Journal of Digital Contents Society Vol. 14 No. 1 Mar. 2013(pp. 41-50)
http://dx.doi.org/10.9728/dcs.2013.14.1.41

소셜 네트워크 분석을 위한 동적 하위 그룹 생성

이현진*

요 약

소셜 네트워크 분석은 l개의 연결을 가지는 n개의 노드를 대상으로 한다. 노드 수 n이 수십 또는 수백 정도의 소셜 네트워크를 분석할 때는 전체 데이터를 대상으로 분석이 가능하지만, 그 이상이 되면 육안으로 분석하기는 어렵다. 따라서 전체 소셜 네트워크를 분리할 필요가 있는데 이 때 사용할 수 있는 방법이 군집화이다. 군집화를 통해 전체 노드를 하위 그룹으로 구성하면, 소셜 네트워크의 특징 분석이나 노드간의 관계 분석을 쉽게 수행할 수 있게 된다. 군집화 기법은 하위 그룹의 개수를 미리 설정해야 하기 때문에 사용자와의 상호 작용이 필요하고, 이렇게 생성된 하위 그룹이 최적의 결과라는 것을 보증할 수 없다. 본 논문에서는 외부 커뮤니티 연관도를 활용한 동적인 하위 그룹 생성 방법을 제안한다. 발견된 하위 그룹의 개수와 하위 그룹 순도를 기준으로 기존의 연구들과 비교하였고, 실험 결과 제안하는 방법의 우수성을 확인할 수 있었다.

키워드 : 소셜 네트워크 서비스, 소셜 네트워크 분석, 하위 그룹 생성, 군집화

Generation of Dynamic Sub-groups for Social Networks Analysis

Hyunjin Lee*

Abstract

Social network analysis use the n nodes with l connections. About dozens or hundreds number of nodes are reasonable for social network analysis to the entire data. Beyond such number of nodes it will be difficult to analyze entire data. Therefore, it is necessary to separate the whole social networks, a method that can be used at this time is Clustering. You will be able to easily perform the analysis of the features of social networks and the relationships between nodes, if sub-group consists of all the nodes by Clustering. Clustering algorithm needs the interaction with the user and computer because it is need to pre-set the number of sub-groups. Sub-groups generated like this can not be guaranteed optimal results. In this paper, we propose dynamic sub-groups creating method using the external community association. We compared with previous studies by the number of sub-groups and sub-groups purity standards. Experimental results show the excellence of the proposed method.

Keywords : Social Networks Services, Social Network Analysis, Sub-groups Discovering, Clustering

1. 서 론

사회적 관계로 성립된 구조의 의미를 분석하는 소셜 네트워크 분석 (Social Network Analysis)은 관계 모형이나 관계의 강약, 밀도의 높고 낮음에 따른 다양한 사회적 역할 및 영향을 분석할 수 있다. 네트워크를 구성하는 노드들 사이의 영향력은 정보를 주고받으면서 발생하는 연결 관계를 분석하여 알 수 있다[1-2]. 컴퓨터 과학 분야에서는 소셜 네트워크 서비스 (Social Networks Services : SNS)에서 사용자 간의 연결을 확장하여 검색을 효율적으로 수행하고자 하는 연구와 사회 현상과 소셜 네트워크상의 현상 분석, 소셜 네트워크의 구성에 관한 연구들이 진행되고 있다[2-8].

네크워크의 기본적인 구성 요소인 사람, 지역,

※ 교신저자(Corresponding Author): Hyunjin Lee
접수일:2013년 02월 25일, 수정일:2013년 03월 13일
완료일:2013년 03월 22일
* 숭실사이버대학교 컴퓨터정보통신학과
Tel: +82-2-708-7863 , Fax: +82-2-708-7749
email: hjlee@mail.kcu.ac

자원과 같은 행위자는 노드 (node)로 표현되고, 이들 노드는 다양한 관계에 의한 연결 형태인 링크 (link)로 나타나게 된다[2]. 이 링크는 노드 사이의 관계 유무, 방향 및 강도 등을 나타낼 수 있다. 아마존(Amazon) 사이트의 추천시스템과 같이 행위자들의 직접적인 연결 관계가 아닌 준 연결망을 통해서도 행위자에 대한 네트워크 분석이 가능하다[5]. 이러한 준 연결망을 통해 협업필터링[6], 관계의 추론[7] 등 다양한 연구가 이루어지고 있다. 또한 네트워크 안에는 다양한 형태의 하위 그룹(sub-group)을 발견할 수 있다. 하위 그룹은 네트워크 내부에서 관찰되는 그룹들로 네트워크의 커뮤니티 구조를 파악하는데 사용된다[8].

네트워크 형태를 잘 이해하는데 있어서 중요한 단계는 그 안에 있는 하위 그룹을 찾는 것이다. 하위그룹은 네트워크 내에서 노드 간의 강한 연결을 보이고, 하위그룹과 하위그룹 사이에는 비교적 약한 연결을 보이는 노드들의 집단이다. 이러한 그룹을 발견하기 위하여 그래프 분할 (graph partitioning), 스팩트럼 분할 (spectral bisection), 계층 트리 (hierarchical clustering)나 군집화 (clustering) 등 다양한 방법들이 제시되고 있다[9].

