트위터 자료의 시간별 분석과 감성 자질을 이용한 핵심 사건 추출

김희환, 촐몽 바야르, 이경순
전북대학교 컴퓨터공학과
e-mail : gabeliel11@naver.com, selfsolee@jbnu.ac.kr

Extracting Core Event Feature Based on Timeline Analysis and Sentiment Feature in Twitter Corpus

Hui-Hwan Kim, Bayar Tsolmon, Kyung-Soon Lee
Dept. of Computer Engineering, Chonbuk National University

요 약

트위터 사용자들은 어떠한 이슈에 대해 트위터를 통해 빠르고 간결하게 다른 사람들과의 지속적인 커뮤니케이션을 원하고, 이러한 특징은 이슈 별 사건에 따라 트윗 개수에 영향을 미치게 된다. 만약 어느 하나의 사회적 이슈에 대해 어떠한 사건이 일어나게 되면 그때의 트윗 개수는 폭발적으로 증가하게 된다. 본 논문에서는 이러한 특징을 이용하여 트위터 자료를 시간별로 분석하여 사건을 인식하고, 감성 자질과 카이제곱 값을 이용해 해당 날짜에 대한 핵심 사건을 추출한다.

1. 서론

사람들이 자신의 의견, 생각, 경험을 서로 공유하기 위해 사용하는 블로그, 미니홈피, 메신저 등을 소셜 네트워크 서비스(Social Network Service ; SNS)라 한다. 트위터(twitter)는 블로그의 인터페이스에 미니홈피의 인적 네트워크 형성, 메신저의 신속성을 한데 모아놓은 소셜 네트워크 서비스라고 볼 수 있다. 하나의 트윗(tweet)을 작성시 트위터는 140자 이내 단문으로 한정 지어놓아 짧은 문장 내에 자신의 의견이나 생각을 포함하도록 유도하고 있다. 더구나 스마트 폰의 빠른 보급화로 인해 트위터의 사용자가 급증하면서 트위터는 기존의 언론 미디어보다 더 빠르게 정보를 파급시키는 효과 또한 가지고 있다. 실례로 뉴욕 허드슨강 여객기 불시착 사건, 강남 파이낸스센터 화재사건 등은 트위터가 언론보다 더 빠르고 정확하게 정보를 전달한 사례이다.

이러한 트위터의 특성 들로 인해 학계의 관심이 증가하고 있고, 트위터에 대한 연구 논문의 수도 점점 증가하고 있는 추세이다. Popescu[1]는 트위터에서 논란이 되는 이슈를 발견하기 위해서 3-회귀 기계 학습 모델(3-regression machine learning models)을 사용하였고, Park[2]은 오피니언 마이닝(opinion mining)을 위해 트위터 데이터 셋을 자동으로 구축하는 방법을 소개하고, 구축된 데이터 셋을 이용하여 긍정과 부정을 분류하는 분류기를 제안하였다. 또한 Sayyadi[4]는 사건을 인식하기 위해 키워드(keyword) 그래프를 사용했고, Zhao[5]는 사회적인 이슈와 특별한 토픽(topic)사이의 관계에서 이벤트를 인식하였다. 박지혜[6]는 Cytoscape 플랫폼을 사용하여 트위터 사용자들의 관계를 시각적으로 표현 할 수 있는 시스템을 개발하였으며, 성병기[7]는 이슈 키워드 추출 및 트위터와 유튜브에 기반한 실시간 검색 시스템을 구현하였다. 이 시스템은 최근의 신문 기사들의 제목과 스니핏을 이용하여 이슈가 되는 키워드를 실시간으로 추출한 뒤 사용자들에게 제공해주고, 유투브와 트위터의 OpenApi 를 이용해 추출된 키워드에 대한 컨텐츠들을 사용자들에게 실시간으로 제공해준다.

