트위터에서 팔로워의 행태분석 모델

정광용, 설재욱, 이경순
전북대학교 컴퓨터공학부/영상정보신기술연구센터
e-mail : {kyjeong0520, wodnr754}@naver.com, selfsolee@chonbuk.ac.kr

Modeling Twitter Follower's Behavior Analysis

Kwang-Yong Jeong, Jae-Wook Seol, Kyung-Soon Lee
Dept. of Computer Science&Engineering, Chonbuk National University

요 약

소셜 네트워크 서비스의 하나인 트위터는 팔로우를 통하여 사용자 간의 관계를 맺을 수 있다. 트위터 사용자들은 다양한 팔로워들이 존재한다. 이 팔로워들은 사용자에 대한 호감을 가지고 팔로우 하거나, 맹목적으로 추종하거나, 부정적인 의견을 지니고 사용자의 행동과 글을 관찰하기 위해 팔로우할 수도 있다. 본 논문에서 사용자에게 팔로워들이 어떠한 목적으로 그 사용자를 팔로워의 행태를 분석하는 모델을 제안한다. 대상사용자의 영향력 있는 팔로워를 추출하고, 팔로워의 리트윗 정보, 프로파일, 최신 트윗의 감정분석을 통해 지지자, 중립, 비지지자로 분류한다. 제안 방법의 유효성을 검증하기 위해 트윗 데이터에서 정치인과 언론인 5 명의 팔로워들 중 무작위로 3 만명을 추출하여 실험하였다. 실험 결과 영향력 있는 사용자 추출을 통한 지지 팔로워 추출이 효과적임을 알 수 있다.

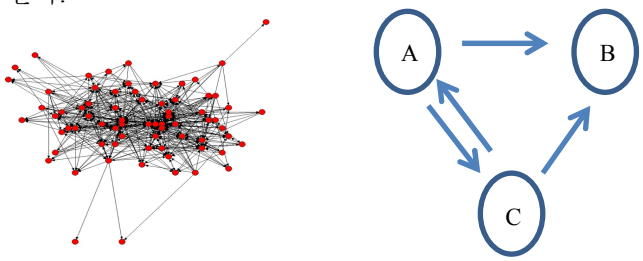
1. 서론

트위터(twitter)는 대표적인 소셜네트워크 서비스(Social Network Service)로 자신의 생각이나 정보 등을 표현한다. 자신과 관계 없는 타인과 정보를 교환할 수 있을 뿐만 아니라 인터넷 카페나 블로그(blog)보다 정보의 확산이 빠르기 때문에 트위터에 대한 연구가 활발하게 진행되고 있다.[1,2,3,4]

트위터를 사용하는데 있어 중요한 요소 중 하나는 팔로우를 하는 것이다. 팔로우를 함으로써 사람들과 정보를 교환할 뿐만 아니라 일상적인 담화를 나누며 소셜 네트워크 서비스의 목표인 지인과의 관계를 강화하고자 한다.

트위터에서 한 사용자가 다른 사용자의 트윗을 구독하는 행위를 "팔로우(follow)"라고 한다.

그림(1)은 팔로우 네트워크 관계를 표현한 그래프이다. A 는 B 를 팔로우 하고 있으며, A 는 B 의 트윗을 구독할 수 있다. 이때 B 는 A 의 "팔로잉(following)"이라 하고, A 는 B 의 "팔로워(follower)"라 한다. A 와 C 는 서로 팔로우하고 있으므로 "맞팔" 관계에 있다고 한다.



그림(1) 트위터에서 팔로우 관계 네트워크

트위터에서 서로간의 팔로우한 사용자를 맞팔 관계라 하는데, 한국인 트위터 네트워크의 맞팔률은 68.2%로 높은 편이다[5].

트위터에서 팔로워의 유형을 분석하는 연구가 있다[6]. 팔로워의 유형으로는 팔로워를 늘리기에만 관심이 있는 그룹, 전형적인 리더그룹으로 트위터 상에서 지식을 공유하고 팔로워들의 글을 관심 깊게 보면서 리트윗하는 그룹 등 다양한 그룹이 존재한다.

또한 해당 텍스트 문서를 작성한 글쓴이의 감정을 파악하는 것이 연구되고 있다[7].

