정보확산 감쇠모델에서 정보근원 추정방법 연구

최재영*,
*호남대학교 미래자동차 공학부

e-mail: jychoi@honam.ac.kr

A Study on Information Source Estimation under Diffusion Decaying Models

Jaeyoung Choi*,
*Dept of Automotive Engineering, Honam University

요 약

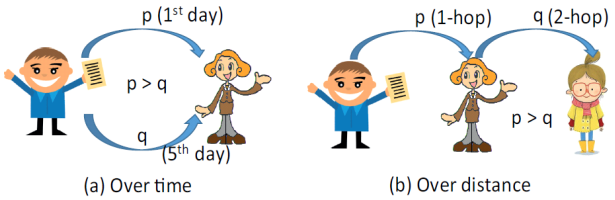
인터넷의 급속한 발전은 네트워크를 통해 많은 정보를 흘러가도록 만들었다. 그 정보들 중에는 다른 시스템이나 사용자에게 루머나 바이러스와 같은 좋지 않은 정보를 포함하는 경우도 있는데, 이에 대비해서 이런 악성정보가 더 많이 퍼져나가지 않도록 다양한 보안방법들이 제안되었다. 본 논문에서는 이를 위해 정보를 처음 퍼뜨린 근원을 추정해 나가는 문제를 고려한다. 특히, 기존의 이 문제를 다룬 많은 연구에서와는 다르게 본 연구에서는 정보가 시간이 지남에 따라, 혹은 근원으로부터 멀어짐에 따라 그 전파력이 감쇠되는 여러 가지 정보확산 모델을 고려한 후, 각각의 모델에서 트리기반 추정법을 사용하여 정보의 근원을 얼마나 잘 탐지해 내는지에 대해 살펴본다.

1. 서론

최근 인터넷을 통해 페이스북이나 카카오와 같은 소셜 서비스 플랫폼이 급속하게 발전이 되었다. 이를 통해 많은 사용자들은 자신과 연결이 되어 있는 주위 사용자들에게 정치적 뉴스와 보도 같은 공공정보 뿐 아니라 개인적으로 공유하고 싶은 다양한 정보들을 급속하게 전달한다. 이런 과정에서 바이러스나 루머와 같은 좋지 않은 정보도 빠르게 퍼져나가게 된다. 본 논문에서는 이런 정보 네트워크에서 임의의 정보가 퍼져나간 경우, 그 정보를 처음 퍼뜨린 사용자 혹은 노드가 누구인지를 밝히는 연구를 다룬다.

네트워크에서 퍼진 정보의 근원을 찾는 문제는 [1][2]에서 처음 제안이 되었다. 해당 연구에서는 정보의 근원에 대한 어떠한 사전 정보가 없는 상태에서 가장 좋은 추정법인 최우추정량(Maximum Likelihood Estimator; MLE)을 사용하였다. 그 결과, 네트워크에 아주 많은 노드가 이 정보를 듣게 되면[1], 처음 정보를 퍼뜨린 근원을 찾을 수 있는 확률이 모든 노드가 같은 수의 이웃 노드를 가진 비교적 간단한 가장 간단한 정규트리(Regular tree)에서도 31%를, 그리고 일반적인 그래프에서는 10%를 넘지 못하는 것이 밝혀졌다. 이는 네트워크 관리자는 해당 정보가 언제 퍼지기 시작했는지에 대한 아무런 정보가 없이 다만, 임의의 시점에 네트워크를 관측했을 때, 그 정보가 확산된 형태(Snapshot)만 보고 근원에 대해 추론하는 방법이다. 해당 연구 이후에 다양한 네트워크에서 퍼진 정보의 근원

1) 일반적으로 이 경우를 감염(Infection)되었다고 표현한다.



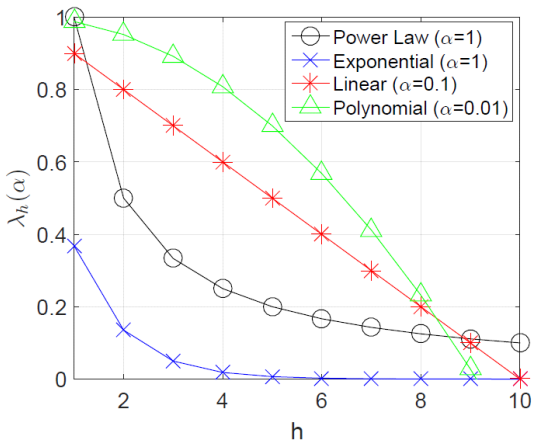
(그림 1) 시간(a) 및 거리(b)에 따른 정보 전달력의 감쇠

을 찾는 문제가 많이 제안되었다[3]-[8]. 그러나 대부분의 경우, 정보의 전달력이 시간이나 또는 정보의 근원과의 거리와 상관없이 동일하다고 가정하였다. 실제로 임의의 정보는 그림 1(a)와 같이, 시간이 지난 후에 그 영향력이 감쇠된다는 사실이 실험적으로 밝혀졌다[9]. 이와 비슷하게 본 연구에서는 그림 1(b)와 같이 정보가 근원지와의 거리가 멀어짐에 따라 그 영향력이 감쇠하는 경우에 대해서 확산 후에 근원을 찾는 문제에 대해서 알아본다.

