시계열 분석을 이용한 커뮤니티 변화에 대한 예측

강현국[1], 나원준[2], 쿠마 아비쉑[1], 양현석[1], 이재길[1]†
[1]한국과학기술원 지식서비스공학과
[2]한국과학기술원 전산학과
{hkang106, ynswon, abhishek, arirang, jaegil}@kaist.ac.kr

Prediction of Changes in Communities By utilizing Time Series Analysis

Hyun-Gook Kang[1], Won-Jun Nah[2], Abhishek Kumar[1], Hyun-Seok Yang[1], Jae-Gil Lee[1]†
[1]Department of Knowledge Service Engineering
[2]Department of Computer Science
Korea Advanced Institute of Science and Technology

요 약

본 연구는 네트워크와 커뮤니티의 변화를 예측하는 방법에 대해 제안한다. 네트워크 및 커뮤니티 변화 예측은 인용 네트워크에서의 연구 추세 변화 등의 다양한 응용 분야에 적용이 가능하다. 본 논문에서는 시계열(Time Series) 예측 모델을 기반으로 네트워크의 정량적 변화를 예측하고, Roulette Wheel Selection 알고리즘을 통해 구조적 변화를 예측하였으며, 예측된 네트워크를 기반으로 커뮤니티를 발견하였다. 또한, 실제 인용 네트워크 데이터 기반 실험을 통해 제안하는 방법의 우수성을 증명하였다.

1. 서 론

네트워크 상에서 형성되는 커뮤니티를 예측하는 것은 미래에 대한 통찰력을 얻을 수 있기 때문에 매우 중요하다[1]. 예를 들어 인용 네트워크에서 커뮤니티를 예측함으로써 세부분야 사이의 상관관계를 파악하여 학문적 융합이 어떻게 이루어 질 것인지를 파악할 수 있다. 또한 소셜 네트워크에서 사람들의 상호작용을 예측해 볼 수 있을 뿐만 아니라 웹에서 관련된 주제 아래 웹 페이지들이 어떠한 상관관계를 가지게 될 것인가에 대한 예측도 가능하다[2].

네트워크의 노드와 링크의 정량적 변화를 측정하고 예측하기 위해 시계열 예측 모델을 사용한다. 또한, 기존의 모델을 개선한 새로운 시계열 예측 모델을 제안한다.

네트워크와 커뮤니티의 정량적 예측을 기반으로 구조적 예측을 수행한다. 정량적 예측은 미래에 얼마나 많은 노드와 링크가 형성 될 것인지를 예측하는 것이며, 구조적 예측은 기존의 네트워크 안에 존재하는 노드 사이의 관계를 예측하는 것을 의미한다.

네트워크의 구조적인 예측 방법에는 Common Neighbors[3], Jaccard's coefficient[4], Preferential Attachment[3], Adamic/Adar[5] 등이 있으며, 본 연구에서는 Roulette Wheel Selection 알고리즘을 사용한다.

커뮤니티는 네트워크 상에서 긴밀한 상관관계가 있는 노드들의 부분 집합이며, 대표적인 커뮤니티 발견 알고리즘에는 응집적(agglomerative) 접근법, 분할적(divisive) 접근법, 점 할당(point assignment) 접근법 등이 있다. 본 연구에서는 응집적 접근 방법 중 하나인 Louvain 알고리즘을 사용한다.

2. 정량적 예측

과거 노드와 링크 개수의 변화 추세를 분석하여 미래의 노드와 링크 개수 변화를 예측한다. 본 논문에서는 Moving Average, Exponential Smoothing, Modified Weighted Average 시계열 예측 모델을 사용한다.

2.1 Moving Average

Moving Average는 과거의 특정 타임스탬프 범위 안에 있는 데이터의 평균값을 계산하여 예측값을 얻어내는 방법이다. Moving Average 예측 모델의 정의는 다음과 같다.

$$Y_{(t)} = \frac{X_{t-1} + X_{t-2} + \cdots + X_{t-n}}{n}$$

Y(t)는 t 타임스탬프에서의 예측값이며 X(t-n) 는 t-n 타임스탬프에서 실제 기록된 관측값이다.(단, t>n)

2.2 Exponential Smoothing

Exponential Smoothing의 정의는 다음과 같다.

$$Y_{(t)} = (1-\alpha) * Y_{(t-1)} + \alpha * X_{(t-1)}$$

Y(t)는 t 타임스탬프에서의 예측값이며 X(t-1)는 t-1타임스탬프에서의 관측값이다. α는0과 1사이의 가중치이다.

