3

소셜 네트워크에서의 인플루엔셜 랭킹

KAIST ▌박호성 · 곽해운
MPI-SWS ▌차미영
KAIST ▌문수복*

1. 서 론

아리스토텔레스는 인간을 사회적 동물이라고 일컬었다. 인간은 서로 서로 관계를 맺으며 사회 활동을 한다. 이러한 사회 구성원 간의 관계를 나타내는 네트워크가 바로 소셜 네트워크이다. 그림 1은 Zachary가 1970년대에 관찰한 대학의 가라데 클럽의 네트워크이다[16].

각 노드는 가라데 클럽의 구성원을 나타내며, 운동 또는 클럽 미팅 시 두 노드 사이에 지속적으로 상호작용이 존재할 경우 그 두 노드들은 엣지로 연결 된다. 이 가라데 클럽은 의견 대립으로 인해 두 개의 파벌로 나뉘어 있었다. 그렇다면 이런 가라데 클럽에서 가장 영향력 있는 사람은 누구일까? 클럽 회장? 가장 덩치가 좋은 사람? 누구나 돌아보게끔 만들만큼 매력적인 여성 회원? 영향력을 어떻게 정의하느냐에 따라 달라지겠지만 그림 1을 보면 노드 16, 노드 17 회원들 보다는 엣지를 많이 갖고 있는 노드 1과 노드 34 회원들이 클럽에 많은 영향을 미칠 것이라고 예상할 수 있다. 실제로 노드 1과 노드 34 회원은 각 파벌의 지도자들 이다. 이렇게 소셜 네트워크에서 영향을 많이 미치는 사람을 인플루엔셜(influential : 영향자)이라고 한다. 모 김치냉장고의 성공 신화를 비롯하여 다양한 성공 사례를 선보이고 있는 입소문 마케팅에서 가장 중요한 것 중 하나가 바로 누가 인플루엔셜인가를 찾아내는 것이다. 영향력 있는 대상에게 마케팅을 집중하는 것이 효과도 좋고 비용도 절감할 수 있기 때문이다. 이 글에서는 소셜 네트워크에서 어떻게 인플루엔셜을 찾고 랭킹을 매기는 지에 대하여 다루고자 한다.

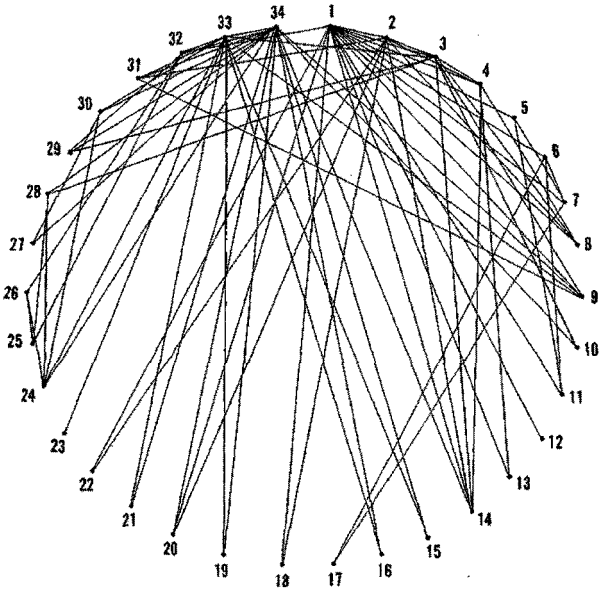


그림 1 가라데 클럽의 네트워크[16]

2. 인플루엔셜을 찾아라

누가 인플루엔셜인지를 알아내고자 하는 연구는 Social Science와 Computer Science에서 꾸준히 진행되어 왔다. 소셜 네트워크에서 인플루엔셜의 정의를 한마디로 나타내기는 힘들다. 네트워크의 종류와 보고자 하는 영향의 종류에 따라 그 뜻이 달라지기 때문이다. 일반적으로 인플루엔셜이라 함은 영향과 권력을 갖고 있고 행사하는 사람을 말한다. 아래에 Social Science와 Computer Science에서 어떻게 인플루엔셜에 대한 연구를 했는지 소개한다.

3. Social Science의 연구

1950년대에 발표된 Katz와 Lazarsfeld의 2단계 유통이론(two-step flow theory)[5]에 따르면 정보나 영향력은 매스미디어에서 수용자로 바로 전달되지 않고 소

* 종신회원

† 본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음[2008-F-016-02, 초정밀 측정 및 분석기술 연구].

