그래프 클러스터링을 이용한 영향력 전파에서의 블로킹 제거 방법

이철기*, 이우기*
*인하대학교 산업경영공학과
e-mail : rich@inha.edu, trinity@inha.ac.kr

Blocking Elimination Method Using Graph Clustering In Influence Propagation

Rich. Chul-Ghi Lee*, Wookey Lee*
*Dept. of Industrial Engineering, Inha University

요 약

영향력 전파 문제는 주어진 네트워크 환경에서 영향력을 최대화 할 수 있는 top-k 노드를 찾는 문제로 데이터 마이닝 분야에서 활발히 연구되어왔다. 본 논문에서는 그래프 클러스터링 기법을 사용하여 영향력을 전파하는 방법을 제안하고자 한다. 이러한 방법에는 두 가지 이점이 있는데 먼저 서로 다른 시드 사이에 영향력이 중복되는 블로킹 현상을 제거하여 수행시간을 단축시킬 수 있다. 다음으로는 유 방향 그래프인 경우 기존의 탐욕 알고리즘보다 더 많은 노드에 전파를 가능하게 한다.

1. 서론

최근 들어 트위터, 페이스북등 여러 거대한 온라인 소셜 네트워크 서비스가 급증하고 있다. 이러한 네트워크를 통해 전달되는 가장 중요한 대상은 '정보'다. 소셜 네트워크를 통한 정보 전달을 설명하는 영향력 전파 모델은 SNS 에서의 마케팅에 적용이 되었다. 그리고 이러한 영향력 전파를 통해 소셜 네트워크에서의 '입소문 효과'를 설명 할 수 있었다. 아날로그에서는 자연 감쇄 효과가 있어서 이러한 정보의 전달에서 목소리가 커도 많은 사람에게 전파가 될 수가 없는 구조였으며, 인맥을 통해 전달해도 정보가 왜곡되기 마련이었다. 그러나 현재의 소셜 네트워크 즉 디지털 환경에서는 이러한 자연 감쇄효과가 없다. 그리 하여 몇 몇 사람에게는 과도한 정보전달의 가능성을 지니며 그렇다고 모든 사람에게 정보가 전달되는 것을 보장하지도 않는다 그러므로 이러한 정보 전달의 중복과 정보의 전파 사이의 적절한 균형이 필요하다. 다시 말해서 소셜 네트워크에서 소수의 influential (영향자)을 이용한 마케팅이 는 것이 오히려 사용자들에게 부정적인 정보로 인식 될 수 있다. 예컨대 나에게 불필요한 정보를 서로 다른 여러 사람이 계속 공급한다면 오히려 그 정보에 대하여 부정적인 영향으로 바뀔 가능성이 커지기 때문이다. 이는 정보이론의 관점에서 살펴보면 정보 과잉으로 설명될 수 있는데 일정 수준이상의 정보를 주게 되면 오히려 역으로 부작용을 야기한다는 것이다. 또한 정보가 중복 될수록 그 정보의 효용은 점점 체감 된다.

본 논문에서는 그래프 클러스터링 방법을 사용하여 블로킹을 제거하여 영향력의 중복을 제거함으로써 정보과잉을 최소화하여 기존의 네트워크 영향력 전파 방식을 통해서도 효과적인 영향력 전파가 가능하게 하였다.

This work was supported by Inha University and the National Research Foundation of Korea(NRF) Grant funded by the Korean Government(MOE) (NRF-2013R1A1A2012887)

2. 관련 연구

정의 1 (네트워크) 네트워크는 유 방향 그래프 $G = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ 로 정의 한다. 집합 $\mathcal{V}$ 는 $v_i \in \mathcal{V}$ 의 $1 \le i \le \mathrm{n}$ 원소를 가진다. 집합 $\mathcal{E}$ 는 $e_k \in \mathcal{E}$ 의 $1 \le k \le \mathrm{m}$ 원소를 가지는데 각각의 원소 e_k는 집합 $\mathcal{V}$의 원소들의 순서쌍 $e_k = \{v_i, v_j\}$로 이루어져 있다. 위의 네트워크를 n 개의 노드와 m 개의 아크로 이루어진 네트워크로 정의한다. 집합 $\mathcal{W}$는 $w_{e_k} \in \mathcal{W}$, 즉 집합 $\mathcal{E}$의 원소들의 가중치를 의미하며 [0, 1] 사이의 가중치 값을 지닌다.

