ISSN(Print): 2287-5905



http://dx.doi.org/10.3745/KTSDE.2012.1.1.069

# Extracting Core Events Based on Timeline and Retweet Analysis in Twitter Corpus

Bayar Tsolmon† · Kyung-Soon Lee††


## ABSTRACT

Many internet users attempt to focus on the issues which have posted on social network services in a very short time. When some social big issue or event occurred, it will affect the number of comments and retweet on that day in twitter. In this paper, we propose the method of extracting core events based on timeline analysis, sentiment feature and retweet information in twitter data. To validate our method, we have compared the methods using only the frequency of words, word frequency with sentiment analysis, using only chi-square method and using sentiment analysis with chi-square method. For justification of the proposed approach, we have evaluated accuracy of correct answers in top 10 results. The proposed method achieved 94.9% performance. The experimental results show that the proposed method is effective for extracting core events in twitter corpus.


**Keywords** : Event Extraction, Timeline Analysis, Retweet, Sentiment Feature, Chi-Square, Twitter Corpus

## 1. 서 론

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[1]. 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[2].

본 논문에서는 트위터 자료의 분석을 통해서 다음 3가지의 특성을 이용하였다. 1) 트윗 개수의 증가 현상이 두드러지는 것이다. 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 때, 특히 사회적인 이슈로 발전되었을 때는 사람들은 그에 대한 관심이 폭발적으로 증가하게 되고 그 결과는 트윗 개수의 급격한 증가로 나타나게 된다. 2) 자신의 의견을 감성 어휘나 기호를 이용하여 직접 표현하는 것이다. 트위터 사용자들은 자신이 관심있어 하는 분야 혹은 이슈(issue)에 대해 지속적으로 자신의 의견을 긍정과 부정 어휘를 이용하여 댓글을 남기고 자신과 공통 분야에 관심이 있는 사람들과 서로 소통하기를 원한다. 트위터 데이터를 관찰해보면 어느 한 이슈에 대해 특정 사건이 발생하게 되면 그 사건에 해당하는 핵심 사건과 함께 감성 자질이 함께 출현하는 경우를 볼 수가 있다. 3) 이슈가 되는 사건에 대한 트윗의 내용에 자신의 공감을 표현하고 전파하기 위해 특정 트윗을 '리트윗(retweet, 퍼 나르기)' 하는 경향이 있다. 리트윗이 많이 된 트윗 내용은 중요시 다룰 만하다고 할 수 있다.

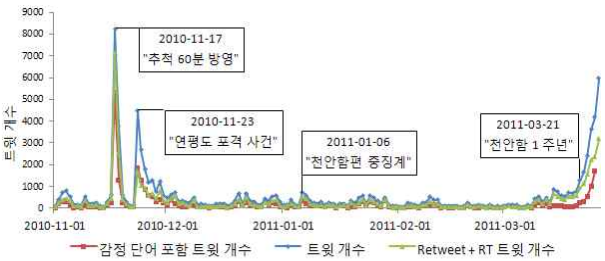


Fig. 1. Daily tweet graph for 'Cheonanham' Twitter data

Fig. 1은 천안함을 질의어로 하여 2010년 11월 1일부터 2011년 3월 26일까지 트위터 내에서 검색된 자료를 수집하여 시간별로 트윗 개수를 그래프화한 자료이다. 질의어에 대한 트윗 중 특정한 사건이 일어나지 않은 대부분의 날에는 트윗 개수가 일정수준 이하로 나타나는 것을 확인할 수가 있다. 하지만 2010년 11월 17일과 같이 천안함에 대해 어떠한 사건이 일어나게 되면 트윗 개수가 급격히 증가되는 것 또한 확인할 수가 있다. 실제로 트윗 개수가 전날 대비 급격히 증가한 2010년 11월 17일과, 같은 달 23일, 2011년 1월 6일, 2011년 3월 21일은 각각 천안함에 대한 추적 60분 방영, 연평도 포격 사건, 추적 60분 천안함편 중징계, 천안함 1주년이라는 사건이 발생했다. 이는 본 논문에서의 관찰한

† 준 회 원 : 전북대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수
논문접수 : 2012년 1월 18일
수 정 일 : 1차 2012년 5월 26일
심사완료 : 2012년 6월 3일
* Corresponding Author : Kyung-Soon Lee(selfsolee@chonbuk.ac.kr)

트위터 자료의 시간상에서의 트윗 개수 변화, 사용자들이 이슈에 대해 감정을 표현한다는 것, 그리고 리트윗(Retweet과 RT) 개수가 서로 관련이 있음을 볼 수 있다.

