

[네트워크 이론] 소셜 네트워크의 성질·동류성(Assortative), 그리고 영향력(Influence)과 동질성(Homophily)

그냥 공부

by 적분 $\int 2tdt = t^2 + c$ · 2017. 4. 5. 15:57

앞서서 네트워크 이론의 기본적인 개념들을 설명했으니, 실제로 우리 주변에서 찾아볼 수 있는 네트워크들에서 어떤 현상이 일어나는지 살펴보도록 하지요.¹

실제 세계 네트워크의 성질

랜덤 네트워크와 다르게 실제 세상의 네트워크들은 빈익빈 부익부 현상이 있습니다. 즉 다른 녀석들과 많이 연결된 애들이 전체 연결의 대부분을 차지하고, 조금 연결된 녀석들이 엄청나게 많습니다. 마치 20%가 80%의 돈을 가져가고, 나머지 80%가 20%의 돈을 가져가는 것처럼 말이죠. 파레토의 법칙이라고도 불리는 이런 현상을 Power-law 분포(멱법칙)라고 합니다.

또한 결집계수(Clustering Coefficient)도 굉장히 높은 편입니다. 즉 내 친구의 친구도 내 친구인 경우가 많다는 거죠. 랜덤으로 그래프를 생성하면 이런일은 잘 발생하지 않지만, 실제 세상에서는 굉장히 자주 있는 일이죠.

마지막으로 노드간 평균 거리가 굉장히 짧다는 특징이 있습니다. 이는 케빈 베이컨의 6단계 법칙(6단계 만 거치면 세상 모든 사람들이 연결될 수 있다는 뭐 그런 얘기...)에서도 알 수 있죠. 노드가 60억이 넘는 그래프에서 두 노드를 아무거나 골랐을 때 그 둘을 잇는 경로가 평균 6 정도 밖에 안된다는 건 굉장히 짧은 겁니다. 신기한 일이죠.

정리하자면, **1.멱법칙, 2.높은 결집계수, 3.짧은 노드 간 평균거리**가 실제 세상의 네트워크에서 보이는 특징이라고 할 수 있습니다. (거꾸로 말해 이런 특징을 만족못하는 네트워크는 원자 자연스럽지 않은 거라고 볼 수도 있겠죠?)

사회적 네트워크의 성질과 그 원인

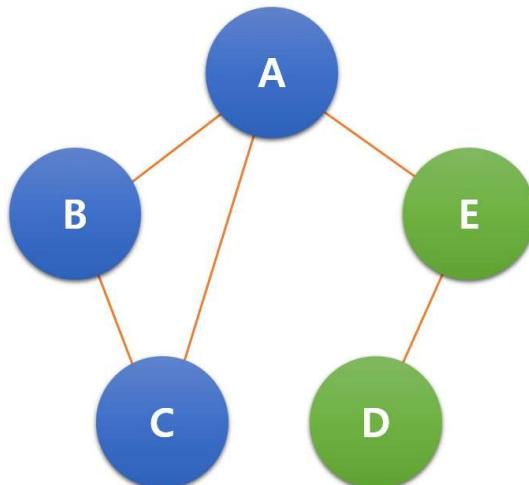
게 만드는 힘으로 보통 다음 3가지를 꼽습니다.

- 영향력(Influence):** 한 노드가 다른 노드와 연결되어 있을 경우, 그 노드의 특성이 다른 노드로 전파되는 경우가 발생하기도 합니다. 이렇게 노드의 연결상태가 노드의 특성에 주는 힘을 **영향력**이라고 정의합니다. 쉽게 말하면 서로 친구이다보니깐 친구의 특성을 닮게 되는 경우를 말하는거죠.
- 동질성(Homophily):** 거꾸로 한 노드가 다른 노드와 유사한 특성을 가지고 있기에 그 둘이 연결되는 경우가 발생하기도 합니다. 이렇게 노드의 특성이 노드의 연결상태에 주는 힘을 **동질성**이라고 합니다. 이 경우는 서로 닮아 있다보니 친구가 되는 경우를 말하겠죠.
- 오염(Confounding):** 네트워크 내부적인 힘이 아니라 외부에서 주는 모든 영향을 통틀어서 오염이라고 합니다. 여기에서는 외부적 요인에 대한 설명은 배제하기로 합니다.