Freeman은 메타데이터(metadata)와 콘텐츠(content)를 기반으로 유사그룹들을 동적으로 식별하는 위상 트리 방법 (topological tree method)을 사용하여, 자동적으로 소셜 네트워크 그룹을 조직하였다[10].

Boulet 등은 배치 커널 자기조직화지도(self-organizing map)와 관련된 라플라시안(Laplacian) 방법을 사용하여 소셜 네트워크 분석을 하는 방법을 제안하였다[11].

Yong-Yeol Ahn 등은 커뮤니티를 링크의 집합으로 재정의 하여 모든 노드들이 여러 커뮤니티에 속하는 상황을 해결하는 방법을 제시하였다[12].

Yoonseip Kang 등은 커뮤니티 발견 문제 해결을 위해 공통이웃 그래프 밀도(Common neighbor-hood sub-graph density)를 정의하여, 유사도 전파(affinity propagation)알고리즘과 결합하는 방법을 제안하였다[13].

Hyunjin Lee 등은 외부 커뮤니티 연관도와 군집화를 결합하여 하위 그룹의 개수를 결정할 수 있는 방법을 제안하였다[4]. 이 방법은 하위 그룹의 개수를 증가시키면서 하위 그룹을 구성하는 방법으로 모든 하위 그룹 개수마다 군집화를 수행함으로써 속도가 느린 단점이 있다.

본 논문에서는 외부 커뮤니티 연관도 기반의 군집화에서 하위 그룹의 구조를 동적으로 변화시켜서 하위 그룹 선택의 정확도와 속도를 향상시키는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 살펴보고, 3장에서는 제안하는 동적 하위 그룹 선택 방법을 살펴본다. 4장에서는 실험과 결과를 분석하고 5장에서 결론을 맺는다.

2. 관련 연구

2.1 관련 연구

네트워크 안에서 하위 그룹을 발견하는 것은 타당한 실용성을 가진다[8]. 특히 소셜 네트워트에서 하위 그룹들은 실제 사회 그룹에서 나타나는 현상으로도 표현될 수 있다[3]. 인용문 네트워크(Co-Citation Networks)의 커뮤니티에서 페이지들이나 단일 토픽 등의 연관성을 나타낼 수 있으며[14], 범죄 수사에서도 범죄 수사의 시각화를 통해서 하위 그룹들의 연관성을 나타내기도 한다[15].

시멘틱 웹 개념의 발달을 통해 소셜 네트워크 모델링과 분석에 대한 접근 방법 중의 하나로 온톨로지 활용이 관심을 얻고 있다[7]. 소셜 네트워크는 사회적 엔티티에 대한 모델로 이루어진 온톨로지 기반 관계에서 추론 메타니즘을 통해 노드 간의 새로운 관계를 발견할 수 있다[16]. 리소스와 리소스의 준 연결망에는 관련된 사회적 노드가 함께 연결되어 있기 때문에, 노드 간의 다양한 의미적 연결 관계에 대한 표현과 추론이 가능하다. 이를 바탕으로 아마존 서점의 추천 시스템 (Recommendation System)이나 협업 필터링 (Collaborative Filtering)을 기반으로 한 신규추천 문제 등이 연구되고 있다[5].

군집화는 집단 사이의 계층적인 의미를 가지는 하위 그룹을 의미하며 광범위하고 다양한 연구가 이루어지고 있다. 포털 사이트에서 서비스되고 있는 뉴스 클러스터링 같은 경우 텍스트 마이닝(Text Mining)을 이용한 뉴스 안의 그룹

발견과 토픽 캡쳐에 의한 관계 연구를 통해 발달해 왔다[17,18]. 커뮤니티 분야에서는 텍스트 마이닝을 사용한 커뮤니티 군집화를 비롯하여 개인적 지역 감지 알고리즘 등의 연구가 이루어지고 있다[19,20].

2.2 외부 커뮤니티 연관도

외부 커뮤니티 연관도는 특정 하위 그룹과 다른 외부 하위 그룹과의 연관도를 이용하여 계산한다[4]. 전체 외부 커뮤니티 연관도 R은 각 하위 그룹의 외부 하위 그룹에 대한 연관도의 평균으로 정의된다. 외부 커뮤니티 연관도는 군집화의 기본 사상 중 군집 외부의 데이터와의 연관도는 최소화 한다는 것을 극대화 하고자 하는 방법이다. 그래서 일반적인 거리 계산 척도인 유클리디언 거리를 사용하였을 때는 거리가 작을수록 연관도가 큰 것이기 때문에 R이 큰 값을 가질수록 좋은 결과를 보인다. 반면에 유사도 계산 척도인 코사인 계수(cosine coefficient), 피어슨 상관계수 (Pearson's correlation coefficient) 등은 값이 클수록 연관도가 높기 때문에, R이 작은 값을 가질수록 군집화 성능은 좋아진다.

$$R = \frac{\sum_{k \in K} R_k}{K} \tag{1}$$

여기서 K는 하위 그룹의 개수이다. 단일 외부 커뮤니티 연관도 R_k는 한 하위 그룹에 속한 노드들의 다른 하위 그룹들과의 유사도에 대한 표준편차로 정의되며 이는 식 (2)와 같다.

$$R_k = \sqrt{\frac{1}{N_k} \sum_{i \in N_k} d^2(r_i, \bar{r})} \tag{2}$$

여기서 , $d(r_i, \bar{r})$는 두 벡터 r_i와 $\bar{r}$ 사이의 거리이고, N_k는 k 하위 그룹에 속한 노드들의 개수이며, $\bar{r}$은 r_i들의 평균이다. r_i는 i번째 노드와 외부 하위 그룹들과의 유사도이다.