본 논문은 이러한 트위터 특성에 초점을 맞추어 핫 이슈들을 대상으로 트위터 데이터를 수집한 뒤, 이를 시간별로 분석하여 각각의 이슈에 사건이 발생함을 인식하고 보다 효과적인 자질 추출을 위해 시간별 어휘 빈도수, 감성 자질, 리트윗 개수, 그리고 시간상에서의 핵심어휘 분포 변화를 반영하기 위해 카이제곱(Chi Square)값을 사용하여 이슈에 대한 핵심 사건을 추출하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 분석된 데이터를 이용하여 핵심 사건을 추출하는 방법을 제안한다. 4장에서는 실험 및 분석 결과를 보여준다. 5장에서는 결론 및 향후 연구에 대해 논하겠다.

## 2. 관련 연구

최근 소셜 네트워크 서비스를 이용하는 사용자의 수가 증가함에 따라 트위터 및 소셜 네트워크 서비스 기반 다양한 연구가 나오고 있다. 본 논문과 관련된 연구로는 사회적인 사건 및 이슈를 인식하는 방법, 시간 분석을 이용한 연구, 감성 분석을 통한 오피니언 마이닝에 관한 연구 및 리트윗 정보를 이용한 연구가 있다.

사회적으로 이슈가 되는 사건을 인식하는 연구로 Popescu[3]는 트위터에서 논란이 되는 이슈를 발견하기 위해 3-회귀 기계학습 모델(3-regression machine learning models)을 사용하였고, Sayyadi[4]는 사건을 인식하기 위해 키워드(keyword) 그래프를 사용하였다. 또한, Benson[5]은 CRF(Conditional Random Field)를 이용하여 트위터 집합에서 이벤트에 대한 지역 및 아티스트 이름 등 정보를 추출하여 이벤트를 인식하는 그래픽(graphical) 모델을 사용하고, 이벤트에 대한 평가를 하기 위해서 도시 가이드 북을 사용하였다. 박지혜[13]는 Cyto-scape 플랫폼을 사용하여 트위터 사용자들의 관계를 시각적으로 표현할 수 있는 시스템을 개발하였으며, 성병기[14]는 이슈 키워드 추출 및 트위터와 유튜브에 기반한 실시간 검색 시스템을 구현하였는데, 이 시스템은 최근의 신문 기사들의 제목과 스니핏을 이용하여 이슈가 되는 키워드를 실시간으로 추출한 뒤 사용자들에게 제공해주고, 유투브와 트위터의 OpenApi를 이용해 추출된 키워드에 대한 컨텐츠들을 사용자들에게 실시간으로 제공해준다.

사회적인 이슈 및 어떤 사건이 일어났을 때 사용자가 자신의 입장하고 생각을 표현하기 위해 감성 단어를 자주 쓰는 경우가 많다. 오피니언 마이닝(opinion mining)에 관한 연구로 Pak[6]은 트위터 데이터 셋을 자동으로 구축하는 방법을 소개하고, 구축된 데이터 셋을 이용하여 긍정과 부정을 분류하는 분류기를 제안하였다. Popescu[7]는 이벤트 및 그 이벤트에 대한 설명을 추출하기 위해 감독 분류기(Supervised Classification method)와 긍정, 부정, 중립으로 구성된 감정 사전을 이용하였다. 본 논문에서는 강한 긍정 및 강한 부정으로 구성되어 있는 감성 어휘 사전 및 감성 기호 리스트를 이용하였다.

시간 분석을 이용한 연구로, 시간 간격을 이용하여 이슈를 인식하는 연구인 Zhao[8]는 시간의 흐름에 따라 일어나는 사회적인 이슈와 특별한 토픽(topic)사이의 관계에서 사건을 인식하였다. Lanagan[9]는 축구, 농구, 야구 등의 어떤 라이브 스포츠의 시작 시간부터 끝 시간 사이에 트위터에 올라온 트윗 및 사용자의 반응을 분석함으로써 사건을 인식하는 방법을 제안하였다.