수의 의견 지도자(opinion leader)를 거쳐 궁극적인 수용자들에게로 전달된다고 한다. 사람들의 태도를 변화시키는데에 매스미디어의 영향 보다 접촉한 의견 지도자의 영향이 더 큰 것이 관찰되었는데 이러한 의견 지도자를 인플루엔셜이라고 볼 수 있다. 1960년대에 발표된 Rogers의 개혁의 확산 이론(diffusion of innovations)[11]에서는 혁신의 수용자를 혁신자(innovators), 초기수용자(early adopters), 초기 다수 수용자(early majority), 후기 다수 사용자(late majority), 지각 수용자(laggards)의 5가지 범주로 나눈다. 이 이론에서 사회적 활동을 많이 하는 소수의 혁신자와 초기 수용자를 인플루엔셜이라고 볼 수 있다. 이러한 이론들은 학계를 넘어 마케팅 비지니스에 적용 되어왔다.

이러한 전통적인 이론과 조금 다른 의견도 있다. 2001년에 Domingos와 Richardson은 새로운 세대들이 인플루엔셜의 의견보다는 동료와 친구들의 의견에 더 귀를 기울이므로 인플루엔셜을 통한 마케팅 보다 협력적 필터링(collaborative filtering) 같은 네트워크에 기반하는 방법이 더 효과적일 것이라 주장하였다[2]. 협력적 필터링 방법의 예는 Amazon.com에서 자신과 비슷한 취향을 가진 사람들이 어떤 책을 샀는지 추천해 주는 시스템을 들 수 있다. 2007년 Watts와 Dodds는 시뮬레이션을 통해 인플루엔셜 뿐 아니라 평범한 사람들의 역할도 강조했다[15]. 인플루엔셜이 평범한 사람들 보다 정보 확산에 자주 큰 영향을 끼치기는 하지만 평범한 사람들의 기여 없이는 모든 확산을 설명할 수 없다는 것이다. 최근의 연구들은 정보 확산이 인플루엔셜에만 의존하지는 않는다는 결과를 보이고 있지만 인플루엔셜의 존재를 부정하지는 않는다.

또한 소셜 네트워크를 그래프로 이해해서 인플루엔셜을 정의하는 개념을 소개하겠다. 이 개념들은 Wasserman과 Faust의 책에 좀 더 자세히 소개되어 있다[14]. 가장 중요한 사람은 대개 그래프에서 전략적으로 좋은 위치에 있는 사람이라고 볼 수 있다. 좋은 위치에 있는 사람은 영향을 끼칠 때 많은 기회를 가질 수 있으며 적은 제약을 받기 때문이다. 그렇다면 좋은 위치란 과연 무엇인가? 이 질문에 대해 딱 떨어지는 정답은 없지만 몇몇 서로 다른 개념들로 정의를 내릴 수 있다. 그림 2는 별 모양으로 생긴 Star 네트워크, 원형의 Circle 네트워크, 선으로 이루어진 Line 네트워크의 예로 이 개념들을 설명하기 위해 사용하겠다.

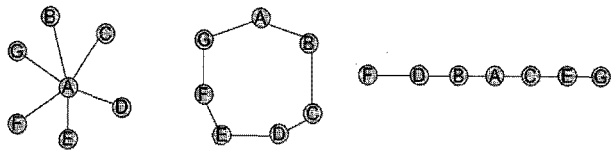


그림 2

첫째 개념은 degree 이다. degree는 어떤 노드가 얼마나 많은 노드와 접해 있는지를 측정하는 개념이다. 그림 2의 (a)에서 노드 A의 degree는 6이고 나머지 노드들의 degree는 1이다. degree 가 높은 노드일수록 정보를 전달할 많은 기회가 있다고 할 수 있다. 그러므로 degree가 가장 높은 노드 A가 가장 좋은 위치에 있다. 그림 2의 (b)에서는 모든 노드의 degree가 2로 같다. 이 네트워크에서는 모든 노드의 위치가 동등한 중요도를 갖는다. 그림 2의 (c)에서는 양 끝의 노드 F,G만 degree가 1 이고 다른 노드들은 degree 2를 갖는다. 양 끝단의 위치만 불이익을 받는 위치인 것을 알 수 있다.