그래프에서 노드에 연결되어 있는 아크의 수를 degree라고 정의하며 유방향 그래프의 경우에는 들어오는 아크와 나가는 아크를 구분하여 $\mathrm{in-degree}$ 와 $\mathrm{out-degree}$로 구분한다. 또한 모든 노드 쌍 사이에 경로가 존재 할 때 그 그래프는 연결되어있다 라고 하고, 만일 어떤 노드에서 경로가 존재하지 않는 노드가 존재한다면 그 그래프는 연결되어 있지 않다고 한다. 네트워크에서 어떤 생각이나 정보가 퍼지는 것에 대한 모델을 생각할 때 우리는 각각의 개별 노드

들이 어떤 정보나 생각에 대해 전파되어 노드가 활성화 되었거나 전파되지 않아 활성화 되지 않거나 이 두 가지 경우만 고려하려고 한다.

가정 1. 노드에서의 영향력은 노드에 연결된 아크를 통해서 이웃 노드 로만 전달된다.

가정 2. 두 노드 사이의 관계에 따라 영향력을 전파하는 서로 다른 아크 가중치가 존재하고 모든 아크의 가중치는 주어진다고 가정한다[7].

이러한 조건에서 영향력을 최대로 전파시키는 top-k 시드 노드를 찾고 영향력을 평가하는 것이 이 문제의 최종 목표이다. 이 문제에 대해 소개하는 기존 연구로는 선형 임계 치 모델(Linear Threshold Model) 다른 모델로는 독립 전파 모델(Independent Cascade)이 있다. IC 모델에서 영향력 전파 모델의 최적 솔루션을 찾는 것은 NP-hard 로 증명되어있다[4]. 영향력 전파 모델의 이러한 점을 극복 하기 위해서 탐욕 알고리즘이 제안 되었다.

새로운 시드 노드가 추가되면 기존의 시드 집합에서 전파되는 영향력과의 중복이 일어나고 이를 블로킹이라고 한다. 예컨대 $\sigma(S)$ 가 시드 셋 S 가 전파시키는 엑티브 노드 라고 한다면 $\sigma^{-1}(v)$ 를 노드 v 에 전파한 시드셋 이라고 정의 할 때 $|\sigma^{-1}(v)| > 1$ 인 경우에 정보과잉 즉 negative 효과가 일어나기 때문에 이를 없애고자 한다. Figure 3 의 경우에는 v_5, v_6, v_7 의 경우에 $|\sigma^{-1}(v)| = 2$ 가 되어 정보 과잉이 일어나게 된다.

앞에서 설명한 영향력 평가의 문제점들을 해결하기 위한 선행 연구도 진행되어 왔다. IPA 알고리즘은 각각의 influence path 가 하나의 영향력을 전파시키는 unit 으로 고려 하는 알고리즘이다. influence path 를 서로 독립이라고 가정 하고 알고리즘을 진행하게 된다. 기본적인 알고리즘의 진행은 CELF 알고리즘을 포함하여 마지널 게인을 최대로하는 시드를 선택한다. IPA 알고리즘은 병렬처리를 지원하여 옆에서 설명한 두 가지 문제점을 해결하려 하였으나 근본적인 문제인 블로킹을 해결하지 못하였다. 영향력 전파의 근본적인 문제는 블로킹 효과로 인하여 마지널 게인을 구하기 어려워 프로세싱 타임이 증가하는 점인데 IPA 알고리즘에서는 influence path 가 독립이라는 가정으로 마지널게인을 쉽게 구하고자 하면서 병렬처리를 통하여 수행시간을 단축시켰다[5]. OGIS 알고리즘은 병렬처리가 아닌 기존의 아크를 통해 이웃에게 영향력을 전파하는 것이 아닌 새로운 greedy 알고리즘이다[6].

다음으로 본 논문에서 하려는 바와 같이 마지널 게인을 더 쉽게 구하려는 연구가 진행 되었다. IC 모델은 각 각 아크의 확률값으로 영향력의 전파가 결정되는데 확률값을 각각의 노드로 전파되는 경우의 수로 미리 계산하여 전파시 한번에 처리하여 수행속도를 개선하는 연구가 진행되었다. 그러나 이 연구는 블로킹이 아닌 단순한 pre-proccesing 으로 수행 속도를 개선한 알고리즘이다[8]. LIDH 알고리즘은 이 CELF 알고리즘을 단순히 개선하여 수행시간을 줄인 알고리즘으로 영향력이 확장되는 거리에 제한을 두어 수행시간을 줄인 알고리즘이다. 이론적으로도 각각의 시드에 대해 영향력이 확장되는 거리에 제한이 크면 클수록 블로킹은 줄어들기 때문에 수행시간이 줄지만 이는 블로킹에 대한 근본적인 해결책이 되지 못하고 limitied propagation distance 값에 deterministic 한 단점이 있다[9].