리트윗을 분석한 연구로 Yang[10]은 트위트에 올라온 글들 중에서 25.5%가 다른 사람의 글을 리트윗(retweet) 한 것이라는 관찰을 통해 사용자의 리트윗 행태를 예측하기 위한 그래프 모델을 제안하였다. Boyd[11]은 사용자들이 트윗 내용을 새로운 사용자들에게 확장해서 널리 전파하기 위해서, 트윗 내용에 자신이 공감하고 있음을 공개적으로 드러내기 위해서, 그리고 다른 사람들의 생각을 검증하기 위해서 리트윗을 한다고 분석하였다.

본 연구에서는 리트윗 행태는 일반 내용에 대한 공감이라기 보다 이슈가 되는 사건에 대한 트윗의 내용에 자신의 공감을 표현하고 전파하기 위한 것으로 보고, 리트윗이 많이 된 트윗 문서에 나타난 어휘들의 중요도를 높여준다.

## 3. 시간상에서 트윗과 리트윗 정보를 이용한 핵심 사건 추출 방법

본 논문에서는 이슈에 대한 감정이 표현된 핵심 사건 어휘를 추출하기 위해 시간상에서 트윗 및 리트윗에 포함된 어휘 분포의 변화를 측정한다.

### 3.1 어휘 빈도수를 이용한 기본 자질 추출

기본 자질 추출을 위해 수집된 트위터 데이터를 형태소 분석기를 이용해 형태소 분석을 한 뒤, 불용어(Stop-Words)와 불필요한 URL정보를 제거하였다. 불용어의 제거는 네이버 실시간 검색어에서 무작위로 100개의 질의어를 추출, 트위터를 통해 검색하여 각 최대 100개의 트윗을 수집하였다. 수집된 트위터 데이터를 형태소 분석 후 어휘 빈도수를 계산하여 총 202개의 불용어 리스트를 만들었다. 정제된 자질들에 대해 하나의 트윗 내에서 거리(window size) 3이내에 있는 어휘들의 바이그램(Bigram)을 핵심 사건 어휘 후보로 추출하였다. 예를 들어, a b c d e 어휘에 대해서, ab, ac, ad, bc, bd, be 등 두 개의 어휘로 이루어진 사건어휘가 후보로 추출된다.

각 시간별로 바이그램으로 추출된 자질들을 어휘 빈도수를 계산하여 큰 값부터 순위화하였다. 빈도수를 이용하여 기본 자질들을 추출한다. 어휘 빈도수에 의한 기본 자질의 값을 $Freq(w, t0)$라 하겠다.

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

여기서 $w$는 바이그램 사건 어휘를 의미하고, t0는 각 날짜를 나타내고, D는 시간 t0에 속하는 트윗 문서를 나타낸다.

### 3.2 감성 자질을 반영한 자질 추출

수집된 트위터 데이터를 관찰해보면 어느 한 이슈에 대해 특정 사건이 발생하게 되면 그 사건에 해당하는 사건 어휘와 함께 감성 자질이 함께 출현하는 경우를 자주 볼 수가 있다. 아래 예는 핵심 사건 어휘와 감성 자질이 함께 나타난 것이다. 즉, '천안함'이라는 이슈에 대한 트윗 중에서 '연평도 포격'이라는 사건이 발생했을 때 '충격'과 같은 감성 어휘가 함께 나타난 것을 보여준다.

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

본 논문에서 사용한 감성 어휘는 윌슨 사전(Wilson lexicon)[12]에서 강한 감성 자질만을 추출하여 구글 번역기 API를 통해 한국어로 번역한 뒤, 사람이 직접 판단하여 한국어 감성 표현에 적합한 강한 긍정과 강한 부정을 나타내는 감성 자질 1192개를 구축하였다. Table 1은 감성 사전에 대한 정보를 나타낸다.

Table 1. Sentiment Lexicon

| 감성 자질 | 개수 | 예 |
|---|---|---|
| 강한 긍정 | 490 | 찬사, 칭찬, 격려, 동의, 신뢰, 옹호 |
| 강한 부정 | 702 | 격노, 냉소, 비난, 반대, 분노, 왜곡 |

감성 기호인 이모티콘(emoticon, 그림말)을 이용하여 사용자들이 트윗 글에 자신의 감정을 표시하기도 한다. 감정이 표현된 트윗에서의 사건어휘 자질 추출을 위해 사용한 감성 기호 리스트는 Table 2와 같다.

