# 사회 연결망에서의 커뮤니티 탐지 정확도 향상을 위한 엣지 웨이팅 기법 성능 비교

강윤석 이준석 김상욱[1]
한양대학교
{dyskang, junseoklee, wook}@hanyang.ac.kr

# Performance Comparison of Edge Weighting Scheme for Accurate Community Detection in Social Networks

Yoonsuk Kang Junseok Lee Sang-Wook Kim
Hanyang University

**요 약**

커뮤니티 탐지는 주어진 그래프에서 유사한 성향을 보이는 노드들을 찾는 방법을 의미한다. 커뮤니티 탐지 알고리즘은 커뮤니티의 특성인 (1) 커뮤니티가 동일한 노드간의 엣지 (i.e., intra-community 엣지) 수는 많고, (2) 커뮤니티가 서로 다른 노드간의 엣지 (i.e., inter-community 엣지) 수는 적다는 것을 이용한다. 많은 실세계 그래프에서 커뮤니티 탐지 정확도가 낮은 문제를 보이는데, 이는 주어진 그래프가 이러한 커뮤니티 특성을 제대로 만족하지 않기 때문이다. 이를 해결하기 위해 주어진 그래프의 엣지에 비중을 부여하는 엣지 웨이팅 (edge weighting) 기법들이 제안되어 왔다. 본 논문에서는 다양한 엣지 웨이팅 기법들을 소개하고 이들의 성능을 비교한다.

## 1. 서 론

다양한 분야에 복잡한 그래프 형태를 보이는 데이터가 많이 존재한다. 이러한 데이터의 각 object는 노드, object와 object를 연결하는 link는 엣지로 표현된다 [1]. 이러한 그래프 데이터를 이해하는데 있어서 사용되는 방법중 하나는 유사한 성향을 보이는 노드들을 하나의 커뮤니티로 묶는 것이다. 사회연결망 그래프에서는 커뮤니티내의 유사한 노드들은 비슷한 성향을 가진 사용자들이 될 수 있고, 논문 인용그래프에서는 비슷한 주제를 다루는 논문들이 될 수 있다. [1]

주어진 그래프에서 토폴로지컬 정보를 이용하여 이러한 커뮤니티 구조(i.e. 커뮤니티들)를 찾는 것을 커뮤니티 탐지라고 한다 [1][2]. 커뮤니티 탐지는 다음과 같은 두가지 커뮤니티 특성을 이용하여 찾는다: (1) 커뮤니티가 동일한 노드간의 엣지 (i.e., intra-community 엣지) 수는 많고, (2) 커뮤니티가 서로 다른 노드간의 엣지 (i.e., inter-community 엣지) 수는 적다.

만약 주어진 그래프가 이 두가지 커뮤니티 특성을 제대로 충족한다면 기존 커뮤니티 탐지 알고리즘들은 높은 정확도로 커뮤니티 구조를 찾을 수 있을 것이다. 그러나 많은 실세계 그래프는 엣지의 수가 매우 적다 (sparse) [2]. 따라서 높은 커뮤니티 탐지 정확도를 보이는 것은 매우 어려운 일이다. Sparse 그래프에서는 intra-community 엣지의 수가 충분히 많지 않고, inter-community 엣지의 수는 그렇게 적지 않기 때문이다.

커뮤니티 탐지 알고리즘을 수행하기 전에, 커뮤니티 탐지의 정확도를 향상시키기 위해 엣지에 비중을 부여하는 다양한 엣지 웨이팅 기법들이 제안되어 왔다 [3][4]. 본 논문에서는 다양한 실세계 그래프에서 다양한 커뮤니티 탐지 알고리즘을 이용하여, 엣지 웨이팅 기법들의 커뮤니티 탐지 성능 개선 정도가 어떤지 확인하고자 한다.