노드와 외부 그룹들과의 유사도는 일반적인 유사도가 아닌 노드가 영향 받는 방향으로 계산한다. 일반적인 유사도는 노드가 속한 하위 그룹을 결정하는데 사용한 척도이다. 이 척도와 하위 그룹들 간의 방향을 이용하여 새로운 노드와 외부 하위 그룹들과의 유사도를 계산한다.

$$r_i = \sum_{k \in K, k \neq j} d_{ik} \bullet \overline{C_j C_k} \tag{3}$$

여기서, j는 i 노드가 속한 하위 그룹의 번호이다. d_{ik}는 i 노드와 k 하위 그룹간의 유사도이고, $\overline{C_j C_k}$는 j 하위 그룹과 k 하위 그룹 사이의 방향을 의미한다.

3. 동적 하위 그룹 선택 방법

3.1 소셜 매트릭스

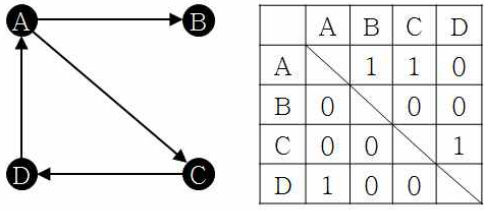
소셜 노드와 노드 사이의 관계를 매트릭스 형태로 나타낸 것을 소셜 매트릭스라고 한다. 소셜 매트릭스는 $n \times n$ (n은 노드의 수) 행렬이다. 한 네트워크에 A,B,C,D 노드가 있을 때, (그림 1, a)와 같이 방향이 있는 관계는 정보나 재화의 흐름을 나타낸다. 방향이 있는 매트릭스에서 행은 정보를 주는 노드이고, 열은 정보를 받는 노드를 의미한다[2]. 소셜 미디어에서 가장 대표적인 방향이 있는 네트워크는 트위터 (twitter)이다. RT로 표현되는 리트윗 (retweet)과 @으로 표현하는 답글 달기 (Reply)를 통해서 트위터 사용자들 사이의 정보의 흐름을 분석 할 수 있다.

(그림 1)의 b와 같이 관계의 방향이 없이 연관성만을 나타내는 네트워크도 존재한다. 이 때 소셜 매트릭스의 행과 열은 같은 정보를 나타낸다[2]. 소셜 매트릭스의 수치를 통해서 관계의 강도 (intensity)를 표현할 수 있다. 대학생들이 수업을 듣는 과목들을 분석해 보면, 과목들 간에는 정보의 흐름이 없기 때문에 방향성은 존재하지 않는다. 하지만, A, C 두 과목을 동시에 수강하는 학생들이 많을수록 두 과목은 서로 연관있는 (같은 전공 내의 과목 또는 인기 교양 강좌 등) 과목으로 판단하면, 이를 표현하기 위하여 관계 사이의 강도를 사용할 수 있다.

3.2 하위 그룹 분석

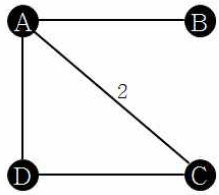
소셜 네트워크의 규모가 작으면, 쉽게 해당 소셜 네트워크의 특징을 분석하는 것이 가능하지만, 소셜 네트워크의 규모가 크면, 해당 소셜 네

(그림 1) 소셜 매트릭스



	A	B	C	D
A		1	1	0
B	0		0	0
C	0	0		1
D	1	0	0	

a) Relation with direction but without strength



	A	B	C	D
A		1	2	1
B	1		0	0
C	2	0		1
D	1	0	1	

b) Relation without direction but with strength

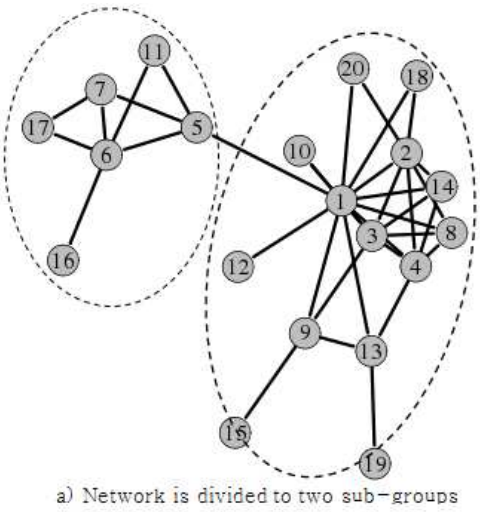
(Figure 1) Social Matrix

트워크를 한 번에 분석하는 것이 불가능하게 된다. 따라서 소셜 네트워크를 비슷한 집단으로 나누는 하위 그룹을 생성하여 분석할 필요가 있다.