트위터 사용자(target user)는 다양한 팔로워들이 존재한다. 이 팔로워들은 사용자에 대한 호감을 가지고 팔로우하거나, 맹목적으로 추종하거나, 부정적인 의견을 지니고 사용자의 행동과 글을 관찰하기 위해 팔로우를 한다. 그렇기 때문에 사용자에게 팔로워들이 어떠한 목적으로 팔로우 했는지를 보여주면 유용하다.

트위터 팔로워들의 행동에 대한 관찰을 통해 다음과 같이 4 가지의 특성을 이용하였다. 1) 사회적으로 영향력 있는 사용자들은 팔로워들 중 사용자를 적극 지지하는 팔로워들이 존재한다. 이들은 대상 사용자와 같은 그룹으로 볼 수 있다. 2) 자신의 타임라인에 올라온 트윗 중에 공감을 표현하고, 전달하기 위해 트윗을 리트윗(retweet) 하는 경향이 있다. 따라서 리트윗이 많이 된 팔로워는 영향력이 있다고 볼 수 있다. 3) 팔로워가 많으면 그 트윗을 보는 사용자가 많다는 것이기 때문에 리트윗이 많고 어떠한 사건을 언급하였을 때 파급력이 크다고 할 수 있다. 4) 프로파일은 주로 트위터 사용자들의 관심도나 선호도를 나

타내는 경향이 있다. 이는 프로파일만을 가지고 사용자의 성향을 파악할 수 있다.

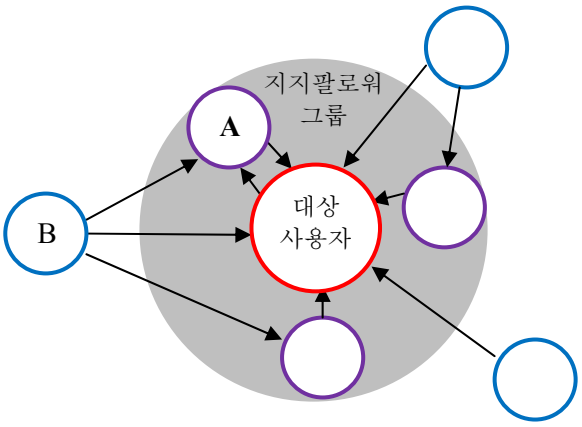
본 논문에서는 팔로워 수와 대상사용자가 언급된 트윗이 리트윗이 많이 된 것을 통해 영향력 있는 지지 팔로워를 추출하고 팔로워의 리트윗 정보, 프로파일, 최신 트윗의 감정분석을 통해 지지자, 중립, 비지지자로 분류하는 방법을 제안한다.

2. 팔로워의 행태 분석 모델

사용자의 행태분석을 위해 사용자 프로파일, 팔로우 관계, 리트윗 정보, 최근 트윗 정보를 감정분석해서 지지자, 중립, 비지지자로 분류한다.

2.1 영향력 있는 지지 팔로워 추출

영향력 있는 지지 팔로워는 대상사용자를 언급한 트윗을 많이 쓰고, 이 트윗들이 많이 리트윗 되고, 팔로워 수가 많은 사용자를 말한다. 여기서 언급한 트윗의 정의는 다른 사용자의 트윗을 리트윗한 트윗이 아닌 사용자가 직접 작성한 글 중 대상사용자를 언급한 트윗을 말한다. 여기서는 *SPFollower(Strong Positive Follower)*로 표기한다.



그림(2) 영향력 있는 지지팔로워 그룹을 반영한 네트워크

그림(2)에서 A 는 *SPFollower* 이고, B 는 대상사용자와 *SPFollower* 를 함께 팔로잉하고 있는 사용자이다.

대상 사용자의 팔로워 무작위 3 만명에 대해서 팔로워 순으로 정렬하고 정렬된 팔로워 상위 1 천명에 대해 **평균 리트윗 수**(*argRT*)과 대상사용자를 **언급한 비율**(*Umention*)을 이용하여 사용자의 **영향력 정도**(*UInfluence*)를 측정한다.