## 2. 관련 연구

### 2.1. 커뮤니티 탐지 알고리즘

다양한 커뮤니티 탐지 알고리즘들이 개발되어 왔고, label propagation [5]과 Louvain [6]은 대표적인 커뮤니티 탐지 알고리즘의 예이다. Label propagation은 각 노드의 커뮤니티를 해당 노드의 이웃 노드들의 커뮤니티를 바탕으로 결정하는 방법이고, Louvain은 modularity 값이 가장 높은 커뮤니티 구조를 주어진 그래프의 커뮤니티 구조로 결정하는 방법이다.

### 2.2. 엣지 웨이팅 (edge weighting) 기법

커뮤니티 탐지 정확도를 향상시키기 위해, 주어진 그래프의 토폴로지컬 정보만을 이용하여 엣지에 비중을 부여하는 엣지 웨이팅 (edge weighting) 기법들이 제안되어왔다 [3][4]. [3]에서는 두 노드의 특징 값들 (common neighbor의 개수, clustering coefficient의 차이, Jaccard index, resource allocation index, Adamic-Adar index, relative degree) 을 aggregate하여 두 노드를 연결하는 엣지의 비중($w_e$)을 결정한다. 이때, 해당 값들을 aggregate하기 위해 <식 1>과 같이 linear regression을 사용한다.

$$w_e = p_0 + \sum p_i x_e^{<i>} \qquad \text{<식 1>}$$

[1] 교신저자

* 이 논문은 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2018R1A5A7059549, No. NRF-2017R1A2B3004581).

[4]에서는 두 노드의 degree 차이를 바탕으로 <식 2>를 이용하여 해당 노드들을 연결하는 엣지의 비중($w_e$)을 결정한다. 이때 $p_i$는 linear regression의 파라미터 값, $x_e^{<i>}$는 i번째 특징 값, v와 u는 엣지 e의 양 끝 노드들, 그리고 deg(v)는 노드 v의 degree를 나타낸다.

$$w_e = 10^{-floor(\log(|\deg(v)-\deg(u)|))} \quad \text{<식 2>}$$

## 3. 성능 비교

### 3.1. 실험 환경

2.2장에서 언급한 두가지 엣지 웨이팅 기법의 커뮤니티 탐지 정확도 개선 정도를 비교하기 위해 우리는 3개의 실세계 네트워크 데이터 (karate, football, DBLP) 를 이용하였다. Karate는 대학교 가라데 클럽의 인간관계 그래프로, 노드는 클럽의 멤버, 엣지는 멤버간의 친분을 나타내고, 커뮤니티는 가라데 클럽을 나타낸다. Football은 미국 축구 리그를 나타내는 그래프로, 노드는 축구팀, 엣지는 팀간의 경기, 커뮤니티는 컨퍼런스 (group) 을 나타낸다. DBLP는 논문간의 인용 그래프로, 노드는 논문, 엣지는 논문간의 인용, 커뮤니티는 학회명을 나타낸다. <표 1>은 각 그래프의 자세한 정보를 나타낸다.

<표 1> 실세계 그래프 정보

| | 노드 수 | 엣지 수 | 커뮤니티 수 |
|---|---|---|---|
| **Karate** | 34 | 78 | 2 |
| **Football** | 115 | 613 | 12 |
| **DBLP** | 13,184 | 47,937 | 5 |

우리는 2.1장에서 언급한 두개의 커뮤니티 탐지 알고리즘인 Label propagation과 Louvain을 사용하였다. 커뮤니티 탐지 정확도를 측정하기 위해 우리는 normalized mutual information (NMI) 을 이용하였다. NMI는 실제 커뮤니티 구조와 커뮤니티 탐지 알고리즘을 통해 얻은 커뮤니티 구조가 얼마나 일치하는지 측정하는 메져로 1에 가까운 수치일수록 알고리즘을 통해 얻은 커뮤니티 구조가 실제 커뮤니티 구조에 근사한 것을 나타낸다.

