데이터마이닝 hw03 네트워크 분석 논문 리뷰

논문: What is Twitter, a social network or a news media?
이름: 정주호
학번: 2022712955
학과: 인공지능융합학과
제출일: 10월 24일 (월)

0. Abstract

이 논문의 목표는 트위터의 위상적인 특성과 새로운 정보 공유 매체로서의 영향력에 관해 연구하는 것이다. 데이터는 전체 트위터 사이트에서 4천170만 명의 사용자 프로필, 14억7천만 개의 소셜 관계, 4,262개의 유행 주제, 1억6천만 개의 트윗을 크롤링하여 사용했다. 이 연구를 통해서 SNS의 특성으로 non-power-law follower distribution, a short effective diameter, and low reciprocity 등을 발견하였다. 또한 트위터에서 영향력 있는 사람들을 식별하기 위해 팔로워 수와 Page Rank별로 사용자 간의 순위를 매겼고, 이 두 순위가 비슷하다는 것을 발견했다. 리트윗별 순위는 앞의 두 순위와 차이가 있어 사용자의 팔로워 수와 사용자의 트윗(게시글)의 인기로 추론한 영향력끼리는 차이가 있음을 보여준다. 이외에도 retweet수, 유명세, 트렌드 등의 주요 지표에 관해 연구해, 전체 트위터의 영역을 아우르고, 트위터에서 정보 확산과 관련하여 정량적으로 연구한 첫 번째 연구로써 그 의의가 있는 논문이다.

1. Introduction

사람들이 트위터에서 어떻게 연결되는지, 트위터상에서 가장 영향력 있는 사람들은 누구인지, 사람들은 트위터에서 무엇에 대해 이야기하는지, retweet을 통해 정보가 어떻게 확산되는지 등 흥미로운 질문에 대한 답을 찾아가면서 저자들은 본 연구의 목적이 새로운 정보 공유 매체로써 트위터의 위상적 특성과 그 영향력에 관해 연구하는 것이라고 이야기한다.

이 논문은 네트워크 분석에서 시작하여 팔로잉과 팔로워의 분포, 팔로워와 트윗 사이의 관계, reprocity(상호관계), degree of separation and homophily(동질감)에 관해 연구한다. 다음으로 팔로워 수, 페이지랭크 및 retweet 수에 따라 사용자들의 순위를 매기고 그들 간의 정량적 비교를 제시한다. 앞에서 언급했듯이 유명한 사람일수록 Retweet되는 수가 많을 것이므로, Retweet에 의한 순위에서는 팔로워 수가 100만 명 미만인 사람들을 팔로워 수가 100만 명 이상인 사람들보다 더 높은 순위를 갖도록 조정한다. 또한 그 시대의 트렌드 주제 분석을 통해 트렌드 주제가 분류되는 범주, 지속 시간, 참여 사용자 수를 정량적으로 보여준다. 마지막으로, 저자들은 retweet을 통한 정보 확산에 관해 연구하였고, 그 각종 지표들과 결과들을 정량적으로 보여준 첫 연구라고 강조한다.

2. Data Collection

데이터는 전체 트위터 사이트에서 4천170만 명의 사용자 프로필, 14억7천만 개의 소셜 관계, 4,262개의 유행 주제, 1억6천만 개의 트윗을 크롤링하여 사용했다.

User profile

전체 공개된 프로필에 한해서, 이름, 사는 지역, 웹페이지, 인적정보, tweet 개수 등의 정보를 포함하고 있다. 그리고 누구를 팔로우 하는지, 누가 팔로우 하는지에 대한 정보도 추출하였다.

Trending topic

Twitter는 가장 자주 언급되는 문구, 단어, 해시태그를 추적해 정기적으로 'Trending topic'이라는 제목으로 게시한다. Twitter는 별도의 설정이 없는 한 기본적으로 모든 사용자의 홈페이지의 오른쪽 사이드바에 현재 유행하고 있는 상위 10개 주제 목록을 보여준다. 이 주제 목록을 수집하여 사용하였다.

3. Basic Analysis

3.3 Reciprocity

논문에서 reciprocity에 대한 용어 설명을 구체적으로 하진 않았지만, 흐름에 의하면 유저끼리 서로 팔로우를 주고 받는 상호성을 의미한다(맞팔로우). 앞에 basic analysis에서 언급했듯이, 트위터에서 팔로우 수에 따른 순위에서 상위에 위치하는 사용자들은 대부분 유명인이나 대중매체기 때문에 자기를 팔로우한 상대를 다시 팔로우하지 않는다. 그러므로 트위터는 실제로 reciprocity가 낮은 경향성을 보여준다. 두 유저 사이에 링크가 있는 쌍은 77.9%가 단방향이며, 오직 전체의 22.1%만이 서로가 서로를 팔로우하는 reciprocate한 상태를 보여준다. 연구진들은 서로를 팔로우하는 상태를 “r-friend”라고 명명하고 트위터가 다른 SNS보다 더 낮은 reciprocity를 보인다고 이야기한다. 또한 논문의 저자들은 67.6%의 사용자들은 트위터에서 팔로잉을 전혀 하지 않고 있는 것을 발견했는데 저자들에 의하면 이러한 사용자들에게 트위터는 소셜 네트워킹 사이트보다 오히려 정보의 원천이라고 추측하고, 이런 형태의 사용자들에 대한 추가 검증은 이 논문의 범위를 벗어났으며 향후 작업을 위해 다루지 않겠다고 이야기한다.