두 번째 개념은 closeness이다. closeness는 어떤 노드가 다른 노드들과 얼마나 가까운지를 측정하는 개념인데 한 노드와 다른 노드들간의 경로의 길이(path length)의 합으로 나타낼 수 있다. 그림 2의 (a)에서 노드 A는 다른 노드들과 모두 거리 1이 떨어져있는 반면 다른 노드는 노드 A를 제외한 다른 노드들과 거리 2가 떨어져 있다. 그러므로 경로의 길이가 가장 짧은 노드 A가 closeness가 가장 좋다고 할 수 있으며 가장 좋은 위치가 된다. 그림 2의 (b)에서는 한 노드에서 다른 노드들까지의 경로의 길이가 각기 다르지만 모든 노드들이 동일한 경로 길이 분포를 갖고 있기 때문에 모든 노드의 closeness는 같다. 그림 2의 (c) 에서는 A의 경로 길이의 합이 12로 가장 짧고 양 끝의 F,G의 경로 길이의 합이 21로 가장 길다. 즉 A의 위치가 가장 좋은 위치인 것이다.

세 번째 개념은 betweenness이다. betweenness는 어떤 노드가 다른 노드 쌍 사이에 위치하는 정도를 측정하는 개념이다. 그림 2의 (a)에서 노드 A가 노드 F에게 접근하기 위해서는 바로 접근 하면 되지만 노드 C가 노드 F에 접근하기 위해서는 반드시 노드 A를 거쳐가야만 한다. 중간에서 정보의 흐름에 가장 잘 개입할 수 있는 노드 A가 가장 betweenness가 높다. 그림 2의 (b)에서는 모든 노드들이 동일한 betweenness를 갖고 있으며 그림 3의 (c)에서는 양 끝단의 노드 F,G는 정보의 흐름을 차단할 능력이 없어서 betweenness가 가장 낮으며 중심에 가까운 노드일수록 betweenness가 높다.

4. Computer Science의 연구

인플루엔셜을 찾기 위한 노력은 Computer Science

커뮤니티에서도 계속되었다. 특히 인터넷의 발달로 전통적인 소셜 네트워크가 확장된 거대한 온라인 소셜 네트워크들이 등장하면서 더욱 활발히 연구되었다. 온라인 소셜 네트워크는 Facebook, Cyworld, Flickr, Twitter, Myspace 등 종류가 많아 전부를 예로 들 수 없을 정도이다. 각기 다른 온라인 소셜 네트워크마다 제공하는 기능과 성격이 다르기 때문에 인플루엔셜을 찾는 유일한 방법이 존재하는 것은 아니다. 하나의 온라인 소셜 네트워크 서비스에서도 인플루엔셜의 정의에 따라 찾아지는 결과가 달라진다. 여기서는 최근 가장 주목을 받고 있는 Twitter를 중심으로 인플루엔셜에 대한 연구를 소개하고자 한다.

Twitter라는 마이크로 블로깅 서비스 에서는 사용자들이 140자 이하의 단문메세지를 간단히 전달할 수 있는 기능을 제공하고 있다. Twitter에서는 다른 사용자를 follow할 수 있으며 follow 함으로써 소셜 네트워크에서의 관계가 맺어 진다. follow한 사용자가 작성한 메시지(앞으로 tweet이라고 일컫겠다)는 자신의 Twitter페이지나 스마트폰 같은 이동통신기기에서 확인할 수 있다. 2009년에 발생한 뉴욕 허드슨강의 비행기 추락 사고를 기존 뉴스 매체보다 Twitter에서 먼저 알려주거나 국가교통정보센터에서 설 연휴 교통정보를 Twitter를 통해 알려주는 사례를 통해 빠른 정보 전달력과 그 유용함으로 주목받고 있는 서비스이다. 일반 사용자 뿐 아니라 사회적으로 유명한 연예인, 스포츠스타, 오피니언 리더, 정치인, 언론사, 기업, 단체 등 각계각층이 소통 및 홍보의 목적으로 즐겨 사용하고 있다(그림 3). 전통적인 소셜 네트워크

그림 3 국가교통정보센터와 김연아의 Twitter

표 1 follower의 수에 의한 랭킹 (top 10)

이름	비고
ashton kutcher	배우
Britney Spears	가수
Ellen DeGeneres	쇼호스트
CNN Breaking News	뉴스
Oprah Winfrey	쇼호스트
Twitter	Twitter
Barack Obama	미국 대통령
Ryan Seacrest	쇼호스트
THE_REAL_SHAQ	스포츠 스타
Kim Kardashian	모델

와 정보를 전달하는 뉴스채널의 성격을 동시에 갖고 있는 Twitter에서 인플루엔셜은 누굴까?

