트위터를 활용한 이벤트 결정 모듈 설계

임준엽, 윤진영, 이범석, 황병연
가톨릭대학교 컴퓨터공학과
e-mail:junyeob1205@catholic.ac.kr

Designing of Event Decision Module using Twitter

Junyeob Yim, Jinyoung Yoon, Bumsuk Lee, Byung-Yeon Hwang
Dept. of Computer Science and Engineering, The Catholic University of Korea

요 약

최근 스마트폰의 보급과 더불어 소셜 네트워크 서비스의 사용자가 급증하였다. 그 중 트위터는 개방적인 네트워크 구조로 인한 정보의 빠른 확산성을 가지고 있다. 또한 트위터 사용자들은 주로 자신들이 경험하거나 겪은 일들을 글로 작성하여 다른 사용자들과 공유한다. 따라서 그들이 남긴 데이터를 수집하고 분석할 수 있다면 트위터를 이벤트 탐지의 도구로써 활용하는 것이 가능하다. 이에 본 논문에서는 트위터를 이용하여 이벤트를 탐지하는 시스템을 제안한다. 실험을 위해 6개월간 수집한 트윗을 이용하였으며 분석을 위해 트윗 발생량에 관한 각종 수치들을 제시하였다. 이를 이용하여 이벤트 후보 지역들을 선별하였고 실험 결과 최종 90%의 탐지율로 이벤트 지역들을 추출하였다.

1. 서 론

소셜 네트워크 서비스(Social Network Service; SNS)는 지인들과의 관계망을 웹(Web)상에 구축하여 서로간의 교류를 돕는 온라인 서비스이다. 최근 들어 스마트폰의 도입으로 인해 웹 접근성이 확대되었고 SNS를 이용하는 사용자들이 급증하였다. 국내 시장에서 많은 인기를 얻고 있는 SNS로는 트위터(Twitter)나 페이스북(Facebook), 카카오스토리(Kakao Story)등이 있다.

그 중 트위터는 다른 SNS와는 구별되는 여러 특징들이 있다. 가장 주요한 특징은 트위터가 가진 개방적인 네트워크 구조이다. 트위터는 사용자간의 관계형성을 위해 팔로워(Follower)-팔로잉(Following)이라는 개념의 시스템을 이용한다. 이는 두 명의 사용자간에 이루어지는데, 이러한 관계가 성립되어야 비로소 특정 사용자가 남긴 게시글을 볼 수 있다. 다른 SNS의 경우에도 이와 유사하게 사용자간의 관계형성을 위한 시스템들이 있으나, 대부분 상호합의하에 관계가 이루어진다. 그러나 트위터는 한쪽의 일방적인 요청만으로도 관계형성이 가능하다. 따라서 이러한 특징은 사용자간의 네트워크 구조를 보다 개방시켜놓기 때문에 다른 SNS보다 빠른 정보의 확산성을 제공한다. 또한 트위터 내의 사용자들은 140자 제한의 단문인 트윗(Tweet)을 작성함으로서 다른 사용자들과 자신의 의견이나 정보 등을 공유한다. 이는 사용자로 하여금 비교적 가벼운 내용의 게시글을 간편하게 작성하도록 유도한다. 따라서 특정 현상을 경험한 사용자로 하여금 보다 빠른 트윗 작성을 유도한다. 마지막으로 개발자들에게 다양한 API를 제공하여 트윗 코퍼스(Corpus)를 가공하고 처리하는 것이 용이하기 때문에 이미 많은 관련 연구에서 이용되고 있다.

트위터 사용자들은 트윗 작성을 통해 주로 자신의 일상이나 경험, 새로운 정보 등을 다른 사용자들과 공유한다. 트윗의 내용적 분류로는 C. Hong 등의 연구[1]에서 언급한 바와 같이 새로운 정보나 뉴스, 개인의 의견이나 감정, 기업의 광고나 홍보, 캠페인 등이 주를 이룬다. 그 중 새로운 정보나 개인의 경험 등은 이벤트 탐지의 도구로서 활용될 수 있다.

