하게 된다. 따라서 각 노드들의 Vertex betweenness centrality 합의 역수를 지표로 사용한다

.

○ Degree 지표 : $D_D(c_i, c_j) = \max C_D(v), v \in c_i \cup c_j$

그리고 degree 지표 D_D는 에지가 추가되면 커뮤니티의 다른 노드들에 대해 전체적인 영향을 주지 않는다. 하지만 D_D의 경우 특정 노드가 얼마나 많은 다른 노드와 직접적인 연결을 맺고 있는지를 나타내는 지표이므로 추가하였다.

〔그림 3〕의 예에서 커뮤니티 연결성 지표로 노드들의 최단 경로 거리의 합인 closeness 지표 $D_C(c_i, c_j, v) = \sum_{v \in c_i \cup c_j} C_C(v)$를 이용하여 〔표 1〕의 순서에 따라 에지를 추가할 경우, 에지 a를 추가할 때 노드 6과 연결하여 생성할 수 있는 경우는 4가지((1,6), (2,6), (3,6), (5,6))가 있다. (1,6), (2,6), (3,6) 중의 한 에지를 추가할 경우 closeness 지표는 0.576이고, (5,6)을 연결할 경우는 0.578이 되어 에지 (5,6)을 먼저 생성하게 된다. 이는 커뮤니티 A에 노드 6을 연결할 때 통합된 커뮤니티의 연결성 지표를 증가시키기 위해서 A의 중심이라고 할 수 있는 노드 5에 연결하는 것이 바람직하다는 직관과 일치한다. 노드 6에 대한 두 번째 추가 에지는 (1,6)이 된다. 노드 6을 A에 병합하기 위한 추가에지수 2개를 모두 연결한 후, 다음 노드 7에 에지를 추가한다. 노드 7에 대해 동일한 과정을 거치면, 추가에지수가 1인 것을 알 수 있고, 에지 (2,7)이 생성된다. 이후 노드 8, 9, 10은 〔표 1〕에서 추가에지수가 0이므로 에지 추가 과정을 멈춘다. 알고리즘이 종료 한 후의 추가에지가 표시된 그래프는 〔그림 4〕와 같다. 점선 에지는 추가된 에지를 나타낸다.

〔그림 5〕는 〔그림 2〕의 적합도를 이용한 추가에지수 결정 방법과 커뮤니티 연결성 지표를 이용하여 커뮤니티를 통합하는 방법이다.

1) 노드 w와 커뮤니티 c에서 연결되지 않은 노드 쌍 집합 $V_{pair} = \{(w,v) : v \in c, (w,v) \notin E\}$를 구한다.
2) $(w,v) \in V_{pair}$인 노드 w, v를 잇는 에지 $e(w,v)$를 추가한다.
3) 커뮤니티 연결성 지표를 계산한다.
4) V_{pair}의 모든 노드 쌍에 대해 단계 1)부터 단계 3)를 반복하여 가장 큰 커뮤니티 연결성 지표를 갖는 노드 쌍(w,v)을 구한다.
5) 그래프 G에 새로운 에지 $e(w,v)$를 추가한다.
6) 노드 w에 대한 추가 에지 수만큼 단계 1)부터 단계 5)을 반복한다.
7) 다음 병합할 노드에 대해 단계 1)부터 단계 6)을 반복한다.

그림 5. 적합도를 이용한 커뮤니티 통합 방법
Fig. 5. Community integration method using fitness

3. 실험결과 및 분석

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

먼저 Zachary의 가라데 클럽 네트워크을 이용하여 두 개의 커뮤니티를 통합하는데 필요한 에지들을 파악하기 위해 Lancichinetti 커뮤니티 탐색 방법을 이용하여 가라데 클럽

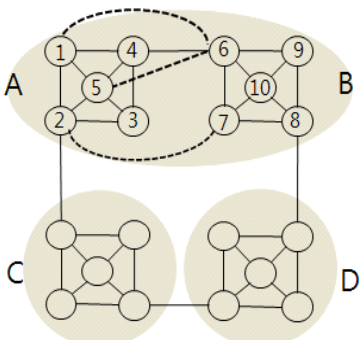

그림 4. 〔그림 2〕 알고리즘과 closeness 지표를 이용하여 통합된 테스트 그래프
Fig. 4. Integrated test graph by the algorithm 〔Fig. 2〕 and closeness index