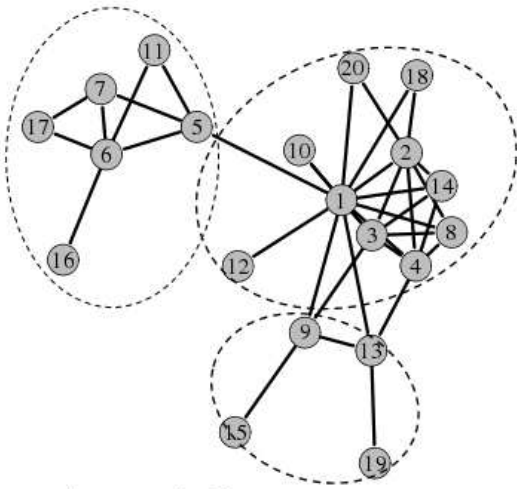
노드들에 대한 다양한 분석을 통하여 소셜 네트워크를 구성하는 하위 그룹을 분석할 때, 노드의 수가 몇 십개이상이 되면 탐색양이 증가하여 노드별로 분석하기 어렵다. 따라서 자동적으로 하위 그룹을 구성하는 방법이 필요하다. 하지만, 소셜 네트워크의 특성상 하위 그룹의 개수를 정확하게 예측하는 것은 불가능하다.

하위 그룹의 개수의 차에 의해서 소셜 네트워크의 구성이 달라지는 현상을 살펴보면 다음과 같다. (그림 2)는 20개의 노드를 가지는 소셜 네트워크이다. 같은 소셜 네트워크에 대해 (그림 2)의 a에서는 2개의 하위 그룹으로 나누었고, (그림 2)의 b에서는 3개의 하위 그룹으로 나누었다. (그림 2)의 b를 보면, 오른쪽 두 개의 하위 그룹 사이에는 각각 2개, 3개의 노드들이 서로 연결되어 있고, 이는 다른 노드들을 보았을 때 하위 그룹으로 분리할 수 있는 결정적인 조건이 될 수는 없다. 하지만, 군집화 알고리즘의 특성상 하위 그룹의 개수를 3개로 지정했기 때문에, 3개의 그룹으로 구성하게 된다. 따라서 (그림 2)의 b를 이용하여 소셜 네트워크를 분석하면, (그림 2)의 a를 이용한 분석에

(그림 2) 20개의 노드를 가진 네트워크에 대한 하위 그룹 선정 결과



a) Network is divided to two sub-groups



b) Network is divided to three sub-groups

(Figure 2) Results of making sub-groups for network with 20 nodes

비하여 유사한 특징을 가지는 하위 그룹들이 나타난다. 이는 소셜 네트워크 분석 결과의 신뢰도를 저하시키는 요인이 된다.

3.3 군집화 알고리즘

대상들을 군집화하는 방법은 다양하지만 공통적인 기본전제는 군집 내의 객체들 간의 유사성을 극대화 하고, 군집간의 유사성은 극소화하는 것이다. 군집화 알고리즘에는 *k*-means 군집화

알고리즘과 같은 분할 기법(partitional)과 계층적(hierarchical)인 알고리즘 등이 존재한다[21].

계층적 군집화 알고리즘은 개별 노드들 간의 거리를 비교하는 방법으로, 항상 계산 결과가 도출되고, 같은 차원의 다른 문제에 대해 항상 계산시간이 일정한 장점이 있지만, 노드의 개수나 노드를 구분하는 특징의 개수가 커지면 전체 계산시간이 기하급수적으로 증가하는 단점이 있다[22]. 소셜 네트워크 데이터에 대한 군집화를 수행할 때는 소셜 네트워크를 구성하는 노드의 개수가 많고 또한, 노드를 구분하는 특징이 소셜 네트워크만큼 존재하기 때문에 대용량 매트릭스를 구성하게 된다. 또한, 네트워크의 노드의 개수가 증가할수록 노드의 특징 벡터도 증가하기 때문에 계층적 알고리즘을 사용하면, 소셜 네트워크의 노드의 수가 증가할수록 분석 시간은 점점 더 많이 소요된다. k-means는 일정 시간 내에 계산이 종료할 수 있고, 노드의 개수나 노드를 구분하는 특징의 개수가 커져도 계산 시간이 점진적으로 증가하는 경향을 보인다.

본 논문에서는 군집을 구성할 때 k-means 군집화 알고리즘을 사용하였으며, 유사성 척도는 코사인 상관계수를 사용하였다. 초기군집형성을 위한 k-means 알고리즘은 <표 1>과 같다.