Table 2. Emoticon list

| 이모티콘 | 개수 | 예 |
|---|---|---|
| 긍정 | 156 | ^o^,(^^),: ),(^.~), <3, (^*^) |
| 부정 | 96 | (T_T),>.<, ㅠㅠ,:-(,(-_-) |

기본 사건 어휘 자질로 추출된 어휘들 중에서 상위 50개에 대해 해당 이슈의 전체 트위터 문서집합에서 감성 어휘 및 기호로 표현된 감성 자질과 함께 출현한 어휘 자질의 빈도수를 OpFreq(w, s)라 하겠다.

$$OpFreq(w,s) = \sum_{s \in D \wedge w \in D} tf(w,s,D) \quad (2)$$

여기서 $tf(w, s, D)$는 어휘 자질 $w$가 감성 자질 $s$과 함께 트윗 $D$에 나타난 빈도수를 나타낸다.

### 3.3 리트윗 정보를 반영한 자질 추출

트위터에서 리트윗은 다른 사람이 올린 글에 대한 공감을 표시하는 의미로 쓰인다. 큰 이슈에 대해서 자신의 생각을 글로 올리는 것보다는 쉬워서 많이 이용되고 있다. 사용자가 리트윗 또는 RT로 표현한 것 또한 공감의 감정을 표현하고 있다고 볼 수 있다.

리트윗(Retweet 및 RT)의 빈도수를 반영한 식은 다음과 같다.

$$RtFreq(w,t0) = \sum_{D \in t0 \wedge D \in Rt} tf(w,D) \quad (3)$$

여기서 $tf(w, D)$는 시간 $t0$에서 리트윗 Rt에 속하는 트윗 문서 $D$에 나타난 어휘 자질 $w$의 빈도수를 나타낸다.

시간 t0에서의 어휘 자질 $w$의 빈도수, 감성 자질 및 리트윗(retweet 또는 RT) 정보를 반영한 수식 OpRtScore(w, t0)은 다음과 같다.

$$OpRtScore(w,t0) = \alpha \cdot Freq(w,t0) + \beta \cdot OpFreq(w,s) + \gamma \cdot RtFreq(w,t0) \quad (4)$$

여기서 파라미터 값은 훈련 이슈에 대한 학습을 통해서 $\alpha$ = 0.2, $\beta$ = 0.5, $\gamma$ = 0.3로 설정하였다.

### 3.4 카이제곱을 이용한 시간별 자질 추출

시간별로 트위터 문서 개수를 분석한 결과 어떠한 이슈에 대해 특정 사건이 발생했을 때 그 날짜의 트윗 개수는 전날에 비해 급격하게 증가하는 현상을 보였다. 만약 그 사건이 그 이전에는 발생하지 않은 새로운 사건 또는 발생한 적이 거의 없는 사건이라면 사건 어휘에 대한 Freq(w, t0) 값이 그 이전날들의 데이터보다 폭발적으로 증가됨을 알 수 있었다. 이러한 특성을 반영하기 위해 본 논문에서는 시간 t0에서 사건 어휘자질 $w$의 중요도를 계산하기 위해 카이제곱을 이용하여 계산하였다. Table 3은 카이제곱 값을 계산하기 위한 분할표이다.

Table 3. Contingency table to calculate term significance on timelines

| | 자질 w가 포함되어있는 트윗 (w∈D) | 자질 w가 포함되어있지 않은 트윗 (w∉D) |
|---|---|---|
| 시간 t0의 트윗들 (D∈$t0$) | $a$ | $b$ |
| 시간 t0 이전 시간 트윗들 (D∉$t0$, D∈$t$, $t$<$t0$) | $c$ | $d$ |

시간 t0에서의 카이제곱 값의 계산 수식은 다음과 같다.

$$ChiSquare(w,t0) = \frac{(a+b+c+d)(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (5)$$

사건 어휘 자질의 순위화에 시간상에서의 특성을 반영한 ChiSquare(*w, t0*)값, 리트윗 및 감정 자질과 함께 출현한 정보를 반영한 OpRtScore(w, t0) 값을 이용한 최종 수식은 다음과 같다.

$$ChiOpRtScore(\mathrm{w},\mathrm{t0}) = \lambda \cdot ChiSquare(\mathrm{w},\mathrm{t0}) + (1-\lambda) \cdot OpRtScore(w,t0) \quad (6)$$

여기서 파라미터 $\lambda$값은 훈련 이슈에서의 학습을 통해서 0.3으로 설정했다.