2. 정보 확산 모델 및 근원 추정 방법

2.1 정보의 확산 모델 (Information Diffusion models)

본 논문에서 고려하고 있는 네트워크는 $G=(V,E)$라는 그래프로 표기될 수 있는데, 여기서 V는 네트워크에 있는 모든 노드(Node)들의 집합이고 E는 각 노드를 연결하는 변(Edge)들의 집합이다. 그리고 이런 노드들을 서로 연결



(그림 2) 거리에 따른 감쇠 모델 예시($\lambda_0 = 1$)

하는 것이 그래프에서 변으로 표현이 된다. 본 연구에서는 앞서 언급한 바와 같이 네트워크에 하나의 정보를 퍼뜨리는 노드가 있는데 이를 정보의 근원이라고 표현하고 v^*로 표기한다. 네트워크에서 이 정보가 근원으로부터 그 이웃에게 전달되어서 퍼져나가게 되는데, 여기서 정보가 퍼지는 현상은 이미 잘 알려진 확산 방법 중 하나인 Susceptible-Infected (SI) 모델을 따른다고 가정한다. 그리고 정보의 전파력이 근원에서의 거리에 따라 감쇠되는 현상을 고려하기 위해, 정보 근원과의 거리 $h \geq 1$ 인 이미 감염된 노드가 이와 연결된 이웃노드에게 전파 비율이 $\lambda_h > 0$ 인 지수분포(Exponential distribution)를 따라서 그 정보를 확률적으로 전달하는 경우를 가정한다. 여기서 $h = 0$ 즉, 근원에서의 전파력은 λ_0로 표기한다. 본 연구에서는 아래와 같이 다양한 정보 감쇠모델을 고려한다. (아래의 모든 정보감쇠 모델은 $\lambda_h(\alpha)$로 표현하고 α는 감쇠모델의 파라미터이고 특히, $\alpha = 0$ 인 경우는 동질(homogeneous)의 확산모델[1][2]와 동일하다. 각각의 감쇠 현상은 그림 2를 참고하면 된다.)

1) Power law(멱급수) 감쇠모델: 본 감쇠모델은 전파력의 함수가 $\lambda_h(\alpha) = \lambda_0 / h^\alpha$ 와 같이 주어진 모델을 말한다.

2) Exponential(지수) 감쇠모델: 지수 감쇠모델은 거리에 따른 전파력의 감소가 지수함수를 따르는 $\lambda_h(\alpha) = \lambda_0 e^{-\alpha h}$ 형태의 감쇠모델이다.

3) Linear(선형) 감쇠모델: 선형감쇠모델은 아래 식과 같이 근원으로부터 멀어질수록 선형적으로 그 영향력이 줄어드는 확산 모델이다. 구체적인 전파력 함수는 $\lambda_h(\alpha) = -\alpha h + \lambda_0$ 같이 표현된다. 위의 두 가지 감쇠 모델과는 다르게 여기서는 특정한 거리 $h = \lceil \lambda_0 / \alpha \rceil$ 이후부터는 더 이상 전파가 되지 않는다.

4) Polynomial(다항) 감쇠모델: 본 감쇠모델은 근원과 가까운 거리에서는 전파력이 크다가 갑자기 전파력이 급감하는 경향성을 가진 모델이다. 전파력 함수는 $\lambda_h(\alpha) = -\alpha h^2 + \lambda_0$ 같이 주어진다. 본 감쇠모델에서도 선형의 경우와 마찬가지로 특정한 거리 $h = \lceil \sqrt{\lambda_0 / \alpha} \rceil$ 이후부터는 더 이상 전파가 되지 않음을 알 수 있다.