<표 1> K-means 알고리즘

step 1:	Initialize the k centers of cluster for $C_i (i \in (1...k))$.
step 2:	Calculate the distance between M input data $X_j (j \in \{1...M\})$ and cluster centers C_i. Allocate the input data to cluster which has smallest distance. $$u_{ij} = \begin{cases} 1 \text{ if } \frac{x_j c_i}{\sqrt{x_j^2 c_i^2}} \leqq \frac{x_j c_k}{\sqrt{x_j^2 c_k^2}} \\ 0 \; otherwise \end{cases} \quad (4)$$ $where\, k \in \{1...k, k \neq i\}, j \in \{1...M\}$
step 3:	Recalculate the center of each clusters $$c_i = \frac{1}{N_i} \sum_k x_k \quad (5)$$ where N_i is a number data in cluster i.
step 4:	Repeat from step 2 to step 3 until belonging cluster of input data is not changed.

<Table 1> k-means algorithm

k-means 알고리즘의 단점은 하위 그룹의 개수인 k를 미리 선정해야 한다는 것이다. 본 논문에서는 외부 커뮤니티 연관도를 사용한 동적 하위 그룹 선택 방법을 사용하여 하위 그룹의 개수를 자동으로 결정한다.

3.4 동적 하위 그룹 선택 방법

군집화 알고리즘을 활용한 동적 하위 그룹 선택 방법은 <표 2>와 같다.

<표 2> 동적 하위 그룹 선택 방법

step 1:	Set an initial number of clusters to be k_0 which should be sufficiently small. Typically k_0 = 2.
step 2:	Apply k-means to data for $k = k_0$. We name the divided sub-groups $C_1, C_2, ..., C_{k0}$.
step 3:	Calculate the external community relationship for all sub-groups. We name the external community relationships $R_1, R_2, ..., R_{k0}$.
step 4:	Select the sub-group C_{km}, which has max external community relationship R_{km} .
step 5:	For a sub-group C_{km}, apply k-means by setting k = 2. We name the divided sub-groups $C_{km}^{(1)}, C_{km}^{(2)}$.
step 6:	Calculate the external community relationship for all sub-groups. We name the external community relationships $R_{km}^{(1)}, R_{km}^{(2)}$.
step 7:	If $R_{km} \geq R_{km}^{(1)}$ and $R_{km} \geq R_{km}^{(2)}$, we accept the two-divided model, and decide to continue; we set $k_0 \leftarrow k_0 + 1$ $R_{km} \leftarrow R_{km}^{(1)}, C_{km} \leftarrow C_{km}^{(1)}$ $R_{k0} \leftarrow R_{km}^{(2)}, C_{k0} \leftarrow C_{km}^{(2)}$. We recalculate the external community relationship for other sub-groups. Go to step 4.
step 8:	If $R_{km} < R_{km}^{(1)}$ or $R_{km} < R_{km}^{(2)}$, we set $O\_R_{km} \leftarrow R_{km}$ $k \leftarrow k_0 + 1$. Execute from step 2 to step 4, which

apply k-means to all data.
step 9: If $O\_R_{km} \geq R_{km}$, we accept the new model, and go to step 5.
step 10: If $O\_R_{km} < R_{km}$, we set $k \leftarrow k_0 - 1$. Execute from step 2 to step 4, which apply k-means to all data.
We decide to stop process.

<Table 2> Dynamic sub-groups discovering method

제안하는 방법은 2차 k-means를 기본으로 특정 분기 시점에서만 전체 k-means를 수행함으로써 순차적으로 증가시키면서 수행하는 방법에 비해 전체 수행 시간을 줄일 수 있다.

<표 2>의 7단계에서 10단계의 외부 커뮤니티 연관도를 비교하는 식은 연관도 계산 척도에 의해 달라진다. 2.2 절에서 기술한 바와 같이 유클리디언 거리 등은 외부 커뮤니티 연관도 R이 클수록 좋은 것이고, 코사인 계수, 피어슨 계수 등은 외부 커뮤니티 연관도 R이 작을수록 좋은 것이다. <표 2>는 코사인계수 척도를 상관 계수로 사용하였을 경우이다. 만약 본 방법을 적용시키는데 있어 유클리디언 거리를 사용하는 경우에는 부등호의 방향을 반대로 하여 적용하면 된다.

4. 실험환경 및 결과

4.1 커뮤니티 평균 순도

하위 그룹을 발견하는 알고리즘의 정확도 기준으로는 발견한 커뮤니티의 평균 순도(average purity)를 사용했다[23]. 그 정의는 다음과 같다.

$$\frac{1}{k}\sum_{k}\sum_{i\neq j \text{and} i,j\in C_k}\frac{\delta(i,j)}{|C_k|^2} \quad (6)$$

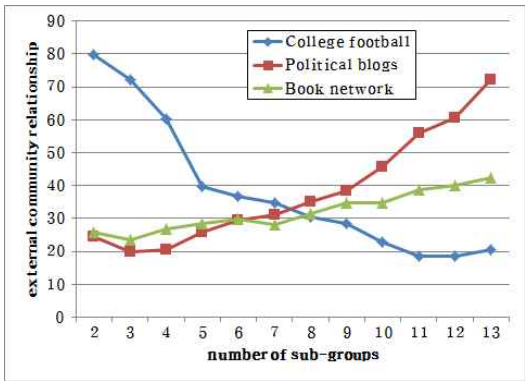
k는 발견된 하위 그룹의 개수이며 $\delta(i,j)$는 노드 i와 j가 실제로도 같은 커뮤니티에 속하면 1, 아니면 0의 값을 갖는다. $|C_k|$ 는 발견된 하위 그룹 중 k번째 것의 크기를 의미한다.