트위터에서 사용자들이 글을 쓰고 리트윗하는 행태를 반영한 본 논문에서의 핵심사건 추출 방법은 수식 (6)에서와 같이 날짜별로 바이그램으로 추출한 어휘의 시간상에서의 사건 어휘를 포함한 트윗 개수의 변화 정도, 리트윗 정도 및 감정 자질과의 공기 정도를 반영한다.

## 4. 실험 및 분석

### 4.1 실험 집합

제안방법의 유효성을 검증하기 위해 네 개의 이슈에 대해 트위터 자료를 2010년 11월 1일부터 2011년3월 26일까지 Twitter API를 이용하여 수집하였다. 각 이슈에 대한 트윗 문서 개수는 Table 4와 같다.

Table 4. Twitter Data Set

| | 천안함 | 김연아 | 박지성 | 지진 |
|---|---|---|---|---|
| 문서 개수 | 84,195 | 26,844 | 131,533 | 46,795 |

Table 5는 각 이슈에 대해 추출한 사건들과 핵심 사건 자질의 상위 10에서의 정답 개수를 나타내고 있다. 트윗 개수가 아주 많거나 카이제곱 계산에서 아주 높은 값을 나타낸 경우를 발생 사건으로 인식하였다. 정답은 네이버 뉴스 검색을 이용하여 사람이 직접 판별하였다.

비교 실험을 위해서 수집된 각 이슈의 트위터 데이터에 대해서 각 날짜별로 기본 자질들을 추출하고, 그 자질들에 대해 어휘 빈도수에 따라 각 자질들을 순위화한 것을 비교 실험의 기준으로 하였다. 어휘 빈도수로 추출한 사건 어휘 후보들 중에서 상위 50개의 자질들을 선택하여 각 방법에 따라 재순위화 하였다. 비교 실험 방법은 다음과 같다.

- *Freq*: 해당 날짜에서 바이그램 어휘의 빈도수를 이용한 방법
- *OpRtScore*: 어휘 빈도수, 리트윗 및 감성 자질을 함께 반영한 방법

Table 5. Event lists and answers for four issues

| 이슈 | 사건 번호 | 발생 사건 | 날짜 | 정답 자질 개수 |
|---|---|---|---|---|
| 천안함 | E1 | 추적 60분 방영 | 2010.11.17 | 2 |
| | E2 | 연평도 포격 사건 | 2010.11.23 | 3 |
| | E3 | 추적 60분 천안함편 중징계 | 2011.01.06 | 5 |
| | E4 | 천안함 1주년 | 2011.03.21 | 4 |
| 김연아 | E5 | 프로그램 발표 | 2010.11.30 | 4 |
| | E6 | 유니세프 친선대사 | 2010.12.02 | 3 |
| | E7 | 김연아 도촬 사건 | 2010.12.27 | 6 |
| | E8 | 김연아 악마가면 | 2011.01.26 | 3 |
| 박지성 | E9 | 박지성 2골 | 2010.11.07 | 5 |
| | E10 | 박지성 시즌 6호 골 | 2010.12.14 | 4 |
| | E11 | 박지성 은퇴 | 2011.01.31 | 7 |
| | E12 | 차범근 고백 | 2011.02.01 | 2 |
| 지진 | E13 | 일본지진 발생 | 2010.11.30 | 3 |
| | E14 | 뉴질랜드 지진 발생 | 2011.02.22 | 6 |
| | E15 | 일본 쓰나미 경보 | 2011.03.09 | 3 |
| | E16 | 일본 대지진 발생 | 2011.03.11 | 1 |

- *ChiSquare*: 시간상에서 어휘를 포함한 트윗 개수의 변화를 계산한 카이제곱을 이용한 방법
- *ChiOpRtScore*: 어휘 빈도수, 리트윗 및 감성 자질과 카이제곱을 함께 반영한 제안 방법

성능 평가는 추출된 사건 어휘들 중에서 상위 10개 자질에 대한 정답포함율로 평가하였다. 예를 들어, 1/3은 상위 10개 자질에서 3개의 정답 자질 중에서 1개를 포함한 것을 나타낸다.