트위터 사용자들은 자신이 관심 있어 하는 분야 혹은 이슈(issue)에 대해 지속적으로 자신의 의견을 남기고 자신과 공통 분야에 관심이 있는 사람들과 서로 소통하기를 원한다. 이러한 특성으로 인해 하나의 사회적 이슈에 대해 트위터 데이터를 시간별로 분석해 보면 그 이슈에 대한 어떠한 사건(event)이 일어나지 않았을 때는 트윗 개수가 어느 수준 이하의 수를 유지하다가, 그 이슈에 특정 사건이 일어났을 때는 그 날의 트윗 개수가 폭발적으로 증가하는 것을 관찰 할 수가 있다.

본 논문은 이러한 트위터의 특성에 초점을 맞추어 몇 개의 이슈들을 대상으로 트위터 데이터를 수집한 뒤, 이를 시간별로 분석하여 각각의 이슈에 사건이 발생함을 인식하고 보다 효과적인 자질 추출을 위해 단어 빈도수와 함께 감성 자질 및 카이제곱(Chi Square)값을 사용하여 그 사건에 대한 핵심 사건을 추출하였다.

본 논문의 구성은 다음과 같다. 2장에서는 트위터 데이터를 시간별로 분석하는 과정을 소개하고, 3장에서는 분석된 데이터를 이용하여 핵심 사건을 추출하는 방법을 제안한다. 4장에서는 실험 및 분석을, 5장에서는 결론 및 향후 연구에 대해 논하겠다.

2. 트위터 데이터의 시간별 분석을 통한 사건 인식

어떠한 이슈에 대해 사람들이 트윗을 작성한다고 생각해 보자. 그 이슈에 아무런 사건이 일어나지 않았을 때는 사람들은 그와 관련된 트윗을 작성하는 횟수가 적다. 하지만 그 이슈에 어떠한 사건이 일어났을 때, 특히 사회적인 이슈로 발전되었을 때는 사람들의 그에 대한 관심이 폭발적으로 증가하게 되고 그 결과는 트윗 개수의 급격한 증가로 나타나게 된다.

(그림 1)은 천안함을 질의어로 하여 2010년 10월 1일부터 2011년 3월 26일까지 트위터 내에서 검색된 자료를 수집하여 시간별로 트윗 개수를 그래프화한 자료이다.

천안함에 대한 트윗 중 특정한 사건이 일어나지 않은 대부분의 날에는 트윗 개수가 일정수준 이하로 나타나는 것을 확인 할 수가 있다. 하지만 2010년 11월 17일과 같이 천안함에 대해 어떠한 사건이 일어나게 되면 트윗 개수가 급격히 증가되는 것 또한 확인 할 수가 있다. 실제로 트윗 개수가 전날 대비 급격히 증가한 2010년 11월 17일과, 같은 달 23일, 2011년 3월 21일은 각각 천안함에 대한 추적 60분 방영, 연평도 포격 사건, 천안함 1주년이라는 사건이 발생했다.

3. 감성 자질과 시간별 분석을 통한 핵심 사건 추출

기본 자질 추출을 위해 수집된 트위터 데이터를 형태소 분석기를 통해 형태소 분석을 한 뒤, 불용어(Stop-Words)와 불필요한 URL 정보를 제거 하였다. 불용어의 제거는 네이버 실시간 검색어에서 무작위로 100개의 질의어를 추출, 트위터를 통해 검색하여 각 최대 100개의 트윗을 수집하였다. 수집된 트위터 데이터를 형태소 분석 후 단어 빈도수를 계산하여 총 202개의 불용어 리스트를 만들었다.

그 후 정제된 자질들을 하나의 트윗 내에서 바이그램(Bigram)으로 추출하였다. 이때의 윈도우 사이즈는 3 이다.