3. 제안하는 방법

제안하는 알고리즘을 설명하기에 논문에서 사용하는 notation 은 기존의 영향력 전파 연구에서 사용되는 일반적인 notation 을 가져와 사용하였으며 정리하면 다음과 같다.

<표 1> 논문에서 사용되는 Notation

Notation	Explanations
G	graph $G = (V, E)$
V	node set $V = \{v_1, v_2, \dots, v_n\}$
E	edge set $E = \{e_1, e_2, \dots, e_m\}$
S	seed set $S = \{s_1, s_2, \dots, s_k\}$ for s_i seed node
C	Cluster set $C = \{c_1, c_2, \dots, c_k\}$, for c_i cluster
k	The number of seeds
σ_I	expected number of nodes influenced without time constraint
$\hat{\sigma}_I$	estimated value of σ_I
T	$T = \sigma_I(S)$
Δ_I	marginal gain $\hat{\sigma}_I\ (S \cup \{v\}) - \hat{\sigma}_I\ (S)$

다음의 notation 을 이용하여 크게 두 부분으로 알고리즘을 설명하고자 한다. 먼저 클러스터링을 하는 부분을 설명하고 클러스터링 한 다음 영향력을 전파하는 것을 설명할 것이다.

본 논문에서는 앞에서 설명한대로 서로 다른 시드 사이의 블로킹을 없애기 위해 영향력 전파(Affinity Propagation)클러스터링 방법을 사용하여 그래프 클러스터링을 하고자 한다. 물론 본 연구는 어떠한 그래프 클러스터링 방식에 따라 제약을 받지 않는다.

유방향 그래프 $G = (V, E, W)$ 가 주어졌을 때 영향력 전파 모델은 다음의 문제로 정의 된다.

정의 1. 영향력 전파 모델은 주어진 그래프 G 로부터 아래의 식(4)를 만족하는 top-k 시드 노드의 집합을 찾는 문제이다.

$$S = \arg \max_{T \subseteq V, |T| = k} \sigma(T, G) \qquad (1)$$

여기서 $\sigma(T, G): V \times G \rightarrow \mathbb{R}$ 은 전파된 영향력의 노드

집합을 의미한다. 즉 노드 집합 T는 영향력이 전파된 노드의 개수이다.

위의 정의를 만족하는 최적 솔루션을 구하기 위해서는 전체 노드의 개수 N 에서 최적 시드 노드 집합 k 개를 찾아야 하기 때문에 $\binom{N}{k} = \frac{N!}{k!(N-k)!} \approx N^k$ 케이스의 검색공간을 필요로 한다. 앞에서 설명한 CELF 알고리즘 또한 이러한 문제로 인해 greedy 알고리즘으로 문제를 해결 하였고 greedy 알고리즘이 최적 솔루션을 보장하지는 않지만 아래의 조건을 만족한다는 전제 하에 다음의 하한 값을 보장한다[4].

조건 1 Non-negativity

$$\forall(S \subseteq V).\ \sigma(S) \geq 0 \tag{2}$$

조건 2 Monotonicity

$$S \subseteq T \subseteq V, \sigma(S) \leq \sigma(T) \tag{3}$$

조건 3 Sub-modularity

$$For\ S \subseteq T \subseteq V,\ v \in V,\ and\ v \notin T$$
$$\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T) \tag{4}$$

위의 조건을 모두 만족하면서 탐욕 알고리즘을 진행 하게 되면 $1 - \frac{1}{e}$ 의 하한 값 즉 63%의 전파력을 보장하게 된다.
독립전파 모델 중 하나인 CELF 탐욕 알고리즘을 사용하여 영향력 전파를 하여 하한 값을 보장하는 top-k 시드 집합을 구할 수는 있지만 영향력이 얼마나 전파되었는지 평가하는 것은 또 다른 이슈이다. 영향력을 평가하기 위해서는 새로운 시드 노드가 추가될 때 마다 중복이 없는 영향력인 다음의 식(9)을 계산 해야 한다.

$$\hat{\Delta}_I(S, v) = \hat{\sigma}_I\ (S \cup \{v\}) - \hat{\sigma}_I\ (S) \tag{5}$$

먼저 서로 다른 클러스터 사이에는 영향력이 전파되지 않는다고 가정한다. 앞에 예제를 가지고 설명하면 클러스터링 한 후에 가정한대로 클러스터 사이의 영향력을 전파하지 않고 영향력을 전파하게 되면 영향력의 중복 즉 블로킹이 존재하지 않는 전파를 할 수 있다. 즉 서로 다른 클러스터에 영향력을 전파 함으로서 서로 다른 시드 노드 사이의 블로킹을 없애고자 한다. 이론적으로는 제안하는 방법을 사용하면 다음의 정리가 가능하다.