2.3 Weighted Moving Average

이 예측모델은 본 논문에서 제안 하는 모델로서 Weighted Average를 응용한 것으로 최근의 관측값에 큰 가중치를 주는 방법으로 정의는 다음과 같다.

$$Y_{(t)} \begin{cases} \text{if } Y_{(t-1)} < X_{(t-1)}, & \dfrac{\alpha * X_{(t-3)} + \beta * X_{(t-2)} + \gamma * X_{(t-1)}}{\theta} \\ \text{if } Y_{(t-1)} >= X_{(t-1)}, & (\alpha * X_{(t-3)} + \beta X_{(t-2)} + \gamma X_{(t-1)}) * \theta \end{cases}$$

또한, θ값 조절을 통해 예측 정확도를 높인다. 만약에 Y(t-1)이 X(t-1)보다 크다면 예측값이 초과적으로 예측되었다는 것을 의미하고 Y(t)를 계산할 때는 초과적인 부분을 줄여줄 수 있도록 θ로 곱하여준다. 이와 반대로 Y(t-1)이 X(t-1)보다 작다면 실제 관측값보다 예측값이 과소 예측되었다는 것을 의미하고 Y(t)에서 이 과소된 예측을 다루기 위해 θ를 나누어준다. 사용된 모든 매개변수의 범위는 아래와 같다.

$$\alpha \le \beta \le \gamma, \ \alpha + \beta + \gamma = 1, \ 0 \le \alpha, \beta, \gamma, \theta \le 1$$

3. 구조적 예측

정량적 예측을 기반으로 네트워크 구조가 어떻게 변화할 것 인가를 예측하며, 예측된 네트워크를 대상으로 커뮤니티를 발견한다.

3.1 네트워크

Power Law를 근거로 네트워크에서 링크가 많은 노드 일수록 향후에 링크가 더 늘어날 확률이 커질 것으로 가정한다. 이에 따라 Roulette Wheel Selection 알고리즘을 사용하여 노드 사이의 링크가 어떻게 형성될 지를 예측한다.

3.2 커뮤니티

커뮤니티 발견 단계에서는 Map-Reduce 형태로 개발된 Louvain 알고리즘을 활용한다. Louvain 알고리즘은 커뮤니티 발견의 정확도가 우수하며 대용량의 네트워크 데이터의 분석도 가능하다.

4. 실 험

Stanford Large Network Dataset Collection 의 논문 인용 네트워크 데이터셋 (Cit-HepPh)을 사용하였으며 정량예측 및 구조예측의 정확도를 검증하였다[6].

4.1 정량예측 실험결과

평균 절대적 오류(Mean Absolute Error)는 예측값과 실제값의 차이를 나타내는 것으로 이 수치가 낮을수록 더 정확하게 예측을 하였음을 뜻한다. 표 1에서 나타난 바와 같이 Moving Average가 가장 큰 오류를 보였고 반면에 본 논문에서 제안한 Modified Weighted Average가 가장 작은 예측 오류를 보이는 것으로 나타났다.

표 1. 시계열 예측 모델에 따른 연도별 노드 예측의 절대적 오류(Absolute Error)와 평균 절대적 오류(Mean Absolute Error)

Year	Moving Average	Exponential Smoothing	Modified Weighted Average(MWA)
1995	0.620049	0.268784	0.268784
1996	0.41259	0.15786	0.110568
1997	0.265712	0.103771	0.053493
1998	0.164224	0.054189	0.001381
1999	0.13641	0.072564	0.020769
2000	0.105715	0.043414	0.010056
2001	0.044972	0.007413	0.007413
MAE	0.25001	0.101142	0.067495

그림 1는 연도별 실제 커뮤니티 개수와 예측된 커뮤니티의 개수를 비교한 결과를 나타낸다. 파란색 선은 실제 커뮤니티의 개수이고 주황색 선은 예측된 커뮤니티 개수이다. 전체 테스트 기간에서 제안하는 방법으로 예측한 커뮤니티 수가 실제 커뮤니티 수와 유사한 경향을 보임을 알 수 있다.

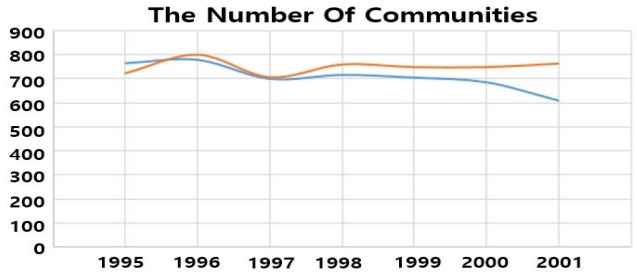

그림 1. 연도별 실제 커뮤니티와 예측된 커뮤니티의 정량적 비교

4.2 구조예측 실험결과

그림 2은 실제 네트워크와 예측된 네트워크의 Out Degree 분포를 나타낸다. 전체적으로 실제 네트워크와 예측된 네트워크의 분포가 유사한 경향을 보임을 알 수 있다.