### 4.2 실험 결과

비교 실험 결과는 Table 6과 같다. 실험 결과에서 *Freq* 방법은 73.9%, OpRtScore 방법은 74.1%으로 비슷한 성능을 보였다. 이는 어떤 이슈에 대해 사건이 발생하면 사람들은 그 이슈와 함께 사건에 대해서도 언급을 하기 때문에 트윗의 개수가 증가할수록 핵심 사건 자질들의 어휘 빈도수도 증가하게 된다. 따라서 리트윗 및 감성 자질이 영향을 미치치 않을 정도로 어휘 빈도수 값이 커지기 때문에 단순한 어휘 빈도수로도 핵심 사건 자질들을 효과적으로 추출할 수 있었던 것이다. 시간상에서 어휘 변화를 반영한 *ChiSquare* 방법은 89.0%의 성능을 나타냈다. 모든 자질을 고려한 제안 방법인 *ChiOpRtScore*은 94.9%로 높은 성능을 보였다. 이는 기준 성능인 *Freq*에 비해 21.0%의 성능 향상을 보인 것이다.

결과 분석을 위한 Table 7는 “천안함” 이슈의 “연평도 포격” 사건이 일어난 날짜의 트위터 데이터에 대해 각 방법에 의해 추출된 핵심 사건 자질들을 상위 10까지 순위화한 것을 보여준다.

실험 결과에서 *Freq*로 순위화했을 때, “연평도 포격”이라는 핵심 사건이 비교적 낮은 순위에 랭크 되어있지만, 리트

Table 6. Performance comparisons (p@10)

| 사건 번호 | *Freq* | *OpRt Score* | *ChiSquare* | *ChiOpRt Score* |
|---|---|---|---|---|
| E1 | 1/2 | 1/2 | **2/2** | **2/2** |
| E2 | 1/3 | 2/3 | 2/3 | 2/3 |
| E3 | 3/5 | 3/5 | **5/5** | **5/5** |
| E4 | 3/4 | **4/4** | 3/4 | **4/4** |
| E5 | 2/4 | 1/4 | 3/4 | **4/4** |
| E6 | 1/3 | 2/3 | 2/3 | **3/3** |
| E7 | 5/6 | 5/6 | 5/6 | 5/6 |
| E8 | **3/3** | **3/3** | **3/3** | **3/3** |
| E9 | 3/5 | 4/5 | **5/5** | **5/5** |
| E10 | **4/4** | **4/4** | 3/4 | **4/4** |
| E11 | 5/7 | 5/7 | **7/7** | 6/7 |
| E12 | **2/2** | 1/2 | **2/2** | **2/2** |
| E15 | **3/3** | **3/3** | **3/3** | **3/3** |
| E14 | **6/6** | 4/6 | 5/6 | 5/6 |
| E15 | 2/3 | 2/3 | **3/3** | **3/3** |
| E16 | **1/1** | **1/1** | **1/1** | **1/1** |
| 평균 | **73.9%** | **74.1%** | **89.0%** | **94.9%** |

Table 7. The top 10 event candidate term extracted by each method

| 순위 | *Freq* | *OpRtScore* | *ChiSquare* | *ChiOpRt Score* |
|---|---|---|---|---|
| 1 | 천안함 사건 | **연평도 포격** | 천안함 연평도 | 북한 도발 |
| 2 | 천안함 연평도 | 민간인 사찰 | 대포폰 사찰 | **연평도 사건** |
| 3 | 천안함 북한 | 천안함 사건 | **연평도 포격** | **연평도 포격** |
| 4 | 천안함 사태 | 현대차 비정규직 | 훈련 북한 | 민간인 사찰 |
| 5 | 민간인 사찰 | 천안함 연평도 | 북한 도발 | 천안함 사태 |
| 6 | 북한 도발 | 천안함 북한 | 사찰 천안함 | 대포폰 사찰 |
| 7 | 대포폰 사찰 | 천안함 사태 | **연평도 사건** | 훈련 북한 |
| 8 | 사찰 천안함 | 북한 도발 | 사건 천안함 | 대포폰 민간인 |
| 9 | **연평도 포격** | **연평도 사건** | 민간인 사찰 | 사건 천안함 |
| 10 | 사건 천안함 | 대포폰 사찰 | 대포폰 민간인 | 사찰 현대차 |

윗 및 감성 자질과 함께 순위화한 *OpRtScore*에서는 새로운 핵심 사건인 "연평도 사건"이 추가되면서 "연평도 포격"은 순위가 올라갔음을 알 수 있다. 시간상에서의 어휘 변화를 반영한 *ChiSquare*에서는 감성 자질과 함께 사용해서 순위화한 결과와 비슷한 결과를 보여주고 있다. 제안 방법인 리트윗 및 감성 자질과 카이제곱 값을 함께 사용한 ChiOpRtScore값으로 순위화한 결과가 가장 좋게 나오는 것을 확인할 수 있다.