4.2 하위 그룹의 개수에 따른 외부 커뮤니티 연관도

하위 그룹의 개수를 변화시키면서 외부 커뮤니티 연관도의 값의 변화를 살펴봄으로써 외부 커뮤니티 연관도가 하위 그룹을 발견하는데 영향을 끼치는 여부를 확인하였다.

실험에는 College football[3], Political blogs[24], Book network[25] 데이터 집합을 사용했다. College football은 12개, Political blogs는 3개, Book network는 2개의 하위 그룹으로 구성되어 있다. College football은 115개의 football team이 존재하고, Political blogs는 1,494개의 blogs, Book network는 46개의 책으로 구성되어 있다.

(그림 3) 하위 그룹의 개수의 증가에 따른 외부 커뮤니티 연관도의 변화

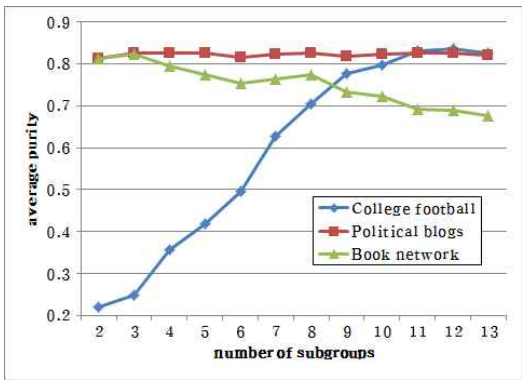


(Figure 3) external community relationship of three data sets with different number of sub-groups

하위 그룹의 개수를 2부터 시작하여 13까지 증가시키면서 세 데이터 집합에 대한 외부 커뮤니티 연관도의 변화를 살펴보면 (그림 3)과 같다. (그림 3)을 보면, College football의 경우 하위 그룹의 개수가 증가함에 따라 외부 커뮤니티 연관도가 지속적으로 감소하는 것을 확인할 수 있다. 하위 그룹의 개수가 11일 때와 12일 때를 살펴보면, 11일 때는 18.6, 12일 때는 18.4로 12일 때 최소값을 보이고, 13일 때는 다시 증가하는 것을 확인할 수 있다. 따라서 College football의 하위 그룹의 개수는 12가 된다. Political blogs와 Book network를 살펴보면, 두 개의 데이터 집합 모두 하위 그룹의 개수가 3일 때, 각각 19.8, 23.5로 외부 커뮤니티 연관도가

최소값을 가지며, 하위 그룹의 개수가 증가할수록 외부 커뮤니티 연관도는 지속적으로 증가하는 것을 확인할 수 있다. 따라서 제안하는 방법에 의해 Political blogs의 하위 그룹의 개수는 3이 되고, Book network의 하위 그룹의 개수는 3이 된다. 데이터 집합에서 제공되는 하위 그룹의 개수는 College football이 12, Political blogs가 3, Book network가 2이고, 제안하는 방법은 각각 12, 3, 3을 하위 그룹으로 선택하고 있다. College football과 Political blogs는 정확하게 하위 그룹을 선택했고, Book network는 1개의 차이를 보이고 있다.

(그림 4) 하위 그룹의 개수의 증가에 따른 커뮤니티 평균 순도의 변화



(Figure 4) average purity of three data sets with different number of sub-groups

하위 그룹의 개수가 증가하면서 세 데이터 집합에 대한 커뮤니티 평균 순도의 변화를 살펴보면 (그림 4)와 같다. (그림 4)를 보면, College football의 커뮤니티 평균 순도는 지속적으로 증가하다가 하위 그룹의 개수가 12일 때 0.8354로 최대값을 보이기 때문에 하위 그룹의 개수가 12일 때 하위 그룹 선택의 정확도가 가장 높았다. Political blogs는 커뮤니티 평균 순도값은 유사하지만, 하위 그룹의 개수가 3일 때와 12일 때, 0.8266으로 최대값을 보인다. 즉, 선택된 하위 그룹의 개수 3이 최적의 값이다. Book network의 경우 하위 그룹의 개수가 3일 때 0.8242로 최대값이다. (그림 3)과 (그림 4)를 살펴보면, 제안하는 방법에 의해서 선택된 하위 그룹의 개수가 실제로 구조적으로 최적의 하위 그룹의 개수를 보이는 것을 확인할 수 있다.

4.3 기존 알고리즘과 성능 비교

본 논문에서 제안하는 방법의 성능을 기존 알고리즘과 비교하였다. 기존 알고리즘은 Newman의 leading eigenvector 방법[26]과 Ghosh 등의 확장된 modularity- maximization 방법[27], Hyunjin Lee 등의 외부 커뮤니티 연관도를 이용한 방법[4]이다. 실험에 사용한 데이터는 앞에서 사용한 데이터와 같은 College football과 Political blogs, Book network의 3개의 데이터 세트를 사용하였다. 실험 결과는 <표 3>과 같다.