3.4 Degree of Separation

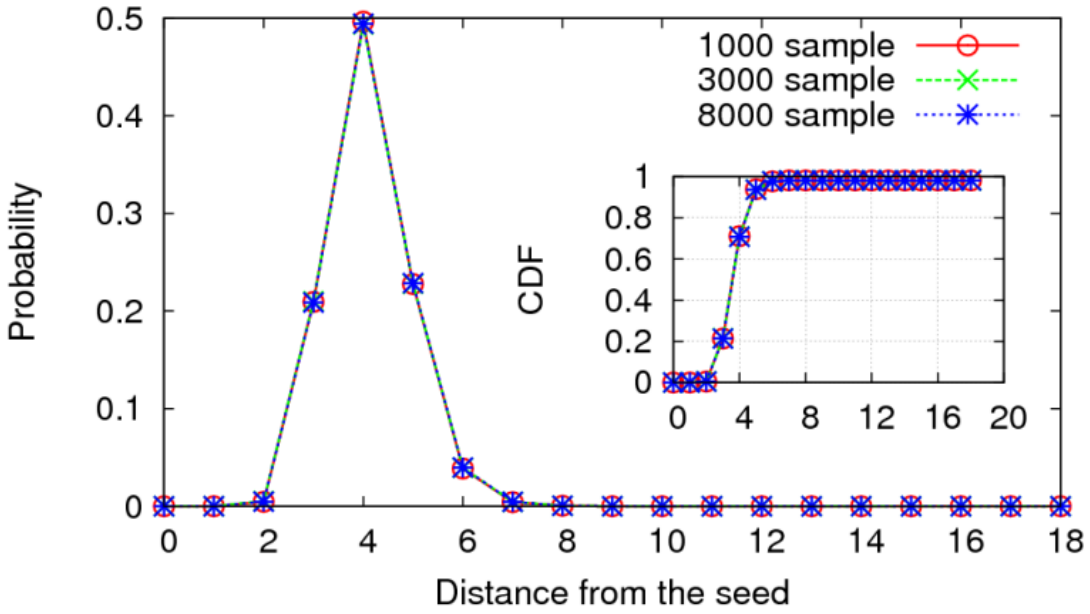


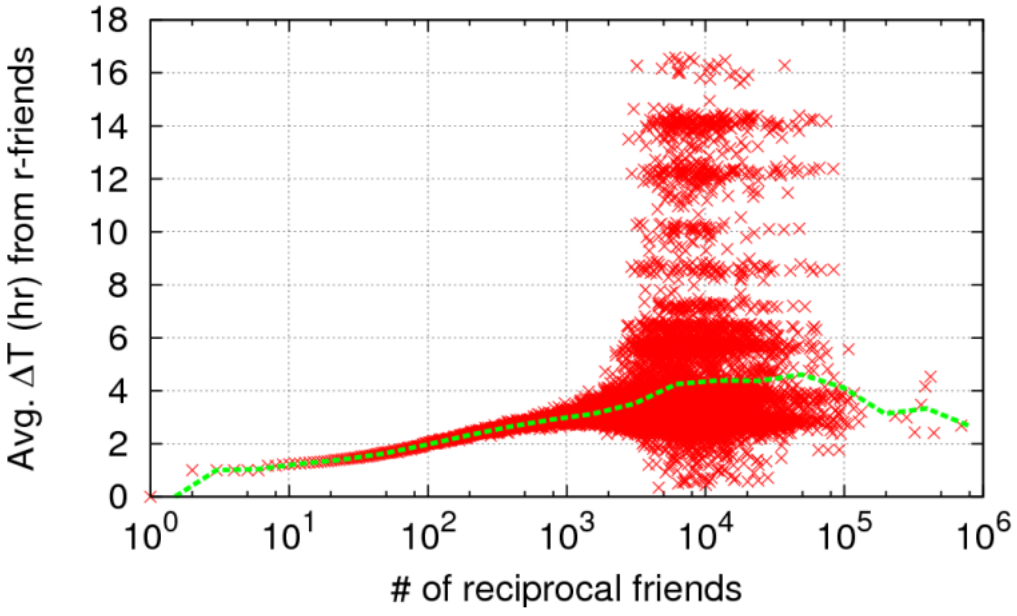
Figure 4: Degree of separation

인간 네트워크에서 굉장히 유명한 이론인 Stanley Milgram's famous 'six degrees of separation' experiment에 따르면, 어떤 사람과도 평균적으로 6단계만 거치면 닿을 수 있다고 한다. 이 부분에서 논문의 저자들은 이 이론을 검증하였는데 개인적으로 재미있었다. 저자들은 앞서 이야기한 인간 네트워크랑 트위터에서의 네트워크는 성격이 다르다고 이야기하는데, 트위터의 관계가 더 직접적이라고 이야기한다. 트위터에서는 다른 사람을 팔로우하는데 (관계를 맺는데) 제약이나 의무가 없기때문에 실제로는 한사람만 팔로우하고 있을 수도 있어서 네트워크를 역추적하기가 어렵다. 또한 앞서 이야기했듯이 전체 사용자 쌍의 22.1%만이 상호적이기 때문에, 저자들은 트위터에서 두 사용자 사이의 평균 경로 길이가 다른 알려진 네트워크보다 길 것으로 예상하였다. 그러나 연구를 진행하고 되게 흥미로운 결과를 발견했는데, 저자들의 가설과는 다르게, 트위터 네크워크에서 평균 경로 길이 4.12면 다른 사람에게 닿을 수 있다는 것이다. 이는 트위터 크기의 네트워크에 비해 상당히 짧은 거리이며, 방향 그래프에 대한 저자들의 예측과는 반대라고 이야기하면서 저자들도 되게 유의미한 발견을 하였다고 주장한다. 나아가 트위터의 역할에 대해 새로운 방향을 제시하는데, 사람들은 단순히 트위터를 SNS의 기능이 아닌 정보를 얻기 위한 수단으로 다른 사람을 팔로우한다는 것이다.

➔ 개인적으로 통상적으로 사회 현실에서 반영되는 이론에 대한 검증을 했다는 점이 인상깊었다. 그리고 그 예측 및 이론과 반대되는 결과에 따른 현상 분석도 타당하여 재미있었다.