3.1 단어 빈도수를 이용한 시간별 기본 자질 추출

바이그램으로 추출된 자질들을 각 시간별로 단어 빈도수를 계산하여 단어 빈도수가 높은 순으로 순위화한 후 기본 자질들을 추출한다. 단어 빈도수로 추출된 기본 자질들의 단어 빈도수 값을 Freq(w, t0)라 하겠다. 여기서 w는 자질을 의미하고, t0는 시간을, D는 트윗 문서를 나타낸다.

$$Freq(w,t0) = \sum_{D \in t0} tf(w,D) \qquad (1)$$

3.2 감성 자질을 이용한 시간별 자질 추출

수집된 트위터 데이터를 관찰해보면 어느 한 이슈에 대해 특정 사건이 발생하게 되면 그 사건에 해당하는 핵심 사건과 함께 감성 자질이 함께 출현하는 경우를 <예 1>과 같이 자주 볼 수가 있다.

자, 생각해보자. ***연평도 포격*** 사건이 **충격**적인 상황에서 천안함, 사대강, 현대자동차, 민간인 사찰 이야기를 멈춰야하는 것은 아니다.

<예 1> 핵심 사건과 함께 감성 자질이 출현한 예

<예 1>에서는 천안함이라는 이슈에 대한 트윗 중에서 '연평도 포격'이라는 사건이 발생했을 때 '충격'과 같은 감성 자질이 함께 출현한 것을 보여준다.

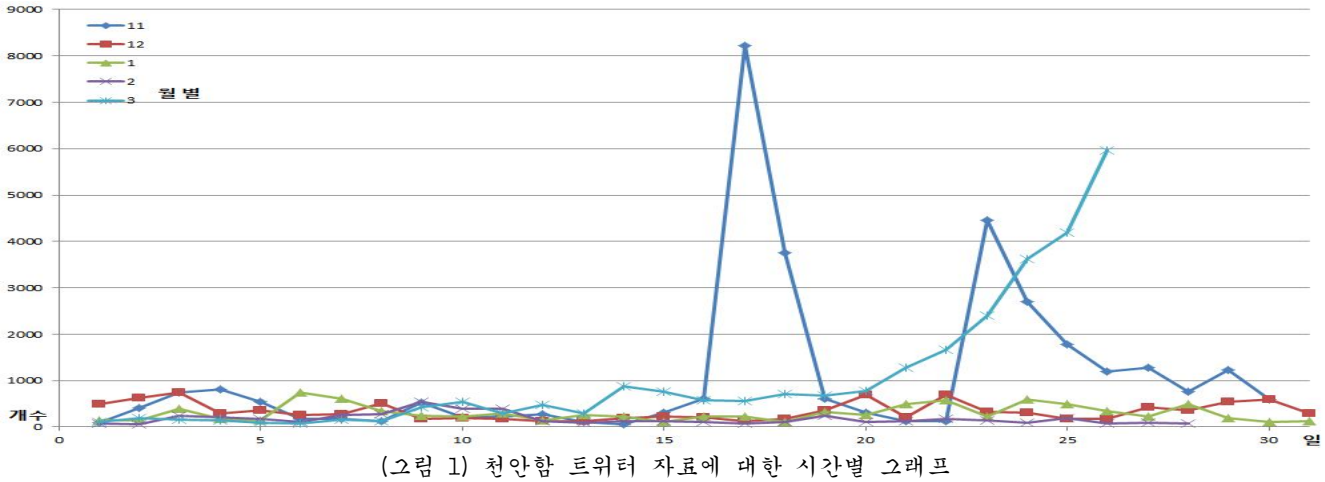
이 정보를 활용하여 기본 자질로 추출된 자질들 중 상위 50개를 대상으로 해당 이슈의 모든 트위터 데이터 내에서 감성 자질과 함께 출현한 트윗의 빈도수를 구하였다. 감성 자질을 사용한 수식은 다음과 같다.

$$OpScore(w,t0) = Freq(w,t0) + 2 \cdot opFreq(w,t0) \qquad (2)$$

여기서 opFreq(w, t0)는 전체 트윗 문서 집합에서 자질 w가 감성 자질과 함께 출현한 트윗의 빈도수를 의미한다. 사용된 감성 자질 사전은 Wilson lexicon[3]에서 강한 감성 자질만을 추출하여 구글 번역기를 통해 번역한 뒤, 사람이 직접 판단하여 감성자질 사전을 구축하였다.