2.2 정보의 근원 추정 (Information Source Estimation)

앞서 주어진 정보감쇠 모델에 대하여 본 논문에서는 일반적인 그래프에서 N개의 감염 노드에 대한 그래프 G_N 이 주어진 경우 다음과 같은 근원 추정방법을 고려한다.

$$\hat{v} = \arg max_{v \in G_N} R(v, T_{bfs}(v)) \qquad (1)$$

여기서 $T_{bfs}(v)$는 G_N에서 노드 v에 Breath-First-Search (BFS)로 만들어진 트리이고 $R(v, T_{bfs}(v))$는 이 경우의 정보 중심성이다. 정보 중심성은 v가 근원이라고 가정했을 때, $T_{bfs}(v)$을 만들어 내는 다양한 종류의 확률 경로(Sample path)가 존재 할 수 있고 이를 더한 값을 의미한다. 자세한 계산 방법은 [1]을 참조하면 된다. 이것은 일반적으로 메시지 교환(Message passing)방법에 의해서 $O(N)$시간 안에 계산이 되는 것으로 알려졌다. BFS는 일반적인 루프(Loop)가 있는 그래프를 트리모양으로 근사시켜주는 잘 알려진 방법 중 하나이다. 따라서 만약, 그래프가 정규 트리인 경우에 (1)에서 정의된 정보(루머) 중심은 [1]에서 소개된 것과 정확하게 일치한다.

3. 분석 및 탐지 실험

본 장에서는 앞서 제시한 정보 감쇠 모델에 기반 하여 시뮬레이션을 통해 실제 그래프에서 정보의 근원을 얼마나 잘 찾을 수 있는가에 대한 실험을 진행 하였다. 실험 Matlab을 사용하였고 각 그래프에서 총 1,000 노드가 감염이 되도록 근원으로부터 정보를 확산시켰고 이런 확산 현상을 총 500회를 진행한 후에 추정방법을 통한 탐지확률에 대한 성능분석을 하였다. 고려된 네트워크와 기본적인 실험 세팅에 대한 설명은 다음과 같다.

3.1 그래프(네트워크) 세팅 및 탐지확률

본 실험에서는 먼저, 가장 간단한 그래프인 정규 트리를 고려하였다. 정규트리인 경우, 차수가 3(d=3)인 경우에 대해서 진행을 하였다. 다음으로, 실제 데이터를 가지고 만들어진 그래프(Real-world graph)로서 페이스북 그래프[10]를 생성시켜서 실험을 하였다. 페이스북 네트워크인 경우는 총 4,039노드가 88,234개의 변으로 구성

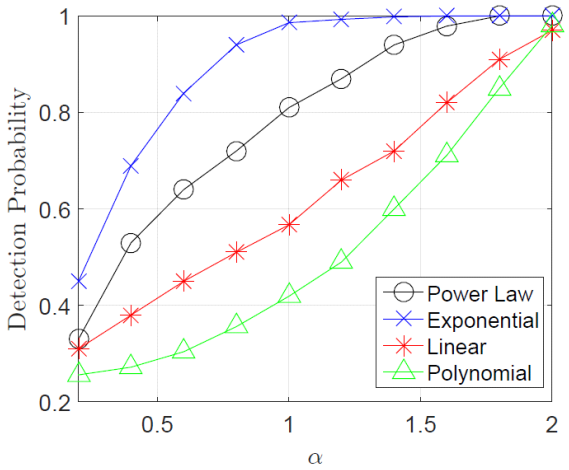
되어 있는 그래프를 몇 개의 집단(cluster)이 그 안에서 긴밀하게 연결되어 있고 (많은 변들이 서로 연결됨) 이런 집단끼리는 각 집단의 내부보다는 덜 긴밀하게 연결이 되도록 형성하였다. (실제 노드들이 연결되어 있는 모양이다). 위와 같이 형성된 그래프 위에서 랜덤하게 정보를 퍼뜨리는 근원을 선택하여 SI 모델에 따라 네트워크에서 전체 확산된 노드의 수가 1,000이 되도록 한 후 제안한 추정법이 실제 근원을 찾는지 아닌지를 확인하는 실험을 진행 하였다.

2) 탐지확률: 탐지확률의 척도는 아래와 같은 식을 사용하였는데 이는 전체 시도된 정보의 확산 중에서 실제로 제안한 추정법이 원래 정보의 근원을 찾는 수로 나눠진 것이다.

$$\text{Detection Probability} = \frac{\#\text{of Detections}}{\#\text{of Trials}} \quad (16)$$