제안하는 방법은 College football 데이터와

<표 3> 제안하는 방법과 기존 방법과의 비교 실험 결과

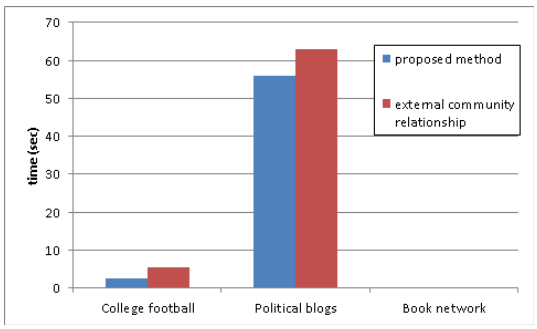
	College football		Political blogs		Book network	
	number of sub-groups	average purity	number of sub-groups	average purity	number of sub-groups	average purity
proposed method	12	0.8354	3	0.8266	3	0.8242
leading eigenvector method	12	0.5865	2	0.8920	6	0.5867
modularity-maximization method	6	0.5808	4	0.6457	3	0.6351
external community relationship	11	0.8350	3	0.8270	3	0.7121

<Table 3> Comparison of experimental results between proposed method and other methods

Book network 데이터에서는 가장 높은 커뮤니티 평균 순도를 기록했으나, Political blogs에 대해서는 세 번째 커뮤니티 평균 순도를 나타냈다. 하위 그룹의 개수를 살펴보면, 제안하는 방법과 leading eigenvector 방법, 외부 커뮤니티 연관도를 이용한 방법과 유사한 정확도로 하위 그룹의 개수를 찾아내는 것을 확인할 수 있다. 하지만 제안하는 동적 하위 그룹 선택 방법은 하위 그룹의 개수와 커뮤니티 평균 순도를 같이 고려했을 때 보다 정확한 하위 그룹의 개수를 발견하면서, 높은 커뮤니티 평균 순도를 보이는 것을 확인할 수 있다.

제안하는 방법의 수행 시간을 3.4절에 기술한 외부 커뮤니티 연관도를 이용한 방법과 비교하였다. 성능을 비교한 실험과 동일하게 3개의 데이터 집합을 사용하였다. 실험 결과는 (그림 5)와 같다.

(그림 5) 실행시간 비교



(Figure 5) Execution Time Comparison

3개의 데이터 집합에 대해 제안하는 방법이 더 빠른 실행 결과를 보였다. College football 데이터의 경우 반복 횟수가 많기 때문에 2배 이상의 속도를 보이고 있다. Political blogs와 Book network는 반복 횟수가 적어서 10% 정도의 성능 향상을 보였다. 하지만, 제안하는 방법은 2개의 초기 군집의 개수에서 시작하지만, 외부 커뮤니티 연관도를 이용한 방법은 3개의 초기 군집의 개수에서 시작한다. 군집을 계산하는 횟수가 제안하는 방법보다 적지만 제안하는 방법이 더 성능이 우수하였다.

5. 결 론

본 연구에서는 소셜 네트워크의 하위 그룹을 분석하기 위하여 군집화 방법을 적용할 때, 동적으로 하위 그룹의 개수를 결정하는 방법에 대해서 연구하였다. 기존에는 하위 그룹의 개수를 임으로 고정하거나 구조 분석을 통하여 설정하였지만, 이는 부정확한 결과를 보이거나 사람의 개입이 많아지는 단점이 있다. 본 연구의 동적 하위 그룹 선택 방법을 사용하여 소셜 네트워크를 분석한다면, 사용자의 개입 없이 소셜 네트워크의 구조를 쉽게 분석할 수 있을 것으로 기대된다. 소셜 네트워크의 하위 그룹을 구성하고, 이를 시각화함으로써, 실무적인 측면에서 전체 소셜 네트워크를 분석하는데 필요한 노력을 줄일 수 있고, 영향력자(influencer)나 악성 소문 유포자 등에 대한 분석이 쉽게 이루어 질 수 있을 것이다.

현재 주류의 소셜 네트워크 서비스는 문자에 특화되어 있지만, 스마트폰의 활성화 등으로 인하여 사진, 동영상 등 멀티미디어를 중심으로 한 서비스인 flickr, pinterest 등의 소셜 네트워크 서비스도 주목받고 있다. 따라서 다양한 미디어를 이용한 소셜 네트워크 서비스를 분석하기 위한 방법에 대해서도 연구가 필요할 것이다.

References

[1] C.T. Butts, “Social network analysis :A methodological introduction”, Asian Journal of Social Psychology, Vol. 11, pp. 13-41, 2008.

[2] Eun-Young Kang, and Kee-Young Kwahk, “Managing Duplicate Memberships of Websites : An Approach of Social Network Analysis,” Journal of Intelligence and Information Systems, Vol. 17, No. 1, pp. 153-169, 2011.

[3] Seong-Hee Kim, and Rho-Sa Chang, “The Study on the Research Trend of Social Network Analysis and the its Applicability to Information Science”, Journal of Korea Society for Information Management, Vol. 27, No. 4, pp. 71-87, 2010.