3.3 카이제곱을 이용한 시간별 자질 추출

시간 별로 트위터 데이터를 분석한 결과 어떠한 이슈에 대해 특정 사건이 발생 했을 경우 그 시간의 트



(그림 1) 천안함 트위터 자료에 대한 시간별 그래프

윗 개수는 증가하게 되고, 트윗 개수가 전날에 대비 급격하게 증가한 경우에는 어떠한 사건이 발생했다고 짐작 할 수 있다. 여기에서 만약 발생한 사건이 그 이전에는 발생하지 않은 새로운 사건이라면 혹은 발생한 적이 거의 없는 사건이라면 사건에 대한 Freq(w, t0)가 그 이전날의 데이터보다 폭발적으로 증가됨을 알 수 있다. 본 논문에서는 시간 t0에서 자질 w의 중요도를 계산하기 위해 카이제곱 값을 이용하였다. <표 1>은 카이제곱 값을 계산하기 위한 분할표이다.

<표 1> 카이제곱값을 계산하기 위한 분할표

	자질 w 가 포함되어 있는 트윗	자질 w 가 포함되 어있지 않은 트윗
시간 t0 의 트윗들	a	b
시간 t0 이전 시간 트윗들	c	d

시간 t0에서의 카이제곱 값의 계산 수식은 다음과 같다.

$$ChiSquare(w,t0) = \frac{(a+b+c+d)(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (3)$$

각 자질 별 ChiSquare(w, t0) 값에 핵심 사건이 감성 자질과 함께 출현하는 경우가 많다는 정보를 활용한 OpScore(w, t0) 값을 이용한 수식은 다음과 같다.

$$ChiOpScore(w,t0) = \lambda \cdot ChiSquare(w,t0) + (1-\lambda) \cdot OpScore(w,t0) \quad (4)$$

여기서 λ는 카이제곱 값에 가중치를 더 주기 위한 파라미터로 본 논문에서는 0.7로 실험하였다.

4. 실험 및 분석

본 논문에서 제안한 방법을 실험 하기 위해 천안함, 김연아, 박지성, 지진 총 4가지의 이슈에 대해 트위터 자료를 2010년 11월 1일부터 2011년3월 26일까지 <표 2>와 같이 수집하였다.

<표 2> 트위터 실험 데이터

	천안함	김연아	박지성	지진
개수	84195	26844	131533	46795

<표 3> 이슈의 대표적인 사건 리스트

이슈	날짜	발생한 사건	사건 번호	정답 자질 개수
천안함	2010.11.17	추적 60 분 방영	E1	2
	2010.11.23	연평도 포격 사건	E2	3
	2011.03.21	천안함 1 주년	E3	4
김연아	2010.11.30	프로그램 발표	E4	4
	2010.12.02	유니세프 친선대사	E5	3
	2011.01.28	김연아 악마가면	E6	3
박지성	2010.11.07	박지성 2 골	E7	5
	2011.01.31	박지성 은퇴	E8	7
	2011.02.01	차범근 고백	E9	2
지진	2010.11.30	일본지진 발생	E10	3
	2011.03.09	일본 쓰나미 경보	E11	3
	2011.03.11	일본 대지진 발생	E12	1

<표 3>은 각 이슈에 대해 발생한 대표적인 사건들과 핵심 사건 자질의 상위 10에서의 정답 개수를 정리한 표이다. 각 사건에 대한 정답은 사람이 직접 판별 하였다.

수집된 각 이슈의 데이터를 시간별로 기본 자질들을 추출하고 그 자질들에 대해 단어 빈도수를 구한다음 단어 빈도수에 대해 각 자질들을 순위화 하였다. 추출된 자질들 중 상위 50개의 자질들을 따로 추출하여 감성자질을 이용한 OpScore 값으로 재순위화 하였고, 또 다른 방법으로 카이제곱만을 이용하여 재 순위화 하였다. 마지막에는 OpScore 값과 카이제곱 값을 함께 이용한 ChiOpScore 값으로 재순위화 하였다.