동류성 측정하기

명목 특성(Nominal Attribute)의 경우

노드가 가지는 특성이 순서를 정할 수 없는 값들인 경우가 있습니다. 친구 관계 네트워크를 예로 들자면, 노드는 각 사람이고, 그 노드의 특성은 사람이 사는 지역이라고 할 수 있겠죠. 지역의 경우 셀수도 없고, 순서를 정할 수도 없기 때문에 명목 특성입니다.



5명의 사람 A,B,C,D,E로 이루어진 네트워크 그래프입니다. 파란색은 서울, 녹색은 수원에 사는 걸 뜻한다고 해보지요. 이때 과연 이 네트워크가 지역별로 얼마나 끼리끼리 어울리는지 어떻게 측정할 수 있을까요? 전체 연결된 간선의 갯수 중와 같은 지역끼리 연결된 간선의 수를 비교해 볼수 있겠습니다. 이 값을 동류 중요성(assortativity significance)이라고 부릅니다.

3

0

을 때만 1이 되고 나머지는 0입니다. 따라서 시그마 전체는 특성 값이 같은 간선을 모두 합한 값이 되겠죠. 그걸 전체 간선의 갯수 m 으로 나누면 **(특성이 같은 노드끼리 연결된 간선) / (전체 간선)** 비율이 나오게 되겠죠? (당연하게도 위와 같이 방향성이 없는 그래프에서 간선의 갯수는 2배를 해서 세주거나, 아니면 인접행렬에서 대칭되는 부분은 빼고 계산해야합니다.)

계산해봅시다. 전체 간선은 총 5개. 특성이 같은 녀석끼리 연결된 간선은 총 4개니깐 $4/5 = 0.8$ 입니다. 계산이 간편하지만 이 값은 크게 의미가 없습니다. 그 이유는 네트워크의 크기나 특성에 따라 그 값이 다르기 때문에 서로 비교가 불가능하기 때문입니다. 만약 5명이 다 서울에 사는 사람으로 구성된 네트워크에서는 값이 항상 1이 나오는 반면, 300명은 서울, 300명은 수원에 살고, 서울사는 애들끼리만 친구하고, 수원 사는 애들끼리만 친구하는데, 서울-수원 친구가 1명이 있는 경우 그 값은 1보다 작게 나와서 전자가 더 끼리끼리 있다고 보이기 때문이죠.

그렇기에 우리는 해당 네트워크에서 예상되는 평균 동류 중요성값과 현재 네트워크의 실제 동류 중요성 값을 비교하는 방식을 사용해야합니다. 그러기 위해 다시 위의 그래프로 돌아가서, 5개의 노드가 있고, A는 3, B는 2, C는 2, D는 1, E는 2개의 노드와 연결된다는 사실을 가지고 한 노드가 다른 노드와 연결될 확률이 얼마인지 계산해봅시다. 먼저 A와 B의 평균 연결정도를 보면, A는 3개와 연결되고, B는 2개와 연결되는데 전체 연결되는 횟수는 총 10개이므로, $3*2/10 = 0.6$ 이 됩니다. 마찬가지로 A-C도 0.6, A-D는 0.3, A-E는 0.6이 되겠죠. 마찬가지로 다른 노드에 대해서도 같은 계산을 한 뒤, 같은 지역에 사는 이들끼리만 연결된 걸 합치면 평균 동류 중요성을 계산할 수 있습니다. 계산식은 생략하고, 실제로 계산해보면 0.58이 나옵니다. 즉, 우리가 구한 0.8에서 0.58를 빼면 0.22가 되겠죠. 이 값을 **모듈러성 (Modularity)**라고 합니다.

모듈러성을 구하는 것을 수식으로 쓰면 다음과 같죠.

먼저 평균 동류 중요성은 다음과 같습니다.

$$\frac{1}{m} \sum_{i,j} \frac{d_i d_j}{m} \delta(t(v_i), t(v_j))$$

평균과 실제의 차이를 보면 되니깐 모듈러성 Q 는

$$\begin{aligned} Q &= \frac{1}{m} \sum_{i,j} A_{ij} \delta(t(v_i), t(v_j)) - \frac{1}{m} \sum_{i,j} \frac{d_i d_j}{m} \delta(t(v_i), t(v_j)) \\ &= \frac{1}{m} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{m} \right) \delta(t(v_i), t(v_j)) \end{aligned}$$

3

0

$$Q = \frac{1}{m} \sum_{i,j}^m B_{ij} \delta(t(v_i), t(v_j))$$

가 됩니다.

실제로 계산해보죠.