이벤트 탐지의 관점에서 트위터를 이용하기 위해서는 우선 특정 이벤트에 대해 평소보다 트윗 발생률이 높아지는 지점을 찾아야 한다. 이후 해당 지점이 실제 이벤트인지 아닌지를 결정함으로서 이벤트의 최종 탐지가 이루어진다. 이에 본 논문에서는 후보 이벤트를 선별하고 최종 이벤트를 결정하는 모듈을 제안한다. 이후 제안하는 시스템의 성능을 평가하기 위해 추출된 최종 이벤트가 실제 이벤트였는지를 검증하였다.

본 논문의 구성은 다음과 같다. 2장의 관련 연구에서 본 논문과 관련된 연구들을 살펴보고 3장에서 제안하는 시스템의 구조와 실험방법을 소개한다. 이후 4장에서 시스템의 성능을 평가하고 5장의 결론과 향후 연구계획에 대해 기술한다.

2. 관련 연구

T. Sakak i등은 트위터를 이용해 현실의 이벤트를 탐지하기 위해 Toretter라는 시스템을 제안하였다[2]. 이 시스템은 트위터 내의 사용자들이 각종 재난상황에서 자신이 겪고 있는 상황을 트윗을 통해 다른 사용자와 공유한다는 사실을 기반으로 연구되었다. 특히 일본에서 발생한 지진이나 태풍에 대해 미리 지정된 키워드를 이용하여 관련 트윗들을 수집하였으며, 이를 기반으로 이벤트가 발생한 지리적 위치를 탐지해 냈다. 이와 더불어 이벤트의 진행방향을 예측하였고 일본의 기상청보다 빠르게 경보를 발생시켰다. 그러나 Torretter시스템은

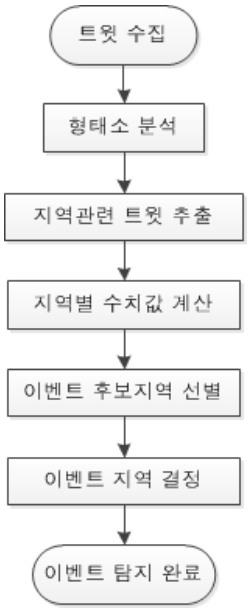
※ 본 연구는 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2011-0009407).

사전에 입력한 키워드가 포함된 트윗들만을 이용한다는 점에서 탐지할 수 있는 이벤트의 범위가 키워드의 내용에 국한된다는 한계를 가지고 있다. 또한 이벤트 관련 트윗의 발생 위치를 판별하기 위해 트윗 내용 외에 따로 저장되는 지역좌표(Geocode)를 이용하였다. 그러나 트윗을 작성한 사용자가 지역좌표를 자동적으로 저장하는 것에 대해 동의를 하지 않는다면 정확한 트윗의 발생위치를 알아내는 것은 쉽지 않다[3]. 따라서 이 경우 지역좌표가 입력되지 않은 트윗은 Toretter시스템에서 이용할 수 없다. 최근 트위터 사용자들이 위치정보를 공개하는 것에 대해 부정적인 반응을 보이는 것을 감안하면 지역좌표에 의존하는 트윗 발생위치 추정방식은 명확한 한계가 있다.

국내에서도 이벤트 탐지를 시도한 연구들이 선행되었다. J. Yim 등의 연구[4]에서는 앞서 소개한 연구와는 다르게 현실에서 발생한 대부분의 이벤트가 물리적인 위치를 가진다는 사실을 기반으로 지명을 포함한 트윗들을 이벤트 탐지의 도구로 활용했다. 따라서 이벤트 관련 키워드를 미리 지정하지 않고도 보다 다양한 이벤트를 탐지하였다. 이러한 방식의 이벤트 탐지는 특정 장소에서 발생하는 이벤트에 대해 다양한 탐지가 가능하다는 장점이 있다. 그러나 이 연구에서는 이벤트의 후보가 되는 지역을 선별할 뿐 최종적인 이벤트의 결정은 수행하지 않았다.

3. 이벤트 결정 모듈

본 논문에서는 현실에서 발생되는 이벤트를 탐지하기 위해 트위터 사용자가 남긴 트윗을 이용하였다. 트윗을 분석하여 이벤트를 추출하는 전체적인 과정은 (그림 1)과 같다.