Cha et al.은 Twitter에서 influence가 어떻게 정의될 수 있는지 알아보기 위해 사용자의 follower의 수, tweet이 retweet 된 횟수, 이름이 mention 된 횟수 등 다양한 기준을 방대한 데이터에 적용하여 비교하였다[1]. 이러한 다양한 측정기법 중 본 글에서는 follower 수의 집계와 retweet된 횟수에 대해 주목해보겠다. 여기서 follower의 수는 앞서 설명한 degree와 비슷한 개념으로 얼마나 많은 사람들이 사용자의 tweet을 직접 전달받게 되는지를 나타내는 것이다.

2010년 Kwak et al.이 발표한 follower 수에 의한 인플루엔셜의 랭킹은 표 1과 같다[8]. 이 기준으로 인플루엔셜을 정의할 경우 팬이나 지지자가 많은 유명인사들이 인플루엔셜이라는 것을 알 수 있다.

하지만 이 랭킹만이 인플루엔셜을 설명한다고 할 수 는 없다. 그래서 사람들이 생각해 본 것이 페이지랭크(PageRank) 알고리즘이다[10]. 구글이 사용하고 있는 이 알고리즘은 여러 링크로 서로 연결된 인터넷 웹페이지들의 네트워크에서 어떤 페이지가 가장 중요한 페이지 인지를 알아내는 알고리즘이다. 페이지랭크 랭킹은 랜덤으로 웹서핑을 하는 사용자가 링크를 따라가면서 웹서핑을 할 때 어떤 웹페이지에 머무르는 시간이 얼마나 긴지를 알아볼 수 있는 랭킹이다. 이 알고리즘을 웹페이지들과 비슷한 링크 구조를 갖는 소셜 네트워크에 적용하면 누가 가장 영향력이 있고 중요한 사람인지 알 수 있다.

페이지랭크는 단순히 웹페이지를 향한 링크의 숫자를 세는 것이 아니라 네트워크에서 웹페이지의 영향이 흐르게 하여 랭킹을 계산한다. 같은 숫자의 링크를 갖고 있는 두 웹페이지가 있을 때 유명한 웹페이지들로부터 링크를 많이 받은 웹페이지가 그렇지 않은 웹페이지보다 랭킹이 높게 된다. 한 페이지의 페이지랭크를 계산 하는 방법은 식 (1)과 같다.

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \quad (1) : \text{PageRank}$$

$PR(pi)$는 pi의 페이지랭크이고 d는 damping factor, N은 웹페이지의 총 개수, $M(pi)$는 pi 페이지를 링크한 페이지들의 집합, $L(pj)$는 pj 페이지에서 밖으로 나가는 링크의 개수이다. damping factor란 웹서핑을 하는 사람이 계속 링크를 따라 클릭하여 이동할 확률이다. 1 − d의 확률로 웹서퍼는 링크를 따라가지 않고 랜덤한 페이지로 점프하게 되며 대개 d의 값은 0.85로 준다. 식 (1)을 살펴보면 한 웹페이지는 자신의 페이지랭크를 자기가 링크하고 있는 페이지에 골고루 나눠준다는 것을 알 수 있다. 링크를 받은 페이지 쪽에서 이렇게 나눠받은 페이지랭크를 Damping factor를 고려해서 합산한 것이 페이지랭크 값이 된다. 이 과정을 페이지랭크 값들이 변동이 크게 없이 수렴할 때 까지 반복하여 최종 페이지 랭크의 값을 구한다. 이 페이지랭크를 Twitter 사용자를 웹페이지로 생각하고 follow 관계를 링크로 생각하여 적용한 결과는 표 2와 같다[8]. follower 수가 많을수록 페이지 랭크가 높은 경향이 있기 때문에 표 1과 비교하여 구성은 비슷하지만 인플루엔셜 의미의 차이에 의해 순서가 일부 바뀌었다는 것을 볼 수 있다.

