문자메시지네트워크의 선형적 특성

한영석 오창근° 박지은
수원대학교 정보미디어학과
yshan@suwon.ac.kr, chang--gun@hanmail.net°, jieun-629@hanmail.net

Linear Property of SMS Network Clusters

Young S. Han, Chang K. Oh°, Jieun Park
Dept. of Information Media, Suwon University

요 약

문자메시지 네트워크는 웹페이지 검색네트워크와는 위상적으로 많이 다르기 때문에 상이한 특성을 가지고 있음을 보였다. 네트워크가 성장함에 따라 강력한 허브를 구성하는 네트워크는 소수의 노드에 연결이 집중되는 특성을 가질 수 있다. 문자메시지 네트워크는 시간이 흐름에 따라서 네트워크의 밀도가 높아지는 속도가 높다는 특징을 가지고 있다. 두 개의 크기가 다른 문자메시지 네트워크에 대한 실험을 통하여 적어도 작은 집단 내의 메시지네트워크는 크기가 증가함에 따라서 허브의 집중도가 power law분포가 아닌 선형적인 특징을 가지고 있음을 보였다.

1. 서론

인터넷을 포함한 네트워크에 대한 연구는 1990년대 중반 이후로 활발히 이루어져서 자유경쟁을 하는 네트워크모델을 취하는 다양한 상황에서 빈익빈 부익부, 양극화, 80:20법칙과 같은 현상 들을 좀더 이해할 수 있게 되었다[1]. 그와 같은 양극화가 극단적으로 증명이 되는 곳은 자유경쟁과 네트워크의 증가가 급속히 일어나는 인터넷에서 특히 분명하게 나타난다.

일부 몇 개의 포털사이트가 인터넷 트래픽의 상당부분을 취하게 됨으로써 콘텐츠사업과 경쟁산업이 전반적으로 부실하게 되는 부작용을 낳고 있다. 양극화는 자유경쟁이 보장된 네트워크에서는 필연적으로 발생하기 때문에 세계화되는 경제와 경쟁이 가장 용이한 인터넷의 근본적인 문제를 해결하기 위한 기초적인 연구가 요구되는 시점이다.

Barabasi[1]에 따르면, 네트워크의 노드가 증가하고 연결에 있어서 자유경쟁이 보장되면, 극단적 양극화가 일어 난다. 인터넷의 경우에 노드는 웹사이트나 웹페이지에 해당하며, 링크는 연결을 의미한다. 노드 수가 정체되어 있으면서 연결선만 증가 한다면, 양극화는 발생하지 않는다. 인터넷에서는 노드에 해당하는 웹페이지들의 평균적인 링크가 증가하는 추세라고 해도 유의미한 수준이 아닐 것으로 추측이 되지만, 네트워크의 증가는 당분간 계속될 것으로 예상된다.

할리우드의 배우들이 같이 출연한 다른 배우들의 수에 대한 분포 역시 power law분포를 비롯하여 사회현상, 자연현상, 생명현상 등 광범위한 영역에서 양극화 현상이 발견이 된다[1]. 자연현상에서 발생하는 양극화는 네트워크상에서 통신의 효율성을 위해서 만들어진 장치이며 이러한 효율화 메케니즘을 통해서 생명체가 진화를 해왔다고 할 수도 있다.

가령, 미국에서 한 개인이 또 다른 개인으로 연결되는 데는 6단계를 거치면 된다고 하는 연구[4]가 있는데, 전체 인구를 고려하면 충분히 경제적인 네트워크라고 할 수 있는데 이는 바로 사회 곳곳에 위치한 허브적 기능을 수행하는 사람들 때문에 가능한 것이다. 웹 네트워크 처럼 집중화가 극단적으로 되는 경우에는 전체 구조가 붕괴될 수 있을 정도로 진행되기도 한다.