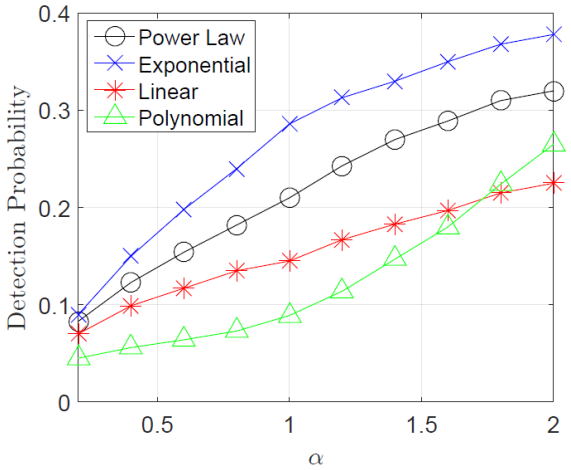
본 실험에서는 총 500번의 확산현상 중에서 정확히 정보의 근원을 찾는 횟수를 나누어 각 경우에 대한 탐지확률을 계산하였다.

5.3 실험 결과



(그림 3) 정규트리에서의 정보근원 탐지확률

먼저, 그림 3 정규트리에서는 모든 경우 감쇠 파라미터인 α가 커질수록 탐지확률이 증가하는 것을 볼 수 있다. 즉, α가 크면 근원을 중심으로 대칭적으로 감쇠현상이 있는 경우, 확산의 모양이 확률적으로 균일한 경향이 크게 되어서 실제 근원을 정보의 중심성으로 어느 정도 잘 찾아질 수 있다는 것을 알 수 있다. 감쇠모델이 지수모델인 경우 가장 높은 탐지확률을 가지는 것을 발견할 수 있었고, 다항감쇠 모델의 경우는, 파라미터가 커질수록 확산이 빨리 멈추는 특성으로 인해 탐지확률이 급격하게 커지는 것을 확인할 수 있었다.



(그림 4) 페이스북 네트워크에서 정보근원 탐지확률

다음으로, 그림 4 실제 페이스북 네트워크에서도 모든 경우 감쇠 파라미터인 α가 커질수록 탐지확률이 증가하는 것을 볼 수 있다. 그리고 비록 근사적인 BFS를 사용하여 정보의 근원을 찾음에도 불구하고 탐지확률이 감쇠모델에서는 높게 나타나는 것을 확인할 수 있었다.

4 결론

본 논문에서는 네트워크에서 임의의 정보가 퍼진 경우 그 근원이 어디인지를 찾아내는 방법에 대한 연구로 특히, 정보의 전파력이 근원으로부터 멀어질수록 감쇠하는 현실적인 모델을 바탕의 연구를 진행 하였다. 이를 위해 여러 가지 정보 감쇠모델을 제안하였고 일반적인 그래프에서 BFS트리를 사용한 추정방법을 통한 탐지확률을 얻었다. 결과적으로, 정보의 전파력이 감쇠되는 정도가 클수록 탐지확률이 커진다는 사실을 다양한 실험을 통해서 보였다.

참고문헌

[1] D. Shah and T. Zaman. Detecting Sources of Computer Viruses in Networks: Theory and Experiment. In Proceedings of ACM SIGMETRICS, 2010.

[2] D. Shah and T. Zaman, “Rumor Centrality: A Universal Source Detector,” In Proceedings of ACM SIGMETRICS, 2012.

[3] J. Choi, S. Moon, J. Woo, K. Son, J. Shin and Y. Yi. Rumor Source Detection under Querying with Untruthful Answers. In Proceedings of IEEE INFOCOM, 2017.

[4] J. Choi and Y. Yi. Necessary and Sufficient Budgets in Information Source Finding with Querying: Adaptivity Gap. In Proceedings of IEEE ISIT, 2018.

[5] W. Dong, W. Zhang and C.W. Tan. Rooting Out the Rumor Culprit from Suspects. In Proceedings of IEEE International Symposium on Information Theory (ISIT),

2013.
[6] Z. Wang, W. Dong, W. Zhang and C.W. Tan. Rumor source detection with multiple observations: fundamental limits and algorithms. In Proceedings of ACM SIGMETRICS, 2014.
[7] J. Choi, J. Shin and Y. Yi. Information Souce Localization with Protector Diffusion in Networks. IEEE/KICS Journal of Communication and Networks, Vol.21, Issue 2, pp 136-147, Apirl, 2019
[8] G. Fanti, P. Kairouz, S. Oh and P. Viswanath. Spy vs. Spy: Rumor Source bfuscation. In Proceedings of ACM SIGMETRICS, 2015.
[9] N. Ohsaka, Y. Yamaguchi, N. Kakimura and K. Kwarabayashi. Maximizing Time-Decaying Influence in Social Networks. Springer International Publishing,, 132-147, 2016.
[10] J. Leskovec and J. McAuley. Learning to discover social circles in ego networks. In Proceedings of NIPS, 2012.