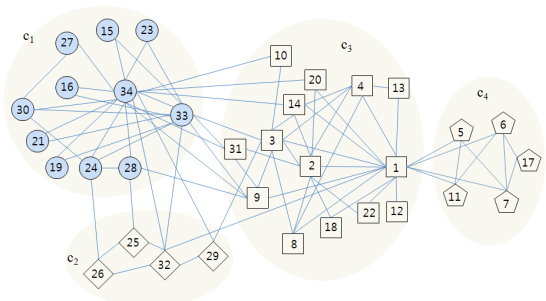

그림 4. Lancichinetti 방법으로 탐색된 가라데 클럽 네트워크
Fig. 6. Karate club network detected by Lancichinetti method

네트워크의 커뮤니티를 탐색한다. 이 네트워크는 4개의 커뮤니티로 구성되어 있다〔그림 5〕. 알고리즘에서 α는 1의 값을 사용하였다.

그리고 〔그림 2〕의 알고리즘을 이용하여 각 커뮤니티간 통합에 필요한 추가에지수를 계산하였다. 다음 〔표 2〕는 적합도를 이용한 통합 방법에서 필요한 추가에지수 결과를 보여 주고 있다. 테이블의 각 항목은 두 커뮤니티를 통합할 때 추가적으로 필요한 추가에지수이다.

표 2. 가라데 클럽 네트워크 통합 결과
Table 2. Results of karate club network Integration

	c_1	c_2	c_3	c_4
c_1		1	6	13
c_2	1		3	5
c_3	6	3		1
c_4	13	5	1	

〔표 2〕의 결과를 이용하여 3.2절의 커뮤니티 연결성 지표를 이용하여 에지를 차례대로 추가하고 Lancichinetti 커뮤니티 탐색 방법을 통해 커뮤니티의 통합 여부를 확인하였다. 그 결과가 〔표 3〕에 있다. 테이블의 각 항목은 커뮤니티 내의 중요노드를 각 커뮤니티 연결성 지표를 이용하여 〔그림 5〕의 통합 방법에 따라 커뮤니티를 통합할 때, 각 커뮤니티가 통합된 상태로 나타날 때의 추가에지수이다. '-'로 표시된 부분은 〔표 2〕의 필요 에지수까지 통합 알고리즘을 수행하였을 경우 통합되지 않은 경우이다.

표 3. 커뮤니티 연결도 지표에 따른 커뮤니티 통합 에지 수
Table 3. The number of additional edges for community integration by community connectivity measures

c_i	c_j	D_D	D_O	D_T	D_C	D_B
c_1	c_2	-	-	-	1	-
c_1	c_3	2	2	2	-	-
c_1	c_4	1	-	-	2	2
c_2	c_3	-	-	-	1	1
c_2	c_4	1	1	1	1	1
c_3	c_4	-	1	1	1	1

〔표 3〕의 실제 통합에 사용된 에지 수 결과로부터 모든 커뮤니티 쌍이 〔표 2〕의 통합에 필요한 에지수 내에서 통합되었음을 알 수 있다. 따라서 〔그림 2〕의 통합에 필요한 추가에지수 결정 방법은 유효함을 알 수 있다. 하지만 커뮤니티 연결성 지표의 경우는 closeness 지표(D_C)가 가장 많은 커뮤니티 쌍이 통합되었음을 알 수 있다.

IV. 결 론

본 논문에서는 두 개의 커뮤니티를 통합하는데 추가적으로 필요한 에지수 및 에지의 위치를 결정하는 방법을 적합도 함수를 이용하여 제안하였다. 기존 연구에서는 에지수가 결정되어 있거나, 단순히 에지의 개수만을 제공하는 반면, 본 논문에서는 적합도 함수를 이용하여 통합에 필요한 추가에지수를 결정하고, 추가 에지의 위치를 결정하기위해 커뮤니티 연결도 지표를 정의하고 이를 이용하여 커뮤니티를 통합하는 알고리즘을 제안하였다.