팔로워가 대상사용자를 언급한 트윗이 리트윗된 평균을 구하는 수식은 다음과 같다.

$$avgRT(u, Target) = \sum_{i=1}^{N} RT(u, Target\,Tweet_i) / N \qquad (1)$$

여기서 Target 은 대상사용자, u 는 Target 의 팔로워이고, N 은 팔로워 u 의 트윗 중 Target 을 언급한 트윗 수, $TargetTweet_i$ 는 팔로워 u 의 트윗 중 Target 을 언급한 트윗이다. 리트윗이 많이 되면 파급력이 커진다.!!!

대상사용자의 팔로워의 전체 트윗 중 대상사용자를 언급한 트윗의 비율을 구하는 수식은 다음과 같다.

$$UMention(u, Target) = \frac{Tweet(u, Taget\,Tweet)}{Tweet(u)} \qquad (2)$$

여기서 *Tweet(u)*는 팔로워 u 의 전체 트윗을 말하며, *Tweet(u,TargetTweet)* 은 팔로워 u 의 트윗 중 Target 을 언급한 트윗 수를 말한다. 대상사용자를 언급한 비율이 많다는 것은 대상사용자에 대해서 관심도가 크다는 말이다.

.

사용자의 영향력 정도를 구하는 수식은 다음과 같다.

$$UInfluence(u) = avgRT(u, Target) \times UMention(u, Target) \qquad (3)$$

여기서 영향력 정도는 수식 (1)과 수식 (2)를 곱한 값으로 팔로워 u 가 쓴 트윗 중 Target 을 언급한 트윗의 리트윗의 평균값과 사용자 u 가 쓴 트윗 중 Target 을 언급한 트윗의 비율을 곱한 값이다.

영향력 정도가 0.1 이상일 때 팔로워는 모두 지지 팔로워였기 때문에 영향력 정도의 임계치(threshold)를 0.1 로 정하여 영향력 정도가 0.1 이상인 사용자만 *41 'Â¸Â¸ ½* 로 선택하였다.

영향력 정도가 높은 사용자 중에서 대상 사용자를 리트윗 하지 않은 팔로워는 지지자로 보기 어렵다. 따라서 대상사용자의 트윗을 한 번도 리트윗하지 않았다면 *SPFollower* 로 분류하지 않는다. <표 1>은 문재인 팔로워 무작위 3 만명에 대해서 팔로워 순 상위 1 천명의 팔로워에 대해 영향력 정도를 기준으로 순위화된 팔로워를 보여준다. 여기서 문재인의 트윗을 리트윗 한 횟수가 0 인 "변희재"와 "오마이 뉴스"는 *SPFollowers* 로 분류하지 않는다.

<표 1> 문재인 팔로워의 영향력 정도 순위와 지지 분류 예

순위	팔로워 이름	문재인의 트윗을 리트윗 한 횟수	영향력 정도	분류
1	변희재	0	22.60	모름
2	문재인 24 시	18	10.56	SP Follower
3	조국	6	7.46	SP Follower
4	문재인 TV	14	5.65	SP Follower
5	레인메이커	3	5.33	SP Follower
6	우금	2	4.35	SP Follower
7	김용민	3	3.29	SP Follower
8	젠틀재인	215	3.08	SP Follower
12	오마이뉴스	0	1.29	모름

2.2 리트윗 정보 및 프로파일과 지지 팔로워를 나타내는 어휘를 이용한 지지 분류

(1) 리트윗은 공감을 나타내므로 2.1 의 과정을 통해

추출한 *41 'Â¸Â¸ ½* 가 작성한 트윗 중 대상사용자를 언급한 트윗을 리트윗한다면 지지 팔로워로 분류한다.

(2) 트위터에서 사용자들이 프로파일에 관심, 선호, 취미 등을 키워드 형식으로 표현하는 경향이 있다. 그래서 프로파일에서 대상 사용자가 언급이 되는 경우에는 지지 팔로워로 분류한다.

(3) 팔로워는 정치인이나 연예인들을 지지하는 표현을 쓴다. 대상 사용자의 이름과 함께 <표 3>과 같은 지지하는 단어로 나타낼 경우 지지 팔로워로 분류한다.