Twitter가 제공하는 기능 중 retweet이라는 기능이 있다. 다른 사람으로 부터 들은 tweet을 자신의 follower들도 알 수 있도록 전달하는 기능으로 원래의 tweet을 그대로 전달하거나 자신의 의견을 덧붙여서 전달할 수 도 있다. 어떤 사용자의 tweet이 얼마나 retweet되었나를 집계하면 그 tweet 이 얼마나 유명한지 알 수 있으며 그 사용자가 얼마나 영향을 미칠 수 있는지를 알 수 있다. 이러한 집계에 의한 랭킹의 결과가 표 3에 나와 있다[8]. 이 랭킹을 보면 표 1, 표 2와는 다르게 상위 랭킹에 연예인이나 정치인 같은 유명인사들의 수가 줄고 대신 뉴스매체들이 많이 등장했음

표 2 페이지랭크에 의한 랭킹 (top 10)

이름	비고
ashton kutcher	배우
Barack Obama	미국 대통령
CNN Breaking News	뉴스
Ellen DeGeneres	쇼호스트
Britney Spears	가수
Oprah Winfrey	쇼호스트
THE_REAL_SHAQ	스포츠 스타
John Mayer	가수
Twitter	Twitter
Ryan Seacrest	쇼호스트

표 3 retweet되는 수에 의한 랭킹 (top 10)

이름	비고
Mashable	뉴스
BNO News	뉴스
TweetMeme	뉴스
oxfordgirl	저널리스트
CNN Breaking News	뉴스
TechCrunch	뉴스
Fabolous	가수
The New York Times	뉴스
lil duval	코미디언
Iran	이란에 관한 내용

을 알 수 있다. 특히 Mashable이나 TweetMeme, TechCrunch같은 소셜 미디어나 IT기술에 관한 특정주제에 관한 뉴스매체들이 많은 영향을 끼치는 것이 특징이다. 이 랭킹에서의 인플루엔셜은 여러 사람에게 전달될 수 있도록 retweet될 만한 가치가 있는 소식을 많이 전해주는 사람들이라고 할 수 있겠다.

페이지랭크를 Twitter에 바로 적용하지 않고 Twitter에 맞게 수정하는 방법도 있다. Daniel Tunkelang은 tweet이 retweet될 확률을 고려하여 TunkRank를 제안하였다[12]. TunkRank에서는 영향(influence)을 식 (2)와 같이 정의한다. Influence(X)는 사용자 X가 작성한 tweet을 읽을 사람들의 수의 기대값이다. 이 때 이 tweet의 retweet으로 읽은 사람의 수도 포함한다.

$$Influence(X) = \sum_{Y \in Followers(X)} (1 + p * Influence(Y)) / \| Following(Y) \|$$

(2) : TunkRank

Followers(X)는 X를 follow하는 사용자의 집합이고 Following(Y)는 Y가 follow 하는 사용자의 집합이다. Y가 X의 follower 라면 1 / || Following(Y) ||의 확률로 X가 작성한 tweet을 읽는다고 가정한다. Y가 X의 tweet을 읽을 때 p의 확률로 retweet을 한다. 이런 가정 아래 Influence(X)는 X가 쓴 tweet을 읽게 되는 사람의 수를 나타내게 된다. 이 알고리즘을 구현하여 랭킹을 실시간으로 조회해 볼 수 있는 웹사이트[13]도 있으며 상위 랭킹은 표 4와 같다. 사용자의 tweet이 얼마나 읽히는지를 보여주는 인플루엔셜 랭킹이다. retweet되는 확률 p를 사용자마다 같게 두었기 때문에 실제로는 다른 사용자 보다 retweet을 좀 더 많이 받는 뉴스 매체들이 표 3과는 다르게 상위권에 잘 보이지 않는다.

Kwak et al.은 정보의 확산에 있어서 정보를 수용하는 순서가 중요하다고 생각하여 Twitter에 유효수용자(effective reader)라는 개념을 도입하여 인플루엔

표 4 TunkRank 랭킹 (top 10)

Rank	User	TunkRank Score	Raw Score
1.	BarackObama	100	10597.6950
2.	kevinrose	100	10197.8000
3.	macrumors	96	7307.4300
4.	leolaporte	95	6684.5200
5.	Oprah	95	6441.6700
6.	lancearmstrong	94	6358.2900
7.	stephenfry	94	6272.7000
8.	aplusk	94	6110.9700
9.	britneyspears	93	5569.9300
10.	ricksanchezcnn	93	5444.4054
11.	nytimes	92	5249.3200
12.	THE_REAL_SHAQ	92	5198.7100
13.	trent_reznor	91	4803.3200
14.	TheOnion	91	4774.5800
15.	NathanFillion	90	4303.8500
16.	algore	90	4276.9500
17.	MarthaStewart	88	3525.7100
18.	perezhilton	88	3493.4400
19.	richardbranson	88	3480.5000
20.	twibbon	88	3345.8100
21.	sitepointdotcom	87	3295.6400
22.	google	87	3224.7400
23.	ijustine	87	3220.9000
24.	mrskutcher	87	3086.9200
25.	SonyPlayStation	86	3016.2700