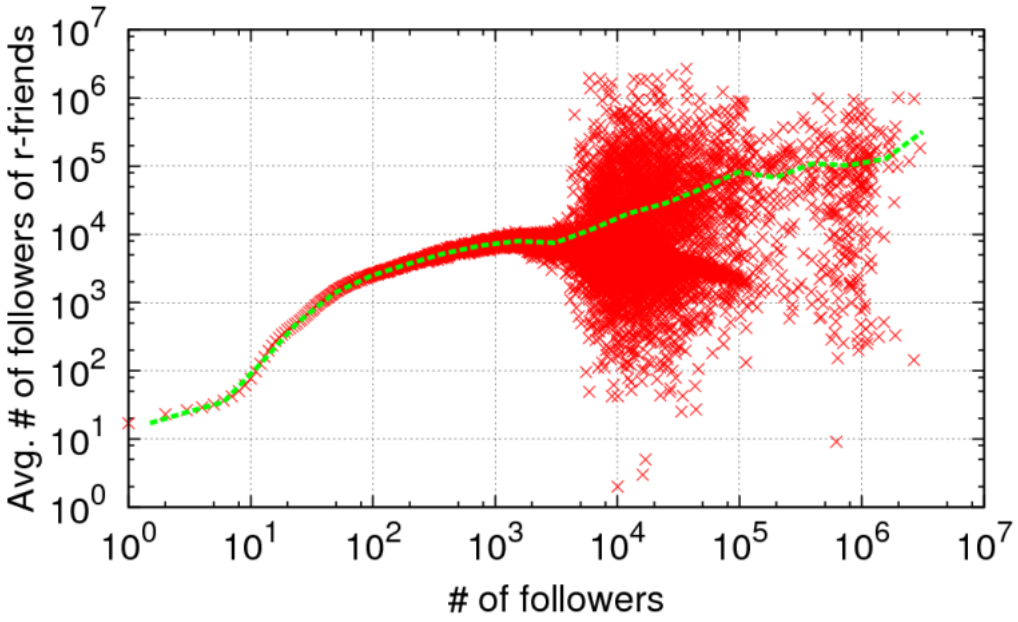
3.5 Homophily

Homophily란 비슷한 "사람들끼리의 연결"을 의미한다. 이는 비슷한 성향의 사람들끼리 연결될 경향성이 더 높다는 가정에 의거한다. 이 부분에서 저자들은 Homophily를 두가지 컨셉으로 조사하는데, 먼저는 지리적 위치이고, 다음은 유명세이다. 먼저 지리적 특성의 경우 트위터를 사용하는 사용자들이 스스로 자신의 위치를 공개하는데 저자들은 오차범위 시간대를 정하고 그 시간대동안 사용자가 스스로 보고한 위치에 사용자가 있다고 가정하고 연구한다.



지리적 위치와 Homophily와의 관계를 비교한 그래프이다. 결과는 사용자와 r-friend의 평균 시간 차이는 3시간 미만이다. 나아가 50명 이하의 r-friends를 가진 사람들의 평균 시차는 약 1.07시간 밖에 되지 않는다. 저자들은 2,000명 미만의 상호 관계를 가진 트위터 사용자가 지리적으로 가까울 가능성이 높다고 주장한다.

다음으로는 사용자의 유명세와 homophily와의 관계를 비교하는데, 이때 유명세는 그 사람의 팔로워 수로 간주하였다. 그 다음 연구진들은 "Does a user of certain popularity follow other users of similar popularity and they reciprocate?"라는 질문에 대답을 하면 이 둘 간의 관계를 비교할 수 있다고 이야기하는데 이부분도 되게 흥미로웠다.



이 부분에서는 사용자들 간의 지리적 위치와 r-friends의 팔로워 수라는 두 가지 관점에서 homophily에 대해 알아보았다. 팔로워가 1,000 이하인 사용자는 지리적으로 r-friend와 가까울 가능성이 높고 r-friend와도 비슷한 popularity(팔로워 수)를 보유하고 있다는 것을 발견하였다.

3장에서 저자들은 트위터의 소셜 네트워킹 측면을 살펴보았고 어느 정도 Homophily를 발견했다고 이야기한다. 그러면서 트위터는 소셜 네트워크의 잘 알려진 특징과 다르다고 주장하는데, 팔로워의 분포는 power-law가 아니며, degree of separation 정도가 예상보다 짧으며, 대부분의 링크는 reciprocate(맞팔)하지 않는다는 점이 그것이다. 하지만 reciprocate한 r-friend에 한정하여 살펴본다면, 그들 사이에는 어느 정도 homophily를 보인다고 결론 지으며 마무리한다.

4. Ranking Twitter Users

트위터 사용자의 인기는 그 사람의 팔로워 수로 쉽게 추정할 수 있다. 그러나 팔로워 수만으로는 사용자의 게시글(트윗)이 여러 번 retweet되거나 단순히 다른 사용자에게 미치는 영향력을 반영하지 못한다. 그래서 저자들은 이부분에서 PageRank 알고리듬과 retweet 수에 따라 사용자를 순

위를 매기고 그 결과를 비교하였다.

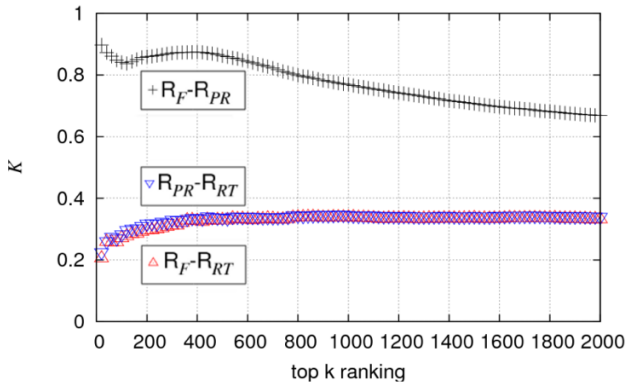
4.1 By PageRank

먼저 팔로잉과 팔로워의 네트워크에 PageRank를 적용한다. 이 네트워크에서 Node는 사용자를 의미하고, 모든 directed edge는 그 사용자가 팔로우하는 사용자에게 매핑된다. 상위 20명의 순위가 매겨진 사용자는 아래 그림에 나와있다. 팔로워 수에 따른 ranking과 PageRank에 따른 ranking이 정확히 일치하지는 않지만, 사용자들은 팔로워 수와 페이지랭크에 의해 비슷하게 순위가 매겨진다는 점을 발견해 냈다.