<표 3> 지지 팔로워를 나타내는 어휘

"지지", "후원", "필승", "존경", "성공", "파이팅", "만세", "파이팅", "승리", "희망"

2.3 팔로워의 최신 트윗에 대한 감정 분석을 통한 지지, 비지지 분류

대상 사용자에 대한 오래 전의 감정보다 최신 글을 통하여 최신 감정을 알아보고자 하기 때문에 팔로워의 최신 600 개의 트윗을 이용하여 내용 분석한다.

트윗에서 RT 는 리트윗과 달리 전달하고자 하는 트윗 앞에 의견을 덧붙일 수 있다. RT 앞에 의견이 있다면 내용 분석을 통하여 리트윗 한 트윗이 긍정이면 +1, 부정이면 -1 의 가중치를 부여한다. RT 분석에 적용되지 않은 트윗은 대상 언급을 한 트윗의 감정을 정치 분야 감정 단어 사전을 기반으로 지지이면 +1, 비지지이면 -1 의 가중치를 준다. 트윗의 가중치 합이 -2 ~ 2 이면 중립, 3 이상이면 지지, -3 이하이면 비지지로 분류한다. 정치분야 감정 단어 사전 구축은 2012 년 9 월 1 일부터 21 일까지의 데이터 중 키워드(새누리당, 민주통합당, 문재인, 박근혜, 정치)가 들어간 모든 트윗을 수집하였다. 수집한 트윗에서 키워드를 기준으로 윈도우사이즈 3 으로 하여 포함하는 모든 단어(약 15 만개)를 빈도수로 정렬하였다. 정렬한 단어 상위 20,000 위 내 감정 단어를 추출하였다. 긍정 단어와 부정 단어의 개수는 <표 4>와 같다.

<표 4> 감정 단어 사전

감정	단어 예	개수
긍정	지지, 중요, 환영 …	589
부정	부패, 분노, 욕설 …	782

3. 실험 및 결과

3.1 트윗 실험 집합

제안한 방법의 유효성을 검증하기 위해 수집 날짜는 2012 년 8 월 1 일부터 30 일까지 대상사용자 5 명(문재인, 박근혜, 유시민, 정봉주, 조갑제)을 선택하고 트위터 API[8]와 twitter4j[9]를 이용하여 각 대상사용자의 팔로워 중 무작위로 최대 3 만명의 사용자 기본 정보를 수집하였다. 수집된 사용자들을 팔로워 순으로 정렬하여 상위 1 천명의 팔로워의 트윗을 최근 작성된 트윗을 최대 600 개 수집하였다. 실험 대상은 각 대상사용자의 팔로워 200 명으로 한다. <표 5>

<표 5> 트위터 실험 데이터

대상 사용자	총 팔로워 수	무작위 추출 최대 3 만명 중 상위 팔로워순으로 선택된 팔로워 수	선택된 팔로워의 총트윗 수
문재인	261,105	1,000	418,626
박근혜	223,070	1,000	381,954
정봉주	394,317	1,000	324,315
유시민	510,351	1,000	352,516
조갑제	17,272	1,000	423,591
	1,406,115	5,000	1,901,002

3.2 실험 결과

<표 6> 팔로워 200 명 분류 결과

	문재인	박근혜	유시민	정봉주	조갑제	평균
정확률	85.3%	75.8%	68.1%	69.6%	60.0%	71.7%
재현율	60.0%	66.8%	59.6%	79.3%	56.6%	63.9%
F_1 score	70.8%	71.0%	62.0%	74.1%	58.2%	67.2%
정확도	73.0%	74.0%	70.0%	79.0%	73.5%	**73.9%**

대상사용자 5 명의 각 팔로워 200 명에 대해 지지, 중립, 비지지로 분류한 결과는 <표 6>과 같다. 대상사용자 5 명의 팔로워를 지지, 중립, 비지지로 분류하는데 정확도가 73.9% 의 성능을 보여주고 있다.

문재인 팔로워 200 명 분류 결과는 <표 7>과 같다. 정답 집합은 대상사용자의 비지지팔로워는 적게 나왔고, 중립팔로워가 많이 나왔다.