셜을 찾았다[7]. 유효수용자는 같은 문맥의 정보를 이전에 접해보지 않은 새로운 정보수용자를 뜻한다. 예를 들어 동계올림픽 결과에 관한 tweet을 이미 읽어서 정보를 알고 있는 정보수용자에게 같은 문맥의 새로운 tweet이 전달된다면 이 정보수용자는 유효수용자가 아니다. follower의 수와 유효수용자의 수를 비교해보면 80%의 Twitter 사용자가 자신의 follower의 20%만을 유효수용자로 가지고 있다고 한다. 이 결과는 follower의 수만 많다고 인플루엔셜이라고 할 수 없다는 사실을 뒷받침한다. 그래서 Kwak et al.은 정보수용 순서를 고려하여 아래와 같은 유효사용자에 기반한 인플루엔셜 랭킹을 제안하였다.

전체 사용자를 U라고 할 때 S(u)는 사용자 u의 상태를 나타낸다. 사용자의 상태는 0과 1 두 가지로 상태 0은 같은 문맥의 정보를 아직 받지 못한 상태를 나타내고 상태 1은 이미 관련 정보를 알고 있는 상태를 나타낸다. 모든 사용자의 상태는 상태 0으로 초기화 된다(식 (3)).

$$\forall u \in U,\ S(u) = 0 \tag{3}$$

사용자 u가 작성한 tweet w에 대한 유효사용자 ER0(w)는 식 4와 같이 u의 follower 중에 상태가 상태 0 인 사용자의 집합이다.

$$ER_0(w) = \{v \mid v \in follower(u) \text{ and } S(v) = 0\} \tag{4}$$

사용자 u의 영향 IF0(u)는 사용자 u가 작성한 모든 tweet의 유효사용자 수의 합이 된다. 식 (5) T(u)는 사용자 u가 작성한 모든 tweet의 집합이다.

$$IF_0(u) = \sum_{w \in T(u)} \| ER_0(w) \| \tag{5}$$

여기에 사용자가 tweet을 읽을 확률과 사용자의 기억력까지 고려한 모델로 부터 얻은 인플루엔셜 랭킹은 표 5와 같다.

표 5 유효사용자를 고려한 랭킹 (top 10)

이름	비고
Mashable	뉴스
SiliconAlleyInsider	뉴스
The New York Times	뉴스
El Online	뉴스
BNO News	뉴스
TechCrunch	뉴스
NPR Politics news	뉴스
CNN Breaking News	뉴스
Guardian Tech	뉴스
NBA	뉴스

이 랭킹에 나타난 대부분의 인플루엔셜은 뉴스미디어로 표 1, 2의 결과와는 큰 차이를 보인다. 정보가 퍼지는데 있어서 유효수용자에 큰 영향을 미치는 인플루엔셜은 뉴스미디어라는 사실을 알 수 있다.

한편 Huberman et al.은 Twitter에서의 관계가 단순한 follow로 이루어진 follower/followee 관계와 메시지를 보내는 등 실제적인 상호작용이 있는 친구 관계로 나뉘어 진다고 보고 친구 관계 네트워크가 더 많은 영향을 미치는 네트워크라고 하였다[4].

5. 인플루엔셜 랭킹의 비교

지금까지 인플루엔셜을 찾는 여러 방법을 소개해왔다. 이러한 방법으로 찾아진 인플루엔셜 랭킹을 어떻게 비교해야 할까? 아래에 몇 가지 방법을 소개한다. 이 방법들로 두 랭킹이 얼마나 비슷하거나 다른지를 알아볼 수 있다.

두 랭킹을 각각 R_1, R_2이라고 하고 랭킹의 길이를 l이라 하자. 우선 두 랭킹이 얼마나 겹치는지를 알아볼 수 있다. 두 랭킹의 겹침(overlap) O는 식 (6)과 같이 나타낼 수 있다.

$$O = \frac{|R_1 \cap R_2|}{l} \tag{6}$$

O는 두 랭킹의 공통인 구성원이 얼마나 많은지에 대한 수치로 구성원의 순서에는 영향을 받지 않는다.