Rank	Ranking by # of followers			Ranking by PageRank in the following/follower network			Ranking by # of retweet in the diffusion network		
	ID	Name	Remark	ID	Name	Remark	ID	Name	Remark
1	aplusk	ashton kutcher	actor	aplusk	ashton kutcher	actor	mashable	Pete Cashmore	news on social media
2	britneyspears	Britney Spears	musician	BarackObama	Barack Obama	president of U.S.	BreakingNews	BNO News	news
3	TheEllenShow	Ellen DeGeneres	show host	cnnbrk	CNN Breaking News	news	tweetmeme	TweetMeme	news on Twitter
4	cnnbrk	CNN Breaking News	news	TheEllenShow	Ellen DeGeneres	show host	oxfordgirl	oxfordgirl	journalist
5	Oprah	Oprah Winfrey	show host	britneyspears	Britney Spears	musician	cnnbrk	CNN Breaking News	news
6	twitter	Twitter	subject of this paper	Oprah	Oprah Winfrey	show host	TechCrunch	Michael Arrington	news on technology
7	BarackObama	Barack Obama	president of U.S.	THE_REAL_SHAQ	THE_REAL_SHAQ	sports star	myfabolouslife	Fabolous	musician
8	RyanSeacrest	Ryan Seacrest	show host	johncmayer	John Mayer	musician	nytimes	The New York Times	news
9	THE_REAL_SHAQ	THE_REAL_SHAQ	sports star	twitter	Twitter	subject of this paper	lilduval	lil duval	comedian
10	KimKardashian	Kim Kardashian	model	RyanSeacrest	Ryan Seacrest	show host	IranRiggedElect	Iran	about Iran
11	johncmayer	John Mayer	musician	lancearmstrong	Lance Armstrong	sports star	espn	ESPN Sports News	news
12	mrskutcher	Demi Moore	actress	jimmyfallon	Jimmy Fallon	actor	persiankiwi	persiankiwi	about Iran
13	iamdiddy	iamdiddy	musician	iamdiddy	iamdiddy	musician	aplusk	ashton kutcher	actor
14	jimmyfallon	Jimmy Fallon	actor	mrskutcher	Demi Moore	actress	StopAhmadi	Raymond Jahan	about Iran
15	lancearmstrong	Lance Armstrong	sports star	PerezHilton	Perez Hilton	power blogger	Alyssa_Milano	Alyssa Milano	actress
16	algore	Al Gore	politician	nytimes	The New York Times	news	huffingtonpost	HuffingtonPost.com	news
17	mileycyrus	Miley Cyrus	actress / musician	mileycyrus	Miley Cyrus	actress / musician	iamdiddy	iamdiddy	musician
18	nytimes	The New York Times	news	stephenfry	Stephen Fry	actor	iranbaan	Fershteh Ghazi	about Iran
19	coldplay	Coldplay	musician	TheOnion	The Onion	news	nprnews	NPR News	news
20	TheOnion	The Onion	news	KimKardashian	Kim Kardashian	model	PerezHilton	Perez Hilton	power blogger

4.2 By Retweet

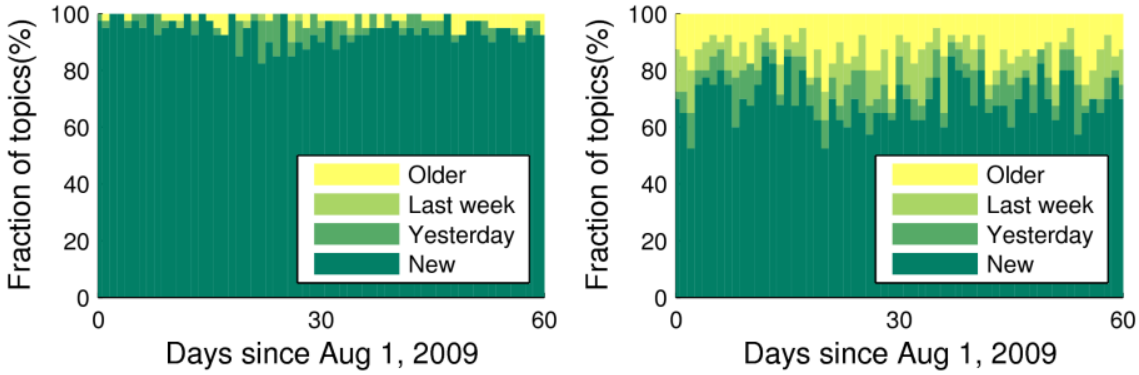
특정 트윗에 대한 retweet 수는 해당 트윗의 인기와 트윗 작성자의 인기를 나타내는 척도이다. 이 부분에서는 그 사용자의 총 retweet 수로 사용자를 순위를 매기는데 이는 우리가 실제 SNS에서 나의 게시글을 사람들이 얼마나 보냐가 아닌 얼마냐 retweet했느냐로 그 사람의 인기를 가늠하는 것과 동일하다. 위 그림의 맨 오른쪽 열은 리트윗 수를 기준으로 상위 20명의 사용자를 나열한 목록이다. Retweet에 의한 순위와 PageRank에 의한 순위에서 동일하게 나타나는 사용자는 페레즈 힐튼만 있다. 나머지는 처음 두 순위 중 어느 쪽에도 속하지 않는다. 재미난 것은, 사용자들을 자세히 살펴보면 2009년 6월 12일이란 선거 기간과 그 이후에 등장한 4명의 사용자가 활발한 트윗으로 유명세를 탔다는 것을 알 수 있다. 저자들은 이 선거 기간에 리트윗에 의해 순위가 상승한 사용자들은 주요 뉴스 매체들이었다는 것을 확인하고, 이 현상을 트위터라는 매체가 정보를 전달하는데 용이하고 그 영향력이 크다는 것을 의미한다고 주장한다. 실제로 retweet별 순위에서 뉴스 매체들이 상승한 것을 보면 트위터가 이 매체의 대체수단이 될 수 있음을 보여준다.