<표 7> 문재인 팔로워 200 명 분류 결과

	정답집합	정확률	재현율
긍정	51 명	80.7%	57.5%
중립	94 명	75.3%	90.9%
부정	1 명	100%	33.3%

3.3 결과 분석

<표 8> ***41 'Â¸Â¸ ½ ,*** **를 통한 지지분류에 대한 평가**

대상 사용자	***! 41 'Â¸Â¸ ½ , 수 !***	***41 'Â¸Â¸ ½*** **의 트윗을 리트윗한 팔로워**	**대상 사용자 트윗을 리트윗한 팔로워**
문재인	15	25	12
박근혜	39	49	24
유시민	6	4	1
정봉주	6	29	19
조갑제	4	9	2

<표 8>은 *SPFollower* 를 고려하였을 때와 대상 사용자만을 고려하였을 때 리트윗한 수의 차이이다. 대상사용자만을 고려했을 때보다 *SPFollower* 를 고려하였을 때 지지팔로워를 분류하는데 재현율이 향상되었다.

실험에서 문재인의 지지팔로워를 분류하는데 *SPFollower* 만을 이용하여 분류하였을 때 전체 재현율이 23%에서 49%로 향상을 보였고, 여기서 *SPFollower* 를 이용한 긍정분류는 98% 의 높은 정확률을 보였다.

<표 9> 프로파일 분석을 통한 지지분류에 대한 평가

대상 사용자	프로파일에 대상사용자를 언급한 팔로워 수	프로파일에 대상사용자를 언급한 팔로워 중 대상사용자를 지지하는 팔로워	정확률
문재인	3/200	3/3	100%
박근혜	4/200	3/4	75%
유시민	1/200	1/1	100%
정봉주	4/200	4/4	100%
조갑제	0/200	-	-

프로파일에서 분석을 통한 지지분류에 대한 평가는 <표 9>와 같다. 프로파일에 대상사용자를 언급한 팔로워의 경우 대상사용자를 지지하는 팔로워로 나타났다. 실험에서 프로파일에 대상사용자를 언급한 팔로워 12 명 중 11 명이 대상사용자를 지지하는 팔로워였다. 여기서 조갑제의 팔로워 중 프로파일에서 조갑제를 언급한 팔로워는 존재하지 않았다.

4. 결론 및 향후 연구

본 논문에서는 트위터 사용자의 팔로워를 *SPFollower* 이용한 대상 사용자의 확장된 개념과 사용자 프로파일, 최신 트윗, 리트윗 정보를 이용하여 팔로워의 행태를 분석하는 모델을 제안하였다.
대상사용자 5 명에 대한 팔로워 트윗 실험 집합에서 영향력 있는 사용자 추출을 통한 지지 팔로워 추출이 효과적임을 알 수 있다.

향후 연구에서는 팔로워가 어떤 유형으로 트윗을 게시하는지 활동 유형을 분류하는 방법으로 확장할 계획이다.

참고문헌

[1] Zhiheng Xu, Yang Zhang, Yao Wu and Qing Yang "Modeling User Posting Behavior on Social Media", SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and de development in information retrieval , pp. 545-554, 2012.

[2] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is Twitter, a Social Network or a News Media?", www '10 Proceedings of the 19th international conference on World wide web pp.591-600, 2010.

[3] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, and Zhaohui Zheng, "Time is of the Essence: Improving Recency Ranking Using Twitter Data," www '10 Proceedings of the 19th international conference on, pp.331-340, 2010.

[4] Jagan Sankaranarayanany, Hanan Samety, Benjamin E. Teitlery, Michael D. Liebermany, and Jon Sperlingz, "TwitterStand: News in Tweets," in Proc. of the 17th ACM SIGSPATIAL international conference on Advances in Geographic Information Systems (GIS '09), pp.42-51, 2009.

[5] 장덕진, "트위터 공간의 한국정치-정치인의 네트워크와 유권자 네트워크", 언론정보연구 제 48 권 제 2 호, pp.80-107, 2011.

[6] 이옥기, "트위터 리더와 팔로우 유형과 특성에 대한 사용자 인식에 관한 연구", 한국소통학회 2011 년 춘계 정기학술대회, pp.9-14, 2011.

[7] 황재원, "감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템", 한국정보과학회, 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제14 권 3 호, pp.336-340, 2008.

[8] https://dev.twitter.com/

[9] http://twitter4j.org/