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} 3 \\ 2 \\ 2 \\ 1 \\ 2 \end{pmatrix}$$

이므로

$$D \cdot D^T = \begin{pmatrix} 9 & 6 & 6 & 3 & 6 \\ 6 & 4 & 4 & 2 & 4 \\ 6 & 4 & 4 & 2 & 4 \\ 3 & 2 & 2 & 1 & 2 \\ 6 & 4 & 4 & 2 & 4 \end{pmatrix}$$

이고, $m = 10$ 이므로 B 는

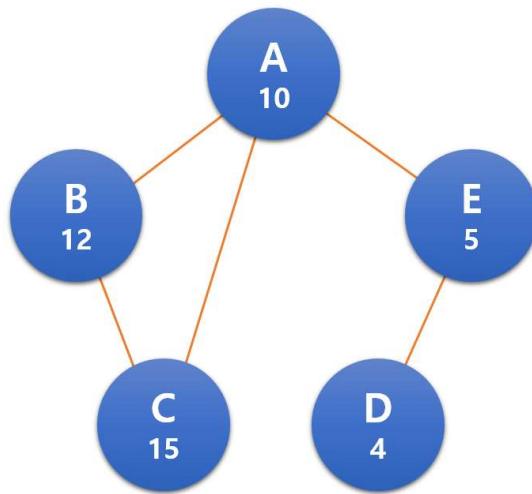
$$B = \begin{pmatrix} -0.9 & 0.4 & 0.4 & -0.3 & 0.4 \\ 0.4 & -0.4 & 0.6 & -0.2 & -0.4 \\ 0.4 & 0.6 & -0.4 & -0.2 & -0.4 \\ -0.3 & -0.2 & -0.2 & -0.1 & 0.8 \\ 0.4 & -0.4 & -0.4 & 0.8 & -0.4 \end{pmatrix}$$

$$\text{따라서 } Q = (-0.9 + 0.4 + 0.4 + 0.4 - 0.4 + 0.6 + 0.4 + 0.6 - 0.4 - 0.1 + 0.8 + 0.8 - 0.4) / 10 = 0.22$$

이 Q 값의 범위는 네트워크에 따라서 달라집니다. 따라서 Q 최대값으로 나눠주면 최대값을 1로 정규화 할수 있겠죠. 동류 중요성의 최대값은 1이고, 평균 동류 중요성은 0.58이었으므로 해당 네트워크에서 가

순서 특성(Ordinal Attribute)의 경우

이번엔 각 노드의 특성이 크기 비교가 가능하고 정도의 차이가 드러나는 순서 특성인 경우를 다뤄봅시다. 위에 처럼 친구 관계 네트워크이지만 이번에는 한달 용돈 수입을 가지고 이들이 얼마나 끼리끼리 노는지 살펴봅시다.



가장 간단한거는 관계를 이루는 두 노드 특성값의 공분산을 살펴볼 수 있겠습니다. 공분산은 두 변수가 얼마나 서로 관련이 있는지 살피는데 유용합니다. 연결된 쌍을 살펴보면 A-B, A-C, A-E, B-C, D-E, (그리고 앞뒤를 뒤집어서 똑같이 또 5개)이 있습니다. 이들의 특성값을 나열해보면

X	Y
10	12
10	15
10	5
12	15
4	5
12	10
15	10
5	10
15	12
5	4

X와 Y의 공분산은 다음과 같이 계산합니다.

(사실 이 공분산 값은 위에서와 마찬가지로 이렇게 계산할 수도 있습니다. 델타함수 부분만 바꿔고 나머지는 동일하죠?)

$$= \frac{1}{m} \sum_{i,j} B_{ij} \cdot t(v_i) t(v_j)$$

즉 X의 편차(원값에서 평균을 뺀값)와 Y의 편차를 곱한 것들의 평균을 구하면 됩니다. 이 식을 잘 정리하면 XY곱의 평균에서 X의 평균과 Y의 평균의 곱을 빼도 공분산이 된다는 것을 증명할 수 있어요. 한 번 구해봅시다. $E(XY) = 104$, $E(X) = E(Y) = 9.8$ 이므로 7.96이 나옵니다.