통신네트워크는 자유경쟁이 비교적 제한적이며, 연결선도 꾸준히 증가하는 특성을 가지고 있다. 물론, 통신네트워크도 경쟁을 하는 대상들이 포함될 경우에는 다르겠지만, 대부분의 통신활동은 비경쟁인 연결선에서 발생한다고 가정해도 무방하다. 문자메시지 네트워크는 비경쟁적인 혹은 경쟁이 높지 않은 네트워크임을 보이고자 하였다. 강한 허브를 가진 네트워크라면, 다음과 같은 power law와 같은 분포에 근접할 것이다.

$$y = ax^k$$

x를 한 노드에 연결된 에지의 수라하고 y를 노드의 빈도라고 한다면, 네트워크의 크기에 따라 집중도가 심화되어야 하지만 본 연구의 메시지 네트워크에서는 평균값이 x의 중간지점에 위치하여 선형에 가깝다는 것을 보이고 자 한다. 클러스터링을 통하여 그룹들간의 특징정보를 이용하여 허브의 존재를 파악하고 이들의 집중도를 측정하고자 한다.

2. 문자 메시지 네트워크

문자메시지 네트워크는 개인이 노드가 되고 전화대전화 및 인터넷대전화 간에 발생하는 단문메시지는 에지가 되는 그래프구조를 지칭한다. 메시지의 발신 혹은 수신 관계에 따라 방향성을 갖게 되며, 메시지 빈도는 에지(edge)의 가중치를 의미한다.

문자메시지는 주로 안면이 있는 경우에 사용된다. 특별한 사회적 상황의 변화가 아니라면, 각 개인과 연계되어 추가되는 노드(메시지 파트너)수는 급격하게 변하지 않는 경향을 보인다. 메시지가 주로 닫혀 있는 그룹 내에서 발생하는지를 확인하기 위해서 대학생들을 대상으로 설문을 실시하였다.

문자메시지 네트워크의 분포특성을 이해하기 위하여 1학년 신입생들을 대상으로 설문을 실시하였다. 입학한지 2주가 지난 상태임으로 서로간의 직접적 연결관계가 성숙되지 않은 상태여서 이미 잘 알려진 학생을 중심으로 허브를 구성할 것으로 가정하였다. 학생들이 지난 한달 동안 같은 샘플 안의 누구와 얼마 정도의 문자메시지를 주고 받았는지에 대해서 설문하였다. 샘플크기가 두 배정도 커졌을 경우에 분포특성은 어떻게 달라질 것인지 알기 위하여 두 개의 샘플을 활용하였다.

특정 사회적 기능을 가진 그룹 내에서의 문자메시지는 사적인 메시지 네트워크와는 달리 허브기능이 유의미한 수준으로 존재한다. 그러나 웹과는 달리 극소수의 노드가 대부분의 연결관계를 독점하고 대부분의 노드는 연결관계를 거의 갖지 못하는 것과는 다를 것이라는 것이 본 연구의 가설이다.

작은 샘플인 경우 (그림 1)에는 선형적인 분포에 근접하고 있음을 알 수 있다.

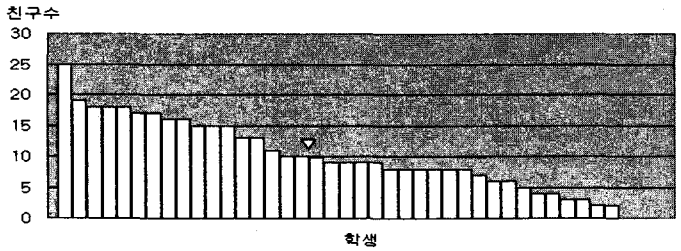

그림1. 각 학생들의 친구 수 분포 (샘플크기: 38명)

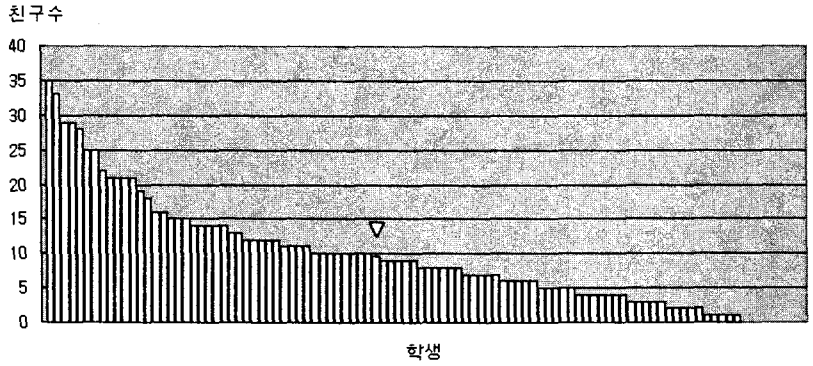

그림2. 각 학생들의 친구 수 분포 (샘플크기: 87명)