이 그래프는 위에서 언급한 number of followers (RF), PageRank (RPR) and the number of retweets (RRT) 세 개의 순위 간의 정량적 비교를 담은 그래프이다. 이 그래프를 분석해보면, 즉, 팔로워수(RF)와 PageRank(RPR)는 유사하지만 retweet수(RRT)는 다르다. 저자들은 이 부분에 집중하는데, 이 리트윗 수(RRT)가 사용자의 팔로워 수와 그 사람의 트윗의 인기도 간의 차이를 나타내는 것이며 이것이 트위터의 영향력에 대한 새로운 관점을 가져왔다고 주장한다

5. Trending the Trend

앞에서는 트위터 네트워크의 위상학적 특성을 살펴보고, 트위터의 낮은 reciprocity에 대해 연구하였다. 저자들은 이번엔 트위터에서 어떤 주제가 유행이 되고, 그 주제가 어떻게 인기를 얻고, 트위터 네트워크를 통해 어떻게 확산되고 사라지는지를 살펴본다. 저자들은 2009년 6월 3일부터 9월 25일까지 4,266개의 고유한 트렌드 주제를 크롤링하였다.

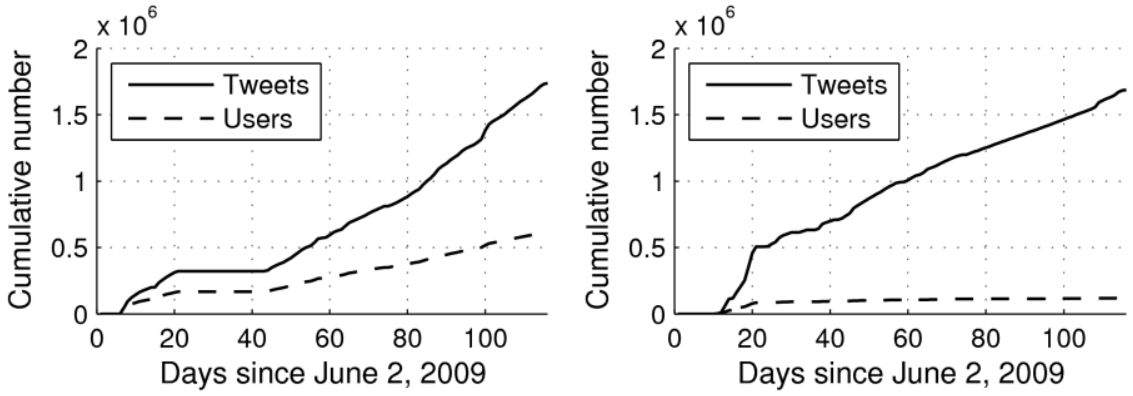


(a) Google (b) Twitter

트위터에서 어떤 주제가 인기 있는지 살펴보기 위해, 저자들은 트위터의 유행 주제를 다른 매체의 주제, 즉 구글 트렌드와 CNN 헤드라인과 비교하였다. 저자들은 트위터 데이터 수집과 같은 기간 동안 구글 트렌드에서 매일 상위 40개의 검색 키워드를 수집했다고 말한다. 또한 동시에 트위터에서 하루에 가장 인기 있는 주제 40개를 추출했다고 한다. 이 둘을 비교할 때 가장 긴 공통 하위 문자열의 길이가 두 문자열의 70% 이상인 경우 검색 키워드와 트렌드 주제를 일치시키는 것으로 간주하였다고 한다. 위 그래프를 보면, 매일 구글의 평균 95%의 토픽이 새로운 반면, 트위터의 72%의 토픽만이 새로운 것이라는 것을 볼 수 있다. 트위터에서는 리트윗, 회신, 언급과 같은 사용자 간의 상호작용 덕분에 구글 검색과 달리 더 널리 퍼질 수 있게 만들며, 이러한 상호작용은 트렌드 주제를 지속시키는 요인이 될 수 있다고 말한다.

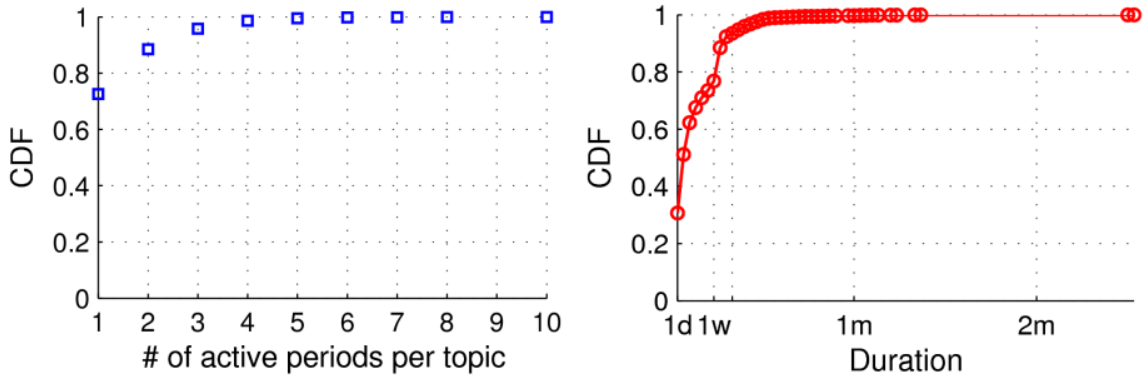
CNN Headline News에서는 일부 뉴스가 CNN보다 먼저 트위터를 통해 터졌고, CNN은 이 뉴스를 보고하는 형태를 띄었다고 이야기한다. 이는 다시 말해, 트위터가 정보를 전달하는 동시에 그 순간에 현상을 실시간으로 중계하는 역할을 하고 있다는 것을 의미한다.