Kendall은 두 랭킹의 차이를 측정하기 위해 Kendall tau 거리($K\tau$)를 제안하였다[6]. $K\tau$는 식 (7)과 같이 나타낼 수 있다.

$$K_\tau(R_1, R_2) = \sum_{r_1 \in R_1, r_2 \in R_2} \overline{K}_{r_1, r_2}(R_1, R_2) \tag{7}$$

이 때 r_1과 r_2가 R_1, R_2에서 같은 순서로 나타나면 $\overline{K}_{r_1, r_2}$는 0의 값을 가지고 r_1과 r_2가 반대의 순서로

나타나면 $\overline{K}_{r_1,r_2}$는 1의 값을 갖는다. 이렇게 구해진 $K\tau$로 두 랭킹의 순서가 얼마나 차이가 나는지를 측정할 수 있다. 그러나 Kendall의 방법은 두 랭킹의 길이가 같고 그 구성원도 동일해야 한다는 한계가 있다.

그래서 Fagin et al.은 $K\tau$의 한계를 극복하기 위해 구성원의 순서가 다르거나 한 쪽 랭킹에만 등장하는 등 두 랭킹의 차이별로 페널티를 주는 방법을 제안하였다[3]. 기본틀은 식 (8)과 같이 Kendall의 방법과 비슷하지만 페널티를 합산하는 방법에서 차이가 난다.

$$F_\tau(R_1,R_2)=\sum_{r_1\in R_1,r_2\in R_2}\overline{F}_{r_1,r_2}(R_1,R_2) \quad (8)$$

$\overline{F}_{r_1,r_2}$은 (i) r_1이 오직 한 쪽 랭킹에만 등장하고 r_2는 다른 랭킹에만 등장하는 경우; (ii) r_1이 한 쪽 랭킹에서 r_2 보다 순서가 앞서고 r_2만 다른 랭킹에 등장하는 경우; (iii) r_1, r_2가 양 쪽 랭킹에 모두 등장하나 그 순서가 다를 경우 에는 1의 값을 갖고, 나머지 경우에는 0의 값을 갖는다. 이 방법을 이용하면 두 랭킹의 구성원이 동일하지 않더라도 길이가 같은 랭킹의 차이를 측정할 수 있다. 식 (9)와 같이 정규화하여 나타내면 K는 K = 0일 경우 두 랭킹이 완전히 다르다는 것을 K = 1일 경우 완전히 일치한다는 것을 나타낸다[9].

$$K=1-\frac{F_\tau(R_1,R_2)}{l^2} \quad (9)$$

그림 4는 위의 K를 이용하여 표 1, 2, 3에서 본 follower 수에 의한 랭킹(RF), 페이지랭킹에 의한 랭킹(RPR), retweet되는 수에 의한 랭킹(RRT)을 상위 20위 랭킹부터 상위 2000위 랭킹까지 비교한 결과이다[8]. RRT 랭킹이 다른 두 랭킹과 차이가 많이 난다는 사실을 알 수 있는데 이는 이 랭킹이 다른 두 랭킹이 찾지 못한 인플루엔셜을 찾는 랭킹이라는 것을 뜻한다.

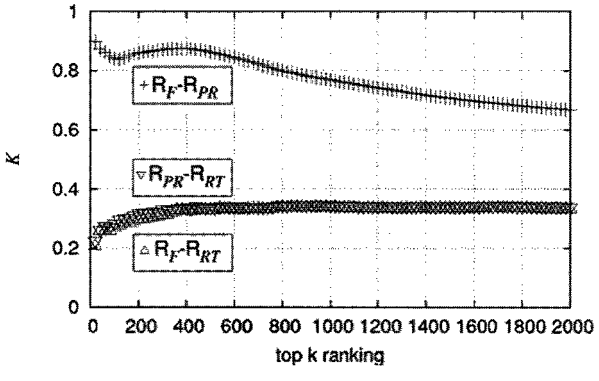


그림 4 K를 이용한 랭킹의 비교[8]

6. 결 론

전문가들은 사람들이 검색엔진에서 직접 정보를 얻는것 보다 친구나 동료의 추천으로 부터 정보를 얻으려는 경향이 증가함에 따라 소셜 미디어 및 소셜 네트워크의 웹 트래픽이 검색 엔진의 트래픽을 초과할 가능성이 있다고 한다. 실제로 지난 12월 Facebook이 Google보다 Yahoo, MSN, AOL 같은 메이져 포털사이트로의 트래픽을 더 많이 발생시켰다[17]. 이런 상황에서 소셜 네트워크에서 인플루엔셜을 찾는 것은 더욱 중요한 문제가 될 것이다.