### 3.2. 성능 분석 결과

<표 2><표 3>는 label propagation, Louvain 커뮤니티 탐지 알고리즘의 탐지 정확도 결과를 나타낸다. 행은 그래프를 나타내고, 엣지 웨이팅 기법을 나타낸다. (i, j)의 값은 i번째 행 그래프에서 j번째 기법을 이용했을 때 label propagation의 NMI를 나타내며, 괄호 안의 값은 original이랑 비교했을 때의 성능 증감 비율을 나타낸다. 이때 original은 비중을 부여하지 않은 그래프에서의 커뮤니티 탐지 정확도를 의미한다.

12가지 조합(2개 커뮤니티 탐지 알고리즘 × 3개 그래프 × 2개 엣지 비중 부여 기법)중 8개에서 성능이 향상되는 것을 보였고 1개는 기존과 동일했으며 3개는 성능이 떨어지는 것을 확인할 수 있었다.

[3]의 엣지 웨이팅 기법은 6개의 조합 (2개 커뮤니티 탐지 알고리즘 × 3개 그래프) 중 5개에서 성능이 향상되는 것을 확인 할 수 있었고 1개는 성능이 떨어지는 것을 확인하였다. 성능이 오르는 경우 original에 비해 평균적으로 16% 향상되는 것을 확인 할 수 있었다. 반면 [4]의 엣지 웨이팅 기법은 6개의 조합 중 3개에서 성능이 향상되는 것을 확인 할 수 있었고, 1개는 기존과 동일했으며, 나머지 2개는 성능이 떨어지는 것을 확인하였다.

<표 2> Label propagation 결과

| | Original | [3] | [4] |
|---|---|---|---|
| **Karate** | 0.70 | 0.84(20%) | 0.70(-) |
| **Football** | 0.83 | 0.91(10%) | 0.89(7%) |
| **DBLP** | 0.43 | 0.36(-16%) | 0.34(-8%) |

<표 3> Louvain 결과

| | Original | [3] | [4] |
|---|---|---|---|
| **Karate** | 0.56 | 0.71(26%) | 0.59(5%) |
| **Football** | 0.90 | 0.91(1%) | 0.86(-4%) |
| **DBLP** | 0.41 | 0.51(27%) | 0.48(17%) |

## 4. 결 론

본 논문에서는 커뮤니티 탐지 정확도 향상을 위해 제안된 엣지 웨이팅 기법들을 소개하였다. 각 기법의 커뮤니티 탐지 성능 개선 정도를 확인하기 위해 다양한 실세계 그래프에서 다양한 커뮤니티 탐지 알고리즘의 정확도를 확인하였다. 실험을 통해 대체로 엣지 웨이팅 기법을 이용하면 타깃 커뮤니티 탐지 알고리즘의 커뮤니티 탐지 정확도가 기존 그래프에서의 커뮤니티 탐지 정확도에 비해 대체로 개선되는 것을 확인하였다. 그러나 몇몇 상황에서는 오히려 정확도가 떨어지는 것을 확인 할 수 있었다. 이후 우리는 정확도가 떨어지는 원인을 분석하고 이를 해결하기 위한 방안을 제안하고자 한다.

## 참 고 문 헌


[1] S. Cavallari et al., “Learning community embedding with community detection and node embedding on graphs,” *ACM CIKM*, 2017.

[2] L. Yang et al., “A unified semi-supervised community detection framework using latent space graph regularization,” *IEEE TOC*, 2015.

[3] X. Lu et al., “Adaptive modularity maximization via edge weighting scheme,” *Information Sciences*, 2018.

[4] M. Ciglan, M. Laclavik, and K. Norvag, “On community detection in real-world networks and the importance of degree assortativity,” *ACM SIGKDD*, 2013.

[5] U. Raghavan, R. Albert, and S. Kumara, “Near linear time algorithm to detect community structures in large-scale networks,” *Physical Review*, 2007.

[6] V. Blondel et al., "Fast unfolding of communities in large networks," *JSTAT*, 2008.