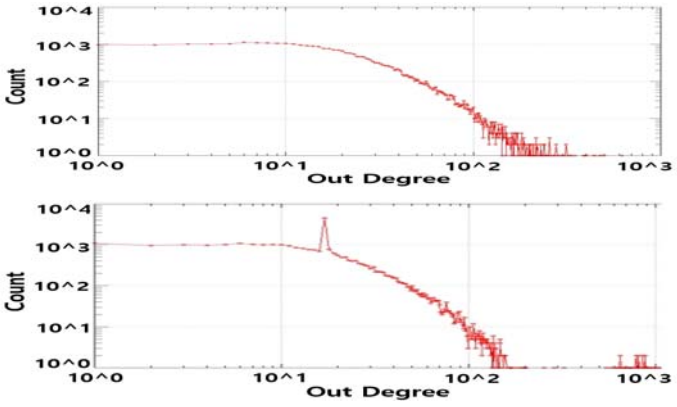

그림 2. 2001년도 실제 네트워크(상)와 예측된 네트워크(하)의 Out Degree 분포

표 2 은 2001 년도에 예측된 네트워크와 실제 네트워크의 Top 10 Modularity 비교를 나타낸다. 전체적으로 유사한 값을 보임을 알 수 있다.

표 2. 2001년도 커뮤니티의 Top-10 Modularity 비교

Label	Actual Network	Label	Predicted Network
2	0.08457451	43	0.151019964
14	0.043685102	2	0.057388499
3	0.035491005	3	0.050362405
4	0.034454924	16	0.026464312
5	0.025950324	7	0.019631662
9	0.023809923	35	0.012754573
22	0.022996611	10	0.011931346
17	0.02066754	15	0.011714163
8	0.020142282	21	0.011069513
28	0.015936186	6	0.008800243

또한, 표 2에서 얻은 Top 10 Modularity로부터 JS-Divergence을 계산한다. JS-Divergence는 실제 네트워크와 예측된 네트워크에서 커뮤니티의 Modularity 분포 차이를 나타낸다. 이 수치가 적을수록 Top 10 community의 구조적 분포를 정확히 예측하였음을 의미한다. 그림 3에 나타난 바와 같이 1996년도에는 JS-Divergence가 0.005로 가장 정확히 구조적 분포를 예측하였고 반면에 1998년도에는 0.048로 가장 큰 오류를 보였다. 하지만 1998년 이후로 수치가 줄어드는 추세를 보이면서 예측의 정확도가 높아졌다.

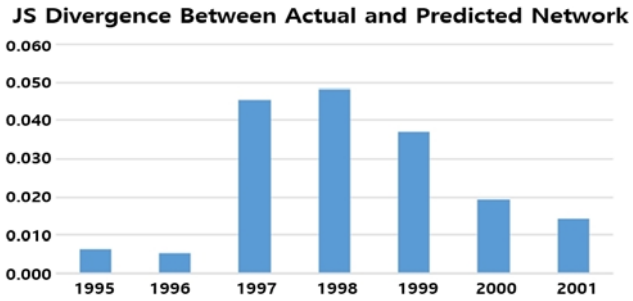

그림 3. Top 10 modularity를 이용한 실제 네트워크와 예측된 네트워크 사이의JS-Divergence

그림 4 는 2001 년도의 예측된 네트워크의 구조를 시각화한 그림이다. 가장 빈번하고 응집성이 강한 주황색 Label 은 Modularity 가 가장 높은 Label 43 이다.

그림 4. 2001년도 예측된 네트워크의 구조 시각화

5. 결 론

본 논문에서는 시계열 예측 모델에 기반하여 네트워크의 정량적 변화를 예측하고 Roulette Wheel Selection 알고리즘을 사용하여 구조적 변화를 예측하였다. 그리고, 예측된 네트워크에 Louvain 알고리즘을 적용하여 커뮤니티를 발견하고 이를 실제 네트워크의 커뮤니티와 비교하였다. 정량적 예측에 적용한 시계열 예측 모델들 중에서 제안하는 방법인 Weighted Moving Average 가 가장 높은 예측정확도를 보이는 것으로 나타났으며, 정량적 예측을 기반으로 수행된 구조적 예측단계에서도 실제 네트워크와 유사함을 확인하였다.

5. 참고문헌

[1] Jung, Shukhwan, and Aviv Segev, “Analyzing Future Growing Citation Networks in Communities." Knowledge-Based Systems (2014).

[2] P. Chen, S. Redner, “Cummunity structure of the physical review citation network” Journal of Informetrics 4, 278-290, (2010)

[3] Newman, Mark EJ. "Clustering and preferential attachment in growing networks." Physical Review E 64.2, 025102, (2001)

[4] G. Salton, M.J. McGill, “Introduction to Modern Information Retrieval”, McGrawHill, Inc., (1986)

[5] L. Adamic, E. Adar, Friends and neighbors on the web, Soc. Netw. 25 (3) (2003).

[6] Arxiv High Energy Physics paper citation network, ”http://snap.stanford.edu/data/index.html#citnets"