이 공분산 값 역시 네트워크에 따라 달라지므로 정규화를 할 필요가 있습니다. X와 Y의 공분산을 X,Y의 표준편차로 나눠주게 되면 이는 피어슨 상관계수(Pearson Correlation)가 되는데, 이 값도 최대값으로 1을 가지므로 정규화된 척도로 유용하게 사용할 수 있죠.

$$\rho(X,Y) = \frac{\sigma(X,Y)}{\sigma(X)\sigma(Y)}$$

X와 Y의 분산은 $110.4 - 96.04 = 14.36$ 입니다. 따라서 X와 Y의 피어슨 상관 계수는 $7.96 / 14.36 = 0.554$ 가 나오게 됩니다. 이 역시도 1에 가까운 편이니 끼리끼리 어울린다고 볼 수 있습니다.

여기까지 네트워크의 동류성을 측정하는 방법에 대해서 정리해보았습니다. 하지만 이 동류성이 영향력에 의해 형성된 것인지, 아니면 동질성에 의해 형성된 것인지는 알 수가 없죠. 각각의 힘이 얼마나 작용하는지는 시간에 따른 네트워크의 변화를 살펴보아야 합니다. 이에 대해서는 다음에 좀더 자세히 다뤄볼래요.

1. 이 글의 내용은 Reza Zafarani, Social Media Mining : An Introduction 의 4장과 8장 내용을 바탕으로 작성되었습니다. [\[본문으로\]](#)
2. Assortative는 종류별로 나뉜다는 의미인데, 통용되는 번역어를 몰라서, 일단 생물학에서 쓰이는 "동류 교배(assortative mating)"에서 동류를 따와 단어를 만들어봤습니다. 학계에서 실제로 쓰이는지는 장담못합니다. [\[본문으로\]](#)

관련글

[더보기](#)

[잠재 디리클레 할당 파해치기] 1.

베이즈 추론

2017.04.20

기존 문헌 간 유사도 계산방식의
한계와 TS-SS 공식

2017.04.14

[데이터 마이닝] 지도 학습
(Supervised Learning) 기법

2017.03.24

[네트워크 이론] 노드 간 유사도
(Similarity) 척도

2017.03.20

댓글 0 개

이름

비밀번호

댓글을 입력해주세요.

비밀글

댓글 남기기

인기글

1

자동 요약 기법의 연구 동향 정리

2018.12.28 03:22

2

[Python] tomotopy로 쉽게 토픽 모델링 실시하기

2019.05.22 17:06

4

AdaGram : 동음이의어를 구분하는 Word2Vec

2018.09.23 03:38

5

상위어 자동 추출(Hypernym Detection) 기법 정리

2018.10.10 02:04

6

[Tensorflow] 문자 인식용 신경망 Python3 코드

2018.11.14 18:28

7

[기계 번역] 이중 언어 데이터에서의 단어 임베딩 (Bilingual Word Embeddings from Non-Parallel Docum...)

2018.11.30 18:19

8

[Kiwi] 지능형 한국어 형태소 분석기 0.6버전 업데이트

2018.12.09 23:23

최신글

[C++11] 멤버 함수 포인터를 일반 함수 포인터로 바꾸기

테크닉

어떤 언어 모델이 좋을까 - 언어 모델을 평가해보자

카테고리 없음

어떤 언어 모델이 좋을까 - 언어 모델의 간략한 역사

그냥 공부

Lamon : 라틴어 품사 태거 개발기

NLP

[C++] EigenRand 0.3.0: 다변량 분포 추가

테크닉

범용적인 감정 분석(극성 분석)은 가능할까

NLP

[C++] EigenRand: Eigen용 Random Library 개발

프로그래밍

글쓴이 $\int 2tdt = t^2 + c$



제가 안 것의 대부분은 인터넷으로부터 왔으니, 다시 인터넷에게 돌려주어야 합니다. bab2min@gmail.com

댓글

12.13 $\int 2tdt = t^2 + c$

로그의 밑이 10이 아니라 2 아닌가요?

12.13 궁금

안녕하세요.. 토픽모델을 수행하려고 하는데 어느...

12.13 학생입니다.

PMI가 양수면 두 사건이 함께 일어날 경향성이 높...

12.10 $\int 2tdt = t^2 + c$

태그

Direct3D 리듬게임 영어구조론 정보조직론 python 한국고대경제와사회 문헌정보통계 포니게임 c++ php pg어 텍스트 마이닝 토픽 모델링 악보 자연언어처리 큰정수 라틴어 우리역사바로알기대회 BigFloat 정보검색론

방문자

오늘	어제	전체
71	538	1,374,815

< 이전 1 ... 92 93 94 95 96 97 98 99 100 ... 540 다음 >

링크

나의 큰 O는 log x야

 bab2min@gmail.com

Skin Images are from [Stinkehund](#).