아래의 <표 4>는 천안함의 연평도 포격 사건이 일어난 날짜의 트위터 데이터에 대해 각 단계별로 추출된 핵심 사건들을 상위 10 까지 순위화한 것이다.

천안함 이슈의 연평도 포격 사건에 대해 실험한 결과 단순히 단어 빈도수 로만 계산 하였을 경우에는 실제적인 연평도 포격이라는 핵심 사건이 비교적 낮은 수준에 랭크 되어있지만, 감성 자질과 함께 순위화한 경우에는 새로운 핵심 사건인 “연평도 사건”이 추가 되면서 “연평도 포격”은 랭크가 2단계 상승했음을 알 수가 있다. 카이제곱으로만 순위화 했을 경우

<표 4> 각 방법을 이용해 추출한 상위 10 자질

	Freq	OpScore	ChiSquare	ChiOpScore
1	천안함 사건	천안함 사건	천안함 연평도	천안함 연평도
2	천안함 연평도	천안함 침몰	대포폰 사찰	천안함 사태
3	천안함 북한	천안함 사태	***연평도 포격***	***연평도 사건***
4	천안함 사태	***연평도 사건***	훈련 북한	***연평도 포격***
5	민간인 사찰	북한 소행	북한 도발	민간인 사찰
6	북한 도발	천안함 연평도	사찰 천안함	대포폰 사찰
7	대포폰 사찰	***연평도 포격***	***연평도 사건***	천안함 사건
8	사찰 천안함	천안함 북한	사건 천안함	천안함 북한
9	***연평도 포격***	민간인 사찰	민간인 사찰	북한 도발
10	사건 천안함	북한 천안함	대포폰 민간인	훈련 북한

에도 감성 자질과 함께 사용해서 순위화한 결과와 비슷한 결과를 보여주고 있다. 그리고 마지막 단계인 감성 자질과 카이제곱 값을 함께 사용한 ChiOpScore 값으로 순위화한 결과가 가장 좋게 나오는 것을 확인할 수가 있다. 하지만 4가지 이슈의 트위터 데이터에서 천안함의 연평도 포격과 지진의 일본 지진 발생이라는 사건 외 몇 가지 경우를 제외하고는 실험 방법에 상관없이 핵심 사건 자질들이 높은 랭크에 위치해 있는 것을 확인 할 수가 있었다. 이는 어느 이슈에 대해 사건이 발생하면 사람들은 그 이슈와 함께 사건에 대해서도 언급을 하기 때문에 트윗의 개수가 증가할수록 핵심 사건 자질들의 단어 빈도수도 증가하게 된다. 따라서 감성 자질이나 카이제곱 값은 영향을 미치치 않을 정도로 단어 빈도수 값이 커지기 때문에 단순한 단어 빈도수로도 핵심 사건 자질들을 효과적으로 추출 할 수가 있었던 것이다.

아래의 <표 5>는 <표 3>에서 제시한 사건 리스트들을 대상으로 실험한 뒤 상위 10 안에 핵심 사건으로 볼 수 있는 자질들이 정답을 얼마나 포함하고 있는지를 나타내고 있다. 즉, 1/3은 3개의 정답 자질중에 1개를 포함한 것을 나타낸다. E는 <표 3>의 사건번호이다.

<표 5> 비교 실험 결과 (상위 10에서의 정답 포함율)

사건 번호	Freq	OpScore	ChiSquare	ChiOpScore
E1	1/2	1/2	2/2	2/2
E2	2/3	2/3	2/3	3/3
E3	3/4	4/4	3/4	4/4
E4	2/4	3/4	3/4	4/4
E5	1/3	1/3	2/3	3/3
E6	2/3	3/3	0/3	0/3
E7	3/5	4/5	5/5	5/5
E8	5/7	5/7	7/7	6/7
E9	2/2	0/2	2/2	2/2
E10	3/3	3/3	3/3	3/3
E11	2/3	1/3	3/3	3/3
E12	1/1	1/1	1/1	1/1
평균	**70.0%**	**71.6%**	**82.0%**	**90.5%**