Fig. 2는 "천안함", "김연아", "박지성", "지진" 이슈에 대한 트위터 자료에서 해당하는 날짜 내에 감성 자질과 카이제곱 값이 높은 자질들을 보여주고 있다. Fig. 1에서 트윗 개수로 파악할 수 있는 사건은 Fig. 2의 "천안함" 부분과 일치하는 것을 볼 수 있다. 여기서 2011-01-06일에 일어난 "추적 60분 천안함편 중징계" 사건은 그림 1의 트윗 개수로는 변별이 되지 않지만 감성 자질과 카이제곱을 이용하는 방법에서 이슈를 탐지할 수 있는 장점이 있음을 알 수 있다.

## 5. 결론 및 향후 연구

본 논문에서는 어느 특정 이슈에 관한 트위터 데이터에서 그 이슈에 사건이 발생 했을 경우 트윗의 개수가 크게 증가하고 감성 표현이 많다는 관찰을 통해 시간 자질과 감성 자질을 이용한 사건 추출 방법을 제안하였다. 핵심 사건을 추출 하는 방법으로는 어휘 빈도수 만을 이용한 방법, 어휘 빈도수, 리트윗 및 감성 자질을 함께 이용한 방법, 카이제곱 만을 이용한 방법, 어휘 빈도수, 리트윗 및 감성 자질과 카이제곱을 함께 이용한 방법으로 비교실험을 하였다.

실험 결과 16개 사건에 대해 상위 10개에서 정답 포함률

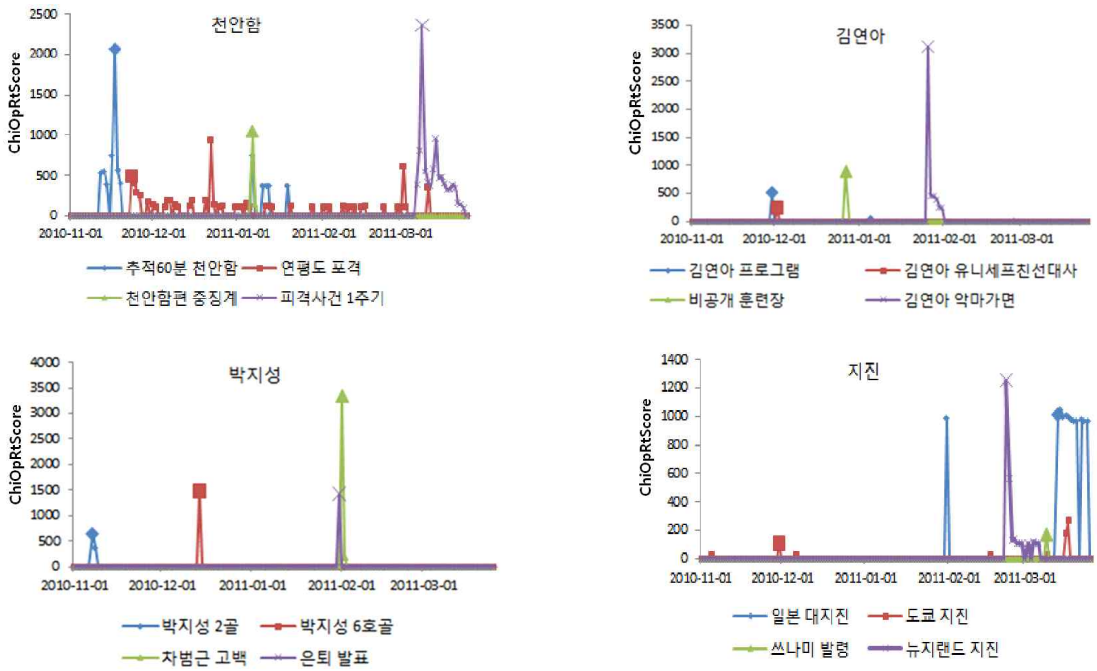


Fig. 2. Event term weights for each issue by ChiOpRtScore

을 평가하였을 때 본 논문에서 제안한 방법이 94.9%의 성능으로 아주 우수한 성능을 보였다. 이를 통해 카이제곱을 이용한 시간 자질과 감성 자질 및 리트윗 자질이 사건 추출에 효과적인 방법임을 알 수 있다.