(그림 1) 이벤트 탐지 과정

본론에 들어가기에 앞서 실험을 위한 트윗 데이터 수집을 위해 트위터에서 제공하는 Streaming API[5]를 이용하여 2013년 8월 1일부터 약 6개월간 트윗을 수집하였다. Streaming API는 트위터 사용자가 작성한 트윗을 실시간으로 전송해주기 때문에 실시간 이벤트 탐지에 용이하다. 그러나 본 논문의 경우에는 실시간적인 시스템의 구축이 목적이 아니라 과거의 이벤트를 대상으로 제안하는 이벤트 결정 방법이 효과적인지를 판단하는 것이 목적이므로 실시간적인 분석은 수행하지 않았다.

트윗의 내용을 분석하여 새로운 정보를 추출하는 대다수의 연구에서는 자연어 처리가 필수적이다. 따라서 이를 위해 루씬(Lucene) 한글 형태소 분석기[6]를 이용하였다. 자연어 처리를 통해 문장으로 작성된 트윗에서 명사들을 추출해내는 것이 가능하며, 추출된 명사들은 이벤트 탐지에 이용될 키워드로서 활용된다.

형태소 분석을 통한 키워드 추출이 완료되면 키워드 중 행정구역명과 지하철역명을 포함하는 트윗들을 추출한다. [4]의 연구에서 언급한 바와 같이 대다수의 현실 속 이벤트들은 특정 장소에서 발생되기 때문에 이벤트가 발생한 지리적 위치를 구분하기 위해 행정구역명을 이용하였다. 이와 더불어 보다 넓은 범위의 지역관련 트윗의 구분을 위해 지역별 랜드마크의 개념으로서 지하철역명을 추가로 추출작업에 이용하였다.

이후 추출된 지역관련 트윗들을 각각의 지명별로 나누어 이벤트 탐지를 위한 해당 지역에서의 수치 값들을 계산한다. 본 논문에서 제안하는 이벤트 탐지를 위한 수치들은 <표 1>과 같다.

<표 1> 트위터 이벤트 탐지를 위한 수치 값

수치 명	설명
TF(Term Frequency)	각 지역별로 1개의 구간에서 발생한 트윗 개수
VT(Variety of Tweets)	각 지역별로 1개의 구간에서 발생한 트윗의 종류 수
DA(Document Average)	각 지역별로 72개의 구간 동안 발생한 트윗의 평균 개수

수집된 트윗을 분석하기 위해 발생 시간에 따라 40분단위로 구간을 나누었다. <표 1>의 TF와 VT는 하나의 구간에서 특정지역을 언급한 트윗의 개수와 종류 수이다. 또한 DA는 72개의 구간인 2일 동안 특정 지역을 언급한 트윗들의 평균 개수이다. 예를 들어 A지역을 언급한 트윗이 한 구간에서 총 100회가 발생됐고, 이 중 중복된 트윗이나 리트윗(Retweet)을 제거한 결과 10개의 트윗이 남는다면 A지역에 대한 해당 구간에서의 TF값과 VT값은 각각 100과 10이 된다. 이때 그 구간을 포함한 2일간의 A지역 관련 트윗이 3600개가 발생되었다면 DA값은 3600을 72로 나눈 50이 된다. 이러한 방식으로 각 지역별 해당구간에서의 이벤트 탐지에 필요한 수치 값들

을 계산한다.

마지막으로 이벤트 탐지를 위해 이벤트가 발생한 구간에서의 트윗 증가량과 DA 및 VT값을 이용한다. 트윗 증가량의 경우 TF에서 DA를 뺀 값으로 계산한다. 이벤트 결정을 위해 이벤트가 발생한 구간에서의 트윗 증가량을 계산하여 이벤트 후보지역들을 추출한 후, DA와 VT값의 비교를 통해 최종 이벤트를 결정한다.