정리 1. 서로 다른 클러스터에 전파된 시드 노드 집합에 대하여 새로운 시드 노드에 대한 영향력 $\hat{\Delta}_I(S, v)$은 다음의 식(6)을 따른다.

$$\hat{\Delta}_I(S, v) = \hat{\sigma}_I\ (v) \tag{6}$$

정리 1은 다음의 식으로 유도 가능하다.

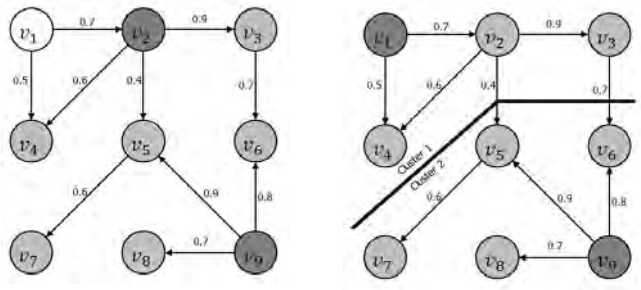
$$\begin{aligned}\hat{\Delta}_I(S, v) &= \hat{\sigma}_I\ (S \cup \{v\}) - \hat{\sigma}_I\ (S) \\ &= \hat{\sigma}_I\ (S) + \hat{\sigma}_I\ (v) - \hat{\sigma}_I\ (S \cap \{v\}) - \hat{\sigma}_I\ (S) \\ &= \hat{\sigma}_I\ (v) - \hat{\sigma}_I\ (S \cap \{v\})\big(\hat{\sigma}_I\ (S \cap \{v\}) = 0\big) \\ &= \hat{\sigma}_I\ (v)\end{aligned}$$

즉 종전의 클러스터링 된 그래프에서 영향력을 전파하여 다음의 영향력 전파 모델 알고리즘을 제안한다.

Algorithm 2 C-CELF greedy(G,σ,k)

```
Require: G : graph, k : the number of seed nodes & cluster
Ensure : # of influence node σ, seed set k
  C ← apcluster(G,k)
  σ(S) ← ∅
  for i = 1 to k do
    S ← argmax(σ(C_i))
  end for
Return S, σ
```

제안하는 방법을 사용하면 정리 1 을 만족하여 블로킹이 일어나지 않아 다항 시간(polynomial time) 내에 영향력을 평가 할 수 있고 기존의 IC 모델을 사용하기 때문에 기존 탐욕 알고리즘의 하한 값을 보장한다. 또한 클러스터 C_i가 다음의 정리 2 를 만족하면 모든 그래프에 영향력 전파가 가능하고 조건을 만족하지 못하더라도 시드 노드로부터 연결되어 있는 클러스터 내의 노드에 따라 하한값 이상의 전파를 하게 된다. 앞에서 설명한 예제로 알고리즘을 진행하면 다음의 (그림 1) 오른쪽과 같다.



(그림 1) CELF 와 C-squre 결과 비교

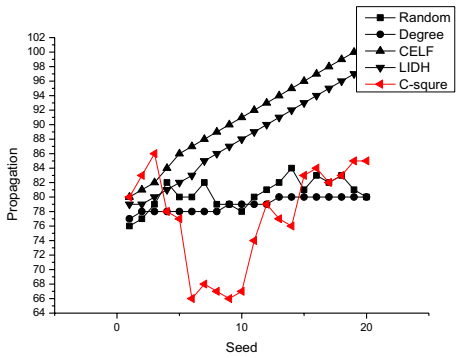
클러스터링 후에 먼저 가장 영향력을 많이 전파시키는 시드인 v_9을 시드로 잡아 영향력을 전파하면 총 5개의 노드가 엑티브 노드가 되고 다음으로 마지널 게인이 높은 v_1를 다음 시드로 잡아 영향력을 전파하여 영향력을 전파하게 된다. 다음의 결과에서 알 수 있듯이 (그림 1)왼쪽의 CELF 알고리즘을 그냥 적용시키는 것 보다 제안하는 방법을 사용하면 더 많은 노드에 전파가 가능함을 확인 할 수 있었다. 또한 정리 2를 통해 $|\sigma^{-1}(v)| = 1$를 만족하게 되어 정보과잉 또한 없음을 확인 할 수 있었다.

