사회연결망에서의 링크 예측 정확도 향상을 위한 전처리 기법

손승범, 최연석, 강윤석, 김상욱[1)]
한양대학교 컴퓨터소프트웨어
{ssbum0925, oppp026, dyskang, wook}@hanyang.ac.kr

A Preprocessing Method for Accurate Link Prediction on Social Networks

Seungbeom Son, Yeonsuk Choi, Yoonsuk Kang, and Sang-Wook Kim
Dept. of Computer Science, Hanyang University

요 약

링크 예측은 주어진 그래프에서 가까운 미래에 발생할 가능성이 높은 새로운 링크를 예측하는 문제이다. 본 논문에서는 유사도 기반 링크 예측의 정확도를 향상시키는 전처리 기법을 제안한다. 제안하는 기법은 유사도 기반으로 예측한 링크들을 그래프에 추가하고, 이 추가된 링크들을 포함하는 그래프를 기반으로 다시 새로운 링크들을 예측하여 추가하는 점진적 추가 방식을 채택한다. 실세계 데이터를 이용한 실험을 통하여, 제안하는 전처리 기법이 기존 링크 예측의 정확도를 향상시킬 수 있는 것을 확인하였다.

1. 서론

실제 사회연결망은 그래프로 표현될 수 있고, 사회연결망에서의 객체와 두 객체 간의 관계는 그래프에서 각각 노드와 링크로 표현된다. 링크 예측은 가까운 미래에 주어진 그래프에서 새로운 링크를 예측하는 문제이다. 이러한 문제를 해결하기 위해 다양한 해결 방법들이 제시되어 왔지만, 링크의 수가 희소(sparse)한 실세계 그래프에서는 링크 예측의 정확도가 낮을 수 있다. 이를 해결하기 위해, 본 논문에서는 링크 예측의 정확도를 향상시키기 위한 전처리 기법을 제시하고 이를 실험을 통해 성능을 입증한다.

2. 관련 연구

2.1. 링크 예측

사회연결망에서의 링크 예측이란 특정 시점을 기준으로 미래의 시점에 새로 생길 링크를 예측하는 문제를 의미한다[1]. 일반적으로 학습 기반 링크 예측과 유사도 기반 링크 예측으로 본 문제를 해결한다. 학습 기반은 기존의 분류모델(classification model)과 같은 머신 러닝 기법을 이용하여 링크를 예측하는 것이다. 유사도 기반은 노드 쌍들의 유사도를 유사도 함수를 통해 측정하고, 측정된 유사도 중 가장 높은 유사도를 보이는 노드 쌍에서 링크가 생길 것이라고 예측하는 것이다. 본 논문에서는 유사도 기반 예측에 초점을 둔다.

2.2. 유사도 함수 (Adamic/Adar Index)

유사도 기반 링크 예측에서 널리 사용되는 대표적인 유사도 함수로 Adamic/Adar Index(AA)가 있다[2]. AA는 수식 (1)과 같이 공통이웃의 degree를 고려하여 유사도를 계산하는 함수로, (1) 공통 이웃의 개수가 많을수록, (2) 그 공통이웃의 친구의 수(degree)가 적을수록 유사도가 높게 계산된다. 이는 어떤 두 사람의 유사도를 계산할 때, 연예인과 같은 유명한 사람을 공통이웃으로 가지고 있어도, 유사도 계산에는 영향을 크게 주지 않기 위함이다.

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log|\Gamma(z)|} \quad (1)$$

$$(where, \Gamma(x) = x\text{의이웃노드})$$

3. 제안하는 방법

유사도 기반 링크 예측은 유사도를 기준으로 top-n개의 링크를 미래에 생성될 링크로 예측한다. 이때, 유사도가 낮음에도 불구하고 top-n에 포함되어 잘못 예측되는 경우가 생길 수 있다. 이럴 경우 링크

1) 교신저자

예측의 정확도가 떨어지게 된다. 실세계 그래프의 경우, 링크의 수가 희소(sparse)하기 때문에 낮은 유사도를 갖는 링크들이 top-n에 포함되어 링크 예측 정확도가 낮아지는 경우가 종종 나타난다.

본 논문에서는 주어진 그래프의 링크를 빽빽(dense)하게 만들어 링크 예측의 정확도를 향상시킬 수 있는 전처리 기법을 제안한다. 이 기법은 n개의 링크를 예측할 때, 확실히 생길 것 같은 m개(m<<n)의 링크들만 예측하고, 예측된 링크들을 다시 그래프에 포함시켜 유사도를 계산하고, 이를 반복함으로써 링크 예측의 정확도를 높인다. 이때, 생성된 링크는 링크가 생성된 시점에 따라 중요성이 다를 수 있다. 사람을 예로 들면, 오랜 친구 사이일수록 더 깊은 유대감을 형성한다고 볼 수 있고, 이를 그래프에서 나타내면 과거에 생긴 링크일수록 최근에 생긴 링크보다 높은 비중을 갖는 것으로 나타낼 수 있다.