또한 사용자가 평균적으로 몇 개의 주제에 참여하는지에 대한 연구에서는 4,100만 명의 트위터 사용자 중 많은 수의 사용자(8,262,545명)가 트렌드 토픽에 참여했으며, 그 중 약 15%가 4개월 동안 10개 이상의 토픽에 참여했다.



(a) Topic 'apple'　　(b) Topic '#iranelection'

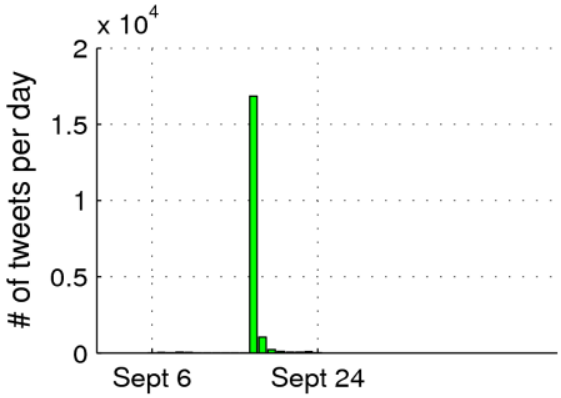
흥미로운 점은 트윗의 수가 지속적으로 증가하는 주제들이 항상 새로운 사용자들의 관심을 사는 건 아니라는 것인데, 위의 그래프를 보면, 'apple'과 '#iranelection'의 두 주제는 비슷한 수의 트윗을 가지고 있지만, 'apple'에 참여하는 사용자의 수는 '#iranelection'의 5배임을 확인할 수 있다. 저자들은 이러한 현상을 통해 특정 트렌드 주제에 대해 오랜 기간에 걸쳐 많은 트윗을 생성하는 핵심 멤버가 존재한다는 것을 발견하였다.



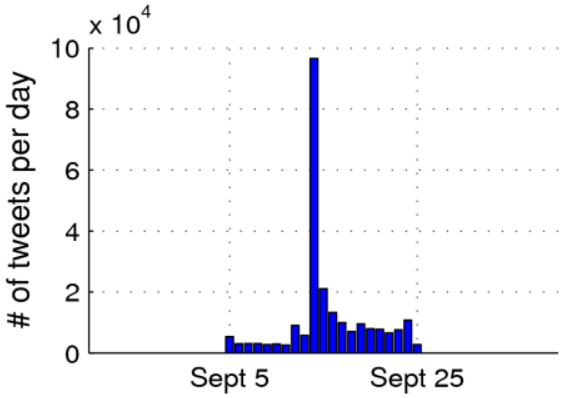
(a) # of active periods / topic　　(b) Duration of active period

Figure 12: Cumulative fraction

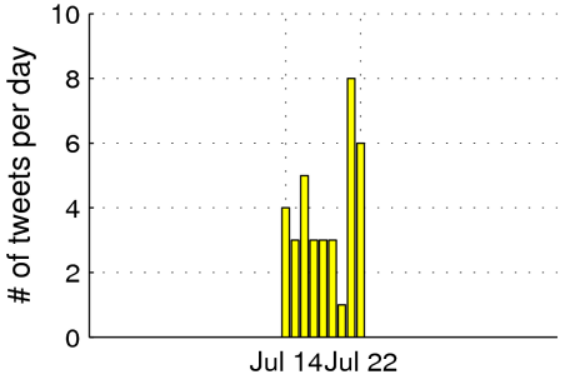
위는 트렌드 주제의 평균 지속 기간을 나타내는 그래프이다. 저자들은 24시간 동안 해당 주제에 대한 트윗이 없으면 그 트랜드 주제가 비활성화하는 것으로 간주하였다. 4,266개의 트랜드 주제에서 6,058개의 활성 기간이 있었다고 한다. 위의 그래프에서 저자들은 활성 기간의 CDF를 그림으로 표시했고 73%의 주제가 단일 활성 기간을 가지고 있다는 것을 발견했다. 또한 대부분의 지속 기간은 일주일 또는 그보다 짧다는 것을 발견했다. 저자들은 이 트렌드 주제들을 하위 카테고리로 나누어 분석하였는데 그 결과는 아래와 같다.



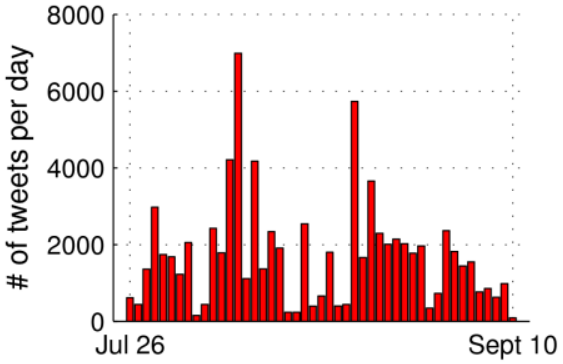
(a) Exogenous subcritical (topic '#backintheday')



(b) Exogenous critical (topic 'beyonce')



(c) Endogenous subcritical (topic 'lynn harris')



(d) Endogenous critical (topic '#redsox')

Exogenous critical에 속하는 주제를 보면 대부분 시기적절한 속보임을 알 수 있으며, 이를 헤드라인 뉴스라고 한다. Endogenous critical 주제는 보다 지속적인 특성을 보이는데, 주로 프로 스포츠 팀, 도시, 브랜드 등이 속한다. Exogenous subcritical 주제는 #동아리에서의 생각들, #싫어하는 것들과 같은 해시태그가 있는데 이는 무의미한 주제로, 사용자의 제한된 관심을 사로잡고 결국엔 사라진다.