정리 2 서로 다른 클러스터에 하나의 시드가 존재한다면 모든 노드 v_i는 $|\sigma^{-1}(v)| = 1$을 만족한다.

증명. 각각의 클러스터에 하나의 시드를 전파시켰을 때 $|\sigma^{-1}(v)| > 1$를 만족하는 노드 v가 있다고 가정하면 2 개 이상의 클러스터에서 v로 연결되어 있는 경로가 존재한다. 따라서 노드 v는 2 개 이상의 클러스터에 속해있다. 이는 클러스터링 가정에 모순된다. 그러므로 서로 다른 클러스터에 하나의 시드가 존재한다면 모든 노드 v_i는 $|\sigma^{-1}(v)| = 1$을 만족한다.

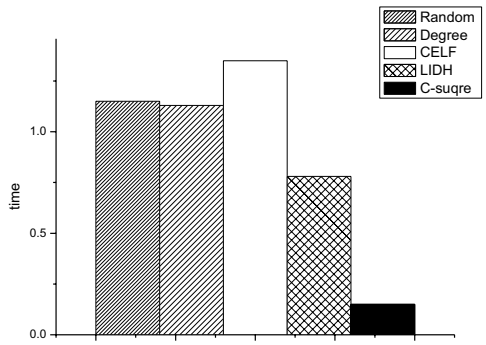
4. 실험

본 논문에서는 제안하는 영향력 전파 기법의 성능을 평가하는 자료로 먼저 서로 다른 분포를 가지는 synthetic 데이터에 대해 실험을 진행한 후에 실제 real 데이터를 이용하여 실험을 진행하였고 페이지 제약으로 모두 수록하지는 못하였다. 먼저 (그림 2)와 같이 Seed 를 증가하며 실험을 진행 한 결과 클러스터의 수가 늘어나게 되면 블로킹을 없애기 위해 서로 다른 클러스터 사이의 영향력이 전파되지 않는 점 때문에 전체 영향력이 떨어지는 것을 확인 할 수 있었다.



(그림 2) 시드에 따른 영향력 전파

또한 예제에서 보인 클러스터의 효과로 인하여 처음에는 기존 방법보다 더 많은 전파를 하는 것을 확인 할 수 있었고 이를 이용하여 블로킹을 제거했을 때의 최적 Seed 를 찾는 실험을 진행하였다. 위의 그림에서의 최적 시드는 3 개이다.



(그림 3) 수행시간 비교

(그림 3) 에서와 같이 수행시간을 비교하였을 때도 제안하는 방법이 가장 빠른 것을 확인 할 수 있었는데 이는 영향력 평가 시 영향력의 중복을 계산할 필요 없기 때문이다.

5. 결론

본 연구를 통해 소셜 네트워크 환경에서 영향력의 블로킹 없이 영향력을 전달하는 방법을 제안하였고, 이를 통해 다항 시간 안에 top-k 시드 노드 집합을 구하고 전파된 영향력을 계산하는 방법을 제안하였다. 기존 연구들과 비교하여 블로킹이 발생하지 않으므로써 영향력 평가 시간을 단축하였고, 최적 시드에서 기존의 알고리즘 이상의 전파를 하는 것을 확인 하였고 추후 연구로는 각 클러스터의 영향력 전파를 병렬처리 한다면 더욱 적은 시간으로 영향력을 전파할 수 있을 것으로 본다.

참고문헌

[2] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." science 315.5814 (2007): 972-976.

[4] Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[5] Kim, Jinha, Seung-Keol Kim, and Hwanjo Yu. "Scalable and parallelizable processing of influence maximization for large-scale social networks?." Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013.

[6] Lee, Jong-Ryul, and Chin-Wan Chung. "A fast approximation for influence maximization in large social networks." Proceedings of the companion publication of the 23rd international conference on World wide web companion. International World Wide Web Conferences Steering Committee, 2014.

[7] Lewis, Ted G. Network science: Theory and applications. John Wiley & Sons, 2011.

[8] Liu, Bo, et al. "Influence spreading path and its application to the time constrained social influence maximization problem and beyond." Knowledge and Data Engineering, IEEE Transactions on 26.8 (2014): 1904-1917.

[9] Lv, Shunming, and Li Pan. "Influence Maximization in Independent Cascade Model with Limited Propagation Distance." Web Technologies and Applications. Springer International Publishing, 2014. 23-34.

[10] Wang, Yu, et al. "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.