이러한 것을 고려한 제안하는 기법은 다음과 같은 과정을 수행한다. 그래프 G가 주어졌을 때, (1) G에 대한 인접행렬(adjaceny matrix) A를 생성한다. (2) 연결되어있지 않은 모든 노드 쌍에 대해 유사도를 계산한다. (3) 유사도가 가장 높은 m개의 링크를 예측하고, 예측된 링크들을 G에 추가하여 인접행렬 A^k를 생성한다. 이때, 새로 추가된 링크는 1이 아닌 w^k로 표기한다. (4) (2)~(3)의 과정을 k번 반복하여 n개의 링크를 예측한다.

4. 실험

본 논문에서는 제안하는 전처리 기법의 성능을 확인하기 위해 실세계 네트워크 데이터인 DBLP(|V|: 13,184, |E|: 47,937, Density: 0.0005)[3]를 이용하였다. DBLP 데이터는 논문 간의 인용 그래프로, 노드는 논문, 링크는 논문 간의 인용을 나타낸다.

우리는 2.2장에서 언급한 Adamic/Adar index를 이용하여 2,000개의 링크를 예측하였다. 정확도 측정 메져로는 precision[4]을 사용하였다. 우리는 보다 더 정밀하게 분석하기 위해 2,000개의 링크를 k등분하여 구간별로 정확도를 확인하였다. k값이 낮을수록 유사도가 높은 링크들이 포함되어 있다. 제안하는 기법에 사용되는 파라미터인 m, k은 각각 200, 10으로 설정하였고, w는 0.8~0.9로 설정하였다.

<표 1>은 DBLP 데이터에서의 baseline과 우리 기법의 링크 precision과 순위(괄호)를 나타낸다. 전체 정확도는 우리 기법이 baseline보다 더 높은 정확도를 보이는 것을 확인 할 수 있었다. 이를 세부적으로 분석해보면, $1 \le k \le 3$ 일 때, 두 기법의 정확도는 같은 것으로 나타났다. 이후, k가 커질수록 우리 기법의 예측 정확도가 baseline에 비해 더 높은 것을 확인 할 수 있었다. 유사도 기반 링크 예측은 top-n 링크 선택 시 하위에 속하는 링크일수록 잘못된 링크가 선택될 가능성이 높은데, 우리 기법을 이용하면 잘못된 링크가 선택될 가능성을 줄여주기 때문에 이러한 결과가 나온 것이다. 본 실험을 통해 우리 기법이 기존 유사도 기반 링크 예측 정확도를 향상시킬 수 있다는 것을 확인하였다.

<표 1> 링크 예측 결과

k	Baseline	Our method (w = 0.8)	Our method (w = 0.9)
total	93.95 (3)	96.25 (1)	96.05 (2)
1	94.00 (1)	94.00 (1)	94.00 (1)
2	94.50 (1)	94.50 (1)	94.50 (1)
3	93.50 (1)	96.00 (1)	96.00 (1)
4	95.00 (2)	95.50 (1)	95.00 (2)
5	96.50 (3)	98.00 (1)	98.00 (1)
6	93.50 (3)	98.00 (1)	96.50 (2)
7	91.00 (3)	97.50 (2)	98.00 (1)
8	95.00 (2)	94.50 (3)	95.50 (1)
9	94.00 (2)	95.00 (1)	94.00 (2)
10	92.50 (3)	99.50 (1)	99.00 (2)

5. 결론

본 논문에서는 유사도 기반 링크 예측 기법의 정확도를 향상시키는 전처리 기법을 제안하였다. 실험을 통해 제안하는 전처리 기법을 이용하면 기존 링크 예측 정확도를 향상 시킬 수 있는 것을 확인하였다. 본 논문에서는 유사도 기반 링크 예측 기법에 대해 연구를 진행하였는데, 향후에는 다양한 링크 예측 기법의 정확도를 향상 시킬 수 있는 방안에 대해 연구를 진행할 예정이다.

감사의 말

본 연구는 (1) 한국연구재단(No.2018R1A5A7059549), (2) 한국연구재단-차세대정보컴퓨터기술개발사업(No.NRF-2017M3C4A7083678), (3) 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업(2016-0-00023)의 지원을 받아 수행된 연구임.

참고 문헌

[1] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Netowrks," In *JASIST*, 2007

[2] L. Adamic and E. Adar, "Friends and Neighbors on the Web," In Social Networks, 2003.

[3] J. Yang and J. Leskovec, "Defining and Evaluating Network Communities based on Ground-Truth," In IEEE ICDM, 2012.

[4] J. Han, P. Pei, and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2011.