샘플 크기가 두 배 정도 증가했을 때는 [그림2] 집중도가 심화되고 있다는 것은 허브기능을 하는 학생들의 역할이 커지고 있음을 나타낸다. 샘플크기가 두배가 허브의 집중도는 증가하였지만 집중도는 두 배가 증가하지 않았다는 것과 평균 친구의 수에 해당하는 부분에 많은 학생들이 존재한다는 점은 통신 네트워크가 극 소수에 몰려 있지 않음을 암시한다.

3. 네트워크 클러스터링

메시지가 교환되는 관계들을 중심으로 허브의 구성 및 역할을 기준으로 네트워크의 크기가 미치는 영향에 대해서 알아 보기 위하여 닫혀 있고 성숙되지 않은 두 개의 샘플 네트워크를 대상으로 실험하였다.

닫혀 있다는 것은 네트워크의 크기가 증가하지 않는다는 뜻이고, 성숙되지 않았다는 것은 시간상에서 에지가 추가될 수 있는 여지가 많다는 것을 의미한다. 성장하지 않는 네트워크가 에지생성이 진행된다면 네트워크의 엔트로피는 증가하고 허브의 기능은 축소되는 방향으로 진행하게 된다. 실제 메시지네트워크의 모집단 안의 구성원들은 웹검색과는 달리 매우 지역적인 파트너와 메시지를 교환하게 되기 때문에 전체 네트워크의 크기에 따라 허브의 존재는 크지 않을 것이라고 추측할 수 있다.

두 개의 미성숙한 네트워크는 같은 조건하에서 네트워크의 크기가 미치는 영향력을 측정할 수 있게 한다. 적어도 개인적인 관계에 있어서 메시지 교환은 네트워크의 크기에 관계없이 허브보다는 지역적으로 이루어진다는 것을 알 수 있게 한다. 본 연구는 네트워크 크기의 차이에 비하여 허브 및 클러스터의 집중화 특성이 제한적으로 강화되는 점을 보이고자 하였다.

아직 네트워크가 완성되지 않은 단계이면서 닫혀 있는 대학 신입생 두 개의 클래스(샘플크기 38명, 87명)를 대상으로 실험하였다. 두 클래스의 전공특성에 따른 샘플자체의 특징도 있을 수 있는 등 여러 요소들은 본 연구에서는 고려하지 않았다.

메시지 트래픽량에 따른 두 노드간의 거리값을 기초로 클러스터링을 하고자 바텀업 방식의 알고리듬[3, 4]을 응용하였다. 개개의 노드가 클러스터라고 가정하고 이들을 통합하는 방식으로 진행된다. 가장 거리가 가까운 관계에 있는 두 개의 클러스터가 합쳐진다. 이후에 클러스터간의 거리를 다시 계산하여 같은 작업을 반복한다. 클러스터간의 통합은 두 클러스터내부의 각각의 응집도와 두 클러스터간의 거리값을 비교하여 응집도에 비하여 거리값이 너무 클 경우에는 통합하지 않는다. 더 이상 통합되지 않는 단계에서 작업을 멈춘다.

계층적 클러스터링 알고리듬

초기화: 각 노드를 중심으로
초기 클러스터 집합을 생성
단계1: 클러스트들 상호간의 거리를 계산
단계2: 거리값이 가장 작은 두 클러스터를 통합
클러스터 통합이 실패한 경우에
두 번째 작은 거리의 클러스터를 시도
단계3: 단계2에서 통합이 발생한 경우에 단계1부터 반복

두 노드(n1, n2)간의 거리식은 다음과 같다.

$$D(n_i, n_2) = 1/(|ie_1| + |ie_2|)$$

$|ie_1|$ 는 노드 1번으로 들어 오는 에지의 가중치 즉 통신빈도를 나타낸다.