제안된 방법을 커뮤니티 탐색에서 대표적으로 사용되는 Zachary의 가라데 클럽 네트워크를 이용하여 추가에지수 결정 알고리즘의 실효성을 입증하였다. 위치 결정을 위한 커뮤니티 연결도 지표의 경우는 제안한 지표들 중 closeness 지표를 이용할 경우 많은 쌍의 커뮤니티가 통합됨을 보였다.

추가적으로 커뮤니티 연결도 지표에서 고유벡터 등 다양한 노드 중앙성 정보를 이용하여 새로운 커뮤니티 연결도 지표를 정의하고 다양한 그래프에 대해 적용하는 실험계산이 필요하다. 또한 본 연구결과를 이용하면 지역 탐색 방법으로 커뮤니티를 탐색할 때 두 커뮤니티가 통합된 것으로 탐색될 것이다. 하지만 다른 커뮤니티 탐색 방법을 사용할 경우, 커뮤니티가 통합된 것으로 탐색될 것인지는 불확실하다. 따라서 추가적인 탐색 방법에도 두 커뮤니티가 통합된 것으로 탐색될 수 있는 일반적인 방법의 연구가 필요하다.

참고문헌

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

〔2〕 S. Fortunato, "Community Detection in Graphs", Physics Reports, Vol. 486, No. 3-5, pp.75-174, Feb. 2010.

〔3〕 A. Clauset, "Finding Local Community

Structure in Networks", Phys. Rev. E 72, 026132, Aug. 2005.
[4] F. Luo, J.Z. Wang, and E. Promislow , "Exploring Local Community Structures in Large Networks", Proceeding WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence pp.233-239, 2006.
[5] J. Bagrow, "Evaluating Local Community Methods in Networks", J. Stat. Mech. P05001, 2008.
[6] J. Chen, O. Zaiane, and R. Goebel, "Local Community Identification in Social Networks", Social Network Analysis and Mining, ASONAM, 2009.
[7] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the Overlapping and Hierarchical Community Structure in Complex Networks", New Journal of Physics, Vol. 11, 033015, Mar. 2009.
[8] H. Tong, B.A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos, "Gelling, and Melting, Large Graphs by Edge Manipulation", CIKM '12 Proceedings of the 21st ACM international conference on Information and knowledge management, pp.245-254, 2012.
[9] B.H. Jun, and C.G. Han, "A method to decide the number of additional edges to integrate the communities in social network by using modularity", Journal of The Korea Society of Computer and Information, Vol. 18, No. 7, pp.101-109, Jul. 2013.
[10] B.H. Jun, and C.G. Han, "A study on a community integration algorithm using vertex betweenness centrality", Proceedings of the Korea Information Processing Society Conference, Vol. 19, No. 2, pp.323-325, 2012.
[11] L. Freeman, "Centrality in social networks conceptual clarification". Social Networks, Vol. 1, No. 3, pp.215-239, 1979.
[12] T. Zhou, L. Lü, and Y. Zhang, "Predicting missing links via local information". The European Physical Journal B, Vol. 71, No. 4, pp.623-630, Oct. 2009.

저 자 소 개

전 병 현

1996: 경희대학교 전자계산공학과 공학사.
1998: 경희대학교 컴퓨터공학과 공학석사.
2014: 경희대학교 컴퓨터공학과 공학박사.
현 재: (주)아이컨택트 책임연구원.
관심분야: 알고리즘, 유전자알고리즘, 빅데이터, 커뮤니티통합
Email : bhjun@icontact.co.kr

이 상 훈

2010: 경희대학교 컴퓨터공학과 공학사.
2012: 경희대학교 컴퓨터공학과 석사
현 재: 경희대학교 컴퓨터공학과 박사 과정.
관심분야: 알고리즘, 그래프 이론, 메타휴리스틱 알고리즘
Email : a01b01c01@khu.ac.kr

한 치 근

1983: 서울대학교 산업공학과 공학사.
1988: 펜실베니아주립대학교 Computer science. 이학석사.
1991: 펜실베니아주립대학교 Computer science. 이학박사.
현 재: 경희대학교 컴퓨터공학과 교수
관심분야: 알고리즘, 계산이론, 유전자알고리즘, 커뮤니티통합
Email : cghan@khu.ac.kr