실험 결과에서 E1, E2, E4, E5, E7를 포함한 대부분의 경우 단어 빈도수만을 이용한 방법보다 감성 자질이나 카이제곱, 또는 그 둘을 함께 이용했을 경우에 성능이 좋았다. 특히, 감성 자질과 카이제곱 값을 함께 이용한 방법의 평균값은 90.5%로 다른 방법들에 의한 결과보다 월등히 높은 것을 알 수 있다. 하지만 E6의 경우 카이제곱을 이용한 방법이나 감성 자질과 카이제곱을 함께 이용한 방법이 상위 10에는 들지 못하고 상위 50개안에만 드는 것을 확인했다.

5. 결론 및 향후 연구

본 논문에서는 어느 특정 이슈에 관한 트위터 데이터가 그 이슈에 사건이 발생 했을 경우 트윗의 개수에 영향을 미친다는 정보를 이용하여, 총 4 가지의 이슈에 대해 트위터 데이터를 모은 다음 시간별로 분석하여 핵심 사건을 추출하는데 목적을 두었다. 핵심 사건을 추출 하는 방법으로는 단어 빈도수 만을 이용한 방법, 단어 빈도수와 감성 자질을 함께 이용한 방법, 카이제곱만을 이용한 방법, 단어 빈도수와 감성 자질과 카이제곱을 함께 이용한 방법으로 비교실험을 하였다.

그 결과 4가지 이슈에 대해 각 사건별로 상위 10의 정답 포함율을 따져 보았을 때, 단순한 단어 빈도수로 추출 했을 경우보다 본 논문에서 제시한 방법들이 상위 10 안에 보다 많은 핵심 사건 자질들이 포함 되어지는 것을 확인 할 수가 있었다.

향후 연구에서는 트위터 데이터를 통합하여 특정 자질의 단어 빈도수 값에 대한 의존도를 줄이고, 기본 자질을 추출 할 때에도 바이그램 보다 트라이그램(trigram)을 사용하여 보다 효과적인 핵심 사건 자질들을 추출하도록 연구를 진행할 계획이다. 또한, 본 논문에서는 사용하지 않았지만 특정 트윗의 리트윗(retweet) 정보와 댓글 정보 또한 이용할 계획이다.

감사의 글

본 논문은 한국전자통신연구원 지식경제 기술혁신 사업의 위탁연구과제로 수행한 연구 결과입니다.

참고문헌

[1] A.-M. Popescu and M. Pennacchiotti, “Detecting Controversial Events from Twitter.” In Proceedings *of CIKM*, 2010.

[2] A. Park, P. Paroubek, “Twitter as a Corpus for Sentiment AnalysE and Opinion Mining.” In Proceedings *of LREC*, 2010

[3] T. Wilson, J. Wiebe, and P. Hoffmann. “Recognizing contextual polarity in phrase-level sentiment analysE.” In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 347–354. Association for Computational LinguEtics, 2005.

[4] H.Sayyadi, M. Hurst, and A. Maykov. “Event Detection and Tracking in Social Streams.” In Proceedings of ICWSM, 2009

[5] Q. Zhao, P.Mitra, and B. Chen. “Temporal and information flow based event detection from social text streams.” In Proceedings of WWW,2007

[6] 박지혜, 김보현, 이명준, 권영근, “TwitNet : 트위터 사용자들의 관계를 시각적으로 나타내는 Cytoscape 플러그인 개발” , 한국정보과학회 2010, 한국컴퓨터 종합 학술발표논문집, 제 37 권 제 1 호.

[7] 성병기, 오진영, 차정원, “LiveTwitter: 트위터 기반 핫이슈 검색 시스템”, 한글 및 한국어 정보처리학회, 제 22 회 한글 및 한국어 정보처리 학술대회 발표논문, pp. 179-182, 2010.