이벤트가 발생할 경우 해당 지역을 언급하는 트윗의 양은 평소보다 증가하는 추세를 보인다. 그러므로 트윗의 발생량이 증가한 지역들을 이벤트 후보지역으로써 선별한다. 이때 트윗의 발생량이 기본적으로 불규칙적인 지역들인 경우 이벤트의 발생여부와 관계없이 후보지역에 선별될 수 있다. 따라서 그러한 지역을 배제하기 위해 트윗 증가량이 최소 10건 이상인 지역들을 이벤트 후보지역으로 선별한다.

최종 이벤트를 결정하기에 앞서 이벤트 후보지역에는 포함되었으나 실제로는 이벤트가 발생하지 않은 지역들의 특징을 살펴볼 필요가 있다. 대부분의 경우가 다량의 중복된 트윗으로 인한 순간적인 트윗 발생량이 증가하여 이벤트 후보지역에 포함된 지역들이었다. 따라서 이와 같은 지역들을 선별된 후보지역에서 제거하기 위해 발생된 트윗의 종류가 얼마나 다양한지를 고려한다. 이를 위해 해당 지역의 DA와 VT를 비교하여 VT가 DA보다 큰 지역만을 추출한 후 최종 이벤트 지역으로 결정한다.

4. 성능 평가

본 논문에서 제안하는 시스템의 성능을 평가하기 위해 트윗 수집기간 중 발생한 실제 이벤트 10건에 대해 탐지율과 정확도를 계산해 보았다. 실험을 위해 사용된 이벤트는 <표 2>와 같다.

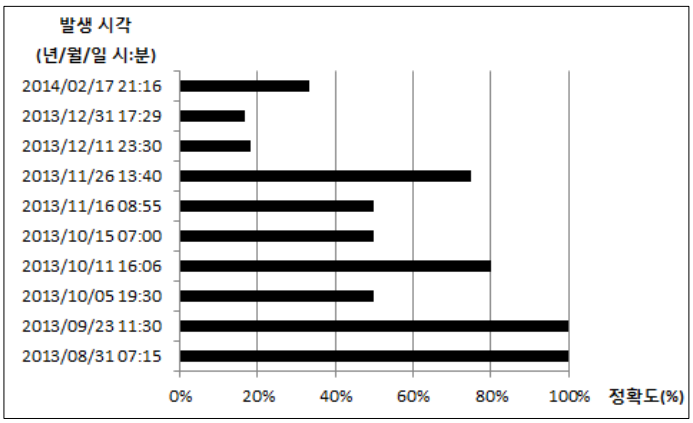
<표 2> 이벤트 탐지율

발생 시각 (년/월/일 시:분)	발생 위치	이벤트 내용	탐지 여부
2014/02/17 21:16	경주시	리조트 붕괴	O
2013/12/31 17:29	서울역	분신	O
2013/12/11 23:30	화명동	화재	O
2013/11/26 13:40	구로디지털단지역	화재	O
2013/11/16 08:55	삼성동	헬기 추락	X
2013/10/15 07:00	사당역	지하철 고장	O
2013/10/11 16:06	영덕군	지진	O
2013/10/05 19:30	여의도역	불꽃 축제	O
2013/09/23 11:30	대명동	화재	O
2013/08/31 07:15	대구역	기차 충돌	O

<표 2>에서 탐지된 이벤트 수를 실험에 쓰인 전체 이벤트 수로 나눈 후 100을 곱하면 전체적인 시스템의 탐지율을 구할 수 있으며, 식 (1)과 같다. 여기서 D_n은 탐지된 이벤트 수를 의미하고, U_n은 실제 이벤트이지만 탐지가 안 된 이벤트 수를 의미한다.

$$Detection\,rate(\%) = \frac{D_n}{D_n + U_n} \times 100(\%) \quad (1)$$

식 (1)에 의해 계산된 시스템의 탐지율은 90%였다. <표 2>의 삼성동에서 발생한 헬기 추락사고의 경우 기존에 "삼성"이라는 단어가 이미 많이 발생되어 헬기 추락으로 인한 트윗의 증가가 미비했던 것으로 나타났다.