[4] Hyunjin Lee, and Taechang Jee, “Social Networks Analysis using External Community Relationship”, Journal of Digital Contents Society, Vol. 12, No. 1, pp. 69-75, 2011.

[5] Jong Hak Park, Yoon Ho Cho, and Jae Kyeong Kim, “Social Network : A Novel Approach to New Customer Recommendations”, Journal of Intelligence and Information Systems, Vol. 15, No. 1, pp. 123-140, 2009.

[6] Hyoung-Do Kim, “Collaborative Filtering by Consistency Based Trust Definition”, Journal of Society for e-Business Studies, Vol. 14, No. 1, pp. 1-11, 2009.

[7] Seung-Hoon Lee, Ji-Hyeok Kim, Heung-Nam Kim, and Geun-Sik Jo, “Inferring and Visualizing Semantic Relationships in Web-based Social Network”, Journal of Intelligence and Information Systems, Vol. 15, No. 1, pp. 87-102, 2009.

[8] M. Girvan and M.E.J. Newman, “Community structure in social and biological networks”, Proceedings of the National Academy of Science, Vol. 99, No. 12, pp. 7821-7826, 2002.

[9] M.E.J. Newman, “Detecting community structure in networks”, Eur. Phys. J. B., Vol. 38, No. 2, pp. 321-330, 2004.

[10] Richard Freeman, “Topological Tree Clustering of Social Network Search Results” in Proceedings of the Eight International Conference on]Lecture Notes in Computer Science (LNCS 4481), Springer, pp. 760-769, 2007.

[11] Romain Boulet, Bertrand Jouvea, Fabrice Rossi, Nathalie Villa, "Batch kernel SOM and related Laplacian methods for social network analysis,"Neurocomputing, Vol. 71, PP. 1257-1273, 2008.

[12] Yong-Yeol Ahn, James P. Bagrow, Sune Lehmann, " Link communities reveal multiscale complexity in networks, "nature, vol. 466, pp.761-764. 2010.

[13] Yoonseip Kang, and Seungjin Choi, “Social Network Analysis using Common Neighborhood Subgraph Density,” Journal of KIISE:Computing Practices and Letters, Vol. 16, No. 4, pp. 432-436, 2010.

[14] C. Chen, “Visualizing Semantic Spaces and Author Co-Citation Networks in Digital Libraries,” Information Processing Management, Vol. 35, No. 3, pp. 401-420, 1999.

[15] F. Kerschbaum and A. Schaad, “Privacy-Preserving Social Network Analysis for Criminal Investigations,” In WPES, Alexandria, Virginia, 2008.

[16] P.O. Wennerberg, “Ontology Based Knowledge Discovery in Social Networks,” Final Report, JRC Joint Research Center, 2005.

[17] D. Joshi and D. GaticaPerez, “Discovering Groups of People in Google News,” Proceedings of the 1st ACM international workshop on Human-centered multimedia, pp. 55-64, 2006.

[18] Hyunjin Lee, and Taechang Jee, “A Study on Optimizing the Number of Clusters using External Cluster Relationship Criterion”, Journal of Digital Contents Society, Vol. 12, No. 3, pp. 339-345, 2011.

[19] P. Velardi, R. Navigli, A. Cucciarelli and F. D’Antonio, “A New Content-Based Model for Social Network Analysis,” Proceedings of the 2008 IEEE International Conference on Semantic Computing, pp. 18-25, 2008.

[20] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar and L. Terveen, “Discovering Personal Gazetterrs : An Interactive Clustering Approach,” Proceedings of ACM GIS, 2004.

[21] Earl Gose, Richard Johnsonbugh and Steve Jost, (1996) Pattern Recognition and Image Analysis. Prentice Hall.

[22] Douglass R. Cutting, David R. Karger, Jan O. Pedersen and John W. Tukey, "Scatter/Gather: a cluster-based approach to browsing large document collections", In Proc. of the 15th annual international ACM

SIGIR, June, pp. 318-329, 1992.

[23] Y. Wang, H. Song, W. Wang and M. An, "A microscopic view on community detection in complex networks," Proceeding of the 2nd PhD Workshop on Information and Knowledge Management, New York, USA, pp. 57-64, 2008.

[24] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," Proceedings of the 3rd International Workshop on Link Discovery, Chicago, Illinois, pp. 36-43, 2005.

[25] Valdis Krebs, "Working in the Connected World: Book Network,", IHRIM Journal, pp. 87-90, 2000.

[26] M. E. J. Newman , "Finding community structure in networks using the eigenvectors of matrices," Physical Review E, Vol. 74, No. 3, 19 pages, 2006.

[27] R. Ghosh and K. Lerman, "Structure of Heterogeneous Networks," International Conference on Computational Science and Engineering, vol. 4, pp.98-105, 2009

이 현 진

1996년: 순천향대학교 전산학 학사
1998년: 연세대학교 대학원 컴퓨터과학 석사
2002년: 연세대학교 대학원 컴퓨터과학 박사
2003년~현재: 숭실사이버대학교 컴퓨터정보통신학과 부교수
관심분야 : 이러닝, 기계학습, 데이터마이닝