향후 연구에는 트위터 데이터를 통합하여 특정 자질의 어휘 빈도수 값에 대한 의존도를 줄이고, 기본 자질을 추출할 때 핵심어구(key-phrase)을 사용하여 보다 효과적인 핵심 사건 자질들을 추출하는 연구가 필요하다.

## 참 고 문 헌


[1] Naver Knowledge Dictionary, "Twitter", http://terms.naver.com/

[2] Duhwan Lee reporter, "Revolution of 140 characters is shaking the Korea, 'the power to change the world, Twitter" http://crepasnews.com

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

[4] H.Sayyadi, M. Hurst, and A. Maykov. "Event Detection and Tracking in Social Streams", In P*roceedings of ICWSM*, 2009.

[5] E.Benson, A.Haghighi, and R.Barzilay, "Event Discovery in Social Media Feeds" In *Proceedings of ACL*, 2011

[6] A.Pak, P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In *Proceedings of LREC*, 2010.

[7] A.-M. Popescu, M.Pennacchiotti, Deepa Arun Paranjpe. "Extracting events and event descriptions from Twitter", In *Proceedings of WWW*, 2011.

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

[9] J.Lanagan and Alan F. Smeaton, "Using Twitter to Detect and Tag Important Events in Live Sports", In *Proceedings of AAAI*, 2011.

[10] Z.Yang, J.Guo, K.Cai, J.Tang, J.Li, L.Zhang, and Z. Su, "Understanding retweeting behaviors in social networks" In Proceedings of CIKM, 2010.

[11] D.Boyd, S.Golder and G.Lotan. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter", In Proceedings of HICSS - 43 IEEE, 2010.

[12] T. Wilson, J. Wiebe, and P. Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis", In *Proceedings of HLT/EMNLP*, 2005.

[13] J. H. Park, B. H. Kim, M. J. Lee and Y. K. Kwon, "TwitNet : Cytoscape Plugin for Visualizing Relation betweens Twitter Users", In proceedings of Korean Institute of Information Scientists and Engineers (KIISE-2010), Vol.37, No.1(D), pp.316-321, June, 2010.

[14] B. K. Sung, J. Y. Oh and J. W. Cha, "LiveTwitter: Hot Issue Search system Based on Twitter", In proceedings of HCLT2010, 2010, pp.179-182.



**Bayar Tsolmon**

e-mail : bayar_277@chonbuk.ac.kr

2012년 전북대학교 컴퓨터공학부(학사)

2012년~현 재 전북대학교 컴퓨터공학과 석사과정

관심분야 : 정보검색, 정보마이닝



**이 경 순**

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과(학사)

1997년 한국과학기술원 전산학(공학석사)

2001년 한국과학기술원 전산학(공학박사)

2001년~2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원

2004년~현 재 전북대학교 컴퓨터공학부/영상정보신기술연구센터 부교수

관심분야 : 정보검색, 정보마이닝


# 트위터 문서에서 시간 및 리트윗 분석을 통한 핵심 사건 추출


Bayar Tsolmon† · 이 경 순††



## 요 약

인터넷 사용자들은 어떠한 이슈에 대해 소셜 네트워크 서비스를 통해 빠르고 간결하게 다른 사람들과 지속적인 커뮤니케이션을 원한다. 사회적 이슈에 대해 어떠한 사건이 일어나게 되면 그날의 트윗 글과 리트윗 개수에 영향을 미치게 된다. 본 논문에서는 트위터 자료에서 사회적인 핵심 사건을 추출하기 위해 시간 분석과 감성 자질 및 리트윗 정보를 이용하는 방법을 제안한다. 제안 방법의 유효성을 검증하기 위해 비교 실험으로 어휘 빈도수를 이용하여 핵심 사건을 추출하는 방법, 어휘 빈도수와 감성 자질을 함께 이용한 방법, 시간 분석을 반영하기 위해 카이제곱만을 이용한 방법과 제안 방법인 어휘 빈도수, 감성 자질, 리트윗 및 카이제곱을 함께 이용한 방법으로 성능을 비교하였다. 성능 평가를 위해서는 추출된 사건리스트에서 상위 10개 결과에서 정확도를 계산하였는데, 제안 방법이 94.9%의 성능을 보였다. 실험을 통해 제안한 방법이 핵심 사건 추출에 효과적인 방법임을 알 수 있다.



**키워드 : 사건 추출, 시간 분석, 리트윗, 감성 자질, 카이제곱, 트위터 코퍼스**