이 글에서는 소셜 네트워크에서 서로 다른 의미를 갖는 다양한 관점의 인플루엔셜을 찾을 수 있다는 것을 소개하였다. 인플루엔셜의 정의가 다양한 만큼 랭킹의 종류도 다양한데 한 랭킹이 다른 랭킹보다 더 좋다고 단언하기는 힘들다. 어떤 인플루엔셜로 부터 새로운 정보를 얻는 것이 가장 좋겠는가? 어떤 인플루엔셜에게 마케팅 캠페인을 실행하면 가장 효과가 좋겠는가? 이와 같은 참조자의 목적에 알맞은 랭킹을 참조하는 것이 중요하다. 각종 정보 미디어와 인터넷 네트워크가 연결되고 네트워크의 다양한 특성을 정량화하는 것이 쉬워짐에 따라 앞으로 Computer Science에서 이 분야에 더욱 큰 기여를 할 것으로 기대된다.

참고문헌

[1] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy, In Proc. of International AAAI Conference on Weblogs and Social Media (ICWSM), 2010

[2] P. Domingos and M. Richardson. Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 57-66. ACM NewYork, NY, USA, 2001

[3] R. Fagin, R. Kumar and D. Sivakumar. Comparing top k lists. In SODA'03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, pages 28-36, Philadelphia, PA, USA, 2003

[4] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. arXiv:0812.1045v1, 2008

[5] E. Katz and P. Lazarsfeld. Personal influence : The part played by people in the flow of mass communications. Free Press, 1955

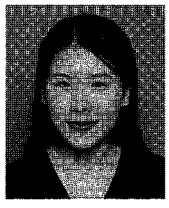
[6] M. Kendall. A new measure of rank correlation. Biometrika, 30(1-2):81?93, 1938
[7] H. Kwak, C. Lee, H. Park, and S. Moon. Finding influentials based on temporal order of information adoption in twitter. WWW'10 poster session, 2010
[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? WWW '10: Proceedings of the 19th international conference on World wide web, 2010
[9] F. McCown and M. L. Nelson. Agreeing to disagree : search engines and their public interfaces. In JCDL'07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pages 309-318, New-York, NY, USA, 2007
[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking : Bringing order to the web. 1998
[11] E. Rogers. Diffusion of innovations, New York:Free Press. 1962
[12] D. Tunkelang. A Twitter Analog to PageRank. http://thenoisychannel.com/2009/01/13/a-twitteranalog-to-pagerank/, 2009
[13] Tunk Rank. http://tunkrank.com, 2009
[14] S. Wasserman and K. Faust. Social network analysis : Methods and applications. Cambridge Univ Pr, 1994
[15] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. Journal of Consumer Research, 34(4):441-458, 2007
[16] W. Zachary. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33(4):452-473, 1977
[17] Facebook Beats Google in Directing Web Portal Traffic. http://www.commercetuned.co.uk/news/facebook-beats-google-in-directing-web-portaltraffic-053.php, 2010

박 호 성

2009 KAIST 전산학과 학사
2009~현재 KAIST 전산학과 석사과정 재학중
관심분야 : 소셜 네트워크, 웹서비스, 클라우드 컴퓨팅
E-mail : hosung@an.kaist.ac.kr

곽 해 운

2006 KAIST 전산학과 학사
2007 KAIST 전산학과 석사
2007~현재 KAIST 전산학과 박사과정 재학중
관심분야 : 소셜 네트워크, 웹서비스 사용자 행태 분석, 추천 시스템
E-mail : haewoon@an.kaist.ac.kr



차 미 영

2002 KAIST 전산학과 학사
2004 KAIST 전산학과 석사
2008 KAIST 전산학과 박사
2008~현재 MPI-SWS 박사후 연구원
관심분야 : 대규모 네트워크 시스템, 멀티미디어 스트리밍 시스템, 온라인 소셜 네트워크
E-mail : mcha@mpi-sws.org

문 수 복

1988 서울대학교 컴퓨터공학과 학사
1990 서울대학교 컴퓨터공학과 석사
2000 University of Massachusetts at Amherst, 컴퓨터공학과 박사
현재 KAIST 전산학과 부교수
관심분야 : 미래인터넷, 소셜 네트워크, 복잡계 네트워크, 테스트베드
E-mail : sbmoon@kaist.edu