	Subcritical		Critical	
Exo.	31.5%	(1,905)	54.3%	(3,290)
Endo.	6.9%	(419)	7.3%	(444)

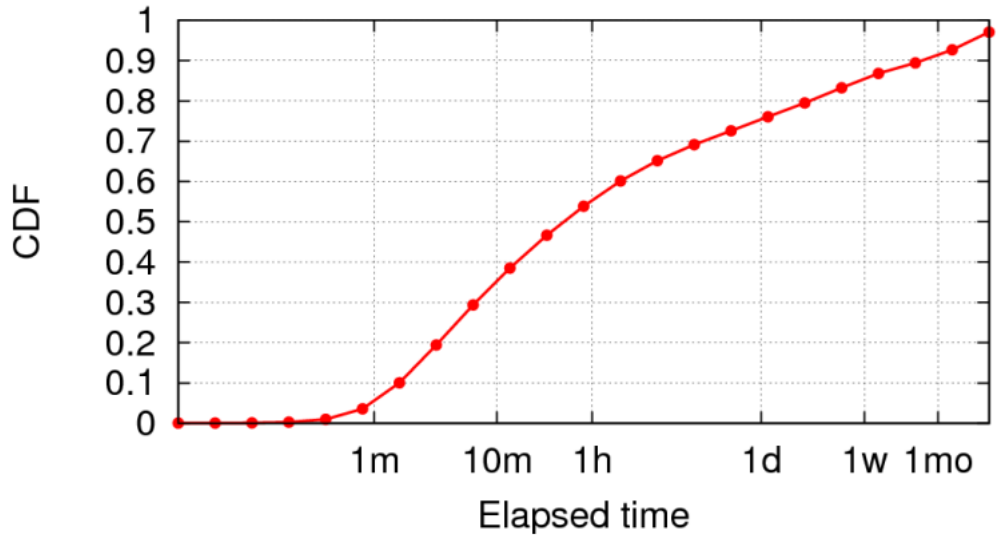
Table 1: # of topics in each category

위는 각 클래스별 활성 기간 수와 백분율를 나타낸 표다. 가장 많은 Exogenous critical 주제에 속한다. 저자들은 이 연구를 통해 트위터 사용자들이 헤드라인 뉴스의 주제에 대해 이야기하고 새로운 뉴스에 반응하는 경향이 있다고 주장한다.

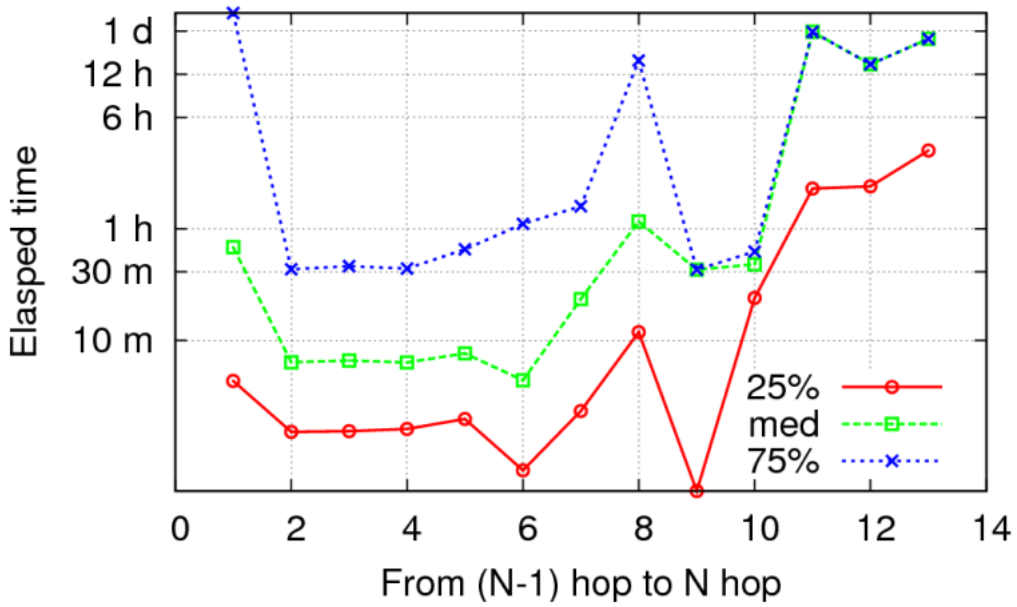
6. Impact of Retweet

이전까지는 어떤 주제가 유행하는지, 어떻게 인기를 얻고 언제 사라지는지에 대해 연구하였다면, 이번에는 트위터에서 정보가 어떻게 확산되는지에 대해 연구한다. Retweet은 단순히 맞팔을 맺고 있는 사람끼리를 넘어 정보를 전달하는 효과적인 수단임을 앞선 연구들에서 밝혀냈다. 그래서 저자들은 트렌드 주제별로 retweet을 파고들어 정보 확산에 영향을 미치는 요인에 대해 조사하였다. 저자들은 어느 한 사용자가 트윗을 올리면 단순히 그 사람의 팔로워만 그걸 보는 것이 아니라 retweet된 것을 본 그 팔로워의 팔로워까지 영향을 주기 때문에 사용자의 팔로워수에 대한 영향력이 아닌 가중치를 부여하여 그 영향력을 계산했다고 이야기한다. 그 결과 사용자가 많든 적든 한번 retweet이 퍼지기 시작하면 그 트윗이 일정 수의 사용자들에게 전달될 가능성이 높고, 이는 정보 확산에 retweet이 강력한 힘을 갖고 있다는 것을 의미한다. 저자들은 트위터상에서도 retweet을 통해 개인들이 어떤 정보가 중요하고 확산해야 하는지를 알고 있고 이는 사회에서 발생하는 집단지성과 비슷한 형태를 띄고 있다고 설명한다.