클러스터간의 거리(D_C)계산 식은 다음과 같다.

$$D_C(C_1, C_2) = (\Sigma_i \Sigma_j D(n_i,n_j))/N$$
$$\text{where } n_i \in C_1,\ n_2 \in C_2$$

N은 C_1의 노드 개수를 나타낸다.

결정한다.

$$\text{merge}(C_1,C_2) = 1 \quad \text{if } CC(C_1) + CC(C_2) < D_C(C_1, C_2)$$
$$= 0 \quad \text{otherwise}$$

클러스터링한 결과는 표1과 표2와 같다.

클러스터내의 평균거리 CC()값을 통합여부를

그룹	인원	학생 ID	내부 평균거리	entropy	최대 entropy	relative entropy
1	2	19,25	0.78	0.347	0.347	1.000
2	3	2,6,13	7.50	0.343	0.363	0.945
4	3	4,9,32	12.01	0.303	0.363	0.833
3	6	8,14,15,17,21,28	10.80	0.242	0.300	0.800
5	11	3,5,7,12,18,20,26,30,31,33,38	15.43	0.181	0.216	0.837
6	13	1,10,11,16,22,23,24,27,29,34,35,37,39	19.38	0.169	0.197	0.851

표1. 샘플1(38명)에 대한 클러스터링 결과

그룹명	인원	학생 ID	내부 평균거리	entropy	최대 entropy	relative entropy
1	1	3	0.00	0.000	-	1.0
2	1	20	0.00	0.000	-	1.0
3	1	33	0.00	0.000	-	1.0
4	1	35	0.00	0.000	-	1.0
5	1	36	0.00	0.000	-	1.0
6	1	67	0.00	0.000	-	1.0
7	10	13,14,15,16,17,27,48,54,61,69	7.62	0.190	0.230	0.826
8	11	2,7,9,19,21,31,37,55,60,68,77	14.56	0.169	0.216	0.772
9	12	5,10,32,41,42,46,47,49,64,65,74	7.11	0.177	0.208	0.851
10	24	1,6,8,11,12,23,25,28,30,34,38,40,44,45,52,56,57,58,59,63,66,71,76,78	18.59	0.104	0.132	0.787
11	24	4,18,22,24,26,29,39,43,50,51,53,62,72,73,75,79,80,81,82,83,84,85,86,87	19.20	0.088	0.132	0.667

표2. 샘플2(88명)에 대한 클러스터링 결과

클러스터의 크기도 샘플크기의 증가율만큼 증가하고 있음을 알 수 있다. 각 클러스터의 평균내부거리는 각 소속 노드들이 클러스터 내의 다른 노드들과의 거리값 들의 평균값을 의미함으로 값이 작을수록 클러스터 내부의 결속력이 강하고 허브의 기능이 약하다는 것을 의미한다. 클러스터의 크기가 커질수록 허브의 기능이 강화된다고 할 수 있음으로 거리값은 증가하게 된다. 샘플1의 최대 클러스터에서 중심노드인 37번 노드가 전체 샘플 내 노드간의 평균거리는 17.73이였다. 샘플2에서는 중심 노드가 63번 79번이였으며 각각 샘플 내 노드들과의 평균 거리는 각각 18.56과 17.88이였다. 거리의 최대값이 20인 점에 비추어 봤을 때, 허브가 샘플 전체에 미치는 접근성은 강한 편이 아니다. 샘플크기의 증가에 따라서도

허브의 장악력에 큰 변화를 가져오지 않았다. 그러나 실제로는 허브노드가 전체 노드에 대해서 갖는 거리값이 샘플크기에 따라 큰 변화가 없다는 것은 샘플크기의 증가에 따라 허브노드가 선형적으로 많은 링크의 증가를 가지고 있음을 알 수 있게 한다.