(그림 2) 이벤트 탐지 정확도

한편 같은 시간대에는 하나의 이벤트만이 아닌 여러 지역에서 여러 이벤트들이 발생할 수 있다. 따라서 시스템에서 특정 시간대에 반환된 다수의 지역들이 실제 이벤트 지역이었는지에 대한 평가 기준으로서 시스템의 시간대 별 정확도를 계산하였다. 계산을 위해 특정 시간대에 반환된 지역들 중 실제 이벤트가 발생한 지역의 수를 총 반환된 지역의 수 $T_n + F_n$으로 나누었다. 이는 식 (2)와 같다. 여기서 T_n은 실제 이벤트가 발생했던 지역의 수를 의미하고, F_n은 이벤트가 발생하지 않은 지역의 수를 의미한다.

$$Accuracy(\%) = \frac{T_n}{T_n + F_n} \times 100(\%) \quad (2)$$

전체 실험데이터에 대해 평균 57%의 정확도를 보였으며 개별적인 정확도는 (그림 2)에 정리하였다. (그림 2)에서 2013년 9월 23일에 발생한 대명동 화재의 경우 같은 시각에 청주와 전주에서도 다른 이벤트가 발생했었다. 즉, 추출된 모든 지역이 이벤트 지역이었다. 따라서 해당 시간대에서의 이벤트 탐지 정확도는 100%이다. 반면에 2013년 12월 31일 17시 29분에는 총 6개의 지역이 이벤트 지역으로 반환되었으나 실제 이벤트가 발생한 지역은 서울역뿐이었다. 함께 반환된 지명으로는 중계, 대화, 남성, 달성, 강화 등이 있었다. 이는 지역을

뜻하는 단어가 아닌 해당 지역명과 같은 발음을 가진 동음이의어들이다. 따라서 이와 같은 단어들을 노이즈로 간주하여 배제시키고, (그림 2)에 기재된 정확도를 다시 계산하면 평균 93%가 나온다.

5. 결론 및 향후 연구 계획

본 논문에서는 이벤트 탐지를 위해 이벤트를 결정하는 시스템을 제안하였다. 이벤트를 추출하기 위해 트윗을 수집하고 형태소 분석을 통해 키워드를 추출하였다. 이후 현실에서 발생한 이벤트를 추출하기 위해 특정 지역명을 포함한 트윗들을 선별하였다. 마지막으로 이벤트 후보지역을 추출한 후 최종 이벤트 지역을 결정하였다. 제안하는 시스템으로 이벤트를 탐지해 본 결과 90%의 탐지율로 이벤트를 추출하였으나, 탐지된 결과의 정확도는 57%로 비교적 낮았다. 이는 이벤트 후보로 추출된 지역명이 일반적으로 자주 쓰이는 단어와 동음이의어 관계였기 때문이다. 따라서 이와 같은 단어들을 노이즈로 보고, 이를 개선하기 위한 방안을 향후 연구과제로 정한다. 이와 더불어 이벤트가 발생한 지역만을 탐지하는 것이 아닌 해당 지역에서 어떤 이벤트가 발생하였는지를 검색하기 위한 방법을 연구할 계획이다.

참고 문헌

[1] C. Hong, H. Kim, “Effective Feature Extraction for Tweets Classification,” Proc. of 2011 Korea Computer Congress, pp. 229-232, 2011.

[2] T. Sakaki, M. Okzaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” Proc. of the 19th Int'l Conf. on World Wide Web, pp.851-860, 2010.

[3] B. Lee, S. Kim, B.-Y. Hwang, “Analyzing the Credibility of the Location Information Provided by Twitter Users,” Journal of Korea Multimedia society Vol.15, No.7, pp.910-919, 2012.

[4] J. Yim, S. Kim, J. Yoon, P. Oh, B. Lee, B.-Y. Hwang, “Detecting Local Events Using Twitter,” Proc. of 2013 Korea Computer Congress, pp.248-250, 2013.

[5] Twitter. (2012, Sep. 24). The Streaming APIs|Twitter Developers [Online]. Available: https://dev.twitter.com/docs/streaming-apis

[6] S. Lee. (2008, Oct. 18). Lucean Korean Morph Analyzer [Online]. Available: http://cafe.naver.com/korlucene