*Retweet Tree: 해당 retweet이 트위터 상에서 얼마나 멀리 확산되는지를 나타내는 트리. 모든 retweet tree는 트위터 네트워크의 하위 그래프이다.



이 그래프를 보면 트윗에서 리트윗까지의 시간 차이를 알 수 있는데, 리트윗의 절반은 1시간 이내에, 75%는 하루 이내에 발생한다. 하지만 리트윗의 약 10%는 한 달 후에 일어나는 것을 볼 수 있다.



이번 그래프는 리트윗 트리의 두 노드 사이의 시간 지연을 나타내는 그래프이다. 리트윗 트리는 대부분 1홉 깊이로 되어 있어 첫 번째 홉의 시차가 퍼져 중앙값이 1시간 미만이고 사분위간 범위가 몇 분에서 하루 이상으로 확대된다. 흥미로운 점은 두 번째 홉부터 시작해서 두 번째 홉 이상 떨어진 리트윗이 훨씬 더 반응적이며, 최대 5홉 떨어진 곳에서 다시 발생한다는 것이다. 저자들은 정보 확산을 위한 매개체로서 트위터의 강점은 리트윗의 속도에서 두드러진다고 강조한다.

7. Related Work

최근 온라인 소셜 네트워킹 서비스의 인기가 증가함에 따라 SNS특성에 대한 연구가 많이 증가하고 있다. 최근의 연구들은 단순히 데이터를 크롤링하는 것을 넘어 그 특성에 대해 깊이 연구하는 것에 초점을 맞추고 있다. 이 논문을 기준으로 트위터는 만들어진 지 3년도 채 되지 않았지만, 지난 2년 동안 많은 관심을 끌었다고 한다.

2007년에 트위터에 대한 예비 분석한 연구가 있는데 데이터 세트는 약 76,000명의 사용자와 100,000개의 게시물을 사용했다. 그 연구에서는 clique percolation methods를 활용하여 주제에 대한 사용자의 의도와 사용자 클러스터를 찾는다. 또한 팔로워 수와 팔로잉수의 관계를 통해 사용자 특성을 분석하는 연구도 있고, 입소문 브랜드에 대한 예비 분석 연구도 있다. 저자들은 자신들의 연구가 트위터 영역 전체를 처음으로 살펴본 것이라고 강조한다. 또한 저자들은 리트윗 트리를 정보 확산의 커뮤니케이션 채널로 취급하고, 리트윗이라는 것이 많은 청중에게 도달하고 빠르게 퍼지는 현상을 관찰하였다(Information cascades).

8. Conclusion

이 논문에서 저자들은 트위터에서 4천170만 명의 사용자 프로필, 14억7천만 개의 소셜 관계, 4,262개의 트렌드 주제, 1억6천만 개의 트윗을 크롤링하여 수집하였다. Follower-following topology 분석에서 저자들은 일반 사회 네트워크에서 나타나는 멱함수가 아닌 팔로워 분포와 짧은 거리의 낮은 reciprocity를 발견했다. 이는 기존에 알려진 인간 소셜 네트워크의 특성과는 다른 경향성을 보임으로 의미가 있다. 또한 서로 맞팔을 맺은 사이에서는 어느정도 homophily를 발견했는데 이는

위치적으로 가깝거나, 유명세에 따른 homophily이다.

저자들은 트위터에서 영향력 있는 사람들을 식별하기 위해 팔로워 수와 페이지랭크별로 사용자를 순위를 매겼고 두 순위가 비슷하다는 것을 발견했다. 만약 리트윗 수로 순위를 매긴다면, 그 순위는 이전의 두 순위와 다르며, 이는 사용자의 팔로워 수와 사용자의 트윗의 인기 간의 차이를 나타낸다. 저자들은 리트윗에 의한 순위는 새로운 관점에서 다른 매체의 영향력을 드러낸다고 주장한다.

마지막으로 저자들은 상위 트렌드 토픽의 트윗을 분석하고 트렌드 토픽의 지속, 확산 시간과 사용자들의 참여에 대해 분석하였다. 그 결과 대부분의 주제(85% 이상)가 헤드라인 또는 지속적인 뉴스임을 밝혀냈고, 이는 트위터가 새로운 정보 전달의 매체로서 역할을 하고 있다는 점을 밝혀냈다. 흥미로운 점은 리트윗을 자세히 살펴보면, 리트윗된 트윗은 원래 트윗의 팔로워 수가 얼마이든 간에 평균 1,000명에 도달하는 것이었는데, 이는 한 번 리트윗된 트윗이 2일, 3일, 4일 홉 떨어진 시점에서 거의 순식간에 리트윗되는 특성과 합쳐져 1차 리트윗 이후 빠른 정보 확산의 원인들을 규명하였다.

9. 가장 인상깊었던 분석 인사이트

1) 개인적으로 3.1 Degree of separation에서 언급했듯이, 통상적으로 사회 현실에서 반영되는 이론에 대한 검증을 했다는 점이 인상깊었다. 현실 인간 네트워크에서는 평균적으로 누구와도 6단계만 거치면 닿을 수 있는데, 트위터에서는 4단계만 거치면 닿을 수 있다는 점이 흥미로웠다. 실제 유명한 예측 및 이론과 반대되는 결과에 따른 현상 분석도 타당하여 재미있었다.

2) 리트윗이라는 것이 정보 확산에 가장 중요한 역할을 하고, 팔로워 수가 얼마이든 간에 한 번 리트윗이 되면 평균 1000명에게 그 트윗이 전달된다는 점을 밝혀낸 것이 인상 깊었다.