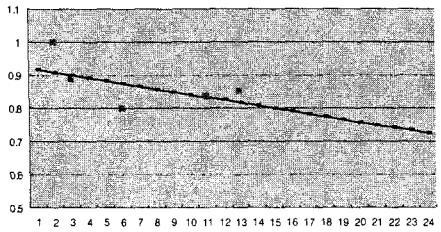

그림3. 샘플1의 상대엔트로피값의 선형분석 (기울기: -0.0083, 상대엔트로피평균: 0.877)

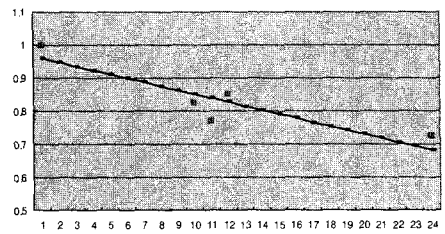

그림4. 샘플2의 상대엔트로피값의 선형분석 (기울기: -0.012, 상대엔트로피평균: 0.78)

엔트로피 값은 클러스터 내의 각 노드가 여타 노드 등과의 거리평균값 들의 분포특성을 의미한다. 엔트로파 값이 최대 엔트로피 값에 근접할수록 분포가 고르다는 것을 의미하며, 허브의 존재가 약하다는 것을 암시한다. 클러스터 엔트로피 값들을 서로 비교하기 위하여 각 클러스터의 엔트로피를 최대엔트로피로 정규화하여 상대엔트로피를 구한다.

그림3과 4는 각각 샘플1과 샘플2의 상대 엔트로피를 클러스터 크기상에 배열한 것이다. 만일 본 연구에서 다루는 메시지 네트워크가 power law분포를 따른다고 한다면, 네트워크의 크기 변화에 따라 중간 클러스터의 급속한 소멸과 소수의 클러스터로 집중화 되며, 클러스터내 엔트로피도 급속히 낮아져야 할 것이다. 또한 클러스터의 크기의 증가에 따라서도 급격한 엔트로피의 감소가 이루어 져야 할 것이다. 왜냐하면 허브의 기능은 샘플의 크기가 증가할수록, 클러스터의 크기가 증가할수록 power law에 따라 확대되기 때문이다.

그림3과 4에 의하면 그와 같은 급격한 구조적 변화를 예측하기 힘들다.

5. 결론

문제메시지네트워크의 크기가 두 배 정도 증가하더라도 클러스터의 특성은 선형적 증가에 그쳤다. 본 연구는 네트워크의 연결선(에지)가 성숙한 수준으로 많아지기 이전의 단계에서의 샘플 크기의 비교이어서 시간상에 늘어나는 에지가 미치는 결과와는 다른 것이다. 노드의 개수가 정해진 상태에서 시간상에 늘어나는 에지로 인하여 허브의 약화가 일어날 것이 분명하며 어느 시점에서는 수렴할 것이다. 본 연구는 같은 수준의 성숙도를 가진 크기가 다른 네트워크에서는 큰 네트워크에서 집중화가 일어나지만 네트워크의 크기에 따른 선형적인 증가라는 것을 보였다. 네트워크 크기가 작게 되면 인간관계의 형성이 좀더 원활하게 되는 사회심리적인 요소와 같은 여러 변수들이 메시지네트워크 특성의 이면에 있음도 유의할 필요가 있다. 본 연구의 한계는 샘플의 크기가 메시지 당사자가 인식할 수 없는 범위 수준으로 커질 경우에 나타날 수 있는 현상을 포함하지 않는 데 있다. 샘플크기가 충분히 커질 경우에 소속하는 노드들은 허브를 통하여 소통하게 되겠지만, 그러한 수요는 메시지 네트워크에서는 크지 않다고 가정한다면, 개인간의 메시지 연결선은 본 연구에서 제시하는 수준을 다소 넘어설 것임으로 평균적인 극단적 빈곤화는 발생하지 않을 것이다.

본 연구는 메시지 네트워크와 같은 덜 양극화된 구조가 극단적인 양극화 네트워크와의 결합을 통하여 좀더 생산적인 구조로의 변화를 이끌어 낼 수 있는 가능성을 예시하고 있다.

참고문헌

[1] Albert L. Barabasi. Linked. A Plume Book. 2003.

[2] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSourze, Jenifer C. Lai, Robert L. Mercer. «Class-based n-gram Models of Natural Language », Computational Linguistics. 18(4): 467-479. 1992.

[3] Ido Dagan, Lillian Lee, and Fernando Pereira. "Similarity-based methods for word sense disambiguation". Proceedings of the 35th annual meeting of the ACL. 1997.

[4] Stanley Milgram. "The Small World Problem," Physiology Today 2. 60-6